ISyE 6740 - Spring 2025
Project Final Report

Team Member Names: Goyeun Yun (gyun37)

Project Title: Precision Medicine in Obesity Using Clustering and RAG-LLM

Contents

1 Problem Statement

1.1 Obesity

Obesity is a complex and multi-factorial disease that is influenced by genetic, environmental, and lifestyle factors. According to the World Health Organization (WHO), obesity rates have more than tripled since 1975, contributing to a rise in metabolic disorders such as type 2 diabetes, cardiovascular diseases, and some cancers [1].

The traditional approach to obesity treatment primarily relies on Body Mass Index (BMI) for classifica- tion, though BMI alone does not always accurately predict the metabolic risk associated with obesity. Treatment typically begins with lifestyle and behavioral modifications, such as dietary changes and increased physical activity, followed by medications if necessary. If these methods prove ineffective, bariatric procedures, including endoscopic interventions or surgery, may be considered[2]. However, this standard, one-size-fits-all approach often falls short in achieving long-term success. It fails to ac- count for the diverse range of obesity phenotypes, varying metabolic responses, and individual genetic predispositions, which significantly influence treatment outcomes[3].

The paper "Factors affecting weight loss variability in obesity" highlights the role of genetics, metabolic efficiency, hormonal regulation, and gut microbiota in influencing weight loss outcomes. It emphasizes the need for personalized approaches to optimize obesity treatment and improve weight management[4]. Also, the paper "Obesity: a gender-view" emphasizes that obesity affects men and women differently due to variations in hormonal regulation, genetic predisposition, metabolic characteristics, and lifestyle factors[5]. Together, there is a need for individualized treatment plans to address obesity effectively.

1.2 Precision Medicine in Obesity

Precision medicine offers personalized preventive, diagnostic, and therapeutic strategies that improve disease classification and optimize treatment efficacy by considering individual variability[6, 7]. In obesity management, multi-omics is pivotal in advancing precision medicine by integrating genomic, epigenomic, transcriptomic, proteomic, and metabolomic data to customize treatments according to an individual's unique biological profile[7]. While this approach is still in its early stages, recent years have seen growing research into AI and machine learning-based methods for classifying obesity subtypes and proposing personalized interventions.

1.3 Objective of the Project

This project aims to enhance the precision medicine approach in obesity management by utilizing data clustering techniques, Retrieval-Augmented Generation (RAG), and Large Language Models (LLM) to develop personalized treatment plans. By leveraging these advanced models, the project seeks to ensure more accurate and customized intervention strategies based on an individual's unique biological, genetic, and lifestyle profile. These technologies will facilitate the classification of obesity subtypes and provide tailored solutions, ultimately improving treatment outcomes and supporting better management of obesity in diverse patient populations.

2 Data Collection and Processing

2.1 Data Source

This study utilizes both structured obesity data and unstructured clinical guideline documents. The primary dataset used in this project is the Obesity Levels Dataset from Kaggle. The dataset

includes 16 different features such as various lifestyle and biometric features such as age, gender, caloric intake, physical activity frequency, and medical background to assess obesity levels with 2111 entries. In addition, three clinical practice guidelines were used as knowledge sources for the RAG-LLM treatment generation [8, 9, 10].

2.2 Data Processing

BMI Calculation

Body Mass Index (BMI) was calculated using the standard formula:

$$BMI = \frac{Weight(kg)}{Height(m)^2}$$

This derived feature was added to the dataset and used prominently in clustering.

Exploratory Data Analysis (EDA)

The data distribution was visualized through histograms and count plots in Figure 1 and Figure 2. A correlation heatmap was generated among continuous features. Notably, BMI had high correlation with Weight and moderate correlation with dietary behavior such as CH2O (water intake) and FCVC (vegetable consumption) shown in Figure 3.

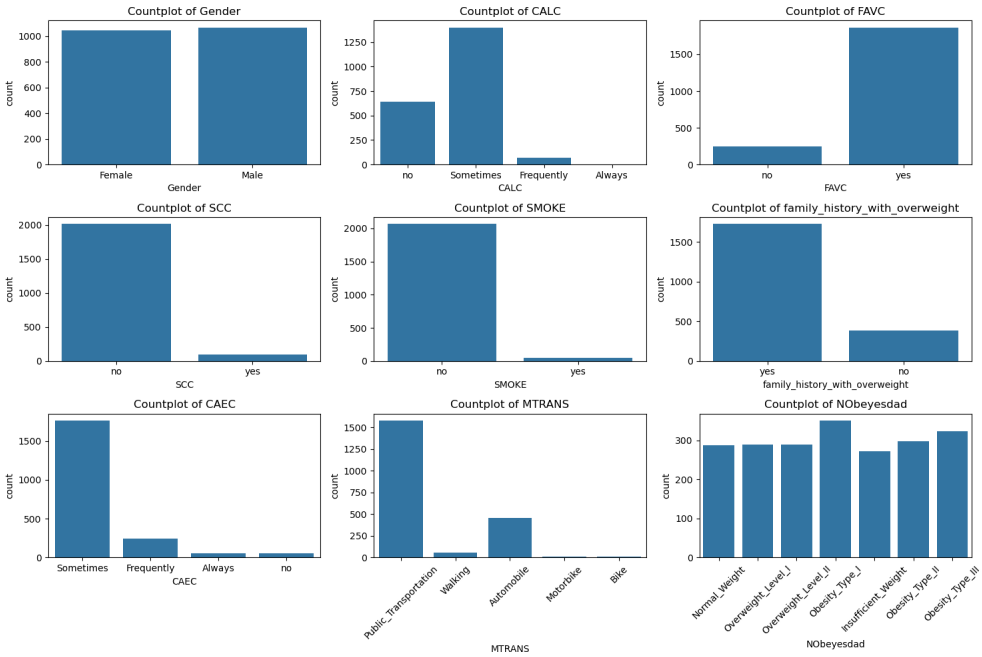

Figure 1: Histograms and Count Plots of Variables: Categorical Variables

2.3 Feature Importance Analysis

To identify the most influential factors contributing to obesity classification, a feature importance analysis was performed using a Random Forest Classifier. The dataset includes both numerical and categorical variables. Prior to model fitting, categorical features were encoded using label encoding, and numerical features were standardized via z-score normalization to ensure uniformity in feature scales. The Random Forest model was trained on all available predictors, excluding the target variable NObeyesdad. Feature importance was then extracted based on the model's internal Gini importance metric.

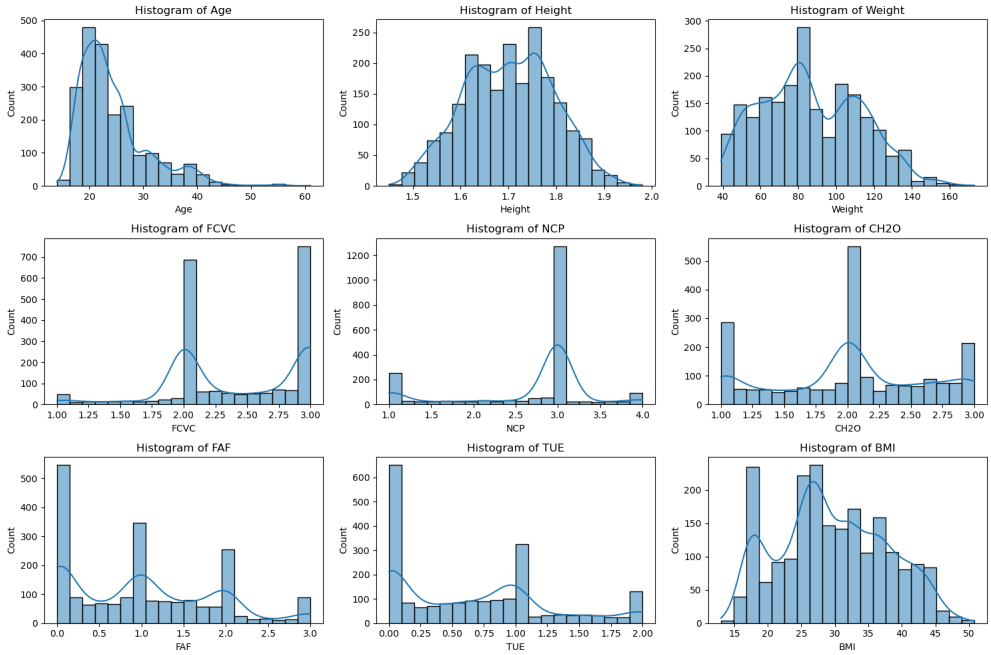

Figure 2: Histograms and Count Plots of Variables: Continuous Variables

Figure 4 presents the importance scores for all 17 features in ascending order. The top predictors identified were:

- **BMI (Body Mass Index)**: By far the most influential variable, accounting for approximately 42% of total importance.
- **Weight**: Closely related to BMI, weight showed substantial predictive power.
- **Gender**, **FCVC (Vegetable Consumption Frequency)**, **Height**, and **Age** followed, each contributing moderately to the classification task.

These results informed our variable selection for downstream tasks, particularly for clustering analysis. The top 6–8 features were selected based on importance scores to ensure both model interpretability and computational efficiency in the clustering and RAG-LLM phases.

3 Methodology

3.1 Clustering for Patient Segmentation

To enable personalized obesity treatment strategies, I first segmented the dataset into representative patient groups using unsupervised clustering. This step allows downstream RAG models to make recommendations tailored to the characteristics of each group.

Feature Selection: A total of 13 features were selected based on domain knowledge and prior variable importance analysis (see Section 2.3). Categorical variables (e.g., Gender, CAEC, CALC) were label-encoded, and continuous variables were standardized using z-score normalization.

Clustering Algorithm (K-Medoids): I used K-Medoids clustering, which uses actual data points as cluster centers and is less sensitive to outliers. To determine the optimal number of

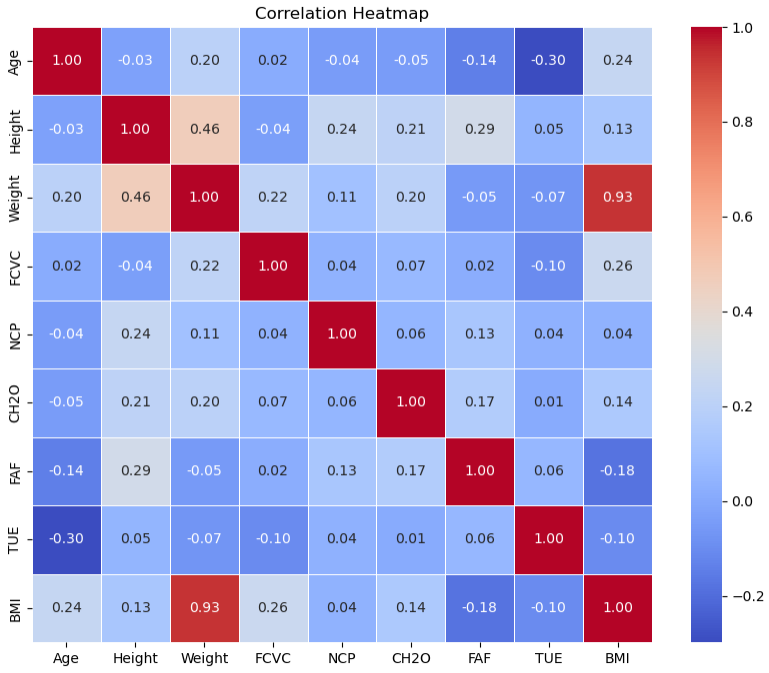

Figure 3: Correlation Heatmap

clusters, the silhouette score was computed for k ranging from 2 to 20. The highest silhouette score of 0.1086 was observed at $k = 10$, which I selected for final clustering.

Cluster Assignment and Summary: Each individual was assigned to one of 10 clusters based on similarity in the selected features (See Appendix B).

For each cluster, the mean of all features was computed to construct a cluster summary profile. These structured profiles were later used as inputs for the LLM in the RAG framework to generate personalized treatment strategies.

3.2 Obesity Level Prediction Model Using Logistic Regression

To assess the individual-level risk of obesity, a binary classification model was constructed that predicts whether a person is obese or not. The target variable is_obese was derived by grouping the original multi-class obesity levels into a binary label: obese (Obesity_Type_I, Obesity_Type_II, Obesity_Type_III) and non-obese (all others). The logistic regression model was chosen for its interpretability and suitability for binary classification tasks.

The dataset was split into training (80%) and testing (20%) sets with stratification to preserve class balance. Feature importance was interpreted through the model coefficients, revealing that BMI, Weight, and dietary behavior variables such as FCVC and CH2O had the highest impact on prediction. The model achieved strong predictive performance as shown in its classification report and accuracy score.

To make the model practically useful, I developed a user-facing obesity risk prediction function. This function takes structured input (e.g., age, weight, food consumption habits) and returns a

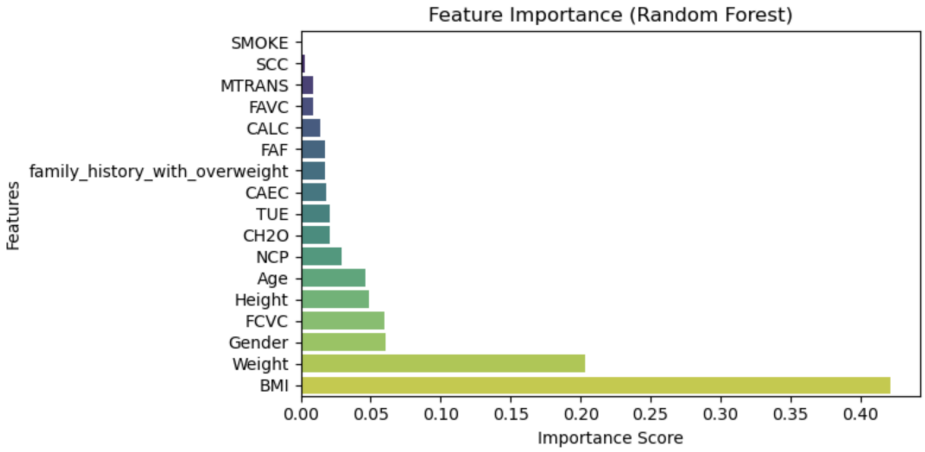

Figure 4: Feature Importance Scores from Random Forest Classifier

personalized risk score between 0 and 1, along with an interpretation label. The classification is defined as follows:

- **High Risk** score (`score` $\geq$ `0.9`): Strong indicators of obesity.
- **Medium Risk** (`0.7` $\leq$ `score` < `0.9`): Behavioral intervention recommended.
- **Borderline** (`0.4` $\leq$ `score` < `0.7`): Lifestyle improvements advised.
- **Low Risk** (`score` < `0.4`): Maintain current habits.

This risk prediction module complements the cluster-based treatment framework by allowing quick individual-level assessments grounded in clinical features and behavioral patterns.

3.3 LLM Integration and RAG System

1) Embedding Clinical Guidelines

To incorporate domain-specific medical knowledge into the system, three authoritative clinical practice guidelines for obesity management were selected [8, 9, 10]. The documents were processed by extracting raw text from both PDF and TXT files. Each document was then segmented into chunks of two sentences, balancing contextual integrity and input size limitations. These chunks were embedded using a pre-trained Sentence-BERT model, producing a vector space of 6,551 semantically meaningful representations. Both the embeddings and their associated text were stored for downstream semantic retrieval in the RAG framework.

2) Retrieval-Augmented Generation (RAG) Design

The RAG pipeline was designed to enrich LLM responses with medically grounded information. When a query, such as a patient-level profile or a cluster-level summary, is received, it is first embedded into the same semantic space as the guideline chunks. Using cosine similarity, the system retrieves the top-k most relevant chunks. These retrieved texts, along with the query, are assembled into a comprehensive prompt and passed to the Gemini 1.5 Pro model for content generation. This approach ensures that the LLM produces responses grounded in established clinical recommendations.

3) Cluster-Based Prompting with Gemini

Rather than treating each patient in isolation, individuals were grouped into ten clusters using the K-Medoids algorithm based on key features such as BMI, physical activity, diet, and family history. The average characteristics of each cluster were used to generate representative profiles. These profiles served as input prompts for the LLM, enabling the generation of population-level treatment strategies that reflect the common traits and needs of each subgroup.

4 Result & Evaluation

4.1 Cluster Summary & Interpretation

To support group-level personalization, I analyzed the mean characteristics of each of the 10 clusters derived from the K-Medoids algorithm. The summary statistics, presented in Appendix B, reveal distinct lifestyle and health patterns across the population.

For example, Cluster 0 represents borderline overweight individuals with a moderate BMI (26.6), low physical activity (FAF 0.54), and insufficient water and meal intake, suggesting the need for basic lifestyle interventions. In contrast, Cluster 9 includes individuals with severe obesity (BMI 42.1) and a 100% family history of overweight, indicating a combination of genetic and behavioral risk that may warrant aggressive intervention.

Cluster 1 also exhibits severe obesity (BMI 34.6) and high body weight, paired with poor dietary habits such as low vegetable intake. Meanwhile, Cluster 3 includes individuals with near-normal BMI and healthy hydration, minimal screen time, and overall balanced behavior. Interestingly, Cluster 5 stands out as a group with a normal BMI but very low physical activity and high screen time, highlighting the presence of potentially "hidden risk" individuals.

On the opposite end of the spectrum, Cluster 7 is characterized by a normal BMI, regular eating patterns, and very high physical activity (FAF 2.12), reflecting a healthy and active lifestyle. This diversity in cluster profiles supports the use of tailored strategies for different patient segments in the subsequent RAG-based treatment generation pipeline.

4.2 LLM-based Recommendation Evaluation

To assess the consistency and personalization of treatment recommendations generated by the RAG-LLM system, two types of responses were compared: one based on the summary of a cluster and the other based on individual patient data from the same cluster. The goal was to determine whether the average profile of a patient group can reliably represent the needs of its individual members in clinical decision-making.

Three clusters were selected for evaluation: Cluster 1, Cluster 4, and Cluster 9. From each cluster, it was randomly sampled five patients. Using their personal health and behavioral data, individualized treatment recommendations were generated with the RAG-LLM model. Separately, one recommendation was also generated using only the cluster's summary profile.

To evaluate the similarity between the summary-based recommendation and each of the five individualized responses, the BERTScore F1 metric was used, a widely used semantic similarity measure in natural language processing[11].

The results showed consistently high levels of agreement between the cluster summary and individual recommendations:

- **Cluster 1:** Average F1 = 0.8723

- **Cluster 4:** Average F1 = 0.8691
- **Cluster 9:** Average F1 = 0.8724

Individual BERTScore F1 values ranged between 0.8591 and 0.8984 across all samples, indicating that the recommendations generated using cluster summaries were largely consistent with those based on detailed personal data.

These results suggest that cluster-level summaries can serve as effective surrogates for individualized input, particularly when detailed personal data is unavailable. While there were minor variations, the high similarity scores support the robustness of the cluster-based RAG approach. In future work, prompt engineering or hierarchical clustering could be explored to further enhance personalization.

In addition, to assess how distinct the RAG-generated responses were across different clusters, the answers generated from the summaries of three distinct clusters (Cluster 1, 4, and 9) were compared against the reference summary of Cluster 1 using BERTScore. The resulting F1 scores were 0.8221, 0.8202, and 0.8190, with an average of 0.8204. These results suggest that while the language model produced somewhat distinct recommendations based on each cluster's summary, the responses still shared a high level of semantic similarity. This indicates potential overlap in suggested treatment strategies across different patient groups, warranting further investigation into how well cluster-specific nuances are captured.

Conclusion and Future Work

This project explored the integration of machine learning, retrieval-augmented generation (RAG), and large language models (LLMs) to provide personalized treatment recommendations for obesity. By segmenting individuals into ten distinct clusters based on clinical and behavioral characteristics, it was able to generate group-level treatment summaries that capture the diversity of obesity profiles. Additionally, a binary obesity risk prediction model was developed using logistic regression and implemented a patient-level risk scoring system.

Despite the promising results, several limitations remain. First, the evaluation is based on automatic metrics such as BERTScore, which may not fully capture clinical accuracy or relevance. Second, the system is sensitive to the quality of guideline chunking and embedding, and the LLM's responses may vary depending on prompt phrasing or API conditions. Lastly, the system has not yet been tested with real-world clinical input or expert validation.

Future Work: Building upon this foundation, the next step is to develop an interactive patient-facing application. In this system, users will be able to enter their own health and lifestyle information (e.g., age, weight, dietary habits, physical activity), and the system will:

- Predict the individual's obesity risk score using the trained logistic regression model.
- Assign the user to the most similar patient cluster based on clinical features.
- Automatically retrieve and display a personalized treatment recommendation using the RAG-LLM pipeline.

This interactive tool would enable real-time, accessible, and personalized health guidance based on both structured data and clinical guidelines. Further improvements may include integrating additional medical literature, experimenting with alternative LLMs and embedding models, and incorporating iterative feedback from healthcare providers to ensure clinical validity and usability at scale.

Appendix A: Feature Description

Feature	Description
Age	Age of the individual
Gender	Gender of the individual
Height	Height (in meters)
Weight	Weight (in kilograms)
CALC	Frequency of alcohol consumption
FAVC	Frequent consumption of high-calorie food
FCVC	Frequency of vegetable consumption
NCP	Number of main meals per day
SCC	Consumption of sugary drinks
SMOKE	Smoking habits
CH2O	Daily water intake
family_history_with_overweight	Family history of overweight
FAF	Weekly physical activity frequency
TUE	Time spent using technology (hours per day)
CAEC	Consumption of food between meals
MTRANS	Transportation method
NObeyesdad	Obesity level (target class)

Appendix B: Cluster-Level Summary

Cluster ID (Size)	Key Characteristics
Cluster 0 (n=259)	Borderline overweight (BMI 26.6), low exercise (FAF 0.54), low water intake, below-average meals.
Cluster 1 (n=241)	Severely obese (BMI 34.6), high weight (112kg), high family history (0.99), low vegetable intake.
Cluster 2 (n=200)	Overweight (BMI 28.1), moderate activity and healthy eating patterns, mid-20s age group.
Cluster 3 (n=177)	Near-normal weight (BMI 25.2), high water intake, very low screen time, balanced lifestyle.
Cluster 4 (n=319)	Normal weight (BMI 23.6), high vegetable intake and physical activity, youngest age group.
Cluster 5 (n=139)	Normal BMI (23.0), very low exercise (0.3), high screen time, sedentary behavior dominant.
Cluster 6 (n=119)	Late-stage overweight (BMI 29.3), high family history (0.92), low physical activity.
Cluster 7 (n=198)	Normal weight (BMI 24.4), very high activity (2.12), regular meals, healthy lifestyle.
Cluster 8 (n=130)	Severely obese (BMI 34.8), high weight, moderate diet quality, low physical activity.
Cluster 9 (n=329)	Most obese (BMI 42.1), perfect family history (1.0), regular habits but low activity.

References

[1] World Health Organization. *Obesity and Overweight*. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[2] W. Chung, J. W. Lim, and J. H. Lee. "Obesity: A gender-view". In: *Frontiers in Endocrinology* 12 (2021), p. 706197. DOI: 10.3389/fendo.2021.706197.

[3] R. Franceschi. "Precision medicine in diabetes, current research and future perspectives". In: *Journal of Personalized Medicine* 12.8 (2022), p. 1233. DOI: 10.3390/jpm12081233.

[4] P. Boutin and J. Le. "Title of the article". In: *Metabolism* 107 (2020), pp. 1541–1538. DOI: 10.1016/j.metabol.2020.154138.

[5] G. Muscogiuri et al. "Obesity: A gender-view". In: *Obesity Reviews* 24.9 (2023), e13530. DOI: 10.1111/obr.13530.

[6] A. M. D. Hurtado and A. Acosta. "Precision medicine and obesity". In: *Gastroenterology Clinics of North America* 50.1 (2021), pp. 127–139. DOI: 10.1016/j.gtc.2020.10.005.

[7] L. Cifuentes et al. "Precision medicine for obesity". In: *Digestive Diseases and Interventions* 5.3 (2021), pp. 239–248. DOI: 10.1055/s-0041-1729945.

[8] W Timothy Garvey et al. "Comprehensive clinical practice guidelines for medical care of patients with obesity". In: *Endocrine Practice* 22.3 (2016), pp. 1–203.

[9] Eduardo Grunvald et al. "AGA clinical practice guideline on pharmacological interventions for adults with obesity". In: *Gastroenterology* 162.7 (2022), pp. 2061–2077.

[10] Sean Wharton et al. "Obesity in adults: a clinical practice guideline". In: *CMAJ: Canadian Medical Association Journal* 192.31 (2020), E875–E891.

[11] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations (ICLR)*. 2020. URL: https://arxiv.org/abs/1904.09675.

[12] M.-A. Cornier. "A review of current guidelines for the treatment of obesity". In: *Managed Care Journal* 28.15 (2022). December 14.

[13] J.-P. Després et al. "Obesity phenotypes, lifestyle medicine, and population health: Precision needed everywhere!" In: *Journal of Obesity & Metabolic Syndrome* 34 (2025), pp. 4–13. DOI: 10.7570/jomes24043.

[14] M. Li et al. *BiomedRAG: A retrieval augmented large language model for biomedicine*. arXiv preprint. 2024. URL: https://arxiv.org/abs/2405.00465.

[15] M. Jin, Q. Yu, D. Shu, et al. *Health-LLM: Personalized retrieval-augmented disease prediction system*. arXiv preprint. 2024. URL: https://arxiv.org/abs/2402.00746.

[16] Z. Zhan, J. Wang, S. Zhou, et al. *MMRAG: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning*. arXiv preprint. 2025. URL: https://www.arxiv.org/abs/2502.15954.