**Clustering Approaches in Dental Treatment Data Analysis**

**Abstract**

This literature review explores how clustering techniques like KMeans are used in healthcare, focusing on dental treatment data. Many recent studies apply unsupervised learning to find patterns in how patients behave, what treatments they receive, and their health outcomes. Some papers also try to include time-based information to make the clusters more meaningful. For example, one study by Peng et al. (2021) used cost data over time to find different types of dental care usage in children, showing that including timing helps create better clusters. Another review by Liu et al. (2023) pointed out that KMeans is the most common algorithm used for segmenting patients, but a lot of studies stop after building the clusters and do not go further to apply them in real healthcare situations. Aljohani (2024) looked at several clustering methods and explained that picking the right one depends on things like how easy it is to understand and how well it performs on different data. One major gap in current research is that most work does not include detailed timelines of treatments or follow how procedures change with age. Our work tries to fill this gap by grouping patients based on their dental treatment history over time, which helps create clearer and more useful patient groups for dentists to understand and use in practice.

**Introduction**

As healthcare becomes more focused on personalized treatment and better overall value, using data to group patients has become important. One way to do this is through clustering, which is a type of unsupervised machine learning that helps find hidden patterns or subgroups in big sets of patient data. In dentistry, where people often receive treatments over many years, it’s helpful to look at how care changes over time. Understanding these patterns can make it easier to plan treatments and catch issues early. This review looks at how clustering has been used in healthcare with dental data and explains how our approach fits in by using patient treatment history across different ages to create meaningful and understandable groups.

**Clustering in Healthcare Applications**

Clustering is commonly used in healthcare to find patterns in patient records. According to Liu et al. (2023), KMeans is the most used clustering method in healthcare studies, mainly because it’s simple to use, works fast, and is easy to understand. Peng et al. (2021) also used KMeans to group children based on their dental care patterns over time by using cost data. They found one group of kids, called the “early-onset” group, that made up a big part of the total treatment cost. This showed that clustering can help spot high-need patients early on. Aljohani (2024) compared KMeans with other methods like DBSCAN, GMM, and SOM, and looked at how stable, accurate, and clear each method was. KMeans did well in terms of speed and simplicity, but Aljohani pointed out that the best algorithm depends on the type of data and what the project is trying to achieve.

**Segmentation Data and Temporal Models**

A temporal model is one that considers not just what happened, but when it happened. This time-aware perspective is especially important in healthcare data, where treatment patterns unfold gradually over many years. In clustering, using temporal features can help us better capture real changes in patient behavior or treatment needs over time.

In the work by Peng et al. (2021), they created time-based feature vectors for each patient by combining cumulative costs of different types of dental visits, such as those with or without operating room involvement and preventive care. These vectors allowed them to track how treatment evolved and identify groups of patients with similar care paths using distance-based clustering. This demonstrated the power of including temporal structure in segmentation.

In our project, we built a similar idea into the feature engineering stage. Specifically, we segmented the patient dataset based on their visit continuity and treatment duration using four visit history types:

* V1: Consistent, long history (max gap ≤ 2 years, total span ≥ 3 years)
* V2: Consistent, short history
* V3: Inconsistent but long history (gap ≥ 2 years, span ≥ 7 years)
* V4: Inconsistent and short history

To focus on patients with rich treatment timelines, we selected only V1 and V3 groups for clustering. This choice ensures that the clusters are formed from patients with meaningful longitudinal data, either consistently or across a long span, which is crucial for reflecting true care trajectories. We then engineered features like procedure age (the age when each treatment happened), visit frequency, and treatment types, organized by age windows. This setup allowed our model to not just recognize what procedures a patient received, but also the timing and progression of those procedures. As a result, the clusters produced from our model better reflect real-life treatment journeys rather than just static snapshots of care.

**Clustering Algorithm Selection**

Checking if the clusters make sense is an important part of any clustering project. Peng et al. (2021) used both the elbow method and silhouette scores to measure how well their clusters were separated and if they were meaningful. Using both methods helped them feel more confident about the groups they found. In our case, we use the elbow method through a tool called KneeLocator, and we also look at visualizations of each cluster to see how different they are. To get more meaningful statistics from our cluster groups, we also calculated a customized mean value to reflect the average treatment intensity per age group. This mean is defined as:

This calculation helps normalize the procedure frequency across age bins, so we’re not just looking at raw totals but understanding how consistently a procedure occurs for different patients within that age range. Liu et al. (2023) pointed out that many papers only do basic validation and don’t test how useful or consistent the clusters are in real clinical settings. Aljohani (2024) also mentioned that when comparing different clustering methods, we should look at how stable the results are and how easy they are to interpret. Based on all this, one of the next steps for our work could be to try out different algorithms and include more advanced validation techniques, like checking if the clusters stay similar over time or comparing them with real-world clinical outcomes.

**Statistical Insights and Cluster Interpretation**

Peng et al. (2021) discovered that a small group of children those who started dental treatments early ended up being responsible for most of the total costs. That insight helped support the idea of starting preventive care sooner for high-risk groups. Our clustering work follows a similar idea by analyzing which types of procedures show up the most in each group. For example, we look at patterns like how often patients get root canals, crowns, or preventive treatments, and how those patterns change with age.

A graph of a patient cluster

AI-generated content may be incorrect.Fig. 1

In Fig. 1, we show a PCA projection of the clusters for patients aged 45–50, including their treatment histories. Each color represents a different group. From this view, we can see clear separation between clusters, which helps us understand that they represent different types of treatment behaviors. A graph of a number of people

AI-generated content may be incorrect.

Fig. 2

A graph of blue bars

AI-generated content may be incorrect.

Fig. 3

A graph of blue bars

AI-generated content may be incorrect.

Fig. 4

Figures 2, 3, and 4 break down meanings of the clusters more by showing the total number of a specific procedure (code 23322) per age group for each cluster.

By summarizing the number of procedures and their age distribution within each cluster, we can calculate simple statistics like averages and ranges to help understand each group's dental care load. Liu et al. (2023) highlighted how turning clusters into real recommendations is important, and Aljohani (2024) also pointed out that understanding what each group represents can help when choosing the right model. Our approach aims to do both interpret the clusters clearly and use the results to better support care planning.

**Gap Analysis**

One major issue is that most studies don’t make full use of long-term, procedure-level data, which could offer much deeper insights into how patients move through different types of care over time. For trajectory-based clustering, where patient patterns are studied over a timeline, is not commonly used in dental research.

Another limitation in existing studies is that many focus mainly on the technical side like building the clustering model or evaluating accuracy but don’t always connect their results back to actual public health applications or strategies. They also often skip testing how stable or reliable their clusters are when applied to new data.

Our project tries to fill these gaps in a few key ways. First, we include time-aware features like the age of the patient at the time of treatment and their visit frequency, which help capture care patterns across the years. We also create visualizations of treatment trends across clusters, making it easier to understand and compare different groups. Lastly, we plan to go a step further by testing how consistent the clusters are when we change the data and exploring other clustering algorithms like GMM and DBSCAN to make sure our approach is both accurate and reliable.

**Conclusion**

Clustering is a useful tool for understanding different types of patients in dental care, especially when we include information about how treatments change over time. Our project builds on earlier research by using K-Means clustering on patient treatment data grouped by age. This helps us create clusters that not only make sense statistically but also tell us something meaningful from a clinical point of view. Looking ahead, there’s a lot of room to improve by testing out more clustering methods, adding stronger validation steps, and finding ways to apply these insights in real healthcare settings where they can actually support better decisions and outcomes.

**References**

Aljohani, A. (2024). Optimizing patient stratification in healthcare: A comparative analysis of clustering algorithms for EHR data. *International Journal of Computational Intelligence Systems, 17*(1), 173.

Peng, J., Zeng, X., Townsend, J., Liu, G., Huang, Y., & Lin, S. (2021). A machine learning approach to uncovering hidden utilization patterns of early childhood dental care among Medicaid-insured children. *Frontiers in Public Health, 8*, 599187.

Liu, P., Wang, Z., Liu, N., & Peres, M. A. (2023). A scoping review of the clinical application of machine learning in data-driven population segmentation analysis. *Journal of the American Medical Informatics Association, 30*(9), 1573-1582.