Language Model-based Approach For Extracting Comorbidities And Complications In Fabry Disease Clinical Notes

UNIVERSITAT DE BARCELONA

AUTHORS

Manuel Cossio1,2 & Ramiro Gilardino1

HE-XPERTS CONSULTING

AFFILIATIONS

1. *HE-Xperts Consulting LLC, Miami, FL, USA.*
2. *Universitat de Barcelona, Barcelona, Spain*

Keywords: Digital health, Comorbidities, Chat GPT, Rare diseases

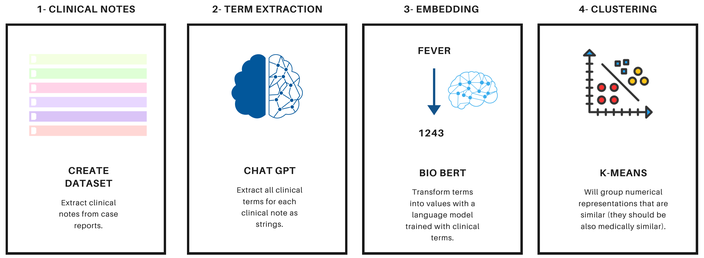

Figure 1. Workflow illustrating the stepwise process of medical terms extraction. *The diagram depicts the extraction of medical terms using Chat GPT, followed by the generation of embeddings through the BIO BERT model. Subsequently, the clustering of these embeddings is performed utilizing the K-means algorithm..*

01. Introduction

In the realm of healthcare, the pressing need to alleviate the burden on healthcare systems through automation has never been more critical. Rare diseases, in particular, pose a daunting challenge due to their limited patient population, making efficient management a complex task. Extracting comorbidities and complications from clinical data not only aids in understanding their frequency but also plays a pivotal role in prevention strategies. Recognizing the potential of language models, this study endeavors to harness their power, aiming to automate the intricate task of analyzing clinical notes. By doing so, it not only addresses the unique challenges posed by rare diseases but also represents a significant leap toward precise and proactive healthcare management, heralding a new era of efficiency and effectiveness in the medical field (1, 2).

02. Objective

Rare disease morbidity burdens healthcare systems. Equipping teams to manage rare diseases can alleviate strain. This study develops an automated system using language models to extract complications and comorbidities from clinical notes of Fabry disease patients.

References

1. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., ... & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine, 6(1), 135.
2. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine, 4(1), 86.

03. Methods

Clinical notes were analyzed using prompt engineering with the chat GPT API in Google Colaboratory with Python 3.9. Extracted terms were embedded using the BIOBERT model and subjected to K-means clustering. GPT chat extracted two terms per cluster. Different cluster numbers (CN) were tested, evaluating mean and standard deviation (SD) of terms per cluster, total clusters (including outliers and those with representative terms), and clusters with specific Fabry disease-related terms (Fig. 1).

04. Results

The extraction and transformation of terms into embeddings were consistently completed within an average of 7 seconds. A comprehensive series of 17 tests was conducted, ranging from 5 to 30 clusters (Fig. 2, 4). As the cluster number (CN) increased, there was a systematic decrease in both the mean and standard deviation of the number of cluster terms (Fig. 2). Outliers exhibited fluctuations across all tests, particularly at lower values (below 5), observed within the 15 to 30 cluster range (Fig. 4). Disease-specific terms exhibited a progressive increase until reaching 16 clusters (Fig. 4). Notably, it was observed that beyond this 16-cluster threshold, there was no significant rise in pathognomonic representative terms specific to Fabry disease, such as proteinuria, acroparesthesia, vasculopathy, and angiocheratoma. Instead, the representative terms expanded to include nonspecific factors like dizziness and nausea, highlighting the crucial importance of cluster optimization for precise term extraction.

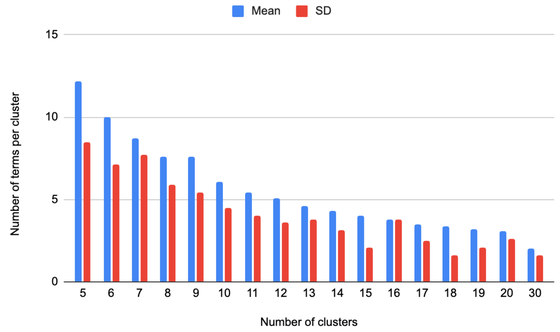

Figure 2. Bar plot illustrating the impact of varying cluster numbers on term grouping using k-means. *Each bar represents the mean and standard deviation (SD) of different cluster number.*

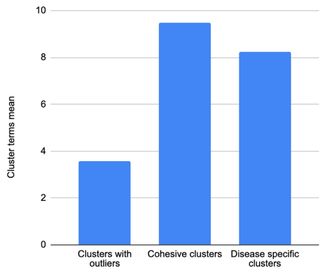

Figure 3. Bar plot depicting the analysis of cluster configurations. *Each bar represents showcases the mean value of clusters with outliers, clusters with cohesive terms, and clusters with disease-specific terms.*

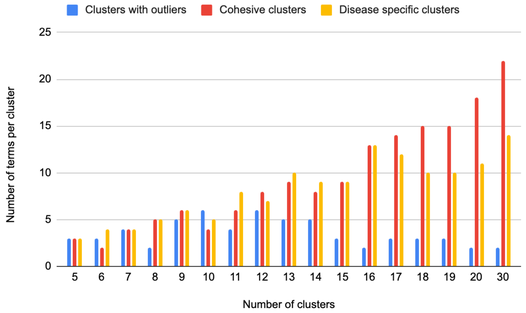

Figure 4. Bar plot displaying cluster characteristics after t-SNE analysis. *Each bar corresponds to a specific number of clusters, illustrating the count of clusters containing outliers, cohesive clusters, and disease-specific clusters. This visualization provides a detailed insight into the distribution of clusters with varying attributes, offering a comprehensive understanding of the clustering outcomes.*

05. Conclusion

This research successfully achieved the extraction of comorbidities and complications from a dataset of 20 clinical notes in an average processing time of approximately 7 seconds. Although effective in clustering terms that possess characteristics related to Fabry disease, there is a need for further investigation to reduce the inclusion of nonspecific terms in forthcoming analyses.