

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 details of the related research. The limitations of RCTs in evaluating drug efficacy will be determined and the benefit of drug reviews as a data source for drug performance will be further demonstrated. Besides that, the use of LLMs and clustering techniques in text analysis will be outlined.

#### **2.2 Drug Efficacy Evaluation in Randomized Controlled Trials (RCTs)**

RCTs are important for medication regulatory approval as well as the development of medical knowledge and policy (Liakos et al., 2024). They are considered the gold standard in clinical research because of the ability to remove bias and distract factors that will cause unreliable results (Liakos et al., 2024). In order to make a reliable results, the participants are randomly allocated to various research groups to ensure that possible factors are uniformly distributed throughout the groups (Liakos et al., 2024). According to the research done by Liakos et al. (2024), RCTs are complex and expensive to conduct because they are strictly controlled by a regulatory standards. However, well designed RCTs are important to help medical professionals to make clinical decision making. This is due to the ability of RCTs that enable medical professionals to derive the cause and effect about drug efficacy. Additionally, the rules and standards that managing RCTs are essential for obtaining drug approval (Liakos et al., 2024).

Noninferiority RCT is conducted to evaluate the efficacy and safety of drugs (Kim, Chan, Belley-Côté, & Drucker, 2022). In the study carried by researchers, noninferiority RCT will compare the performance of a new treatment with an existing treatment (Kim et al., 2022). According to Kim et al. (2022), noninferiority RCT

provide 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 (Kim et al., 2022). Normally, noninferiority RCT will serve as the evidence that support the introduction of new drug treatments (Kim et al., 2022). The trial also increase the available treatment options by allowing for a more personalized treatment plans that meet patient needs (Kim et al., 2022). 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 are no standard treatment exist (Kubota & Narukawa, 2023). Furthermore, RCTs can be used to describe the cause-and-effect relationship between drugs and outcomes (Kubota & Narukawa, 2023). 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 (Kubota & Narukawa, 2023). 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 (Newell, 2020). Today, there is little relationship between RCTs and policy such as cost-effectiveness and medical efficacy even though most RCTs focus on medical treatments (Newell, 2020). This is because the study population should be clearly defined with specific diagnostic criteria and limitations (Newell, 2020). According to Jiang, Lai, Yang, Gao, and Zhou (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 to accurately represent the features in real-world scenarios (Jiang et al., 2024). Therefore, in order 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 (Jiang et al., 2024).

### **2.2.1 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 commented by patients provide insights on their experiences with medicine, including its efficacy and side effects (Dinh et al., 2020). With the rise in online reviews, large amounts of data on the Internet allows comprehensive evaluations and pattern identification (Dinh et al., 2020). The study by Dinh et al. (2020) stated that advanced analytics techniques can potentially extract valuable insights from unstructured data. Thus, reviews based on efficacy and side effects can effectively categorized and used to guide medical professionals. Overall, real-time drugs reviews including different patient groups with a wide range of health issues and characteristics are helpful in determining specific effectiveness and side effects in separate patient populations.

A sentiment analysis had been done on an online medical platform to investigate the patient review as the real-world data source in understanding patient experiences. According to the authors, researchers can gain a deep understanding of various disease states in the effectiveness of treatments by analyzing patient reviews (Cimino, Culbertson, Watkins, Li, & Wangeshi, 2024). The approaches used by authors are natural language processing (NLP) and machine learning algorithm to analyze patient experience. NLP was used to process and analyze the text data in understanding the expression in patient feedback and assign the sentiment scoring to quantify the text data (Cimino et al., 2024). Then, machine learning algorithms such as support vector machines (SVM) and random forest had been applied to classify the text data (Cimino et al., 2024). In conclusion, the study highlighted the importance of real-world drug reviews data in gaining insights into disease states and patient preferences that are not covered in RCTs. The sentiment analysis of drug reviews

provide an effective methodology for identifying patient experiences and treatments which will contribute in enhancing the patient care (Cimino et al., 2024).

According to the study on the analysis of disease specific programs (DSPs), real-world data is important in understanding disease management and informing medical professionals in healthcare decisions (Anderson et al., 2023). DSPs is a multi-perspective real-world data source that gathering the information from patients, caregivers and physicians into treatment patterns, patient reported outcomes and the patient experience (Anderson et al., 2023). Real-world data allow the inclusion of diverse patient populations compared to traditional clinical trials (Anderson et al., 2023). As mentioned by Anderson et al. (2023), 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 group 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 (Anderson et al., 2023).

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.

### **2.3 Sentiment Analysis of Drug Reviews**

Sentiment analysis of drug reviews by deep learning method had been carried out (Al-Hadhrami, Vinko, Al-Hadhrami, Saeed, & Qasem, 2024). Sentiment analysis is a subfield of NLP that identify and extract the meaningful information from text data (Al-Hadhrami et al., 2024). The experiment started with the data collection that consists of patient experiences, satisfaction levels and the side effects of the consumed drugs. After preprocessing and word embedding process, two deep learning methods, bidirectional long short-term memory (Bi-LSTM) and hybrid model (Bi-LSTM-CNN)

had been developed to enhance the understanding of reviews data. From the findings, the proposed method Bi-LSTM-CNN can achieve a high accuracy and F1 scores at 97 percent for negative sentiments, 91 percent for neutral sentiments and 87 percent for positive sentiments (Al-Hadhrami et al., 2024). The high accuracy and F1 scores indicated that deep learning methods have a good performance in classifying sentiments accurately. Thus, sentiment analysis of drug reviews can be act as a method for analyzing the effectiveness of drugs to make a better treatment decision.

The findings of the experiment from aspect-based sentiment analysis of drug reviews showed that sentiment analysis can effectively extracted the overall patient reviews with high accuracy (Gräßer, Kallumadi, Malberg, & Zaunseder, 2018). Logistic regression model was applied to the extracted features and the results reached 92.24 percent accuracy and 83.99 percent Cohen's Kappa score (Gräßer et al., 2018). Cohen's Kappa score is a statistical measure that used to evaluate the level of agreement between two raters (Gräßer et al., 2018). The study showed that by using sentiment analysis to extract the meaningful features from drug reviews that comment by patients and accurately classified the sentiments that related to drug efficacy have the potential to gain insights into real-world drug efficacy that is beyond the RCTs results.

According to Bu, Liu, and Ju (2024), sentiment analysis is used to discover the hidden pattern in unstructured data such as review data to analyze the expression of the users for a specific events. Sentiment analysis can be divided into four levels which are document, sentence, phrase and aspect levels (Alqaryouti, Siyam, Abdel Monem, & Shaalan, 2024; Bu et al., 2024; Jim et al., 2024). Sentiment analysis at document level and sentence level was treating document as whole data or basic information unit while sentences were considered as a short document and classifying the sentiments into positive, negative and neutral expressions (Al-Hadhrami et al., 2024). Meanwhile, phrase level sentiment analysis is the evaluation the emotion of the certain phrases (Jim et al., 2024). Phrase level sentiment analysis is more useful than document and sentence level sentiment analysis in the terms of expression variation within the sentence (Jim et al., 2024). The well-known sentiment analysis tasks is Aspect Based Sentiment Analysis (ABSA) which used to extract the aspects and find the viewpoints

to support its sentiment polarity (Bu et al., 2024). In this study, aspect-based sentiment analysis will be utilized to extract the important reviews data for pattern recognition. This is because aspect-based sentiment analysis or sometimes will be said as feature extraction that identified the reviews' characteristics through the comments to identify the aspects (Alqaryouti et al., 2024). Aspects are referred as attributes, characteristics or features of something (Alqaryouti et al., 2024). Thus, by analyzing the drug reviews by aspect-based sentiment analysis, the patterns in patient reviews of the drugs such as effectiveness, side effects and patient needs can be identified.

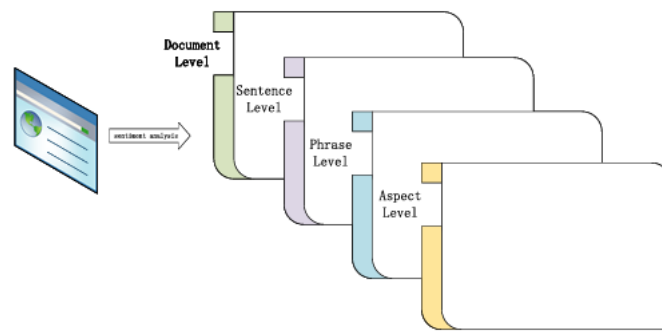


Figure 2.1: Levels of Sentiment Analysis (Bu et al., 2024)

### 2.3.1 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, Yang, Hu, & Chen, 2024). The pre-training allowed LLMs to perform well on a variety of downstream tasks (Ampel et al., 2024). For your information, downstream task is a task that depends on previous output.

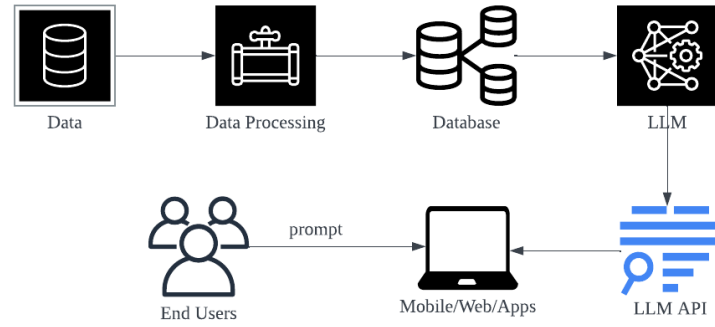


Figure 2.2: LLM Flow

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 (Yao et al., 2024). 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 Rangapur and Rangapur (2024). By evaluating the potential of LLMs in performing the conversational question answering task, ChatGPT showed the higher accuracy, relevance and consistency in generating the relevant response compared to others (Rangapur & Rangapur, 2024). Thus, the ability of ChatGPT in producing a 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 became more powerful and able to produce more reliable information when there are lots of data been inputted into LLMs (Tai et al., 2024). LLMs such as ChatGPT which developed to learn from human feedback and conduct the conversations and solve mathematical problems (Tai et al., 2024). ChatGPT had been used in various field which included write clinical information about patients in medical field and summarize text from academic paper in academic field (Tai et al., 2024). 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 (Tai et al., 2024). With extensive training data for LLMs to learn the pattern, the accuracy of code identification and interpretation able to be further improved (Tai et al., 2024).

ChatGPT is a new chatbot that developed by OpenAI to answer questions on various topics (Belal, She, & Wong, 2023). ChatGPT has the ability to write code, generate the phrases and sentences and perform arithmetic solutions (Belal et al., 2023). According to the study done by Belal et al. (2023) 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 including user-friendly interface, easily accessible to non-experts in interpreting text data and the adaptability to perform various tasks (Belal et al., 2023). However, the results that produced by ChatGPT were dependent on the prompt used in the analysis and had potential bias (Belal et al., 2023). The bias was due to the training of ChatGPT with vast amounts of data that available on the internet (Belal et al., 2023).

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 using ChatGPT for evaluation the language understanding ability is because of its performance and low cost (Wang et al., 2023). 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 the consistency by evaluating the output generated from the same input (Reiss, 2023).

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 (Reiss, 2023). The study also demonstrated that pooling output by obtaining the important features from previous features map can improve the reliability of ChatGPT.

Table 2.1: Summarization of Advantage and Disadvantage of ChatGPT

Advantageous of ChatGPT	Disadvantageous of ChatGPT
User-friendly interface (Belal 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)	
Make a reliable prediction without labeled data for training (Wang et al., 2023)	

### 2.3.2 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 (Oyewole & Thopil, 2023). Identification of patterns, measurement of similarities, grouping of data and the outcomes were the process in clustering algorithms (Oyewole & Thopil, 2023). The authors suggested that pattern representation was referred as feature selection where only the useful information that will be recognized (Oyewole & Thopil, 2023). The similarity between two data had been computed in clustering process to group the data into different groups (Oyewole & Thopil, 2023). Furthermore, according to authors, optimum number of clusters was important and made impact on the output of data (Oyewole & Thopil, 2023). 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 (Oyewole & Thopil, 2023). 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 (Oyewole & Thopil, 2023).

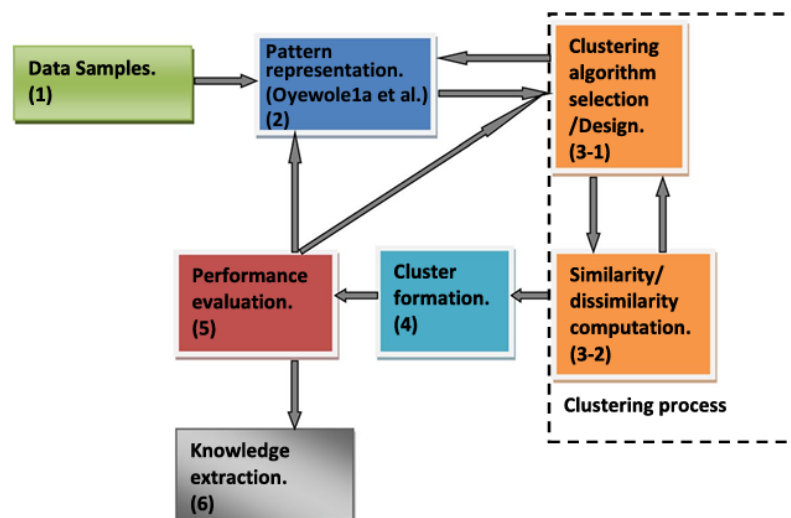


Figure 2.3: Clustering Steps (Oyewole & Thopil, 2023)

To understand the relationship among data points and discover the valuable insights in the dataset, cluster analysis was playing an crucial role in the data mining (Hu, Jiang, Dong, Liu, & He, 2024). This is because clustering techniques had the ability in partitioning data into different groups based on their characteristics (Hu et al., 2024) which facilitate the decision making process. Clustering algorithms considered all the features in the dataset and utilized the learning process to generate the clustering results (Hu et al., 2024). Clustering techniques can be interpreted in three ways, in-clustering, pre-clustering and post-clustering (Hu et al., 2024). Pre-clustering focused on the feature extraction and feature selection to ensure the capture of significant characteristics in the dataset (Hu et al., 2024). Meanwhile, in-clustering was illustrated the clusters with the selecting models that applied to the features (Hu et al., 2024). Lastly, post-clustering was the interpretation of the generated outcomes (Hu et al., 2024). The interpretation of in-clustering and post-clustering was based on the applied models which are decision tree, rules, prototype, convex polyhedral and description (Hu et al., 2024). 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 (Hu et al., 2024). The authors believed that interpreting clusters were important to ensure the reliable and consistent result (Hu et al., 2024). Therefore, interpreting the generated clusters by understanding the context of models is crucial in the decision-making process.

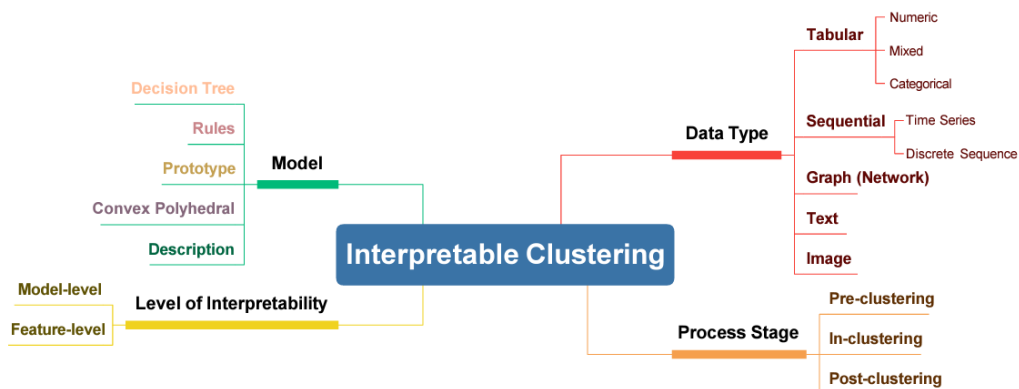


Figure 2.4: Criteria in Interpreting Clustering (Hu et al., 2024)

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, Gu, & Ji, 2024). According to Xu et al. (2024), 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 diseases or help in analysing drugs performance (Xu et al., 2024). Thus, the clustering results able to assist medical professionals in the diagnosis and develop a new drug (Xu et al., 2024).

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, Kim, & Shin, 2022). Authors highlighted that applying text mining approach in biomedicine field can analysed the large amounts of biomedical data effectively (Lee et al., 2022). 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 (Lee et al., 2022).

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 further analysed to select the best approach to be used in this thesis.

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

Density-based clustering techniques grouped the data points with their the ability in capturing the arbitrary shape of clusters (Hahsler, Piekenbrock, & Doran, 2019). The authors suggested that density-based clustering will exclude the noise data from the clusters (Hahsler et al., 2019). According to Hahsler et al. (2019), density-based clustering started by defining the structures for underlying density from the dataset. Unlike the others clustering methods that using parametric distribution or variance in grouping the data points, density-based clustering approaches had the capability to find the arbitrary shape clusters, handle noisy data and no need to specify the optimum number of clusters (Hahsler et al., 2019). The most popular density-based clustering technique is DBSCAN. The workflow of DBSCAN was identified a core points based on their density within a defined neighborhood (Hahsler et al., 2019). 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 (Hahsler et al., 2019). The algorithm stopped when there are no more data points can be linked as the density reachable points and a cluster will be formed (Hahsler et al., 2019).

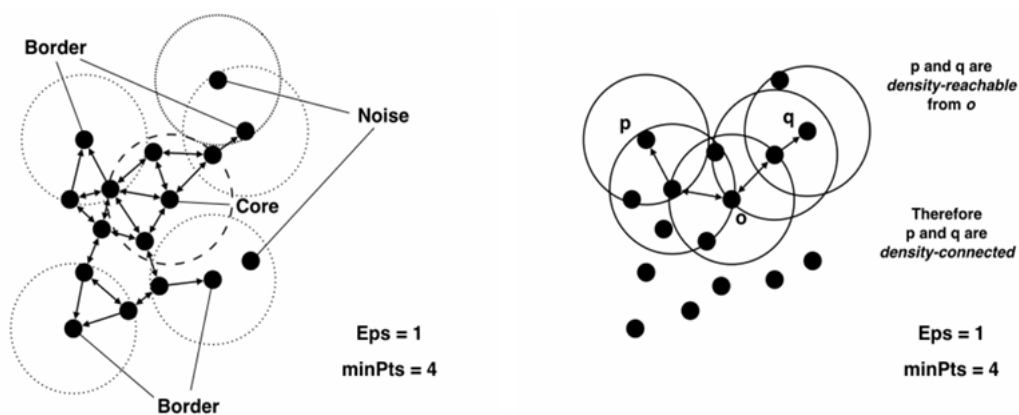


Figure 2.5: Concept of DBSCAN (Hahsler et al., 2019)

The figure 2.5 was further explained as below. There are two important parameters in DBSCAN,  $\epsilon$  (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  $\epsilon$ -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:

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

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, Pandey, & Dahiya, 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.

### 2.3.2.2 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 (Nielsen & Nielsen, 2016). 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 (Nielsen & Nielsen, 2016). 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 (Nielsen & Nielsen, 2016).

Single linkage defined as:

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

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:

$$L(R, S) = \max(D(i, j)), i \in R, j \in S \quad (2.3)$$

Where:

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

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

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



Average linkage defined as:

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

Where:

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

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

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

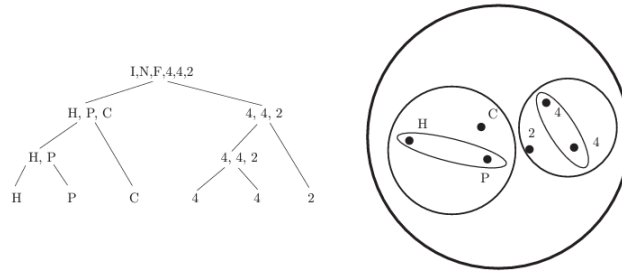


Figure 2.6: Dendrogram (left) and Venn Diagram (right) for Visualization (Nielsen & Nielsen, 2016)

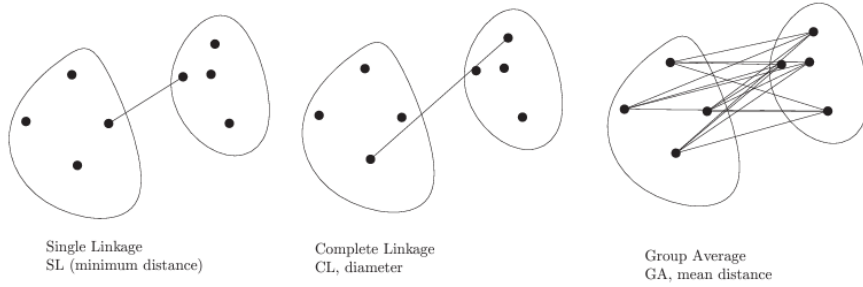


Figure 2.7: Linkage Strategies to Define Distance (Nielsen & Nielsen, 2016)

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 (Oti & Olusola, 2024). 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.

### 2.3.2.3 K Means

K Means is partitional clustering algorithm that partitions dataset into smaller groups based on the distance between the centroid point (Ikotun, Ezugwu, Abualigah, Abuhaija, & Heming, 2023). With the increasing number of clusters, K means algorithm able to achieve the decreasing in the square error (Ikotun et al., 2023). 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 (Ikotun et al., 2023). 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 should be considered in 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 (Ikotun et al., 2023). This is because the performance of the clustering depend on the number of clusters while different initial centroids can produce different resulted clusters (Ikotun et al., 2023).

K Means algorithm defined as:

$$D(C_k) = \sum |x_i - \mu_k|^2 \quad (2.5)$$

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 (R. Liu, 2022; Pratama, Hidayah, & Avini, 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 (R. Liu, 2022; Pratama et al., 2023). However, K means algorithm was sensitive and needed to be carried out carefully at the initial stage. This is because the performance of 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 (Chong, 2021; Ikotun et al., 2023; R. Liu, 2022).

Table 2.2: Summarizing the Performance of Clustering Approaches

Clustering Approaches	Advantageous	Disadvantageous
DBSCAN	<ul style="list-style-type: none"> <li>• Insensitive to noisy data</li> <li>• No predefined number of clusters is required</li> <li>• Identify the clusters in various shapes and sizes</li> </ul>	<ul style="list-style-type: none"> <li>• Performance depends on the parameters</li> <li>• High computational cost</li> </ul>
Agglomerative Hierarchical	<ul style="list-style-type: none"> <li>• Graphical representation</li> <li>• Robust to noise and outliers</li> <li>• No predefined number of clusters is required</li> </ul>	<ul style="list-style-type: none"> <li>• Issues in identifying the clusters with different densities</li> <li>• Computational complexity</li> </ul>
K Means	<ul style="list-style-type: none"> <li>• Easy implement</li> <li>• Scalable and Flexible</li> </ul>	<ul style="list-style-type: none"> <li>• Performance depends on the initial parameters</li> <li>• Sensitive to outliers and noisy data</li> </ul>

## **2.4 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. Besides that, sentiment analysis that applied to the drug reviews enabled the author to identify the patterns of patient opinions and allow the understanding of drug performance in the real-world scenarios. The experiment started with applying LLM, ChatGPT model to extract 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

- Al-Hadhrami, S., Vinko, T., Al-Hadhrami, T., Saeed, F., & Qasem, S. N. (2024). Deep learning-based method for sentiment analysis for patients' drug reviews. *PeerJ Computer Science*, 10, e1976.
- Alqaryouti, O., Siyam, N., Abdel Monem, A., & Shaalan, K. (2024). Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics*, 20(1/2), 142-161.
- 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*. Paper presented at the 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.
- Bu, K., Liu, Y., & Ju, X. (2024). Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning. *Knowledge-Based Systems*, 283, 111148.

- 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*. Paper presented at the SAS 2020 Global Forum.
- Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). *Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning*. Paper presented at the Proceedings of the 2018 international conference on digital health.
- 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*. Paper presented at the 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*.
- Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. (2024). Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 100059.
- 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.
- Lee, D.-g., Kim, M., & Shin, H. (2022). *Drug Repositioning with Disease-Drug Clusters from Word Representations*. Paper presented at the 2022 IEEE International Conference on Big Data and Smart Computing (BigComp).
- Liakos, A., Pagkalidou, E., Karagiannis, T., Malandris, K., Avgerinos, I., Gigi, E., . . . 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.

- 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.
- Pratama, M. A. Y., Hidayah, A. R., & Avini, T. (2023). Clustering K-Means untuk Analisis Pola Persebaran Bencana Alam di Indonesia. *Jurnal Informatika Dan Teknologi 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*.
- Rathod, D., Patel, K., Goswami, A. J., Degadwala, S., & Vyas, D. (2023). *Exploring drug sentiment analysis with machine learning techniques*. Paper presented at the 2023 International Conference on Inventive Computation Technologies (ICICT).
- 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*. Paper presented at the 21st international scientific conference AMSE applications of mathematics and statistics in economics.
- Shahapure, K. R., & Nicholas, C. (2020). *Cluster quality analysis using silhouette score*. Paper presented at the 2020 IEEE 7th international conference on data science and advanced analytics (DSAA).
- Shahid, T., Singh, S., Gupta, S., & Sharma, S. (2022). *Analyzing Patient Reviews for Recommending Treatment Using NLP and Deep Learning-Based Approaches*. Paper presented at the International Conference on Advancements in Interdisciplinary Research.
- 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.
- 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.