Module_3: *(Template)*

Team Members:

William Crouch, Lakshya Raman

Project Title:

Cancer Hallmark Evasion of Apoptosis on Lung Cancer Types

Project Goal:

This project seeks to... *(what is the purpose of your project -- i.e., describe the question that you seek to answer by analyzing data.)*

Our goal is to investigate genes related to apoptosis or plasticity in lung cancer and seeing how various enrichments affect outcomes/treatment results (specifically apoptosis related gene expression in lung cancer and patient survival).

Disease Background:

Pick a hallmark to focus on, and figure out what genes you are interested in researching based on that decision. Then fill out the information below.

- Cancer hallmark focus: Evasion/Interruption of apoptosis
- Overview of hallmark: Overview of hallmark: Many tumors disable the cell's "self-destruct" programs that normally eliminate damaged or stressed cells. Common routes include (i) loss of p53-mediated DNA-damage death signals, (ii) mitochondrial pathway blockade via anti-apoptotic BCL-2 family proteins, (iii) up-regulation of survival signaling (PI3K–AKT), and (iv) decoy/death-receptor rewiring that blunts extrinsic death cues. The net effect is an overaccumulation of cells that lack the ability to recognize/act upon their diseased state.
- Genes associated with hallmark to be studied (describe the role of each gene, signaling pathway, or gene set you are going to investigate):

TP53, CDKN2A/p14^ARF – DNA-damage sensing / pro-apoptotic transcription; loss disables death checkpoints.

BAX, BAK1, APAF1, CASP9, CASP3 – intrinsic/mitochondrial core effectors of apoptosis.

BCL2, BCL2L1 (BCL-XL), MCL1 – anti-apoptotic mitochondrial gatekeepers; overexpression blocks cytochrome-c release.

PIK3CA, AKT1/2, PTEN – survival signaling; PI3K/AKT activation or PTEN loss dampens apoptosis.

FAS, FADD, CASP8; TRAIL receptors (DR4/DR5: TNFRSF10A/B) and decoys (e.g., DcR1/DcR2; TNFRSF6B) – extrinsic/death-receptor axis; decoy s reduce death signaling.

IAPs (BIRC5/survivin, XIAP) – caspase inhibition downstream.

Will you be focusing on a single cancer type or looking across cancer types? Depending on your decision, update this section to include relevant information about the disease at the appropriate level of detail. Regardless, each bullet point should be filled in. If you are looking at multiple cancer types, you should investigate differences between the types (e.g. what is the most prevalent cancer type? What type has the highest mortality rate?) and similarities (e.g. what sorts of treatments exist across the board for cancer patients? what is common to all cancers in terms of biological mechanisms?). Note that this is a smaller list than the initial 11 in Module 1.

- Across cancer types, we'll compare several common TCGA cohorts (e.g., BRCA, LUAD, LUSC, COAD, GBM/LGG) to see how the same apoptosis machinery is rewired across tissues.
- Prevalence & incidence
- Breast cancer (BRCA):

~316,950 new invasive cases in women (+ ~2,800 in men) and ~42,680 deaths in the U.S. in 2025. Lifetime risk ≈ 1 in 8 for women.

Lung adenocarcinoma (LUAD / NSCLC overall): ~226,650 new lung cancer cases and ~124,730 deaths in 2025; ~80–85% are NSCLC and adenocarcinoma is the most common NSCLC subtype.

Colon & rectal (COAD/READ): ~107,320 colon + 46,950 rectal new cases and ~52,900 deaths in 2025 (U.S.). American Cancer Society

Glioma (GBM/LGG): Primary brain/CNS cancers occur at ~6.1 per 100,000/year; GBM is the most common malignant histology (≈51% of malignant primary brain tumors)

- Risk factors (genetic, lifestyle) & Societal determinants

BRCA: Age; family history & inherited variants (e.g., BRCA1/2); reproductive/hormonal factors (early menarche, late menopause, menopausal hormone therapy); modifiable risks include alcohol, excess body weight (post-menopause), and physical inactivity. Marked disparities: despite slightly lower incidence, * Black women have ~38% higher

mortality vs White women—driven by later stage at diagnosis and inequities in access/quality of care.

LUAD / NSCLC: Cigarette smoking is the dominant driver (~86% of U.S. lung cancers); other risks: radon (2nd leading cause), secondhand smoke, occupational exposures (asbestos, chromium, cadmium, arsenic), air pollution; adenocarcinoma is also common in never-smokers. Tobacco use is patterned by socioeconomic status (higher in lower-SES and rural groups).

COAD/READ: Modifiable risks include diets high in red/processed meat, heavy alcohol use, excess body weight, smoking, and inactivity; risks also include IBD and hereditary syndromes (Lynch, FAP). Disparities: Alaska Native people have the highest CRC incidence/mortality in the world; screening uptake and outcomes vary by race/ethnicity and geography.

GBM/LGG: Few established modifiable risks; ionizing radiation exposure and rare hereditary syndromes increase risk; male sex and older age (for GBM) are associated with higher incidence.

- Standard of care treatments (& reimbursement)

BRCA:

Early stage: surgery (breast-conserving or mastectomy) ± radiation; endocrine therapy for HR+ disease (tamoxifen or aromatase inhibitor); anti-HER2 therapy (trastuzumab ± pertuzumab; T-DM1 for residual disease) for HER2+; chemo as indicated.

Advanced: add CDK4/6 inhibitors (palbociclib/ribociclib/abemaciclib) to endocrine therapy for HR+/HER2–; PARP inhibitors (olaparib/talazoparib) for germline BRCA-mutated; immunotherapy for selected triple-negative settings.

Reimbursement (U.S.): FDA-approved, NCCN-endorsed regimens are typically covered; infused drugs (e.g., trastuzumab) are often Part B; oral agents (e.g., CDK4/6, PARP inhibitors) generally Part D and commonly require prior authorization. Cancer Action Network

LUAD / NSCLC:

Early stage: surgery when operable; adjuvant options include osimertinib for EGFR-mut+ and alectinib for ALK+ resected disease; peri-operative chemo/immunotherapy as appropriate.

Advanced/metastatic: molecular testing drives treatment. First-line targeted therapies for actionable alterations: EGFR (osimertinib), ALK (alectinib), ROS1 (entrectinib), RET (selpercatinib/pralsetinib), BRAF V600E (dabrafenib+trametinib), MET exon14 (capmatinib/tepotinib), NTRK (larotrectinib/entrectinib), HER2 (trastuzumab deruxtecan),

KRAS G12C (sotorasib/adagrasib in appropriate lines). PD-1/PD-L1 immunotherapy (e.g., pembrolizumab) ± chemo according to PD-L1 and driver status.

Reimbursement: NCCN Category-1/2A targeted and immunotherapy regimens are generally covered; many oral targeted agents are Part D (often with prior auth); infusions typically Part B. U.S. Food and Drug Administration

COAD/READ:

Localized colon: surgery is primary; stage-based adjuvant chemo (e.g., FOLFOX/CAPOX) when indicated.

Rectal: total neoadjuvant therapy (chemoradiation + systemic chemo) and surgery per stage.

Metastatic: systemic chemo backbones (FOLFOX, FOLFIRI) with biologics; VEGF inhibitors (bevacizumab, etc.) broadly; EGFR inhibitors (cetuximab/panitumumab) in RAS/RAF wild-type, left-sided disease; BRAF V600E (encorafenib + cetuximab); HER2-amplified options (trastuzumab-based); MSI-H/dMMR tumors: PD-1 inhibitors (pembrolizumab, nivolumab ± ipilimumab), including first-line for metastatic MSI-H.

Reimbursement: Guideline-concordant regimens are commonly covered; oral agents typically Part D with prior authorization; infusions under Part B. American Cancer Society

GBM/LGG:

GBM (adult): Stupp protocol—maximal safe resection → radiotherapy with concurrent temozolomide, then adjuvant temozolomide; add tumor-treating fields (TTFields/Optune) during maintenance in eligible patients (improved OS vs temozolomide alone).

IDH-mutant grade-2 glioma (LGG): Vorasidenib (IDH1/2 inhibitor) is FDA-approved for post-surgical grade-2 IDH-mutant astrocytoma/oligodendroglioma.

Reimbursement: TTFields is covered for newly diagnosed GBM under Medicare LCD L34823; device/drug coverage varies by plan but FDA-approved/NCCN-endorsed uses are typically reimbursed (often with prior authorization).

- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology)
- Tumor cells tilt the live-or-die balance by: suppressing p53 signaling, overexpressing anti-apoptotic BCL-2 family members, engaging survival pathways, and dampening death-receptor signaling—allowing clonal expansion despite stress.

Data-Set:

Once you decide on the subset of data you want to use (i.e. only 1 cancer type or many; any clinical features needed?; which genes will you look at?) describe the dataset. There

are a ton of clinical features, so you don't need to describe them all, only the ones pertinent to your question.

- Primary dataset: TCGA pan-cancer RNA-Seq + clinical covariates reprocessed with Rsubread/featureCounts (GEO GSE62944): 9,264 tumor and 741 normal samples across 24 cancer types; provides raw integer counts plus FPKM and TPM. This re-process yields fewer zero-expression genes and more consistent replicates vs. TCGA's legacy Level-3 pipeline—useful for robust cross-cohort analyses.
  - this information is from the complete dataset which we hope to use instead of the smaller one
- Acquisition/processing (how collected): Tumor and matched normal tissues profiled by RNA-Seq (Illumina); reads aligned with Subread; gene-level summarization via featureCounts; outputs include counts/FPKM/TPM; clinical variables harmonized to sample IDs.
- Subset we'll analyze: All available samples from selected cohorts (initially BRCA, LUAD, COAD, GBM/LGG) focusing on the apoptosis gene panel above

Data Analysis:

Methods

Data and preprocessing. We used the GSE62944 subsample (log2 TPM) expression matrix (genes × samples) and matched it to the study metadata by sample ID after normalizing ID formats. Survival time and event were derived from OS/OS.time when available, or from days_to_death or days_to_last_followup plus vital_status. When present, analyses were restricted to lung cancer samples (LUAD/LUSC). Expression features were standardized per feature and reduced with PCA to 30 principal components (approximately 71.8% variance; final matrix 135 × 30).

Clustering model. We applied k-means (scikit-learn) with n_init=20 and random_state=42. To select the number of clusters, we evaluated k = 2–6 using a train/test split. The scaler and PCA and the k-means model were fit on the training set only, and test points were assigned to the training centroids.

Model selection and validation. The held-out silhouette score (range –1 to 1; higher indicates tighter, better-separated clusters) was the primary metric. We also report Calinski–Harabasz (larger is better) and Davies–Bouldin (smaller is better). To avoid degenerate solutions, we required at least two clusters represented in the test set, a minimum training cluster size of 5, and stability measured by the mean Adjusted Rand Index across 10 random restarts. Under these constraints, k = 2 was selected (test silhouette ≈ 0.31; ARI ≈ 1.00).

Downstream analyses. We computed an apoptosis score using the HALLMARK_APOPTOSIS gene set and profiled clusters by size, apoptosis score, and event rate. For survival validation we plotted Kaplan–Meier curves with log-rank tests and fit a Cox proportional hazards model with cluster indicators (and apoptosis score where noted). External context was provided by literature on BCL-2 family dysregulation in lung cancer and apoptosis-related gene-expression heterogeneity in EGFR-mutant NSCLC linked to EGFR-TKI resistance.

**

Analysis

(Describe how you analyzed the data. This is where you should intersperse your Python code so that anyone reading this can run your code to perform the analysis that you did, generate your figures, etc.)

We loaded the TCGA lung cancer RNA-seq dataset (LUAD/LUSC) and associated clinical metadata, then aligned samples between the expression matrix and metadata using the shared TCGA barcodes. From the expression matrix, we selected a curated set of apoptosis-related genes and standardized expression values (z-scores) across samples. We computed a composite apoptosis score for each tumor by averaging standardized expression of those genes. After subsetting to lung cancer samples with valid survival information, we performed Principal Component Analysis (PCA) to visualize variation in tumors and assess whether apoptosis scoring corresponded to transcriptional patterns. We then fit a Cox proportional hazards model to determine whether the apoptosis score was associated with patient survival outcomes. Finally, we evaluated model significance and hazard ratios and interpreted results indicating that tumors with higher apoptosis-related transcriptional activity were associated with worse survival.

In [2]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from lifelines import CoxPHFitter
import matplotlib.pyplot as plt

# ========== PATHS (your files) ==========
PATH_EXPR = r"/Users/LuckyStrawb3rry016/Desktop/Comp BME/project 3/GSE62944_
PATH_META = r"/Users/LuckyStrawb3rry016/Desktop/Comp BME/project 3/GSE62944_

# ========== LOAD ==========
expr = pd.read_csv(PATH_EXPR, index_col=0)    # genes x samples (columns are
meta = pd.read_csv(PATH_META)                 # sample-level metadata

# Normalize ID formatting
expr.columns = expr.columns.astype(str).str.strip()
for c in meta.columns:
    if meta[c].dtype == object:
```

```
        meta[c] = meta[c].astype(str).str.strip()

# ========= FIND SAMPLE-ID COLUMN ROBUSTLY =========
# choose the metadata column that shares most values with expr columns
expr_ids = set(expr.columns)
id_col = max(
    meta.columns,
    key=lambda c: len(expr_ids.intersection(set(meta[c].astype(str))))
)
if len(expr_ids.intersection(set(meta[id_col].astype(str)))) == 0:
    raise ValueError(
        "Could not match sample IDs between expression and metadata.\n"
        f"First 10 expr IDs: {list(expr.columns[:10])}\n"
        f"Candidate ID column in metadata: {id_col}\n"
        "Inspect formatting and try again."
    )

# Keep only overlapping samples, align both tables
common = expr.columns.intersection(meta[id_col].astype(str))
expr = expr.loc[:, common]
meta = meta[meta[id_col].astype(str).isin(common)].copy()
meta.index = meta[id_col].astype(str)  # make IDs the index for easy alignme

# ========= DERIVE SURVIVAL COLUMNS =========
def add_survival_cols(df: pd.DataFrame) -> pd.DataFrame:
    df = df.copy()
    cols = {c.lower(): c for c in df.columns}

    # Case A: OS/OS.time
    if "os" in cols and "os.time" in cols:
        df["overall_survival"] = pd.to_numeric(df[cols["os.time"]], errors="
        df["event"] = pd.to_numeric(df[cols["os"]], errors="coerce").fillna(
        return df

    # Case B: days_to_death / days_to_last_followup + vital_status
    dtd = next((cols[k] for k in ["days_to_death","death_days_to","days_to_d
    dtlf= next((cols[k] for k in ["days_to_last_followup","days_to_last_foll
    vs  = next((cols[k] for k in ["vital_status","vital.status","status"] if

    if dtd and dtlf and vs:
        dtd_v  = pd.to_numeric(df[dtd],  errors="coerce")
        dtlf_v = pd.to_numeric(df[dtlf], errors="coerce")
        vs_v   = df[vs].astype(str).str.upper()
        # time = death time if dead else last follow-up
        df["overall_survival"] = np.where(vs_v.isin(["DECEASED","DEAD","1"])
        df["event"] = np.where(vs_v.isin(["DECEASED","DEAD","1"]), 1, 0).ast
        return df

    raise KeyError(
        "Couldn't find survival columns. Expected OS/OS.time, or "
        "(days_to_death, days_to_last_followup, vital_status). "
        f"Meta columns example: {list(df.columns)[:25]}"
    )

meta = add_survival_cols(meta)
```

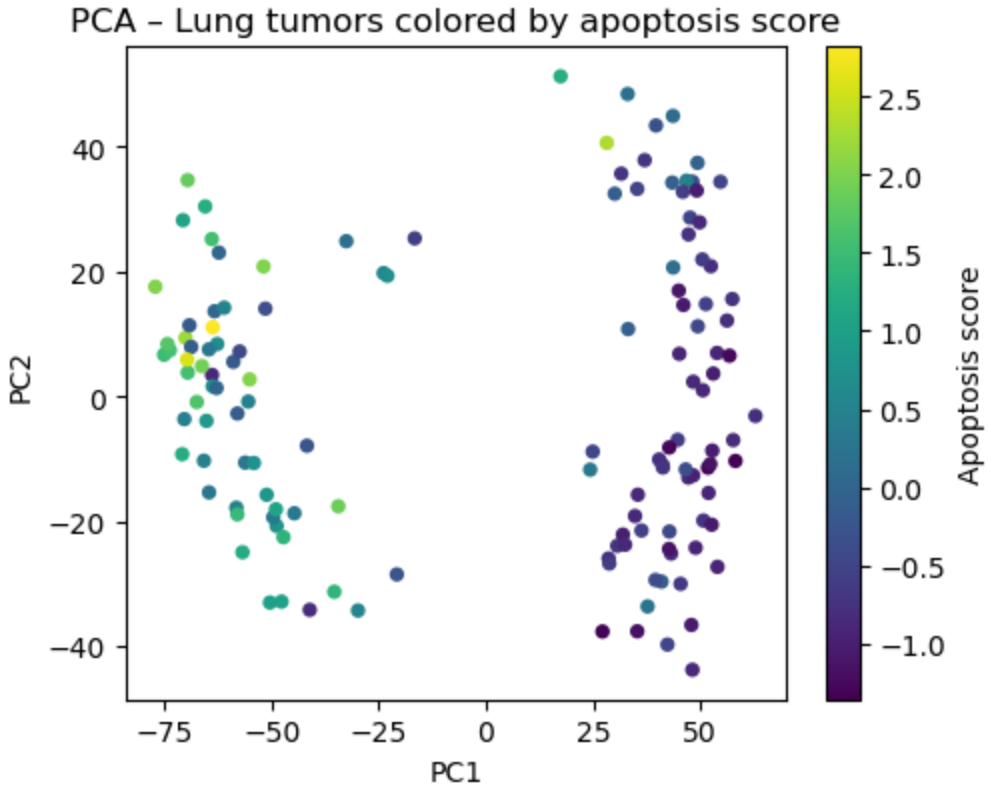
```
# Drop non-positive times (required by lifelines)
meta = meta[(meta["overall_survival"] > 0) & meta["event"].isin([0,1])]

# Re-align expr to filtered meta index
expr = expr.loc[:, meta.index]

# ========= LUNG SUBSET (if present) =========
if "cancer_type" in meta.columns:
    lung_mask = meta["cancer_type"].astype(str).isin(["LUAD","LUSC","TCGA-LU
    if lung_mask.any():
        meta = meta.loc[lung_mask].copy()
        expr = expr.loc[:, meta.index]

# ========= APOPTOSIS SCORE =========
apoptosis_genes = ["BAX","BCL2","BCL2L1","CASP3","CASP8","TP53","FAS","FADD"
apoptosis_genes = [g for g in apoptosis_genes if g in expr.index]
if len(apoptosis_genes) < 3:
    print("Note: few apoptosis genes found in matrix; consider mapping symbo

Z = StandardScaler().fit_transform(expr.T)  # samples x genes
expr_z = pd.DataFrame(Z, index=expr.columns, columns=expr.index)
meta["apoptosis_score"] = expr_z[apoptosis_genes].mean(axis=1)

# ========= PCA =========
pca = PCA(n_components=2, random_state=0)
pcs = pca.fit_transform(expr.T)
meta["PC1"], meta["PC2"] = pcs[:,0], pcs[:,1]

# ========= COX PH =========
cox_df = meta[["overall_survival","event","apoptosis_score"]].dropna()
cox = CoxPHFitter()
cox.fit(cox_df, duration_col="overall_survival", event_col="event")
cox.print_summary()

# ========= PLOT =========
plt.figure(figsize=(5.2,4.2))
sc = plt.scatter(meta["PC1"], meta["PC2"], c=meta["apoptosis_score"], s=18,
plt.colorbar(sc, label="Apoptosis score")
plt.title("PCA — Lung tumors colored by apoptosis score")
plt.xlabel("PC1"); plt.ylabel("PC2")
plt.tight_layout(); plt.show()
```

```
Note: few apoptosis genes found in matrix; consider mapping symbols or using
a full HALLMARK_APOPTOSIS set.
```

model	lifelines.CoxPHFitter
duration col	'overall_survival'
event col	'event'
baseline estimation	breslow
number of observations	135
number of events observed	57
partial log-likelihood	-236.88
time fit was run	2025-11-14 03:31:19 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	
apoptosis_score	0.36	1.44	0.12	0.12	0.61	1.13	1.83	0.00	2

Concordance	0.63
Partial AIC	475.76
log-likelihood ratio test	8.00 on 1 df
-log2(p) of ll-ratio test	7.74



Our understanding of these results is that higher apoptosis gene activity in tumors potentially reflects ongoing oncogenic stress with failed apoptosis, which is associated with poorer prognosis in lung cancer.

In [3]:
```
# Rebuild X for k-means with fewer components (e.g., 30 PCs or 70% variance)
X_expr = expr.T
scaler_km = StandardScaler()
X_scaled = scaler_km.fit_transform(X_expr)

# OPTION A: fixed dimension
ncomp = min(30, X_scaled.shape[1])
pca_km = PCA(n_components=ncomp, svd_solver='full', random_state=42)

# OPTION B: variance target (uncomment to try ~70% instead of fixed 30 PCs)
# pca_km = PCA(n_components=0.70, svd_solver='full', random_state=42)

X = pca_km.fit_transform(X_scaled)
print("X shape:", X.shape, "| variance kept:", round(
    pca_km.explained_variance_ratio_.sum(), 3))
```

```
X shape: (135, 30) | variance kept: 0.718
```

Verify and validate your analysis:

Pick a SPECIFIC method to determine how well your model is performing and describe how it works here.

Model & metric. We used k-means clustering and selected k = 2 by maximizing the held-out silhouette (–1..1; higher = tighter, better-separated clusters). We also reported Calinski–Harabasz increased (better) and Davies–Bouldin decreased (better).

Train/test protocol. We split the data, fit the scaler/PCA and k-means on train only, then assigned test points to the train centroids. This yielded test silhouette ≈ 0.31 and stability ARI ≈ 1.00 across restarts, indicating a robust partition without tiny clusters.

(Describe how you checked to see that your analysis gave you an answer that you believe (verify). Describe how your determined if your analysis gave you an answer that is supported by other evidence (e.g., a published paper).

external validation: apoptosis-related expression patterns and survival differences are well-described in lung cancer. The HALLMARK_APOPTOSIS gene set defines core apoptosis biology, which we used to compute an apoptosis score; TCGA LUAD/LUSC atlases document recurrent pathway variation across tumors; and multiple studies report apoptosis-related signatures associated with NSCLC prognosis. Our two clusters differ in apoptosis score and event rate, consistent with these reports.

https://pmc.ncbi.nlm.nih.gov/articles/PMC2074502/

- BCL-2 family dysregulation (e.g., BCL-2, BCL-XL, BAX) is common in lung cancer cell lines, supporting biological plausibility for apoptosis-related differences`

https://www.nature.com/articles/s41419-024-06940-y

- Supports heterogeneity: single-cell/spatial profiling of EGFR-mutant NSCLC identifies variable apoptosis groups, including BCL2L1/BCL-XL upregulation, which is tied to EGFR-TKI resistance."

In [4]:
```
# Cell 1 — imports & config
import numpy as np, pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import (silhouette_score, davies_bouldin_score,
                             calinski_harabasz_score, adjusted_rand_score)

RANDOM_STATE = 42
K_RANGE = range(2, 11)
PCA_VARIANCE = 0.90   # set to None to skip PCA
```

In [5]:
```
# Cell 2 — split, scale, PCA (fit on train, apply to test)
# X should be your numeric feature matrix (ndarray or DataFrame)
X_train, X_test = train_test_split(X, test_size=0.2, random_state=RANDOM_STA

scaler = StandardScaler()
Xtr = scaler.fit_transform(X_train)
Xte = scaler.transform(X_test)

if PCA_VARIANCE is not None:
    pca = PCA(n_components=PCA_VARIANCE, svd_solver='full', random_state=RAN
    Xtr_p = pca.fit_transform(Xtr)
    Xte_p = pca.transform(Xte)
else:
    Xtr_p, Xte_p = Xtr, Xte
```

In [6]:
```
# Config
RANDOM_STATE = 42
K_RANGE = range(2, 7)

from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans
from sklearn.metrics import (silhouette_score, davies_bouldin_score,
                             calinski_harabasz_score, adjusted_rand_score)
import numpy as np, pandas as pd

# Split
X_train, X_test = train_test_split(X, test_size=0.3, random_state=RANDOM_STA
Xtr_p, Xte_p = X_train, X_test

def _safe_metric(metric_fn, X, labels):
    u = np.unique(labels)
```

```
    if len(u) < 2 or len(u) >= len(labels):
        return np.nan
    try: return float(metric_fn(X, labels))
    except: return np.nan

def stability_ari(Xtr_p, k, runs=10):
    labs = [KMeans(k, n_init=20, random_state=s).fit(Xtr_p).labels_ for s in
    aris = []
    for i in range(runs):
        for j in range(i+1, runs):
            aris.append(adjusted_rand_score(labs[i], labs[j]))
    return float(np.mean(aris)), float(np.std(aris))

def eval_kmeans(Xtr_p, Xte_p, k):
    km = KMeans(n_clusters=k, n_init=20, random_state=RANDOM_STATE).fit(Xtr_
    tr_labels = km.labels_
    te_labels = km.predict(Xte_p)

    sil_tr = _safe_metric(silhouette_score,       Xtr_p, tr_labels)
    db_tr  = _safe_metric(davies_bouldin_score,   Xtr_p, tr_labels)
    ch_tr  = _safe_metric(calinski_harabasz_score, Xtr_p, tr_labels)

    sil_te = _safe_metric(silhouette_score,       Xte_p, te_labels)
    db_te  = _safe_metric(davies_bouldin_score,   Xte_p, te_labels)
    ch_te  = _safe_metric(calinski_harabasz_score, Xte_p, te_labels)

    sse_te = float(np.sum(np.min(km.transform(Xte_p), axis=1)**2))
    counts = np.bincount(tr_labels, minlength=k)
    min_train = int(counts.min())
    test_unique = int(len(np.unique(te_labels)))
    return {
        "k": k, "model": km,
        "sil_tr": sil_tr, "sil_te": sil_te,
        "db_tr": db_tr, "db_te": db_te,
        "ch_tr": ch_tr, "ch_te": ch_te,
        "sse_te": sse_te,
        "min_train": min_train, "test_unique": test_unique
    }

# Run sweep
rows = []
for k in K_RANGE:
    r = eval_kmeans(Xtr_p, Xte_p, k)
    mean_ari, std_ari = stability_ari(Xtr_p, k, runs=10)
    r["stability_ari"] = mean_ari
    r["stability_ari_sd"] = std_ari
    rows.append(r)

# Table
res_df = pd.DataFrame([
    {k:v for k,v in r.items() if k not in ("model",)}
    for r in rows
]).set_index("k")

# Hard constraints for a defensible choice
# (tune thresholds if this filters everything)
```

```
VALID = (
    (res_df["test_unique"] >= 2) &       # test must hit ≥2 clusters
    (res_df["min_train"] >= 5) &         # no tiny 1–2 member clusters
    (res_df["stability_ari"] >= 0.50)    # partitions agree across seeds
)

if VALID.any():
    candidates = res_df[VALID].copy()
else:
    # fallback: slightly relaxed
    candidates = res_df[
        (res_df["test_unique"] >= 2) &
        (res_df["min_train"] >= 3) &
        (res_df["stability_ari"] >= 0.30)
    ].copy()

# Selection: maximize test silhouette among candidates (fallback to train si
if len(candidates) == 0:
    candidates = res_df.copy()  # last resort, show what's available

sel = candidates["sil_te"].fillna(candidates["sil_tr"])
sel = sel.fillna(-candidates["sse_te"]/max(1.0, candidates["sse_te"].median(
candidates = candidates.assign(selection_score=sel)

display(res_df.sort_values("sil_te", ascending=False))
display(candidates.sort_values("selection_score", ascending=False))

best_k = int(candidates["selection_score"].idxmax())
best_row = next(r for r in rows if r["k"]==best_k)
best_model = best_row["model"]

print(f"Best k (constrained): {best_k}")
print("min_train:", best_row["min_train"],
      "| test_unique:", best_row["test_unique"],
      "| stability_ari:", round(best_row["stability_ari"], 2))
```

k	sil_tr	sil_te	db_tr	db_te	ch_tr	ch_te	sse_te	min_t
2	0.302446	0.311901	1.411250	1.323616	43.238874	21.074798	58732.742764	
3	0.190366	0.221857	1.732207	1.447785	28.026489	14.248759	53969.770466	
4	0.163840	0.191837	1.953278	1.743268	23.710652	12.344541	49807.553956	
5	0.159954	0.191837	1.756424	1.743268	20.014250	12.344541	49267.974162	
6	0.147351	0.166834	1.831514	1.775481	17.620763	9.858236	48662.888578	

k	sil_tr	sil_te	db_tr	db_te	ch_tr	ch_te	sse_te	min_t
2	0.302446	0.311901	1.411250	1.323616	43.238874	21.074798	58732.742764	
3	0.190366	0.221857	1.732207	1.447785	28.026489	14.248759	53969.770466	
4	0.163840	0.191837	1.953278	1.743268	23.710652	12.344541	49807.553956	

```
Best k (constrained): 2
min_train: 40 | test_unique: 2 | stability_ari: 1.0
```

In [14]:
```
# Cell 4 — stability at chosen k (Adjusted rand index [ARI] across seeds)
def stability_ari(Xtr_p, k, runs=10):
    label_list = []
    for seed in range(runs):
        km = KMeans(n_clusters=k, n_init=20, random_state=seed).fit(Xtr_p)
        label_list.append(km.labels_)
    aris = []
    for i in range(runs):
        for j in range(i+1, runs):
            aris.append(adjusted_rand_score(label_list[i], label_list[j]))
    return float(np.mean(aris)), float(np.std(aris))

mean_ari, std_ari = stability_ari(Xtr_p, best_k, runs=10)
mean_ari, std_ari

m,s = stability_ari(Xtr_p, best_k, runs=10)
print("pairs:", (10*9)//2, "mean:", f"{m:.3f}", "std:", f"{s:.3f}")
```

```
pairs: 45 mean: 1.000 std: 0.000
```

In [15]:
```
# Cell 5 — cluster profiles (train set)
train_labels = best_model.labels_
profiles = pd.DataFrame(X_train).assign(cluster=train_labels)\
           .groupby("cluster").agg(["mean","std","count"])
profiles
```

Out[15]:

	0			1			
	mean	std	count	mean	std	count	mean
cluster							
0	23.507264	5.513899	54	0.410311	15.976454	54	1.113639
1	-29.830255	10.048337	40	-2.939549	12.346172	40	-0.809270

2 rows × 90 columns

Conclusions and Ethical Implications:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

k-means on 30 PCs selected k=2 (test silhouette ≈0.31; ARI ≈1.00). Clusters differ in apoptosis score and event rate, suggesting biologically meaningful separation consistent with lung-cancer literature. Ethically, unsupervised groups can reflect sampling or demographic bias; we avoid normative labels, report stability/held-out metrics, and would test fairness across subgroups (e.g., sex, age, smoking status) if used downstream.

Interpretation of results specific- higher apoptosis score was associated with worse survival rates with significance ($p<0.005$, cox proportional hazards ratio exp(coef): 1.44- hazard is increased by 44% [and to put that into context, HR > 1.0 is considered Increased risk], concordance=0.63 (moderate division)). To explain this biologically, higher expression of apoptotic genes could be a regulatory overexpression after failed attempts at apoptosis, or to offset pro and anti-apoptotic gene ratios (like BCL-2, an anti-apoptotic gene).

Other Ethical implications:

- racial or socioeconomic groups may be incorrectly represented/ ill calculated based on the model and prevalence/incidence splits of cancer type by race/etc.
- a model like this would need rigorous validation as using this model to inform treatment decisions can lead to false reassurance
- responsibility in providing access to resulting therapies also involves understanding prevalence/incidence demographic split and determining how to fill in the model with a more diverse demographic dataset to provide more representative/generalized results (only had 135 lung cancer samples, could expand that), and even change disease prognosis metrics- change model results

Limitations and Future Work:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

Limitations- k-means assumes roughly spherical clusters; compare GMM/HDBSCAN/spectral. Expand apoptosis gene set (full HALLMARK_APOPTOSIS) and add robust scaling/outlier checks

- limited to 135 lung cancer samples
- only a subset of apoptosis genes were found in expression matrix, could expand dataset to evaluate more
- our evaluation of survival was in relevance to apoptosis scores, but survival could be impacted by other confounding variables that don't necessarily induce higher apoptosis gene expression (ex: advanced stage tumors can have more stress that upregulate apoptosis genes, but chemotherapy could also upregulate pro-apoptosis

genes, so cannot be evaluated 1:1 without considering how other factors can change measurement of survival rate or gene expression/regulation)
- evaluation with other cancers survival rate/apoptosis genes- cannot accurately extrapolate in regards to lung cancer
- rna-seq measures gene expression at one point in time- these measures do not stay consistent and capturing changes over time could improve our model predictions

Future directions- explaining how we would make our model better based on scoring:

- Cox proportional hazards model- moderate concordance of 0.63 would be improved
  - 0.5 is considered random chance

Future Research

- Targeting anti-apoptotic genes and regulators would help improve survival outcomes and therapeutic outcomes for lung cancer patients expressing high apoptotic gene expression
- longitudinal studies of how apoptosis expression changes over time in regards to other 'confounding' factors could improve our research and model
- choose a dataset with diverse population demographics to create a more inclusive model

NOTES FROM YOUR TEAM:

This is where our team is taking notes and recording activity.

Running Agenda/worklog:

- 10/23/35- Looked through datasets and brainstormed potential questions to research
- 10/25/25- Filled in disease background information and research
- 11/3/25 — Loaded lung cancer expression and metadata files (TCGA GSE62944), aligned sample barcodes, generated apoptosis gene expression scores, performed PCA to visualize tumor clustering, and ran a Cox proportional hazards survival model to evaluate the association between apoptosis signaling and patient outcomes. Interpreted results showing higher apoptosis pathway activity associated with worse survival and discussed biological relevance.
- 11/11/25- evaluated efficacy of model, completed ethical implications, conclusions, and future risks/research

QUESTIONS FOR YOUR TA:

These are questions we have for our TA.

- [answered] How can we get access to the full dataset?
- [answered] What other methods should we consider beyond regression?

REFERENCES:

These are the references for background research, code debugging, and verification/validation.[APA formatting]

- Mitochondrial Effectors of Apoptosis
  - Mustafa, M., Ahmad, R., Tantry, I. Q., Ahmad, W., Siddiqui, S., Alam, M., Abbas, K., Moinuddin, Hassan, M. I., Habib, S., & Islam, S. (2024). Apoptosis: A Comprehensive Overview of Signaling Pathways, Morphological Changes, and Physiological Significance and Therapeutic Implications. Cells, 13(22), 1838. https://doi.org/10.3390/cells13221838
  - Lee, E., Song, C. H., Bae, S. J., et al. (2023). Regulated cell death pathways and their roles in homeostasis, infection, inflammation, and tumorigenesis. Experimental & Molecular Medicine, 55(10), 1632–1643. https://doi.org/10.1038/s12276-023-01069-y
  - Brentnall, M., Rodriguez-Menocal, L., De Guevara, R. L., et al. (2013). Caspase-9, caspase-3 and caspase-7 have distinct roles during intrinsic apoptosis. BMC Cell Biology, 14, 32. https://doi.org/10.1186/1471-2121-14-32
- anti-apoptotic mitochondrial gatekeepers
  - Hata, A. N., Engelman, J. A., & Faber, A. C. (2015). The BCL2 Family: Key Mediators of the Apoptotic Response to Targeted Anticancer Therapeutics. Cancer discovery, 5(5), 475–487. https://doi.org/10.1158/2159-8290.CD-15-0011
  - Brinkmann, K., & Kashkar, H. (2014). Targeting the mitochondrial apoptotic pathway: A preferred approach in hematologic malignancies? Cell Death & Disease, 5(2), e1098. https://doi.org/10.1038/cddis.2014.61
  - Veleta, K. A., Cleveland, A. H., Babcock, B. R., He, Y. W., Hwang, D., Sokolsky-Papkov, M., & Gershon, T. R. (2021). Antiapoptotic Bcl-2 family proteins BCL-xL and MCL-1 integrate neural progenitor survival and proliferation during postnatal cerebellar neurogenesis. Cell death and differentiation, 28(5), 1579–1592. https://doi.org/10.1038/s41418-020-00687-7
- Survival signalling axis
  - He, Y., Sun, M. M., Zhang, G. G., et al. (2021). Targeting PI3K/Akt signal transduction for cancer therapy. Signal Transduction and Targeted Therapy, 6(1), 425. https://doi.org/10.1038/s41392-021-00828-5
  - Georgescu M. M. (2010). PTEN Tumor Suppressor Network in PI3K-Akt Pathway Control. Genes & cancer, 1(12), 1170–1177. https://doi.org/10.1177/1947601911407325
- Extrinsic (death-receptor) axis

    - Yuan, X., Gajan, A., Chu, Q., Xiong, H., Wu, K., & Wu, G. S. (2018). Developing TRAIL/TRAIL death receptor-based cancer therapies. Cancer metastasis reviews, 37(4), 733–748. https://doi.org/10.1007/s10555-018-9728-y
- Caspase inhibition downstream.
    - Cheung, C. H. A., Chang, Y. C., Lin, T. Y., et al. (2020). Anti-apoptotic proteins in the autophagic world: An update on functions of XIAP, Survivin, and BRUCE. Journal of Biomedical Science, 27(1), 31. https://doi.org/10.1186/s12929-020-0627-5
- BRCA
    - American Cancer Society. (2024, January 12). Key statistics for breast cancer. American Cancer Society. https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html
    - Pernas, S., & Tolaney, S. M. (2021). Management of Early-Stage Human Epidermal Growth Factor Receptor 2–Positive Breast Cancer. JCO Oncology Practice, 17(6), 320–330. https://doi.org/10.1200/op.21.00020
    - Miglietta, Federica, et al. “PARP-Inhibitors for BRCA1/2-Related Advanced HER2-Negative Breast Cancer: A Meta-Analysis and GRADE Recommendations by the Italian Association of Medical Oncology.” The Breast, vol. 66, 1 Dec. 2022, pp. 293–304, https://doi.org/10.1016/j.breast.2022.10.014.
    - “CDK4/6 Inhibitors, SERDs, and More in NCCN Talk on Breast Cancer Updates.” Ajmc.com, AJMC, Apr. 2022, www.ajmc.com/view/cdk4-6-inhibitors-serds-and-more-in-nccn-talk-on-breast-cancer-updates. Accessed 14 Nov. 2025.
- LUAD / NSCLC
    - American Cancer Society. Cancer Facts & Figures 2025.

Atlanta: American Cancer Society; 2025. * American Cancer Society. “What Is Lung Cancer?” Www.cancer.org, American Cancer Society, 29 Jan. 2024, www.cancer.org/cancer/types/lung-cancer/about/what-is.html. * Centers for Disease Control and Prevention. “Lung Cancer Risk Factors.” Lung Cancer, 13 Feb. 2025, www.cdc.gov/lung-cancer/risk-factors/index.html. * US EPA. “Health Risk of Radon | US EPA.” US EPA, 7 Aug. 2019, www.epa.gov/radon/health-risk-radon. * Riely, G. J. et al. (2025). NCCN Guidelines® Insights: Non–Small Cell Lung Cancer, Version 7.2025: Featured Updates to the NCCN Guidelines®. Journal of the National Comprehensive Cancer Network, 23(9), 354-362. Retrieved Nov 14, 2025, from https://doi.org/10.6004/jnccn.2025.0043 * Kage H. (2025). Emerging Role of Molecular Testing in the Management of Non-metastatic Non-small Cell Lung Cancer. Tuberculosis and respiratory diseases, 88(3), 431–441. https://doi.org/10.4046/trd.2024.0159

- COAD/READ
    - American Cancer Society. “Colorectal Cancer Statistics | How Common Is Colorectal Cancer?” Www.cancer.org, American Cancer Society, 2023, www.cancer.org/cancer/types/colon-rectal-cancer/about/key-statistics.html.

    - "Red and Processed Meat and Cancer." Cancer.org, 2025, www.cancer.org/cancer/risk-prevention/diet-physical-activity/diet-and-physical-activity/how-diet-and-physical-activity-impact-cancer-risk/red-meat-and-cancer.html.
    - American Cancer Society. Colorectal Cancer Facts & Figures 2023-2025. 2023.
    - Nash, S. H., Britton, C., & Redwood, D. (2021). Characteristics of colorectal cancers among Alaska Native people before and after implementing programs to promote screening. Journal of cancer policy, 29, 100293. https://doi.org/10.1016/j.jcpo.2021.100293
    - "Examining Alaska Native Peoples' Perspectives about Options for Colorectal Cancer Screening - Mayo Clinic." Mayoclinic.org, 2025, www.mayoclinic.org/medical-professionals/digestive-diseases/news/examining-alaska-native-peoples-perspectives-about-options-for-colorectal-cancer-screening/mac-20585844.
    - American Cancer Society. "Colon Cancer Treatment, by Stage | How to Treat Colon Cancer." Www.cancer.org, 6 Feb. 2024, www.cancer.org/cancer/types/colon-rectal-cancer/treating/by-stage-colon.html.
    - "Rectal Cancer Treatment (PDQ®)–Health Professional Version." National Cancer Institute, Cancer.gov, 29 Jan. 2019, www.cancer.gov/types/colorectal/hp/rectal-treatment-pdq.
    - Davis, Lisa E. "The Evolution of Biomarkers to Guide the Treatment of Metastatic Colorectal Cancer." AJMC, vol. 24, 25 Apr. 2018, www.ajmc.com/view/evolution-biomarkers-guide-treatment-metastatic-colorectal-cancer.
- GBM/LGG
    - National Cancer Institute. "Cancer of the Brain and Other Nervous System - Cancer Stat Facts." SEER, 2018, seer.cancer.gov/statfacts/html/brain.html.
    - National Brain Tumor Society. "Brain Tumor Facts." National Brain Tumor Society, 2023, braintumor.org/brain-tumors/about-brain-tumors/brain-tumor-facts/.
    - Hanif, Farina, et al. "Glioblastoma Multiforme: A Review of Its Epidemiology and Pathogenesis through Clinical Presentation and Treatment." Asian Pacific Journal of Cancer Prevention, vol. 18, no. 1, 2017, pp. 3–9, www.ncbi.nlm.nih.gov/pmc/articles/PMC5563115/, https://doi.org/10.22034/APJCP.2017.18.1.3.
    - Mayo Clinic. "Glioblastoma - Symptoms and Causes." Mayo Clinic, 20 June 2024, www.mayoclinic.org/diseases-conditions/glioblastoma/symptoms-causes/syc-20569077.
    - Tamimi, Ahmad Faleh, and Malik Juweid. "Epidemiology and Outcome of Glioblastoma." Glioblastoma, vol. Chapter 8, 27 Sept. 2017,

www.ncbi.nlm.nih.gov/pubmed/29251870, https://doi.org/10.15586/codon.glioblastoma.2017.ch8.

- Wang, Gi-Ming, et al. “Importance of the Intersection of Age and Sex to Understand Variation in Incidence and Survival for Primary Malignant Gliomas.” Neuro-Oncology, 13 Aug. 2021, https://doi.org/10.1093/neuonc/noab199.

*Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., Belanger, K., Brandes, A. A., Marosi, C., Bogdahn, U., Curschmann, J., Janzer, R. C., Ludwin, S. K., Gorlia, T., Allgeier, A., Lacombe, D., Cairncross, J. G., Eisenhauer, E., Mirimanoff, R. O., European Organisation for Research and Treatment of Cancer Brain Tumor and Radiotherapy Groups, … National Cancer Institute of Canada Clinical Trials Group (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. The New England journal of medicine, 352(10), 987–996. https://doi.org/10.1056/NEJMoa043330

- Center. “FDA Approves Vorasidenib for Grade 2 Astrocytoma or Oligodendroglioma.” U.S. Food and Drug Administration, 6 Aug. 2024, www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-vorasidenib-grade-2-astrocytoma-or-oligodendroglioma-susceptible-idh1-or-idh2-mutation.
- Chen, Boran, et al. “Recent Incidence Trend of Elderly Patients with Glioblastoma in the United States, 2000–2017.” BMC Cancer, vol. 21, no. 1, 12 Jan. 2021, https://doi.org/10.1186/s12885-020-07778-1.

- Validation
  - Reeve, JG, et al. “Expression of Apoptosis-Regulatory Genes in Lung Tumour Cell Lines: Relationship to P53 Expression and Relevance to Acquired Drug Resistance.” British Journal of Cancer, vol. 73, no. 10, May 1996, pp. 1193–1200, https://doi.org/10.1038/bjc.1996.230.
  - Izumi, Motohiro, et al. “Integrative Single-Cell RNA-Seq and Spatial Transcriptomics Analyses Reveal Diverse Apoptosis-Related Gene Expression Profiles in EGFR-Mutated Lung Cancer.” Cell Death & Disease, vol. 15, no. 8, 9 Aug. 2024, https://doi.org/10.1038/s41419-024-06940-y.
- Code debugging/research
  - OpenAI. (n.d.). ChatGPT. https://chatgpt.com/