UNIVERSITAT ROVIRA I VIRGILI
ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA

Phenotypic Clustering of Differentiated Thyroid Cancer Patients Using *k*-Means

Master in Health Data Science — MHEDAS

Subject: Project and Research Methodologies

Author(s): David Cabezas Antolin, Frances Scarlett Thomas, Ravneet-Rahul Sandhu Singh, Roger Puig I Arxer, Sofía González Estrada

Date: January 20, 2026

UNIVERSITAT ROVIRA i VIRGILI
BIOINFORMATICS BARCELONA
UAB Universitat Autònoma de Barcelona
UNIVERSITAT DE BARCELONA
Universitat de Lleida
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
UNIVERSITAT DE VIC UNIVERSITAT CENTRAL DE CATALUNYA
Universitat de Girona
UGA Université Grenoble Alpes
eit Health
Co-funded by the European Union

Table of Contents

1. Introduction

Differentiated Thyroid Cancer (DTC) represents the most common endocrine malignancy, posing complex challenges for healthcare systems worldwide. While survival rates are generally high following surgical intervention, the clinical course is frequently complicated by the risk of disease recurrence, which remains a significant concern for long-term management. Consequently, post-operative care requires a delicate balance between effective oncological surveillance and the minimization of treatment-related morbidity. This project employs an unsupervised machine learning approach, specifically the *k*-means clustering algorithm, to explore phenotypic patterns in a post-surgical cohort, aiming to refine risk stratification for recurrence and improve patient outcomes.

1.1 Background and Clinical Importance

The global impact of thyroid cancer has risen dramatically over the past decades, with incident cases increasing by 167% [1]. This drastic increase is mostly due to improvements in diagnostic procedures and increased surveillance campaigns, which has led to an increase in overdiagnosis [2], although overall mortality remains low [1]. This is particularly evident in middle-aged women who are diagnosed at disproportionately higher numbers than males due to a varied set of risk factors [3]. One study showed that the difference in diagnosis is not due to inherent sex-based biological differences, but to gender-based differences in detection [4], taking into account care-seeking behaviors and increased likelihood of thyroid function evaluation in women by healthcare professionals. However, once diagnosed, men have higher risk of adverse outcomes and recurrence compared to women [5]. Disparities also exist at the socio-demographic level, with countries with lower socio-demographic index (SDI) showing projected increased burden of thyroid cancer [6].

The majority of patients in high SDI countries undergo thyroidectomy, the complete removal of the thyroid. This leads to lifelong reliance on thyroid medication supplementation and regular dosage management. Moreover, of those diagnosed with thyroid cancer, many receive radiotherapy and neck lymph node dissection, both invasive and uncomfortable procedures. Due to this, the American Thyroid Association (ATA) changed its guidelines from performing thyroid excision or other invasive procedures to those that reflect a watch-and-wait approach for patients with lower-risk thyroid cancer, though long-term recurrence risk remains a significant concern for both survivors and clinicians [7].

Despite shifts in guidelines to more conservative measures, recurrence can still occur years or decades after initial diagnosis. Recurrence within 10 years for those with papillary thyroid cancer, the most common type of thyroid cancer, ranges from 1.6% in those with low-risk types to 22.7% in those categorized as high risk [8]. Although general risk of recurrence and subsequent mortality is relatively low compared to other types of cancers, thyroid cancer survivors experience increased psychological, physical, and financial stress due to the possibility of recurrence [9]. Health systems can be inappropriately strained by providing increased screening and diagnostic services through inefficient treatment

pathways without adequate triaging of thyroid cancer survivors [10]. Accurate, generalizable, and clinically relevant prediction tools provide the most advantageous impact on the health of thyroid cancer survivors and the care provided by healthcare systems. However, some predictive tools are limited in accurately capturing the individual recurrence risk for men, women, or for survivors in low-resource settings [11].

1.2 State of the Art

Recent advances in machine learning have demonstrated the potential of unsupervised methods to reveal latent patient subgroups and risk patterns from clinicopathologic data. In the context of DTC, several studies have established foundational benchmarks for prediction and risk stratification.

Onah et al. (2025) systematically evaluated unsupervised feature engineering pipelines for DTC recurrence prediction, comparing multiple dimensionality reduction techniques including Independent Component Analysis, Factor Analysis, and Truncated Singular Value Decomposition [12]. While their Principal Component Analysis based approach achieved predictive accuracies exceeding 99% when integrated with supervised classifiers such as Random Forest, the transformed features lack direct clinical interpretability. This highlights a fundamental tradeoff: dimensionality reduction can improve predictive performance but often obscures the relationship between features and outcomes. In contrast to pure dimensionality reduction, other recent studies have leveraged unsupervised learning for direct phenotypic discovery. For instance, Jiang et al. (2024) applied *k*-means clustering to a cohort of acute ischemic stroke patients, specifically addressing the challenge of unordered categorical clinical data [13]. Their analysis successfully identified three distinct clinical phenotypes characterized as *Arteriosclerosis*, *Mild stroke*, and *Cardiogenic stroke*, which exhibited significantly different prognostic outcomes and recurrence risks [13]. This work serves as a crucial precedent for the current study, demonstrating that unsupervised clustering can uncover clinically meaningful subgroups with divergent risk profiles.

Beyond feature engineering and prediction, unsupervised clustering offers a complementary approach to identify previously unrecognized patient subgroups with direct clinical interpretability. Ezugwu et al. (2022) provide a comprehensive review of clustering algorithms, emphasizing the continued relevance of partitional methods in modern data analysis [14]. Among these, *k*-means clustering is highlighted as a computationally efficient and scalable algorithm, particularly well suited for tabular clinical data due to its linear time complexity and robust performance on large datasets. The review also discusses that while hierarchical and density-based clustering methods have their applications, partitional approaches remain the state of the art for exploratory analysis.

Despite these methodological advances, a gap remains in applying unsupervised clustering directly to DTC recurrence data to discover novel phenotypic subgroups. Prior work has focused on prediction or feature extraction, but the identification of distinct recurrence risk profiles through clustering has not been thoroughly explored.

1.3 Objectives

Building on these works, this study applies *k*-means based phenotypic clustering on data tracking recurrence among DTC survivors. The aim of this study is to use machine learning to reveal unique phenotypic combinations or recurrence patterns of thyroid cancer survivors that can inform future predictive models.

To achieve this, the specific objectives are defined as follows:

- To preprocess the retrospective clinicopathologic data, transforming categorical features into a numerical format suitable for clustering.
- To determine the optimal number of patient phenotypes using validation metrics such as the Elbow method.
- To characterize the resulting clusters through statistical comparisons of demographic and pathological feature distributions between groups.

1.4 Scope

The scope of this project is defined by the following boundaries regarding the study population, data nature, and methodology:

- **Study Population:** The analysis focuses exclusively on a retrospective cohort of patients diagnosed with differentiated thyroid cancer who have undergone surgical intervention.
- **Data Characteristics:** The dataset is derived from a single medical center and consists exclusively of structured tabular data representing demographic, clinical, and pathological characteristics. It does not include unstructured data types such as raw medical imaging or genomic sequencing.
- **Methodological Boundaries:** The project is strictly limited to unsupervised learning, specifically the application of the *k*-means clustering algorithm. While supervised prediction models and deep learning techniques are prevalent in the literature, they fall outside the scope of this specific phenotypic analysis.

2. Methodology

This section outlines the framework for identifying phenotypic subgroups in Differentiated Thyroid Cancer (DTC) patients. The workflow proceeds sequentially through exploratory data analysis, feature engineering, data preprocessing, k-means clustering with optimal k selection via the elbow method, and cluster characterization to assess the resulting phenotypes and their clinical relevance for recurrence risk stratification.

The data used in this project come from the DTC recurrence dataset, obtained from the UCI Machine Learning Repository [15]. This dataset was originally introduced in the study *Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study*, published in *Head and Neck* (2023) [16].

2.1 Exploratory Data Analysis

The dataset contains 383 patient records with 16 variables describing demographic characteristics, clinicopathologic tumor features, and a binary outcome indicating recurrence of well-differentiated thyroid cancer. The demographic and clinicopathological characteristics of the study cohort are summarized in Annex **Table A.1**, stratified by recurrence status. Statistical comparisons between recurrence groups were performed using appropriate tests based on variable type and distribution. For continuous variables, normality was assessed using the Shapiro-Wilk test. Age, the only continuous variable in the dataset, failed the normality assumption ($p < 0.05$) and was therefore summarized as median [IQR] and compared between groups using the Mann-Whitney U test:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{2.1}$$

where n_1 and n_2 are the sample sizes of the two groups, and R_1 is the sum of ranks for group 1. The test statistic assesses whether the distributions of the two groups differ by comparing the rank sums.

For categorical variables, frequencies and percentages were calculated, and group comparisons were performed using either the Chi-squared test or Fisher's exact test when expected cell counts were less than 5. The Chi-squared test statistic is defined as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.2}$$

where O_{ij} represents the observed frequency in cell (i, j) and E_{ij} represents the expected frequency under the null hypothesis of independence. A p-value < 0.05 was considered statistically significant for all tests.

The cohort was highly skewed toward female patients (312, 81.5%) compared to male patients (71, 18.5%). Of the total cohort, 275 (71.8%) did not experience recurrence and 108 (28.2%) experienced disease recurrence during the 10 to 15 years of follow-up. Very

few patients were current (12.8%) or former (4.2%) smokers, and only 51 (13.3%) exhibited some form of hyper- or hypothyroidism at baseline.

Several baseline characteristics showed statistically significant associations with recurrence risk. Patients who experienced recurrence were significantly older (median age 44.5 years vs 36.0 years, $p < 0.001$), more likely to be male (38.9% vs 10.5%, $p < 0.001$), and had higher rates of current smoking (30.6% vs 5.8%, $p < 0.001$). Clinical examination findings, including the presence of adenopathy ($p < 0.001$) and multinodular goiter ($p = 0.011$), also differed significantly between groups.

Tumor characteristics demonstrated strong associations with recurrence. Multi-focal disease was substantially more prevalent in the recurrence group (64.8% vs 24.0%, $p < 0.001$). Most of the study population was characterized as low risk (249, 65.0%), followed by intermediate risk (102, 26.6%), with only 32 patients characterized as high risk (8.4%). The distribution across American Thyroid Association (ATA) risk categories differed markedly by recurrence status, with 29.6% of recurrent cases classified as high-risk compared to 0% in the non-recurrence group ($p < 0.001$). Notably, 100% recurrence occurred among those with stage III or IV disease and among those labeled as high risk ($p < 0.001$). At baseline, 39.4% of patients had T2 tumor staging with 25.1% having T3a staging. Advanced tumor stage (T3b, T4a, T4b) and distant metastasis (M1) were almost exclusively observed in patients who experienced recurrence. Post-treatment response patterns also strongly predicted outcomes, with structural incomplete response observed in 82.4% of recurrence cases versus only 0.7% of non-recurrence cases ($p < 0.001$).

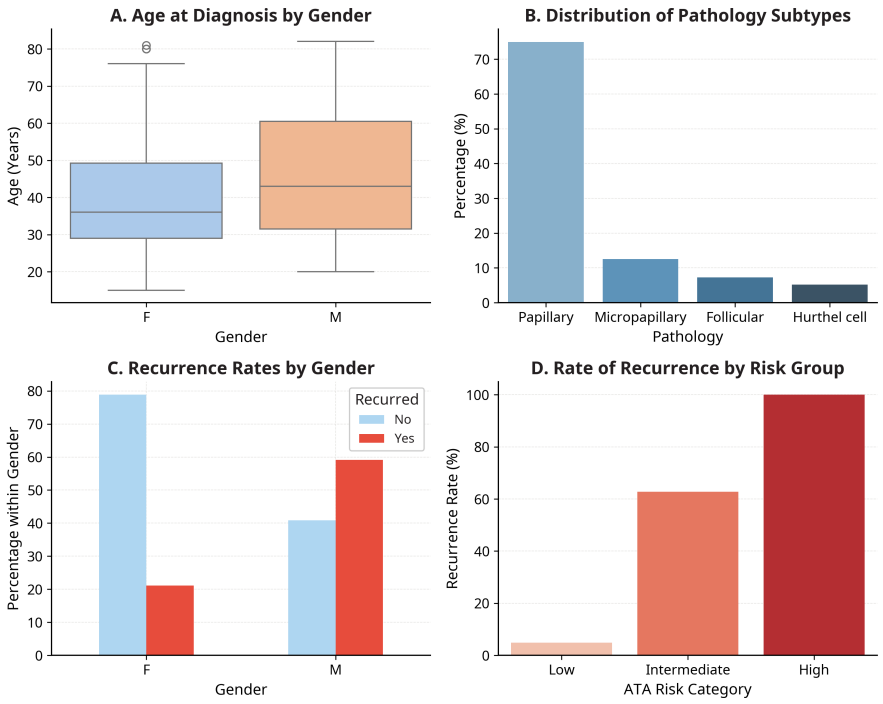

Figure 2.1: Demographic and clinical characteristics of the study cohort. (A) Age distribution at diagnosis stratified by gender. (B) Distribution of pathological subtypes. (C) Recurrence rates by gender. (D) Recurrence rates across ATA risk categories.

Figure 2.1 provides a visual summary of key demographic and clinical patterns. Panel A illustrates the age distribution by gender, showing that male patients tend to be diagnosed at older ages. Panel B displays the distribution of pathological subtypes, with papillary carcinoma being the most common histology (74.9%), followed by micropapillary (12.5%), follicular (7.3%), and Hürthle cell (5.2%) variants. Panel C demonstrates gender-based differences in recurrence rates, with males exhibiting substantially higher recurrence rates compared to females. Panel D shows the relationship between ATA risk stratification and recurrence, revealing a clear gradient: low-risk patients had minimal recurrence (4.8%), intermediate-risk patients showed moderate recurrence (62.7%), and high-risk patients experienced the highest recurrence rates (100%).

2.2 Feature Engineering

Feature engineering was performed to balance analytical simplicity with the preservation of clinically meaningful granularity. The approach prioritized retaining information relevant to phenotypic characterization while removing redundant or confounding variables.

The original dataset contained two separate smoking-related variables: current smoking status and history of smoking. These were combined into a single categorical variable with three levels: *Never*, *Former*, and *Current*. This simplification reduced the number of variables while maintaining the clinically relevant distinction between smoking exposure categories.

Next, specific variables were excluded from the clustering analysis due to their incorporation into composite staging systems. To quantify the degree of association between categorical variables, we employed Cramér's V statistic, defined as:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, r-1)}} \tag{2.3}$$

where χ^2 is the Chi-square statistic, n is the sample size, k is the number of categories in one variable, and r is the number of categories in the other. Cramér's V ranges from 0 (no association) to 1 (perfect association).

The individual TNM (Tumor, Node, Metastasis) components were removed because Stage is a composite metric that integrates T, N, and M information. Retaining any individual component alongside Stage would introduce redundancy (**Figure 2.2**, A). The analysis revealed varying degrees of correlation: M exhibited very strong correlation with Stage (Cramér's V = 0.755), confirming near-perfect redundancy, as M1 disease automatically classifies patients as Stage IV; T showed moderate correlation (Cramér's V = 0.425); and N showed weaker correlation with Stage (Cramér's V = 0.256). The N component was removed in favor of the adenopathy variable (Cramér's V = 0.633), which captures the same underlying nodal disease but provides additional spatial information (bilateral, left, right, posterior, extensive) valuable for phenotypic characterization (**Figure 2.2**, B).

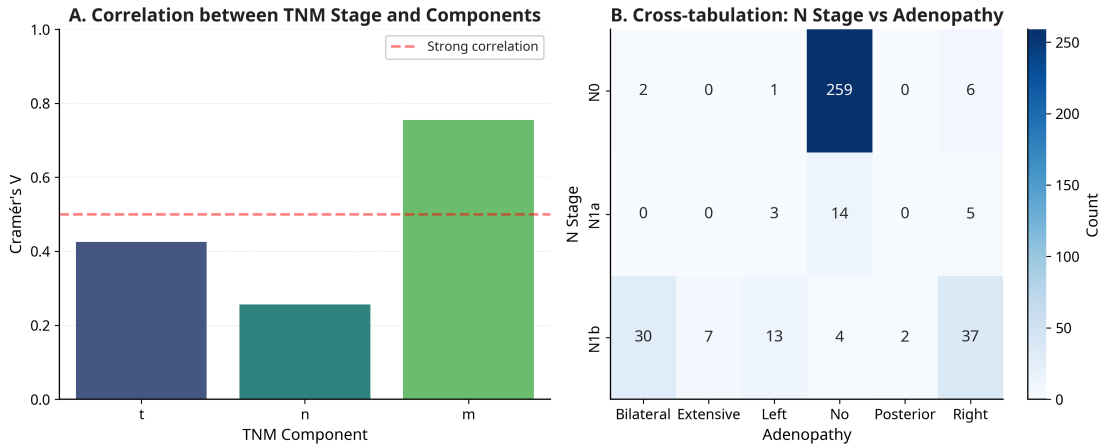

Figure 2.2: Correlation analysis for feature selection. (A) Cramér's V coefficients between TNM Stage and its individual components (T, N, M). (B) Cross-tabulation heatmap showing the relationship between N stage and adenopathy status.

Both Stage and ATA risk stratification were retained in the clustering analysis. TNM Stage is an anatomic staging system that reflects the extent of tumor burden, whereas ATA initial risk stratification incorporates anatomic and histopathologic features, and in some contexts molecular findings, to estimate the risk of persistent or recurrent disease [17]. These variables capture distinct dimensions of disease severity and prognosis.

2.3 Data Preprocessing

Prior to clustering, the dataset underwent standardized preprocessing. First, the recurrence outcome variable was removed from the feature set. As established in the study objectives, this analysis adopts a purely unsupervised approach, treating the data as if recurrence status were unknown. Excluding the outcome variable prevents perfect separation of groups based solely on recurrence status, which would hide the phenotypic heterogeneity we aim to discover.

For the continuous variable, age was standardized using the standard scaling transformation:

$$z = \frac{x - \mu}{\sigma} \tag{2.4}$$

where x represents the original age value, μ is the mean age across the cohort, and σ is the standard deviation.

All categorical variables were transformed using one-hot encoding, which converts each categorical feature with k levels into k binary indicator variables.

2.4 *k*-Means Clustering Algorithm

The k-means clustering algorithm was employed to identify distinct phenotypic subgroups within the DTC patient cohort. The following subsections describe the algorithm mechanics and the method for determining the optimal number of clusters.

2.4.1 Algorithm Overview

The k-means algorithm, first proposed by Lloyd in 1957 [18] and later formalized by MacQueen in 1967 [19], is an unsupervised learning method that partitions n observations into k clusters by iteratively minimizing the within-cluster sum of squares (WCSS). The objective function is defined as:

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2.5}$$

where C_i represents the i-th cluster, μ_i is the centroid (mean) of cluster C_i, and $\|x - \mu_i\|^2$ denotes the squared Euclidean distance between observation x and centroid μ_i. The algorithm proceeds in two alternating steps: (1) assignment, where each observation is assigned to the nearest centroid, and (2) update, where centroids are recalculated as the mean of all observations assigned to each cluster. These steps repeat until convergence, defined as the point at which cluster assignments no longer change between iterations.

The primary tunable parameters in the k-means implementation are k, the number of clusters to form, and `n_init`, the number of times the algorithm is run with different centroid initializations. Since k-means is sensitive to initial centroid placement, multiple initializations help ensure convergence to a global rather than local minimum of the objective function. In this analysis, k was determined using the elbow method (described in the next section), and `n_init` was set to 20 to ensure robust clustering results.

2.4.2 Optimal k Selection: Elbow Method

The elbow method is a heuristic approach for determining the optimal number of clusters by examining the relationship between k and the within-cluster sum of squares (WCSS). As k increases, WCSS necessarily decreases because observations are divided into smaller, more homogeneous groups. However, beyond a certain point, adding more clusters yields diminishing returns in variance reduction. The optimal k is identified at the elbow point, the value where the rate of WCSS decrease sharply diminishes, indicating that additional clusters provide minimal improvement in compactness.

For this analysis, k-means clustering was performed for values of k ranging from 2 to 10, with WCSS calculated for each configuration. The resulting elbow curve is presented in **Figure 2.3**. Visual inspection of the curve reveals a clear inflection point at $k = 4$, where the slope transitions from steep to gradual. This suggests that four clusters provide an optimal balance between model complexity and variance explanation. Values of $k < 4$ would oversimplify the phenotypic heterogeneity in the cohort, while values of $k > 4$ would lead to overfitting with minimal gains in cluster descriptiveness.

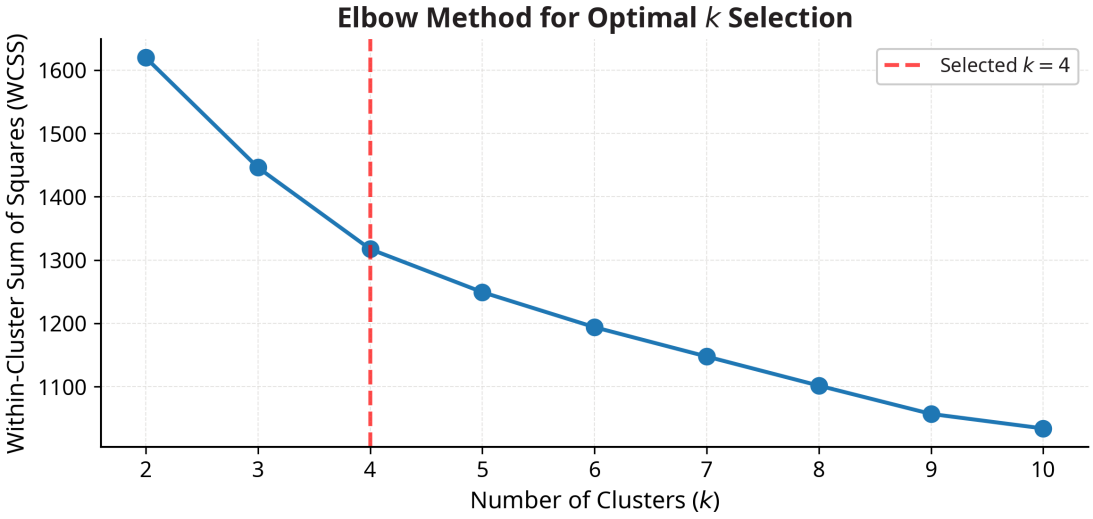

Figure 2.3: Elbow curve for optimal k selection. Within-cluster sum of squares (WCSS) plotted against number of clusters (k).

2.5 Cluster Characterization

Following cluster assignment, a systematic characterization was performed to identify which clinical and pathological features most strongly distinguished between clusters. This analysis proceeded in two stages: first, quantifying the discriminative power of each feature across all clusters through statistical testing; second, conducting pairwise comparisons to determine which specific cluster pairs differed on key features.

2.5.1 Feature Importance Analysis

For the continuous variable (age), the Kruskal-Wallis test was used to assess differences across all clusters. The Kruskal-Wallis test is a non-parametric extension of the Mann-Whitney U test for comparing more than two groups. The test statistic H is defined as:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{2.6}$$

where N is the total sample size, k is the number of clusters, R_i is the sum of ranks for cluster i, and n_i is the sample size of cluster i.

For categorical variables, the Chi-squared test was applied to assess whether the distribution of categories differed across clusters.

To account for multiple testing, the Benjamini-Hochberg procedure controlled the False Discovery Rate (FDR) at $\alpha = 0.05$. This method adjusts p-values to limit the expected proportion of false discoveries while being less conservative than Bonferroni correction.

To facilitate visual comparison of cluster profiles, feature values were standardized using z-score transformation. For each feature, the mean value per cluster was calculated and then

standardized across all clusters, yielding z-scores that represent the number of standard deviations above or below the overall mean. Results were visualized as a spider plot, with each axis representing a feature and each cluster depicted as a distinct layer.

2.5.2 Pairwise Cluster Comparisons

Following the identification of discriminative features, pairwise statistical comparisons were conducted between all possible cluster pairs to determine which specific clusters differed on each feature. For age, the Mann-Whitney U test was applied to each pair of clusters. For categorical variables, the Chi-squared test was used, with contingency tables constructed for each cluster pair separately.

Since multiple pairwise comparisons were performed for each feature, the Benjamini-Hochberg procedure was again applied to control the false discovery rate at $\alpha = 0.05$. The FDR correction was performed separately for each feature to account for the repeated testing across cluster pairs. Results were visualized as a heatmap displaying the adjusted p-values, with asterisks indicating statistically significant differences between cluster pairs.

3. Results

This section presents the findings from the *k*-means clustering analysis of 383 Differentiated Thyroid Cancer (DTC) patients. Following the identification of four optimal clusters (see Methodology), the results are organized as follows: cluster distribution and visualization in principal component space, statistical identification of discriminative features, characterization of phenotypic profiles across all clinical variables, and pairwise statistical comparisons between clusters.

3.1 Cluster Identification and Distribution

Application of the *k*-means algorithm with $k = 4$ resulted in four distinct patient clusters with varying sizes. Cluster 3 was the largest, comprising 169 patients (44.1%), followed by Cluster 1 with 88 patients (23.0%), Cluster 0 with 69 patients (18.0%), and Cluster 2 with 57 patients (14.9%). Principal component analysis (PCA) was applied to visualize the clusters in a two-dimensional space (**Figure 3.1**). The PCA plot demonstrates clear spatial separation between clusters, particularly between Cluster 2 and the remaining groups, suggesting distinct phenotypic patterns. Clusters 0, 1, and 3 show some degree of overlap in the reduced dimensional space, though they remain distinguishable.

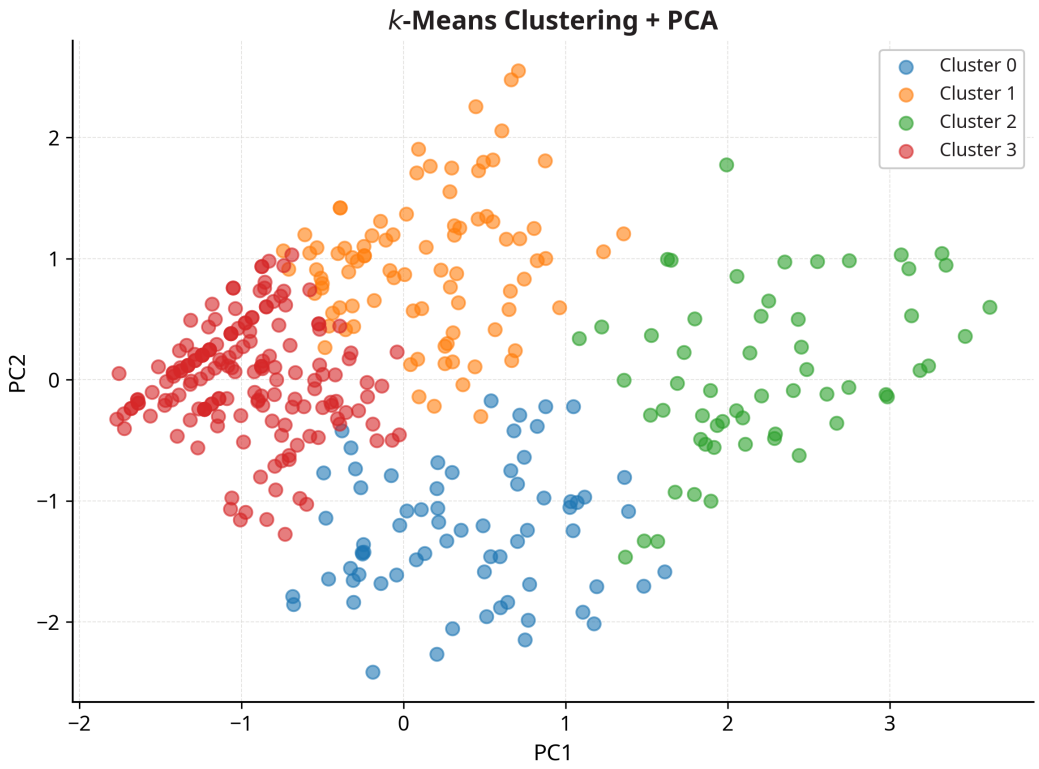

Figure 3.1: *k*-Means clustering visualization in principal component space. The four identified clusters are displayed in two dimensions using the first two principal components (PC1 and PC2).

Comprehensive characterization of clinical and pathological features across clusters revealed distinct phenotypic profiles (**Figure 3.2**). Cluster 0 comprised predominantly young female patients (median age 33 years, 76.8% female) with intermediate risk disease (92.8%),

predominantly multi-focal tumors (59.4%), and poor treatment response (58.0% structural incomplete response). Cluster 1 consisted of middle-aged female patients (median age 51 years, 89.8% female) with predominantly low-risk (94.3%), uni-focal disease (78.4%), minimal adenopathy (98.9% absent), and excellent treatment response (61.4%). Cluster 2 was characterized by older patients with the highest median age (62 years), a higher proportion of males (52.6%), elevated smoking rates (54.4% current smokers), predominantly multi-focal disease (87.7%), advanced stage distribution (49.1% Stage II or higher), and poor treatment outcomes (70.2% structural incomplete response). Cluster 3 represented the youngest patient population (median age 30 years), predominantly female (90.5%), with low-risk disease (95.9%), uni-focal tumors (84.6%), all Stage I (100%), and favorable treatment response (82.8% excellent response).

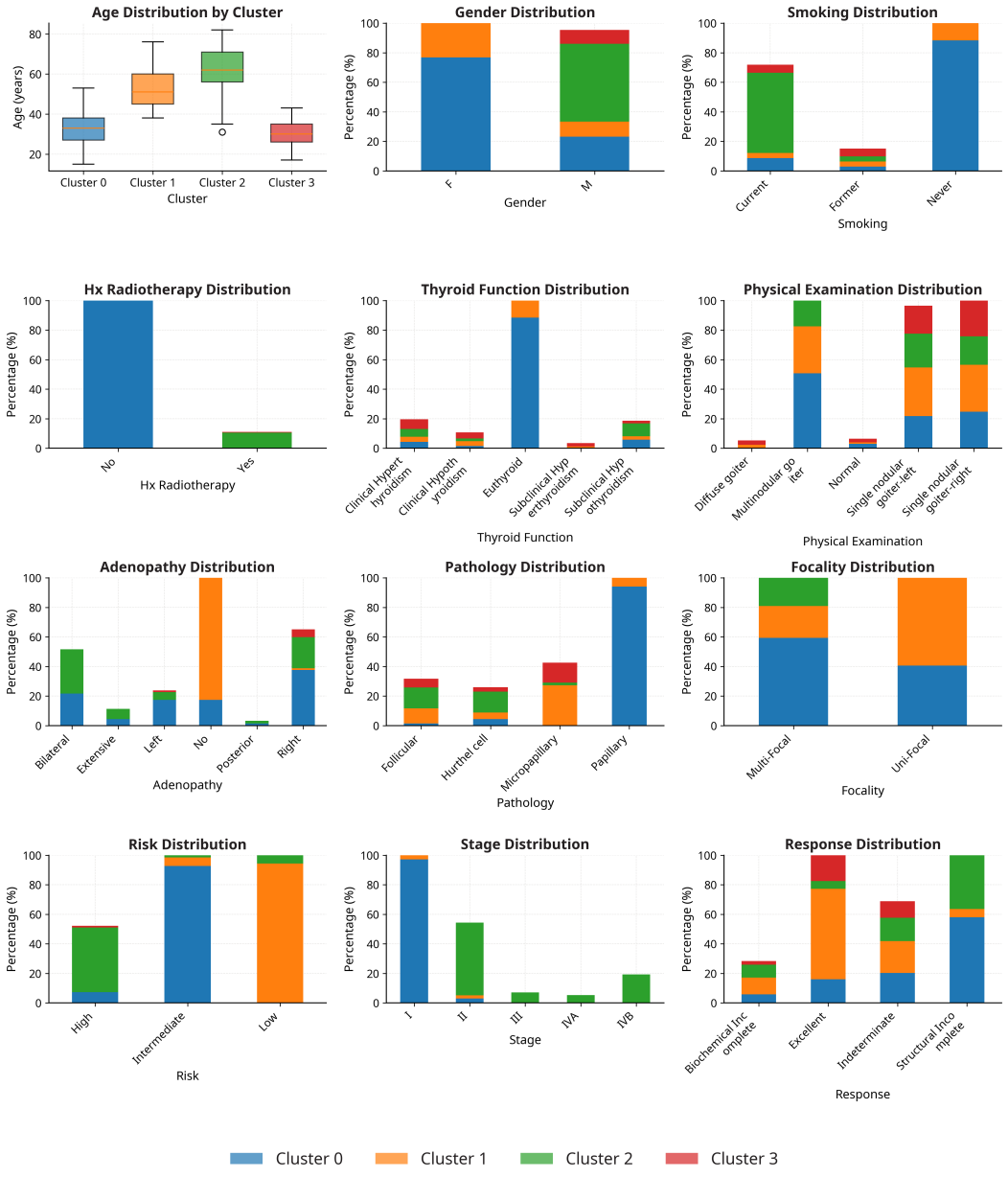

Figure 3.2: Distribution of clinical and pathological features across the four identified clusters. Each subplot shows the distribution of a specific feature, with colors representing different clusters.

3.2 Discriminative Features Across Clusters

To identify which clinical and pathological features most strongly distinguished between the four clusters, statistical testing was performed on all variables. The Kruskal-Wallis test was applied to age (the only continuous variable), while the Chi-squared test was used for all categorical variables. To control for multiple testing, the Benjamini-Hochberg false discovery rate (FDR) correction was applied at $\alpha = 0.05$.

All features except thyroid function demonstrated statistically significant differences across clusters after FDR correction (**Table 3.1**). Risk stratification showed the strongest association with cluster belonging ($\chi^2 = 388.73$, $p < 0.001$), followed by age ($H = 256.13$, $p < 0.001$), disease stage ($\chi^2 = 272.12$, $p < 0.001$), adenopathy status ($\chi^2 = 234.45$, $p < 0.001$), and treatment response ($\chi^2 = 210.51$, $p < 0.001$). Focality ($\chi^2 = 122.41$, $p < 0.001$), smoking status ($\chi^2 = 106.04$, $p < 0.001$), and gender ($\chi^2 = 58.09$, $p < 0.001$) also showed strong discriminative power. Pathology type ($\chi^2 = 55.03$, $p < 0.001$), history of radiotherapy ($\chi^2 = 28.40$, $p < 0.001$), and physical examination findings ($\chi^2 = 42.95$, $p < 0.001$) demonstrated weaker but still significant associations. Thyroid function was the only feature that did not significantly differ across clusters ($\chi^2 = 13.10$, $p = 0.36$).

Table 3.1: Statistical significance of clinical and pathological features in discriminating between clusters. Features are ranked by adjusted *p*-value.

Feature	Test	Statistic	*p*-value	*p*-adj
Risk	Chi-square	388.73	< 0.001	< 0.001
Age	Kruskal-Wallis	256.13	< 0.001	< 0.001
Stage	Chi-square	272.12	< 0.001	< 0.001
Adenopathy	Chi-square	234.45	< 0.001	< 0.001
Response	Chi-square	210.51	< 0.001	< 0.001
Focality	Chi-square	122.41	< 0.001	< 0.001
Smoking	Chi-square	106.04	< 0.001	< 0.001
Gender	Chi-square	58.09	< 0.001	< 0.001
Pathology	Chi-square	55.03	< 0.001	< 0.001
Hx Radiotherapy	Chi-square	28.40	< 0.001	< 0.001
Physical Exam	Chi-square	42.95	< 0.001	< 0.001
Thyroid Function	Chi-square	13.10	0.362	0.362

For visualization and deeper characterization, the top eight most discriminative features (risk, age, stage, adenopathy, response, focality, smoking, and gender) were selected for spider plot analysis based on their statistical significance and clinical relevance. Feature values were standardized using *z*-score transformation to enable direct comparison across different scales, with *z*-scores representing the number of standard deviations above or below the overall mean for each cluster.

The spider plot (**Figure 3.3**) reveals distinct phenotypic profiles for each cluster. Cluster 0 (blue) is characterized by intermediate risk stratification, young age, poor treatment response, and multifocal disease. Cluster 1 (orange) demonstrates low risk stratification, elevated age, excellent treatment response, unifocal disease, predominantly female compo-

sition, and minimal smoking exposure. Cluster 2 (green) shows the highest values for age and stage, with male predominance, heavy smoking exposure, poor treatment response, and multifocal disease. Cluster 3 (red) exhibits low risk stratification, the youngest age, excellent treatment response, unifocal disease, predominantly female composition, and minimal smoking exposure.

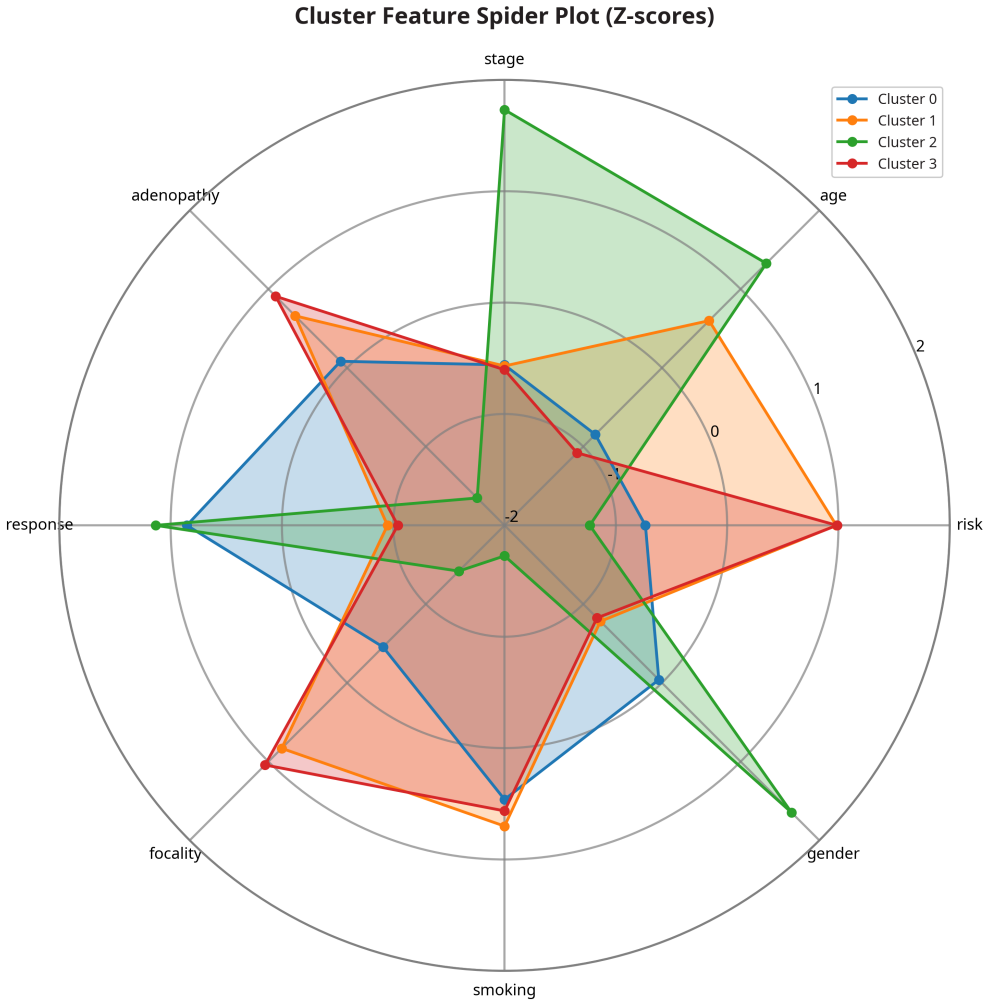

Figure 3.3: Spider plot showing standardized feature profiles (z-scores) across the four clusters. Each axis represents one of the eight most discriminative features, and each colored line represents the profile of one cluster. Positive z-scores indicate above-average values, while negative z-scores indicate below-average values for that cluster.

3.3 Statistical Comparison Between Cluster Pairs

Following the identification of features that discriminate across all clusters, pairwise statistical comparisons were conducted to determine which specific cluster pairs differed significantly on each feature. A total of 48 pairwise comparisons were performed (6 cluster pairs × 8 features), with Mann-Whitney U tests applied to age and Chi-squared tests applied to categorical variables. False discovery rate correction was applied separately for each feature to account for multiple testing. Of the 48 comparisons, 36 (75.0%) showed statistically significant differences after FDR adjustment at $\alpha = 0.05$.

Age emerged as the most discriminative feature, showing significant differences across all six cluster pairs (100% discrimination power). Risk stratification, adenopathy status, treatment response, and focality each differentiated five of six pairs (83% discrimination). Gender showed moderate discrimination (4/6 pairs, 67%), while stage and smoking exposure each distinguished three of six pairs (50% discrimination). The complete pattern of pairwise comparisons is visualized in **Figure 3.4**, which displays FDR-adjusted *p*-values as a heatmap with features ordered by discrimination power.

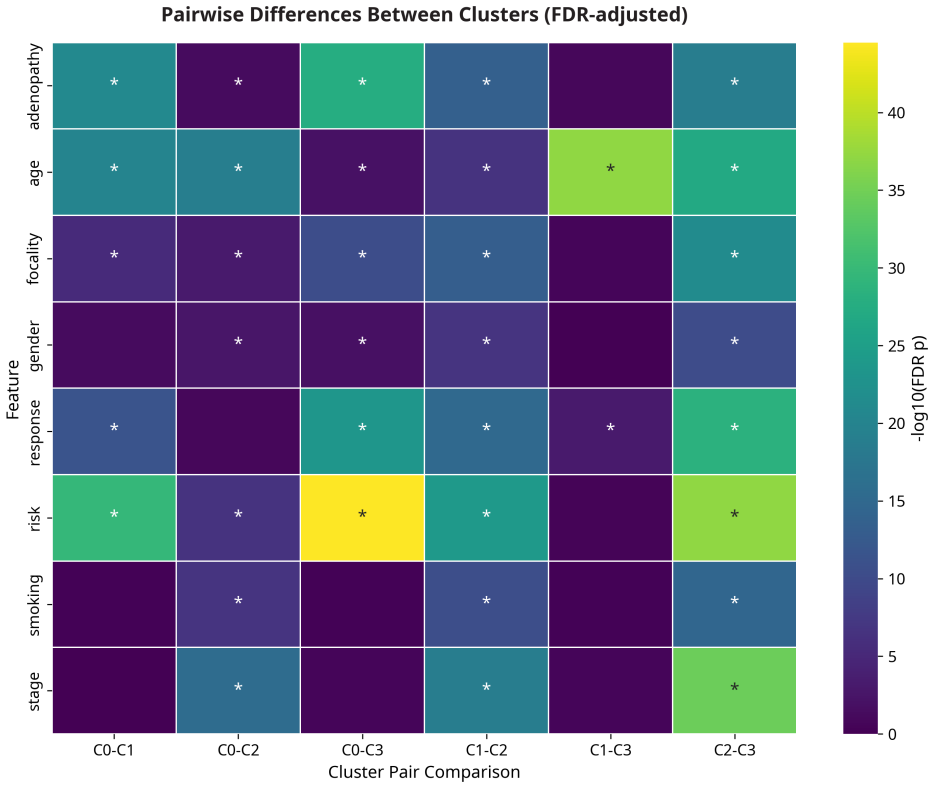

Figure 3.4: Heatmap of pairwise statistical comparisons between clusters. Colors represent $-\log_{10}$(FDR-adjusted p-value), with darker colors indicating stronger statistical significance. Asterisks (*) mark comparisons that remain significant after FDR correction at $\alpha = 0.05$.

The most different cluster pairs were Clusters 0-3 and Clusters 2-3, which differed significantly on multiple features, as evident in the heatmap where these comparisons show bright yellow-green coloring. Cluster 0-3 showed particularly strong divergence on risk stratification (brightest yellow signal), while Cluster 2-3 demonstrated significant differences across age, risk, response, and stage. Cluster 3 consists of the youngest patients with Stage I, low-risk disease and favorable outcomes, whereas Cluster 2 exhibits the most adverse profile across all measured variables: oldest age, male predominance, heavy smoking exposure, multifocal disease, advanced stage, and poor treatment outcomes. Similarly, Cluster 0 represents young patients but with intermediate-risk features and poor treatment response, creating stark contrasts with Cluster 3's favorable phenotype.

Clusters 0 and 2 also demonstrated substantial separation, differing on seven of eight features (87.5%). Although both clusters exhibit poor treatment response (Cluster 0: 58.0% structural incomplete; Cluster 2: 70.2% structural incomplete), they are distinguished by

age (Cluster 0 median 33 years vs. Cluster 2 median 62 years), gender distribution (Cluster 0: 76.8% female vs. Cluster 2: 52.6% male), smoking exposure (Cluster 0: 88.4% never vs. Cluster 2: 54.4% current), and disease stage (Cluster 0: 97.1% Stage I vs. Cluster 2: 49.1% Stage II or higher). The only feature on which these clusters did not differ significantly was treatment response, suggesting that poor outcomes can happen in both young patients with intermediate-risk pathology and older patients with lifestyle risk factors and advanced disease.

In contrast, Clusters 1 and 3 showed the greatest similarity, differing on only two of eight features (25% divergence): age and treatment response. This is clearly visible in the heatmap, where the C1-C3 comparison column shows predominantly dark purple coloring with only two brighter cells, indicating most features are not significantly different. Both clusters are characterized by predominantly female composition (Cluster 1: 89.8%; Cluster 3: 90.5%), low ATA risk stratification (Cluster 1: 94.3%; Cluster 3: 95.9%), unifocal disease, minimal adenopathy, and early-stage presentation (both >97% Stage I). The primary distinction lies in age distribution. Cluster 1 represents middle-aged patients (median 51 years) while Cluster 3 comprises the youngest cohort (median 30 years). Additionally, treatment response rates differ, with Cluster 3 exhibiting a higher proportion of excellent responses (82.8%) compared to Cluster 1 (61.4%). This pattern suggests that within the low-risk disease category, age may modulate treatment outcomes, with younger patients achieving more favorable responses even when pathological features are similar.

4. Conclusions

This study demonstrates that unsupervised machine learning can identify clinically meaningful phenotypes among differentiated thyroid cancer patients that extend beyond conventional risk stratification. Four distinct subgroups were identified with recurrence rates ranging from 4.7% to 82.5%, and statistical validation through pairwise comparisons with FDR correction confirmed that these phenotypes represent significantly distinct patient populations.

Key findings include: (1) Cluster 2, characterized by advanced age (median 62 years), high current smoking prevalence (54.4%), and male predominance (52.6% male), demonstrated the highest recurrence rate (82.5%, 47/57 patients) with all distinguishing characteristics reaching statistical significance compared to other clusters ($p < 0.001$); (2) Cluster 0 revealed a clinical paradox where predominantly Stage I patients (97.1%) exhibited 65.2% recurrence (45/69 patients), significantly differing from other Stage I clusters in treatment response and ATA risk classification, with 92.8% classified as intermediate risk and 58.0% experiencing structural incomplete response; (3) Clusters 1 and 3 demonstrated favorable outcomes (9.1% and 4.7% recurrence, respectively) with significantly different profiles from high-risk clusters, suggesting potential candidates for de-intensified surveillance, though differing primarily in age (median 51 years vs. 30 years) and response rates (61.4% vs. 82.8% excellent response).

These findings support the potential utility of multivariate phenotyping to complement existing prognostic frameworks and guide personalized surveillance strategies. The statistical confirmation of inter-cluster differences, with age discriminating all cluster pairs (100%), and risk, adenopathy, response, and focality each discriminating 83% of pairs, strengthens the clinical relevance of these phenotypes. The identification of a young, intermediate-risk phenotype (Cluster 0) with unexpectedly high recurrence rates despite early-stage disease highlights limitations in current staging systems and suggests that pathological features and treatment response patterns may override age-related prognostic advantages in certain patient subgroups. Conversely, the low recurrence rates observed in Clusters 1 and 3 support the potential for risk-adapted surveillance strategies that could reduce unnecessary follow-up intensity in carefully selected low-risk populations.

Several limitations warrant consideration. First, this analysis was conducted on a single institutional cohort, and external validation across diverse populations is necessary before clinical implementation. Second, the unsupervised nature of k-means clustering does not establish causal relationships; the identified phenotypes are descriptive rather than explanatory. Third, the elbow method for cluster selection involves subjective interpretation, and alternative approaches such as silhouette analysis or gap statistic may yield different optimal k values. Fourth, the 10 to 15 year follow-up period, while substantial, may not capture very late recurrences. Finally, unmeasured confounders such as treatment heterogeneity, surgeon experience, radioactive iodine dosing, and iodine supplementation practices may influence outcomes.

Future directions include external validation of the identified phenotypes in independent cohorts to assess generalizability across different clinical settings and populations. Development of a supervised classification model would facilitate assignment of new patients to phenotypes at diagnosis, enabling prospective clinical application. Integration of additional clinical variables not captured in this dataset, such as detailed treatment protocols and imaging findings, could further refine phenotype characterization. Finally, prospective evaluation of phenotype-specific surveillance protocols could assess whether differential follow-up intensity improves patient outcomes while optimizing healthcare resource allocation.

References

[1] Z. Dou, Y. Shi, and J. Jia, "Global burden of disease study analysis of thyroid cancer burden across 204 countries and territories from 1990 to 2019," *Frontiers in Oncology*, vol. 14, p. 1412243, 2024.

[2] S. Vaccarella, S. Franceschi, F. Bray, C. P. Wild, M. Plummer, and L. Dal Maso, "Worldwide thyroid-cancer epidemic? the increasing impact of overdiagnosis," *New England Journal of Medicine*, vol. 375, no. 7, pp. 614–617, 2016.

[3] D. W. Chen, B. H. H. Lang, D. S. A. McLeod, K. Newbold, and M. R. Haymart, "Thyroid cancer," *The Lancet*, vol. 401, no. 10387, pp. 1531–1544, 2023.

[4] K. Leclair, K. J. L. Bell, L. Furuya-Kanamori, S. A. Doi, D. O. Francis, and L. Davies, "Evaluation of gender inequity in thyroid cancer diagnosis: Differences by sex in us thyroid cancer incidence compared with a meta-analysis of subclinical thyroid cancer rates at autopsy," *JAMA Internal Medicine*, vol. 181, no. 10, pp. 1351–1358, 2021.

[5] S. H. Hsieh, S. T. Chen, C. Hsueh, T. C. Chao, and J. D. Lin, "Gender-specific variation in the prognosis of papillary thyroid cancer tnm stages ii to iv," *International Journal of Endocrinology*, vol. 2012, no. 1, p. 379097, 2012.

[6] S. Hu, X. Wu, and H. Jiang, "Trends and projections of the global burden of thyroid cancer from 1990 to 2030," *Journal of Global Health*, vol. 14, pp. 1–15, 2024.

[7] M. D. Ringel, J. A. Sosa, Z. Baloch, *et al.*, "2025 american thyroid association management guidelines for adult patients with differentiated thyroid cancer," *Thyroid*, vol. 35, no. 8, pp. 841–985, 2025.

[8] W. Chatchomchuan, Y. Thewjitcharoen, K. Karndumri, *et al.*, "Recurrence factors and characteristic trends of papillary thyroid cancer over three decades," *International Journal of Endocrinology*, vol. 2021, pp. 1–7, 2021.

[9] J. Hampton, A. Alam, N. Zdenkowski, C. Rowe, E. Fradgley, and C. J. O'Neill, "Fear of cancer recurrence in differentiated thyroid cancer survivors: A systematic review," *Thyroid*, vol. 34, no. 5, pp. 541–558, 2024.

[10] O. F. Bathe and C. Stretch, "Prognostic biomarkers for papillary thyroid cancer: Reducing overtreatment, improving clinical efficiency, and enhancing patient experience," *Technology in Cancer Research & Treatment*, vol. 24, 2025.

[11] M. Kamińska, M. Trofimiuk-Müldner, G. Sokołowski, and A. Hubalewska-Dydejczyk, "Machine learning in endocrinology: current applications and future perspectives," *Endocrine*, vol. 90, no. 2, p. 357, 2025.

[12] E. Onah, U. J. Eze, A. S. Abdulraheem, U. G. Ezigbo, K. C. Amorha, and F. Ntie-Kang, "Optimizing unsupervised feature engineering and classification pipelines for differentiated thyroid cancer recurrence prediction," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, p. 182, 2025.

[13] Y. Jiang, Y. Dang, Q. Wu, B. Yuan, L. Gao, and C. You, "Using a k-means clustering to identify novel phenotypes of acute ischemic stroke and development of its clinlabomics models," *Frontiers in Neurology*, vol. 15, p. 1366307, 2024.

[14] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, 2022.

[15] S. Borzooei and A. Tarokhian, "Differentiated thyroid cancer recurrence." UCI Machine Learning Repository, 2023.

[16] S. Borzooei, G. Briganti, M. Golparian, J. R. Lechien, and A. Tarokhian, "Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study," *European Archives of Oto-Rhino-Laryngology*, 2023. Published online: 30 October 2023.

[17] B. R. Haugen, E. K. Alexander, K. C. Bible, *et al.*, "2015 american thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer," *Thyroid*, vol. 26, no. 1, pp. 1–133, 2016.

[18] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982. Originally published as a Bell Labs technical report in 1957.

[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, University of California Press, 1967.

Annexes

A. Demographic and Clinicopathological Characteristics

Table A.1: Demographic and clinicopathological characteristics of the study cohort (N=383), stratified by recurrence status. Continuous variables are presented as median [IQR] and compared using the Mann-Whitney U test. Categorical variables are presented as n (%) and compared using the Chi-squared test or Fisher's exact test. A p-value < 0.05 is considered statistically significant.

Characteristic	Overall (N=383)	No Recurrence (N=275)	Recurrence (N=108)	*p*-value
Age, years				<0.001
Median [IQR]	37.0 [29.0, 51.0]	36.0 [28.0, 46.0]	44.5 [31.8, 62.0]	
Gender, n (%)				<0.001
Female	312 (81.5)	246 (89.5)	66 (61.1)	
Male	71 (18.5)	29 (10.5)	42 (38.9)	
Smoking Status, n (%)				<0.001
Current	49 (12.8)	16 (5.8)	33 (30.6)	
Former	16 (4.2)	12 (4.4)	4 (3.7)	
Never	318 (83.0)	247 (89.8)	71 (65.7)	
History of Radiotherapy, n (%)				0.002
No	376 (98.2)	274 (99.6)	102 (94.4)	
Yes	7 (1.8)	1 (0.4)	6 (5.6)	
Thyroid Function, n (%)				0.272
Clinical Hyperthyroidism	20 (5.2)	17 (6.2)	3 (2.8)	
Clinical Hypothyroidism	12 (3.1)	10 (3.6)	2 (1.9)	
Euthyroid	332 (86.7)	234 (85.1)	98 (90.7)	
Subclinical Hyperthyroidism	5 (1.3)	5 (1.8)	0 (0.0)	
Subclinical Hypothyroidism	14 (3.7)	9 (3.3)	5 (4.6)	
Physical Examination, n (%)				0.011
Diffuse goiter	7 (1.8)	7 (2.5)	0 (0.0)	
Multinodular goiter	140 (36.6)	88 (32.0)	52 (48.1)	
Normal	7 (1.8)	5 (1.8)	2 (1.9)	
Single nodular goiter-left	89 (23.2)	63 (22.9)	26 (24.1)	
Single nodular goiter-right	140 (36.6)	112 (40.7)	28 (25.9)	
Adenopathy, n (%)				<0.001
Bilateral	32 (8.4)	5 (1.8)	27 (25.0)	
Extensive	7 (1.8)	0 (0.0)	7 (6.5)	
Left	17 (4.4)	5 (1.8)	12 (11.1)	
No	277 (72.3)	247 (89.8)	30 (27.8)	
Posterior	2 (0.5)	0 (0.0)	2 (1.9)	
Right	48 (12.5)	18 (6.5)	30 (27.8)	
Pathology, n (%)				<0.001
Follicular	28 (7.3)	16 (5.8)	12 (11.1)	
Hurthel cell	20 (5.2)	14 (5.1)	6 (5.6)	
Micropapillary	48 (12.5)	48 (17.5)	0 (0.0)	
Papillary	287 (74.9)	197 (71.6)	90 (83.3)	
Focality, n (%)				<0.001
Multi-Focal	136 (35.5)	66 (24.0)	70 (64.8)	
Uni-Focal	247 (64.5)	209 (76.0)	38 (35.2)	

Continued on next page...

Table A.1: Demographic and clinicopathological characteristics (Continued).

Characteristic	Overall (N=383)	No Recurrence (N=275)	Recurrence (N=108)	*p*-value
Risk, n (%)				<0.001
High	32 (8.4)	0 (0.0)	32 (29.6)	
Intermediate	102 (26.6)	38 (13.8)	64 (59.3)	
Low	249 (65.0)	237 (86.2)	12 (11.1)	
Tumor Stage (T), n (%)				<0.001
T1a	49 (12.8)	48 (17.5)	1 (0.9)	
T1b	43 (11.2)	38 (13.8)	5 (4.6)	
T2	151 (39.4)	131 (47.6)	20 (18.5)	
T3a	96 (25.1)	55 (20.0)	41 (38.0)	
T3b	16 (4.2)	2 (0.7)	14 (13.0)	
T4a	20 (5.2)	1 (0.4)	19 (17.6)	
T4b	8 (2.1)	0 (0.0)	8 (7.4)	
Metastasis (M), n (%)				<0.001
M0	365 (95.3)	275 (100.0)	90 (83.3)	
M1	18 (4.7)	0 (0.0)	18 (16.7)	
TNM Stage, n (%)				<0.001
I	333 (86.9)	268 (97.5)	65 (60.2)	
II	32 (8.4)	7 (2.5)	25 (23.1)	
III	4 (1.0)	0 (0.0)	4 (3.7)	
IVA	3 (0.8)	0 (0.0)	3 (2.8)	
IVB	11 (2.9)	0 (0.0)	11 (10.2)	
Response to Therapy, n (%)				<0.001
Biochemical Incomplete	23 (6.0)	12 (4.4)	11 (10.2)	
Excellent	208 (54.3)	207 (75.3)	1 (0.9)	
Indeterminate	61 (15.9)	54 (19.6)	7 (6.5)	
Structural Incomplete	91 (23.8)	2 (0.7)	89 (82.4)	