IDENTIFYING PATTERNS IN DRUG EFFICACY BY ANALYZING DRUG REVIEWS THROUGH A CLUSTERING APPROACH

GUI YU XUAN

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will discuss the related issues and the previous studies that have been done. The limitations of RCTs were determined and how the drug reviews can enhance the findings of RCTs was discussed in this chapter. Lastly, the use of LLMs and clustering techniques in text analysis were illustrated in this chapter.

2.2 Drug Efficacy Evaluation in Randomized Controlled Trials (RCTs)

RCTs are important for allowing the medicine to be used in the real-world (Liakos et al., 2024). Besides that, RCTs also provide the framework and structure to describe the policy of the medicine and the way to consume the medicine. RCTs can avoid bias in findings and can generate reliable results. This is because selected volunteers are randomly separated into different groups. Therefore, the distribution of people with certain characteristics can be evenly distributed to test the drug performance. The research done stated that RCTs are complex and expensive due to the strict rules to conduct them. These rules and standards aligned with the study of RCTs are to ensure there is no harm to the patient when the drug is in used. Thus, when RCTs are conducted correctly, then healthcare professionals can determine the safety of drug. Indirectly, healthcare professionals can make clinical decisions.

Noninferiority RCT is a method to measure drug efficacy and safety (Kim et al., 2022). In the study carried out by researchers, noninferiority RCT will compare the performance of a new treatment with an existing treatment. According to the author, noninferiority RCT provides a guideline in evaluating drug efficacy through the estimation of the ability of the treatment in maintaining efficacy while have the improvement in safety and side effects. Normally, noninferiority RCT will serve as the

evidence that supports the introduction of new drug treatments. The trial also increases the available treatment options by allowing for more personalized treatment plans that meet patient needs. Thus, it helps in drug development by enhancing the therapy strategies and offers a better treatment plan.

A comprehensive analysis of the role of RCTs in new drug applications for orphan drugs had been carried out (Kubota & Narukawa, 2023). The study highlighted the relationship between RCTs and the severity of disease outcome, the type of drug usage and the nature of primary endpoints. Besides that, the authors indicated that RCTs are essential for generating high quality evidence of drug efficacy and ethical issues especially when there is no standard treatment exist. Furthermore, RCTs can be used to describe the cause-and-effect relationship between drugs and outcomes. Hence, the effectiveness of drug in specific diseases can be determined. Moreover, RCT data was required to support the effectiveness and safety claims for the drug approval process. Thus, RCT in this study showed the importance of RCTs in evaluating drug efficacy, guiding the regulatory process and driving clinical practice.

In general, RCTs involve selecting a group of patients or clients and randomly allocating everyone to a treatment group (Newell, 2020). The measurement and finding will then be collected after treatment and comparing the outcomes statistically. Today, there is little relationship between RCTs and policy such as cost-effectiveness and medical efficacy even though most RCTs focus on medical treatments. This is because the study population should be clearly defined with specific diagnostic criteria and limitations. According to Jiang et al. (2024), the differences between RCTs and real-world research were because of the variability of characteristics between RCT and real-world populations. Normally, the characteristics in the real-world populations are hard to measure directly. Hence, the variance restricts RCTs people from accurately representing the features in real-world scenarios. Therefore, to successfully generalize RCTs results, it is necessary to evaluate the difference between the variables observed in the RCTs sample and the variables observed in the real-world population.

Table 2.1: Summary the Study of Analyse RCTs

References	Experiment	Strength	Limitation
(Kim et al., 2022)	Compare the performance of a new treatment with an existing treatment by noninferiority RCTs	• Provides a guideline in evaluating drug efficacy • Easily to interpret when new treatment outperformed than existing treatment	• Produce unreliable conclusion when the assumptions of drug effect are different from existing • Patient can affect the measurement result
(Kubota & Narukawa, 2023)	Discuss the factors on the necessity of RCTs	• Provides the severity of disease • Describe of the usage type of drug and the clinical outcome	• Not Specified on the limitations of RCTs
(Newell, 2020)	Identify the relationship between RCTs and policy	• Provide with the standard of cost and medical efficacy	• Only select a group of patients
(Jiang et al., 2024)	Analyze on the generalization of RCTs to real world	• Not specified the strength of RCTs	• RCTs population can't fully capture the real-world scenarios.

2.3 Patient Review as A Real-World Data Source

Customers share their opinions about experienced drugs on internet review sites (Dinh et al., 2020). As a result, drugs reviews can be considered as statistical data that enable medical professionals in collecting medical data before making clinical decisions. This is because drug reviews that are commented on by patients provide insights on their experiences with medicine, including its efficacy and side effects. The Internet offered lots of information to enable the analysis on pattern recognition. Research can understand the pattern of the topics by analyzing the data on the Internet.

The study stated that the valuable insights from multimodal data can be visualized by using machine learning algorithms. Thus, healthcare professionals can utilize the tools to categorize drug reviews based on their effectiveness and side effects. Online platforms allow all people to share and comment on their experience. The involvement of different populations with different health status and demographic information in online drug reviews are important to gain insight of a drug performance across diverse populations.

There was a sentiment analysis that had been done to investigate the patient's experiences by studying patient review from an online medical platform. According to the authors, patient reviews help to gain the understanding of drug when used in patients with different diseases. (Cimino et al., 2024). The authors utilized natural language processing (NLP) and machine learning algorithms to analyze patient reviews. NLP was used to process and analyze text data. The patterns in reviews were identified and assigned text data according to their sentiment emotions. Then, the text data was classified by the support vector machines (SVM) and random forest to validate the performance of the model. In conclusion, the study highlighted drug reviews are important because they provide information that involved patients with different disease states. Sentiment analysis that categorizes patient experiences in positive, negative and neutral allow healthcare professionals to identify the drug efficacy effectively.

A study had been done to interpret the real-world data in the disease specific programs (DSPs) analysis. Real-world data allow the healthcare professionals to understand the disease management which help to make clinical decisions (Anderson et al., 2023). DSPs is a multi-perspective real-world data source that gathers the information from patients, caregivers and physicians into treatment patterns, patient reported outcomes and the patient experience. Real-world data allow the inclusion of diverse patient populations compared to traditional clinical trials. As mentioned before, real-world data captured the experiences and outcomes of patient which are important to analyze the effects of drugs. Besides that, the study also stated that the differences between groups of patients able to be identified with the real-world data

and the results from analyzing process will be further enhancing the patient outcomes and quality of life.

However, the researchers did indicate that consumers have difficulty going through all comments due to the unstructured text data (Dinh et al., 2020). Therefore, to ensure that the drug reviews that done by patients able to be understandable by others and assist medical professionals in improving the performance and effectiveness of drugs, some models and algorithms will be carried out to classify the text data into meaningful insights.

Table 2.2: Summary of The Study Done on Drug Review

References	Experiment	Strength	Limitation
(Dinh et al., 2020)	Utilize the text mining and supervised learning to discuss the online drug reviews	• Can considered as statistical data • Involvement of different populations in sharing their experiences	• Not specified the limitations
(Cimino et al., 2024)	Sentiment analysis on patient experiences	• Gain understanding of drug when used in different conditions	• NLP can't understand the meaning of words in different context
(Anderson et al., 2023)	Interpret the involvement of real-world data source with DSP	• Real-world data allow the understanding of disease management • Real-world data illustrated the involvement of people in clinical routine.	• DSPs do not represent the real population

2.4 Large Language Models (LLMs) in Text Analysis

With the growth of technology, there is an increasing number of textual datasets that have been available from digital sources. There is lots of information that

can be obtained from social media posts to online review platforms. However, analyzing the vast amounts of unstructured data to discover the underlying patterns is a complex task because unstructured data do not have a standardized format that enables analysis in a simple way. In addressing these issues, LLMs were recognized for their effectiveness in classification, summarization and generation task. LLMs are advanced deep learning models that are pre-trained on large amounts of textual data to capture the complex language patterns (Ampel et al., 2024). The pre-training allowed LLMs to perform well on a variety of downstream tasks. For your information, downstream task is a task that depends on previous output.

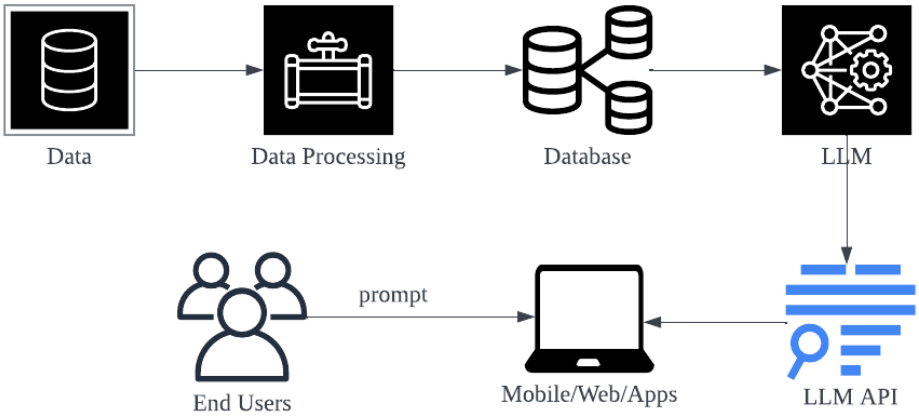

Figure 2.1: LLM Flow

According to Yao et al. (2024), LLM is a language model with a large number of parameters that has been pre-trained for tasks including self-supervised learning to produce and predict the text. The ability of LLMs that help in decision making and problem solving was due to the comprehensive understanding of natural language context, capability in producing human-like text, strong awareness of context and powerful problem-solving skills. ChatGPT, Gemini, Mixtral and Claude are examples of LLMs (Rangapur & Rangapur, 2024). The evaluation of accuracy, fluency and coherence of the generated responses by the LLMs model had been conducted by. By evaluating the potential of LLMs in performing the conversational question answering task, ChatGPT showed higher accuracy, relevance and consistency in generating the relevant response compared to others. Thus, the ability of ChatGPT in producing

relevant and accurate responses makes it became the first option to select as a conversational AI.

LLMs consists of large computing system that take textual data as an input into the Artificial Neural Network (ANN) to transform data into numerical format (Tai et al., 2024). The ANN will become more powerful and able to produce more reliable information when there is lots of data been inputted into LLMs. LLMs such as ChatGPT which developed to learn from human feedback and conduct conversations and solve mathematical problems. ChatGPT has been used in various field which included write clinical information about patients in medical field and summarize text from academic paper in academic field. From the experiment done by the authors in enhancing the coding in qualitative research, they found that LLMs can perform the checking on codes and providing additional knowledge that help authors to understand the steps. With extensive training data for LLMs to learn the pattern, the accuracy of code identification and interpretation can be further improved.

ChatGPT is a new chatbot that developed by OpenAI to answer questions on various topics (Belal et al., 2023). ChatGPT can write code, generate phrases and sentences and perform arithmetic solutions. According to the study that had be done to analyze the use of ChatGPT for data labeling, it showed that ChatGPT able to perform better and achieved 20 percent and 25 percent higher accuracy than other lexicon-based unsupervised methods in Tweets dataset and Amazon Reviews Dataset respectively. The advantages of ChatGPT include user-friendly interface, easily accessible to non-experts in interpreting text data and the adaptability to perform various tasks. However, the results that produced by ChatGPT were dependent on the prompt used in the analysis and had potential bias. The bias was due to the training of ChatGPT with vast amounts of data that available on the internet.

The research in analyzing the sentiment analysis ability of ChatGPT had been conducted (Wang et al., 2023). Sentiment analysis is used to learn the expression patterns in the text. The authors use ChatGPT for evaluation the language understanding ability is because of its performance and low cost. The experiment was started by giving the instruction for each task and evaluate the performance by

accuracy and F1 score. The findings illustrated that ChatGPT is highly competitive sentiment analysis performance and able to make a reliable prediction without labeled data for training. Meanwhile, a study on investigating the reliability and consistency of ChatGPT had been carried out (Reiss, 2023). This study was based on the ability of ChatGPT in classifying websites into News or not News. There are total of 234 websites that had been randomly selected and the website texts were obtained to transform into plain text. Krippendorff's Alpha was used to measure consistency by evaluating the output generated from the same input.

To ensure the consistency and reliability of the classification results, there are several scenarios that were introduced to ChatGPT. The scenarios included using various parameters such as temperature settings, changing the words in provided instruction and repeating the inputs multiple times. Even though there are advantages from ChatGPT, the experiment did conclude that ChatGPT is non-deterministic and inconsistent in outputs. This is due to the temperature settings that had been assigned to control the randomness of generated output (Reiss, 2023). Lowering temperature settings will reduce the randomness of generated text and produce a deterministic output. The study also demonstrated that pooling output by obtaining the important features from previous features map can improve the reliability of ChatGPT.

Table 2.3: Summarization of Advantage and Disadvantage of ChatGPT

Advantageous of ChatGPT	**Disadvantageous of ChatGPT**
Make a reliable prediction without labeled data for training (Wang et al., 2023)	Results that produced depend on the prompt (Belal et al., 2023)
Easily accessible to non-experts in interpreting text data (Belal et al., 2023)	Had potential bias due to pre-training data (Belal et al., 2023)
Adaptability to perform various tasks (Belal et al., 2023)	Non-deterministic and inconsistent in outputs (Reiss, 2023) due to temperature settings
Highly competitive sentiment analysis performance (Wang et al., 2023)	

2.5 Unsupervised Learning in Drug Reviews

Clustering, text mining and Latent Dirichlet Allocation (LDA) had been utilized in the study to discover the knowledge on breast cancer drugs by drug reviews (Nilashi et al., 2024). LDA was used in this study to obtain aspects of patient experiences. Then, clustering techniques were used to group the reviews based on their similarities in effectiveness. The author showcased that the effectiveness of each cluster will experience different side effects. The authors highlighted that the side effects of a drug should be concerned even the drug effectiveness able to perform under specific condition. Besides that, there was a study on applying clustering to the online patient medication reviews to discover the underlying knowledge of patient information and side effects (Yildirim & Kaya, 2019). The patient reviews were collected from medical websites. Then, users' demographic and the side effects of the drug in the patient reviews were identified. K-means clustering algorithm was applied to the cleaned dataset to group the patient into different groups based on their characteristics and experiences. According to the authors, the clustering results are able to tell the relationship between patient demographic and its side effects when using the drug.

There was a study done on applying the k-means clustering algorithm to develop drug recommender system (Posch & Tiwari, 2021). The authors derived the reviews that are related to ADHD disease from drug reviews dataset. Then, the data cleaning and text processing were applied to remove HTML issues and stop words that were presented in dataset. Patient reviews regarding to the ADHD disease had been clustered based on the experiences of patients when consuming drug. Then, the clustering results were used to predict the rating of the drug by neural networks. The authors concluded that predicting the rating of each drug based on the clustering results allows the selection of best drug to different patient groups.

A method that combine of unsupervised learning and supervised learning was carried out to investigate the impact of drug reviews in predicting medical preferences (Allenbrand, 2024). The raw data of reviews had been processed by text mining and bag-of-words to vectorized the textual data into numerical data type. Then, topic

modeling with LDA was conducted to identify the topics for each review that related to side effects and benefits. After that, clustering methods such as k-means, agglomerative and density-based spatial clustering (DBSCAN) were utilized to cluster the review based on identified topics by LDA. Lastly, the classification process was carried out to classify the clustering results by ratings. According to the author, clustering techniques and topic modeling were implemented to further enhance the performance of supervised learning in distinguish the pattern of reviews. The combination supervised and unsupervised learning allow the prediction of patient satisfaction with the drug consumed as the experiment showed that the benefits and side effects found from the feedback was essential to develop the personalized medicine.

The findings from previous studies illustrated the clustering in discovering the underlying insights of side effects and effectiveness of drug by analyzing drug reviews. Categorizing drug reviews based on the topics that related to side effects and effectiveness, discovering the relationship between demographic and side effects of drug and predicting the ratings based on the patient experiences clusters did demonstrated that the clustering approach had the benefit to understand the relationship between drug efficacy and patient satisfaction.

Table 2.4: Summary of the Study on Unsupervised Learning

References	Experiment	Strength	Limitations
(Nilashi et al., 2024)	Discover the knowledge on breast cancer drugs by drug reviews with Expectation Maximization (EM) and LDA	• LDA allow the identification of main aspects in reviews • EM allow the interpretation of side effects segments	• Still exist with irrelevant data in reviews • The discovery of side effects can be further improved
(Yildirim & Kaya, 2019)	Discover the hidden pattern of drug in patient reviews with k-means clustering	• Visualize the drug performance based on the characteristic of patients	• Number of clusters is hard to define

(Posch & Tiwari, 2021)	Drug recommender system for ADHD by k-means	• Identify the best drug to consume based on the clustering results	• Encountered with poor evaluation and runtime issues
(Allenbrand, 2024)	Investigate the impact of drug reviews in predicting medical preferences by LDA, clustering approaches	• Perform better in data preprocessing • Characterize the reviews based on the ratings	• Extraction is needed to view the reviews in a more comprehensive way • Noise should be reduced

2.5.1 Clustering Techniques in Text Analysis

Clustering is a technique that groups the unlabelled data into different class without training and the grouping process was conducted by measuring the similarity between the features (Oyewole & Thopil, 2023). The training of clustering is by analysing the patterns and relationship between features in the dataset. Identification of patterns, measurement of similarities, grouping of data and the outcomes were the process in clustering algorithms. The authors suggested that pattern representation was referred as feature selection where only the useful information that will be recognized. The similarity between two data had been computed in clustering process to group the data into different groups. Furthermore, according to authors, optimum number of clusters was important and made impact on the output of data. In general, Euclidean distance was the most used methods to obtain the similarity between two data while sum of squared error and Silhouette index are the methods that had been used to obtain the optimum number of clusters. Today, clustering techniques had been used in several field including manufacturing, energy and healthcare. Clustering techniques in healthcare field assisted in identifying the diseases, understanding the patterns of data and predicting health issues.

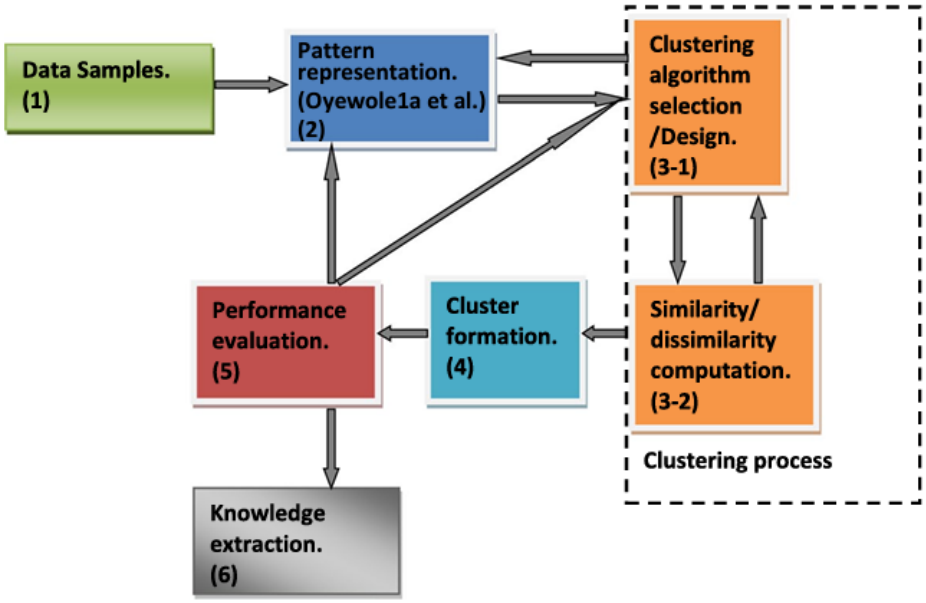

Figure 2.2: Clustering Steps (Oyewole & Thopil, 2023)

Clustering techniques was playing a crucial role in the data mining. This is because clustering techniques can discover the valuable information in the dataset (Hu et al., 2024). Clustering techniques partition data into different groups based on their characteristics. Clustering algorithms generate the clusters by understanding the relationship between data points. Clustering techniques can be interpreted in three ways, in-clustering, pre-clustering and post-clustering. Pre-clustering focused on the feature extraction and feature selection to ensure the capture of significant characteristics in the dataset. Meanwhile, in-clustering was illustrated the clusters with the selecting models that applied to the features. Lastly, post-clustering was the interpretation of the generated outcomes. The interpretation of in-clustering and post-clustering was based on the applied models which are decision tree, rules, prototype, convex polyhedral and description. Decision tree model demonstrated the derived process from dataset into clusters along the path; rules-based model generated rules based on the features; prototype model utilized prototype as the representative of each clusters and group the data points if closely to the prototype; convex polyhedral model defined the boundaries planes to capture the cluster group while description model represented the key features as a description and grouped the features based on the specific concept. The authors believed that interpreting clusters were important to

ensure the reliable and consistent result. Therefore, interpreting the generated clusters by understanding the context of models is crucial in the decision-making process.

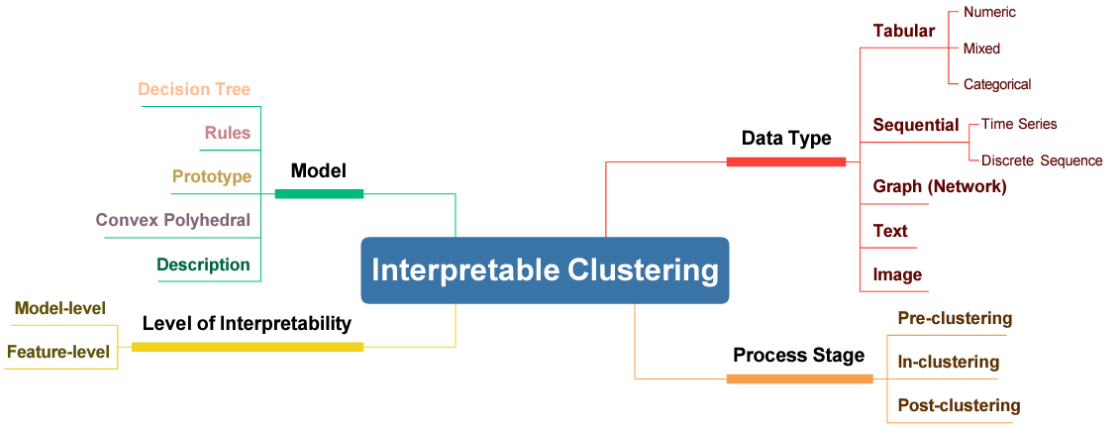

Figure 2.3: Criteria in Interpreting Clustering (Hu et al., 2024)

There is an experiment that utilized deep learning-based text clustering framework to analyse the accuracy and efficiency of text clustering. Clustering of text is a method the grouping text data based on similarity and extracting the important features from unstructured data by classifying the similar text data into same categories (Xu et al., 2024). According to the author, the process of text clustering is mapping texts into a feature vector space and employ the clustering techniques to categorize texts based on their similarity. At the beginning of experiment, several steps had been carried out in the data preprocessing such as tokenization, stop words removal and text normalization. Second, pre-train models (or called as LLMs) were implemented to understand the pattern of information in the text data. Third, deep embedded clustering based on autoencoders was used to extract the meaningful features and apply clustering algorithm to cluster the data. The result showed that with the deep learning-based text clustering framework, the accuracy and efficiency of text clustering can be further improved and a more reliable results can be generated. According to the author, clustering the patient reviews able to classify patient according to diseases or help in analysing drugs performance. Thus, the clustering results able to assist medical professionals in the diagnosis and develop a new drug.

Besides that, there is another research had been studied to improve the drug repositioning performance. Drug repositioning is the investigation of existing drugs

for new discovery strategy based on the analysing of clinical data (Lee et al., 2022). Authors highlighted that applying text mining approach in biomedicine field can analysed the large amounts of biomedical data effectively. Thus, the authors used the word2vec algorithm to generate embedded word vectors for the diseases and drugs to represent the relationship between diseases and drugs. Then, hierarchical clustering method had been applied to the word vectors to group the data based on their similarities. According to authors, the experiment successfully extracting the meaningful features from the dataset where there are 4,163 diseases and 3,930 drugs were extracted from 17,606,652 MEDLINE abstracts. Then, clustering techniques was grouping the extracted features into nine clusters. Therefore, the study that enabled the identification of potential drugs for discovery enhance drug selection process.

In conclusion, the ability of clustering techniques in discovering the underlying patterns of the text data and grouping the data based on their similarities enhancing the medical process. Therefore, the popular clustering algorithms were discussed and the suitable approach will be chosen for the analysis.

2.5.1.1 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density-based clustering techniques have the ability in capturing the arbitrary shape of clusters (Hahsler et al., 2019). Thus, the data points will be grouped by this ability. The authors also suggested that noisy data will be excluded and would not group together with other data points. According to the author, density-based clustering started by defining density of the dataset. There is no predefined number of clusters needed in density-based clustering techniques. This is because density-based clustering techniques captured the clusters by density. Therefore, unlike other clustering techniques that required the predefined parameters, density-based clustering techniques assigned the data points according to the density. The commonly used density-based clustering technique is DBSCAN. DBSCAN identified all data points as core points, border points or noise data and clustered the core points by measuring the density. For your information, the algorithm started with assigning random data points as the central and defining the data points that were closer to the central point. The

algorithm stops when there are no more data points linked as the density reachable points and a cluster will be formed.

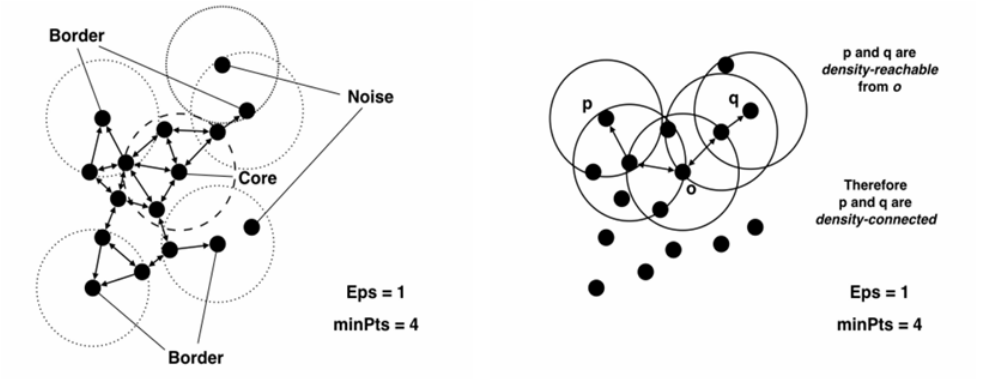

Figure 2.4: Concept of DBSCAN (Hahsler et al., 2019)

The figure 2.5 was further explained as below. There are two important parameters in DBSCAN, Ɛ (radius of neighbourhood) and minPts (minimum number of points to form a cluster) (Hahsler et al., 2019). Let considered there is a dataset to be clustered. The Ɛ-neighbourhood will be assigned with the value to identify to the data points within the radius of the assigned central point. The data points will be classified into core points, density reachable points and outliers. If data point had a distance with the minimum value of minPts will be considered as core point. Meanwhile, density reachable point referred as a data point that was reachable to the core point and is with the assigned radius. Lastly, the data point that does not meet the conditions of core points and density reachable points was clustered as outliers.

It was defined as:

$$N\epsilon(p) = [q \in D \mid d(p,q) < \epsilon] \tag{2.1}$$

Where:

$N\epsilon(p)$: set of points within the radius

$d(p,q)$: measurement of distance

D: dataset

The DBSCAN had the advantage in identifying clusters by effectively removing noise and outliers, do not require prior knowledge of the number of clusters and able to identify the clusters in various shapes and sizes (Bushra & Yi, 2021; Hahsler et al., 2019). However, the performance of DBSCAN depended on the parameters which can lead to misleading results when not specified the parameters correctly and the computational cost was high for distance measurement (Bhardwaj et al., 2022; Bushra & Yi, 2021; Ji & Wang, 2021). Therefore, a few steps on the selection of parameters should be considered to improve the clustering results and optimize the performance of DBSCAN.

2.5.1.2 Agglomerative Hierarchical

Agglomerative Hierarchical clustering is an unsupervised technique that build a binary merge tree that started to store the data into leaves and merge the two closest sets until reach the root of tree (Nielsen & Nielsen, 2016). Hierarchical clustering approach was introduced to have a large number of partitions and each partitions had its own dendrogram. (Murtagh & Contreras, 2017). Dendrogram is the graphical representation of the tree. The agglomerative hierarchical algorithm started by assigning each of the data points as a cluster. Then, for each iterative, the distance between two clusters was calculated and merged the closest pair of clusters to one cluster until single cluster was left. There are three strategies to define the good linkage distance which are single linkage, complete linkage and average linkage. Single linkage calculated the minimum distance between two data points, complete linkage calculated the maximum distance between two data points while average linkage calculate the average distance between all data points in two clusters.

Single linkage defined as:

$$L(R,S) = \min\big(D(i,j)\big), i \in D, j \in D] \quad (2.2)$$

Where:

$L(R,S)$: linkage between two cluster

$\min(D)$: minimum distance between data

$D(i,j)$: distance between two data points

Complete linkage defined as:

$$L(R,S) = \max\big(D(i,j)\big), i \in D, j \in D] \tag{2.3}$$

Where:

$\max(D)$: maximum distance between data

Average linkage defined as:

$$L(R,S) = \frac{1}{n_R \times n_S} \sum_{i=1,j=1}^{n_R,n_S} D(i,j), i \in R, j \in S \tag{2.4}$$

Where:

$\sum_{i=1,j=1}^{n_R,n_S} D(i,j)$: sum distance of clusters

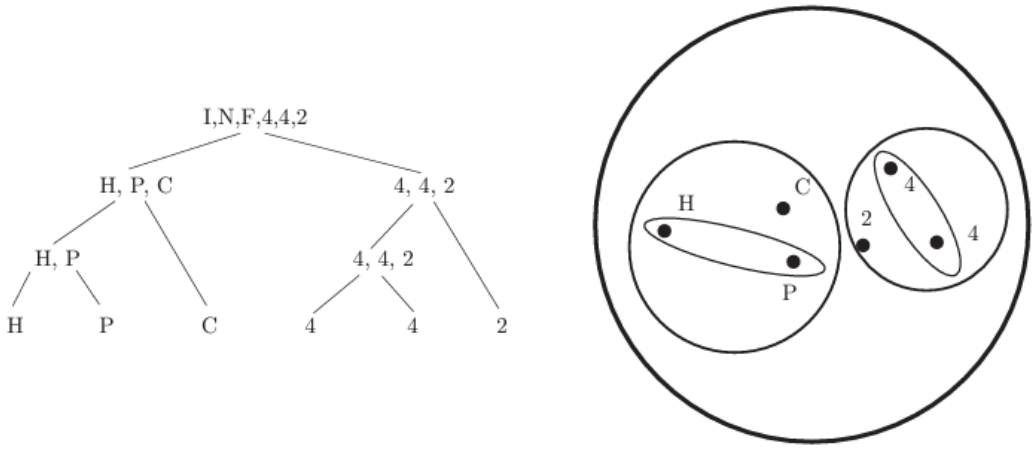

Figure 2.5: Dendrogram (left) and Venn Diagram (right) for Visualization (Nielsen & Nielsen, 2016)

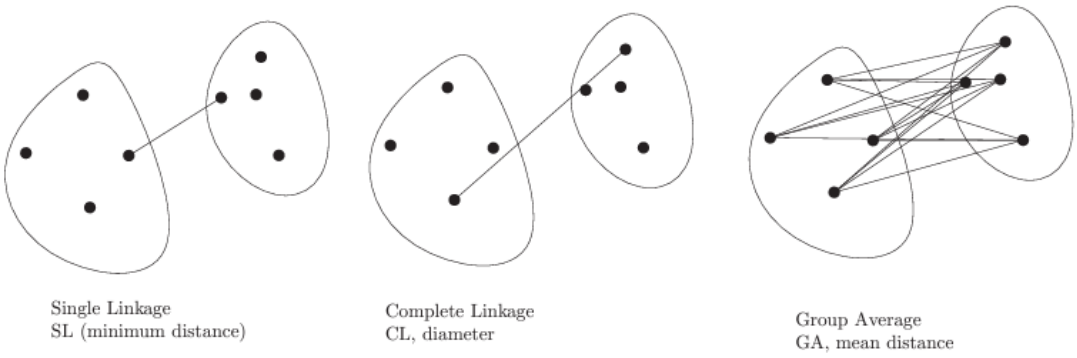

Figure 2.6: Linkage Strategies to Define Distance (Nielsen & Nielsen, 2016)

Agglomerative hierarchical clustering techniques offered several advantages than other clustering algorithms. First, the dendrogram provided graphical representation of the hierarchical structure of data allow the understanding of relationships between clusters (Oti & Olusola, 2024). This is because the graphical allow the researchers to gain the insights into data at various levels. Besides that, agglomerative hierarchical clustering also robust to noise and outliers (Benatti & Costa, 2024). This characteristic allowed the agglomerative hierarchical clustering to perform in high dimensional dataset. Lastly, as agglomerative hierarchical clustering did not require a predefined number of clusters, thus it was flexible in clustering data as the clusters were formed naturally without controlling. However, this situation did rise the issues in identifying the clusters with different densities. Besides that, agglomerative hierarchical clustering was computational complexity because the distance between all data points needed to be calculated.

2.5.1.3 K Means

K Means is partitional clustering algorithm that partitions dataset into smaller groups based on the distance between the centroid point (Ikotun et al., 2023). With the increasing number of clusters, K means algorithm able to achieve the decreasing in the square error. The minimum squared error between data points and the mean of the cluster will be found and assigned the data points to the nearest cluster. The step in K Means algorithm began by randomly selecting a few centroids from dataset. Then, the distance of data points with centroids will be calculated and assigned the data points to the nearest centroids. Lastly, the new centroid value was calculated for the next

iteration. There were three parameters that should be considered in the K means algorithm which are the number of clusters to be formed, the centroid points and the distance metric to be used in the experiment. This is because the performance of the clustering depends on the number of clusters while different initial centroids can produce different resulted clusters.

K Means algorithm defined as:

$$D(C_k) = \sum |x_i - \mu_k|^2 \tag{2.5}$$

Where:

C_k: data points of Cluster k

$\sum |x_i - \mu_k|^2$: distance of data points and centroids

According to Chong (2021), K means clustering was straightforward algorithm that enabled non expert users to partition dataset into the desired number of clusters. The implementation and interpretation of K means approach was easy and widely used for clustering tasks (Liu, 2022; Pratama et al., 2023). Furthermore, the scalability and flexibility of K means algorithms enabled it to perform well in large dataset and work with various type of data such as numerical data and categorical data. However, K means algorithm was sensitive and needed to be carried out carefully at the initial stage. This is because the performance of the K means algorithm was determined by the number of generated cluster, the initial centroids and the outliers or noisy data that presented in the dataset.

Table 2.5: Summarizing the Performance of Clustering Approaches

Clustering Approaches	**Advantageous**	**Disadvantageous**
DBSCAN	• Insensitive to noisy data • No predefined number of clusters is required • Identify the clusters in various shapes and sizes	• Performance depends on the parameters • High computational cost

Agglomerative Hierarchical	• Graphical representation • Robust to noise and outliers • No predefined number of clusters is required	• Issues in identifying the clusters with different densities • Computational complexity
K Means	• Easy implement • Scalable and Flexible	• Performance depends on the initial parameters • Sensitive to outliers and noisy data

2.6 Discussion

RCTs were conducted with a controlled environment by selecting volunteers with specific demographics limiting the involvement of large populations. Even though it can evaluate the drug efficacy without bias but exist with the drug performance gap in real-world scenarios. Patient reviews offer valuable information that captures the experience of patients when consuming the drug. The side effects and effectiveness of the drug when applied to different groups of people can be derived from the patient review. However, bias can occur in the review based on the patient's preferences. To address this issue, text analysis methods such as LLMs were used to retrieve reliable insights from patient reviews.

LLMs such as ChatGPT provide with the techniques that able to understand the sentiment of the words, phrases and sentences. The ability of LLM in analyzing the variations of text data allows the identification of important features in the reviews. Besides that, LLM can perform better than traditional text processing such as LDA due to the ability of LLM in understanding the context and meaning of the reviews. However, LLMs have limitations in terms of identifying the similarities between words. Indirectly, it limits the grouping of similar side effects and effectiveness. Therefore, clustering techniques had been carried out to enhance this study. The grouping of similar features from reviews into different groups enhanced the overall findings of the results.

In conclusion, the combination of LLMs and clustering techniques in this research allows the evaluation of drug efficacy in a more comprehensive way. LLMs with the ability to recognize and analyze the relationship between the characters, words and sentences help in retrieving the meaningful insights of drug review. Meanwhile, clustering techniques with the ability to group similar data points reduce the complexity of the dataset and enable the identification of patterns effectively.

2.7 Summary

As discussed before, RCTs still had limitations in considering diverse patient population in analyzing the drug performance. Thus, drug reviews generated by patients provide valuable information into the patient experience and side effects that was more useful than RCTs. Most of the previous studies implement the traditional text processing such as LDA to identify the relevant features in drug review. However, with the advancement in technology, LLMs can provide a more comprehensive understanding of the drug reviews pattern. Therefore, instead of using traditional text processing, the experiment started with applying LLM, ChatGPT model to retrieve the keyword from the drug reviews. Then, DBSCAN was chosen as the clustering technique to group the keywords due to its ability in handling vary shapes and densities of clusters, able to remove outliers effectively and no predefined number of clusters was required.

REFERENCES

Allenbrand, C. (2024). Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews. *Healthcare Analytics*, *5*, 100288.

Ampel, B., Yang, C.-H., Hu, J., & Chen, H. (2024). Large language models for conducting advanced text Analytics Information Systems Research. *ACM Transactions on Management Information Systems*.

Anderson, P., Higgins, V., Courcy, J. d., Doslikova, K., Davis, V. A., Karavali, M., & Piercy, J. (2023). Real-world evidence generation from patients, their caregivers and physicians supporting clinical, regulatory and guideline decisions: an update on Disease Specific Programmes. *Current Medical Research and Opinion*, *39*(12), 1707-1715.

Belal, M., She, J., & Wong, S. (2023). Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv preprint arXiv:2306.17177*.

Benatti, A., & Costa, L. d. F. (2024). Agglomerative clustering in uniform and proportional feature spaces. *arXiv preprint arXiv:2407.08604*.

Bhardwaj, A., Pandey, A., & Dahiya, S. (2022). Review based on Variations of DBSCAN algorithms. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS),

Biswas, B., Sengupta, P., & Ganguly, B. (2022). Your reviews or mine? Exploring the determinants of "perceived helpfulness" of online reviews: a cross-cultural study. *Electronic Markets*, *32*(3), 1083-1102.

Bushra, A. A., & Yi, G. (2021). Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. *IEEE Access*, *9*, 87918-87935.

Chong, B. (2021). K-means clustering algorithm: a brief review. *vol*, *4*, 37-40.

Cimino, A., Culbertson, C., Watkins, E., Li, J., & Wangeshi, S. (2024). RWD119 A Methodological Approach Using Sentiment Analysis of Online Medical Platforms As a Real-World Data Source of Patient Experiences. *Value in Health*, *27*(6), S381.

Dinh, T., Chakraborty, G., & McGaugh, M. (2020). Exploring Online Drug Reviews using Text Analytics, Sentiment Analysis and Data Mining Models. SAS 2020 Global Forum,

Gruber, J. B., & Votta, F. (2024). Large Language Models.

Gui, C., Han, D., Gao, L., Zhao, Y., Wang, L., Xu, X., & Xu, Y. (2024). Application of Enhanced K-Means and Cloud Model for Structural Health Monitoring on Double-Layer Truss Arch Bridges. *Infrastructures*, *9*(9), 161.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*, 1-30.

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, *125*(13), 1716.

Hu, L., Jiang, M., Dong, J., Liu, X., & He, Z. (2024). Interpretable Clustering: A Survey. *arXiv preprint arXiv:2409.00743*.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178-210.

Ji, Z., & Wang, C.-L. (2021). Accelerating DBSCAN algorithm with AI chips for large datasets. Proceedings of the 50th International Conference on Parallel Processing,

Jiang, K., Lai, X.-x., Yang, S., Gao, Y., & Zhou, X.-H. (2024). A Practical Analysis Procedure on Generalizing Comparative Effectiveness in the Randomized Clinical Trial to the Real-world Trialeligible Population. *arXiv preprint arXiv:2406.04107*.

Kaul, P., Bose, B., Kumar, R., Ilahi, I., & Garg, P. K. (2021). The strength of a randomized controlled trial lies in its design—randomization. *Supportive Care in Cancer*, 1-3.

Kim, K. S., Chan, A.-W., Belley-Côté, E. P., & Drucker, A. M. (2022). Noninferiority Randomized Controlled Trials. *Journal of Investigative Dermatology*, *142*(7), 1773-1777.

Kostis, J. B., & Dobrzynski, J. M. (2020). Limitations of randomized clinical trials. *The American journal of cardiology*, *129*, 109-115.

Kubota, Y., & Narukawa, M. (2023). Randomized controlled trial data for successful new drug application for rare diseases in the United States. *Orphanet Journal of Rare Diseases*, *18*(1), 89.

Kumar, A., & Shekhar, S. (2024). Hybrid model of unsupervised and supervised learning for multiclass sentiment analysis based on users' reviews on healthcare web forums. *J. Artif. Intell*, *7*(4).

Lee, D.-g., Kim, M., & Shin, H. (2022). Drug Repositioning with Disease-Drug Clusters from Word Representations. 2022 IEEE International Conference on Big Data and Smart Computing (BigComp),

Liakos, A., Pagkalidou, E., Karagiannis, T., Malandris, K., Avgerinos, I., Gigi, E., Bekiari, E., Haidich, A.-B., & Tsapas, A. (2024). A Simple Guide to Randomized Controlled Trials. *The International Journal of Lower Extremity Wounds*, 15347346241236385.

Liu, J., Zhou, Y., Jiang, X., & Zhang, W. (2020). Consumers' satisfaction factors mining and sentiment analysis of B2C online pharmacy reviews. *BMC medical informatics and decision making*, *20*, 1-13.

Liu, R. (2022). Data Analysis of Educational Evaluation Using K-Means Clustering Method. *Computational Intelligence and Neuroscience*, *2022*(1), 3762431.

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6), e1219.

Newell, D. J. (2020). Randomised controlled trials in health care research. In *Researching Health Care* (pp. 47-61). Routledge.

Nielsen, F., & Nielsen, F. (2016). Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 195-211.

Nilashi, M., Ahmadi, H., Abumalloh, R. A., Alrizq, M., Alghamdi, A., & Alyami, S. (2024). Knowledge discovery of patients reviews on breast cancer drugs: Segmentation of side effects using machine learning techniques. *Heliyon*, *10*(19).

Oti, E. U., & Olusola, M. O. (2024). OVERVIEW OF AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS. *Technology*, *7*(2), 14-23.

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, *56*(7), 6439-6475.

Posch, A., & Tiwari, P. (2021). Persona-based drug recommender system using online reviews. In.

Pratama, M. A. Y., Hidayah, A. R., & Avini, T. (2023). Clustering K-Means untuk Analisis Pola Persebaran Bencana Alam di Indonesia. *Jurnal Informatika Dan Tekonologi Komputer (JITEK)*, *3*(2), 108-114.

Qiu, K., & Zhang, L. (2024). How online reviews affect purchase intention: A meta-analysis across contextual and cultural factors. *Data and Information Management*, *8*(2), 100058.

Rangapur, A., & Rangapur, A. (2024). The Battle of LLMs: A Comparative Study in Conversational QA Tasks. *arXiv preprint arXiv:2405.18344*.

Reiss, M. V. (2023). Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Řezanková, H. (2018). Different approaches to the silhouette coefficient calculation in cluster evaluation. 21st international scientific conference AMSE applications of mathematics and statistics in economics,

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. 2020 IEEE 7th international conference on data science and advanced analytics (DSAA),

Sridharan, K., & Sivaramakrishnan, G. (2024). Unlocking the potential of advanced large language models in medication review and reconciliation: a proof-of-concept investigation. *Exploratory Research in Clinical and Social Pharmacy*, *15*, 100492.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, *23*, 16094069241231168.

Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

Xu, Q., Gu, H., & Ji, S. (2024). Text clustering based on pre-trained models and autoencoders. *Frontiers in Computational Neuroscience*, *17*, 1334436.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Yildirim, P., & Kaya, A. (2019). Clustering of Phentermine HCL Drug from Online Patient Medication Reviews. *Procedia Computer Science*, *151*, 1146-1151.

Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, solitons & fractals*, *140*, 110121.