

Phenotypic Clustering of Differentiated Thyroid Cancer Patients Using k -Means

David Cabezas Antolin,
Frances Scarlett Thomas,
Ravneet-Rahul Sandhu Singh,
Roger Puig I Arxer,
Sofía González Estrada

Master in Health Data Science — MHEDAS

January 12, 2025



1. Introduction	4
Clinical Context	4
The Problem	5
Research Objectives	6
2. Methodology	7
Dataset	7
Feature Engineering	8
Elbow Method	9
Cluster Characterization	10

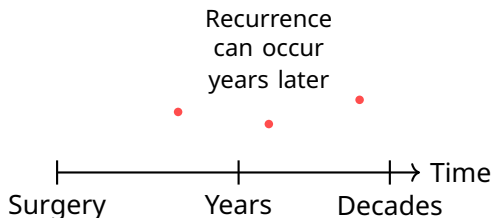
3. Results	11
Cluster Identification	11
Pairwise Statistical Comparisons	13
Recurrence Risk Stratification	14
4. Conclusions	15

Differentiated Thyroid Cancer (DTC):

- Most common endocrine malignancy
- 167% increase in incidence (past decades)
- High survival rates post-surgery
- Recurrence remains a major concern

Key Challenge:

- Balance oncological surveillance
- Minimize treatment morbidity
- Recurrence: 1.6% (low-risk) to 22.7% (high-risk)

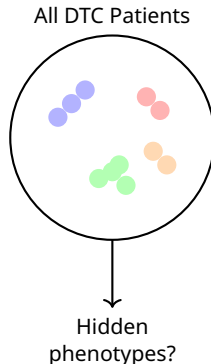


Current Limitations:

- Traditional risk stratification insufficient
- Prediction tools lack generalizability
- Limited accuracy for specific subgroups
- Inefficient surveillance pathways

Impact on Healthcare:

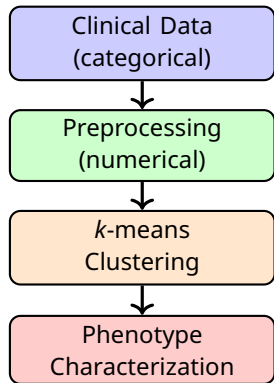
- Psychological & financial burden on survivors
- Strained health systems
- Need for personalized risk assessment



Aim: Apply k -means clustering to discover distinct phenotypic subgroups in DTC survivors

Specific Objectives:

- **Objective 1:** Preprocess clinicopathologic data (transform categorical → numerical features)
- **Objective 2:** Determine optimal number of clusters (Elbow method, Silhouette analysis)
- **Objective 3:** Characterize clusters through statistical comparison & recurrence analysis



Data Source:

- UCI Machine Learning Repository
- 15-year cohort study
- Well-differentiated thyroid cancer

Cohort Characteristics:

- N = 383 patients
- 16 variables (demographic, clinical, pathological)
- 28.2% recurrence rate
- 81.5% female, median age 39 years

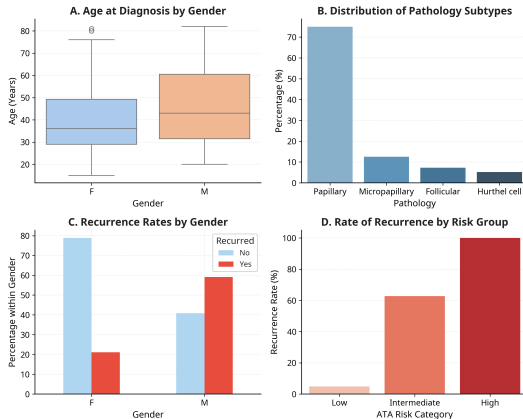


Figure: Demographic distributions of the cohort.

Actions Taken:

- Combined smoking variables (Never/Former/Current)
- Removed TNM components (T, N, M)
- Kept Stage (composite metric)
- Kept ATA risk (distinct from Stage)
- Retained Adenopathy over N stage

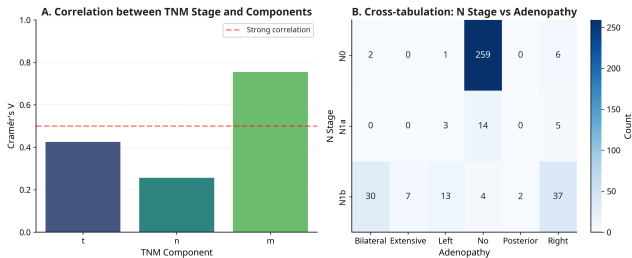


Figure: Feature correlation matrix using Cramér's V.

Algorithm:

- Partitional clustering method
- Minimizes within-cluster sum of squares (WCSS)
- Iterative: assignment \rightarrow update

Objective function:

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

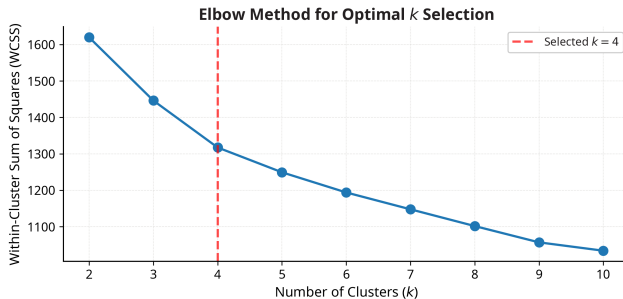


Figure: Elbow plot showing WCSS vs. number of clusters.

Statistical Analysis:

- **Feature importance:** Kruskal-Wallis (age), Chi-squared (categorical)
- **Pairwise comparisons:** Mann-Whitney U (age), Chi-squared (categorical)
- **Multiple testing correction:** Benjamini-Hochberg ($\text{FDR} < 0.05$)

Spider Plot

Standardized cluster profiles

Pairwise Heatmap

FDR-adjusted p -values

Four Distinct Clusters Identified:

- **Cluster 0:** 69 patients (18.0%)
- **Cluster 1:** 88 patients (23.0%)
- **Cluster 2:** 57 patients (14.9%)
- **Cluster 3:** 169 patients (44.1%)

Key Observation:

- Clear spatial separation in PCA space
- Cluster 2 most distinct
- Some overlap between Clusters 0, 1, 3

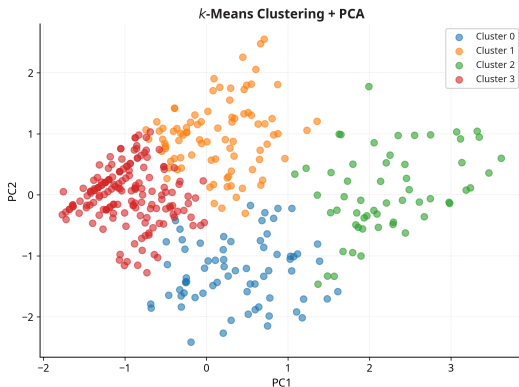


Figure: PCA visualization of the four clusters.

Most Discriminative Features:

- **Risk stratification** ($\chi^2 = 388.73$)
- **Age** ($H = 256.13$)
- **Stage** ($\chi^2 = 272.12$)
- **Adenopathy** ($\chi^2 = 234.45$)
- **Response** ($\chi^2 = 210.51$)
- **Focality** ($\chi^2 = 122.41$)
- **Smoking** ($\chi^2 = 106.04$)
- **Gender** ($\chi^2 = 58.09$)

All $p < 0.001$ after FDR correction

Only thyroid function was non-significant

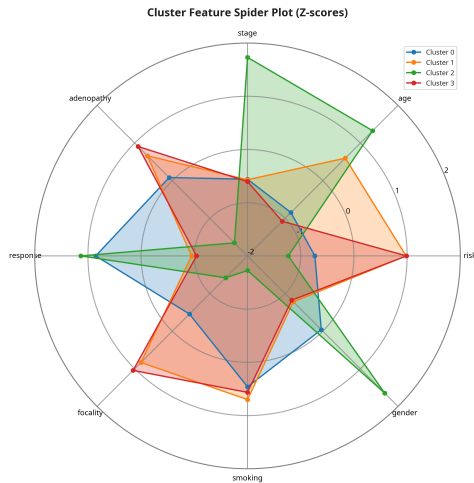


Figure: Spider plot: standardized feature profiles (z-scores) across clusters.

Key Findings:

- **36/48** comparisons significant (75%)
- **Age:** discriminated all 6 pairs (100%)
- **Most different:** C0-C3, C2-C3
- **Most similar:** C1-C3 (only age & response differ)

Clinical Insights:

- Poor outcomes via different pathways
- Age modulates treatment response in low-risk disease
- Clear phenotypic gradient

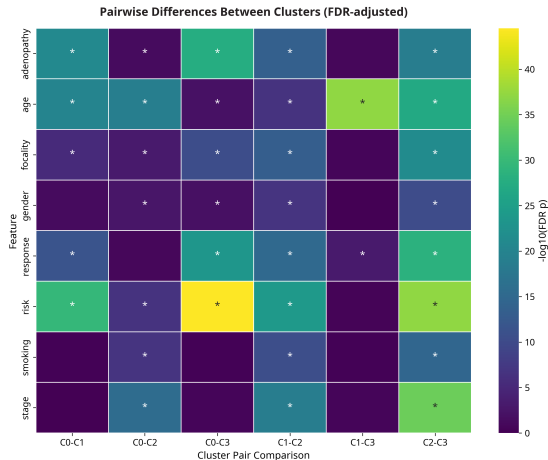


Figure: Heatmap: $-\log_{10}(\text{FDR } p\text{-value})$. Brighter = more significant.



Clear phenotypic gradient: Cluster 3 (youngest, low-risk, excellent response) → Cluster 1 (middle-aged, low-risk) → Cluster 0 (young, intermediate-risk, poor response) → Cluster 2 (oldest, multiple adverse features)

Key Findings

- Identified **4 clinically distinct phenotypes** (recurrence: 4.7%–82.5%)
- **Cluster 2** (older, male, smokers): 82.5% recurrence
- **Cluster 0** (young, Stage I): paradoxical 65.2% recurrence
- **Clusters 1 & 3**: favorable outcomes (9.1%, 4.7%)

Clinical Implications

- Complements risk stratification
- Enables personalized surveillance

Future Directions

- External validation
- Clinical application

Phenotypic Clustering of Differentiated Thyroid Cancer Patients Using k -Means

David Cabezas Antolin,
Frances Scarlett Thomas,
Ravneet-Rahul Sandhu Singh,
Roger Puig I Arxer,
Sofía González Estrada

Master in Health Data Science — MHEDAS

January 12, 2025

