Module 3:

Team Members:

Logan Heselton, Noah Edouard

Project Title:

The Relationship Between Apoptotic-related gene expression and Pathological Tumor Stage in Lung Squamous Cell Carcinoma (LUSC)

Project Goal:

This project seeks to determine how apopotitic gene expression levels influence tumor stage progression in patients with Lung Squamous Cell Carcinoma (LUSC). By analyzing clinical and genomic data, we aim to identify whether altered TP53, BLC-family, and caspase gene expression correlates with more early tumor stages (stages 1-2) or later tumor stages (3-4). Understanding this relationship could provide insight into tumor aggressiveness, potential prognostic markers, and therapeutic targets for improving lung cancer outcomes.

Project Question:

- How do apoptotic-related gene expression patterns (TP53, BCL-family, and caspase genes) differ between early-stage (I–II) and late-stage (III–IV) lung squamous cell carcinoma (LUSC)?

Disease Background:

- Cancer hallmark focus: **Evading Apoptosis**
- Overview of hallmark:
    - Apoptosis is the process of programmed cell death where there is a mass triggering of a variety of physiologic signals to disrupt the cell membrane, break down nuclear skeletons, degrade chromosomes, etc. Although this is a normal physiological process, it was discovered in 1972 that cancer cells can evade apoptosis, serving as a major barrier to cancer. It was found that the bcl-2 oncogene upregulates

anti-apoptotic activity and promotes the formation of B cell lymphomas by enhancing lymphocyte activity. It was also discovered that the p53 tumor suppressor gene can remove a key component of the DNA damage sensor that can induce the apoptotic effector cascade, a necessary component that executes apoptosis. Together, these discoveries involve the PI3 kinase-AKT/PKB pathway, which transmits anti-apoptotic survival signals. The pathways signals were found to mitigate apoptosis in a substantial fraction of human tumors that were activated by extracellular factors such as IGF-1/2 or IL-3. This hallmark has led to many open questions across all cancer types such as how this changes antitumor therapies and what new technologies can restore apoptotic defense mechanism. It has also led to a major question of what physiological pathways in cancer still operate effectively in reactivating cell death in cancer cells.

- Genes associated with evading apoptosis:
    - **What we are going to investigate**
        - TP53 (Tumor Protein p53):
            - Role: Master tumor suppressor; activates transcription of pro-apoptotic genes in response to DNA damage or stress.
            - Signaling Pathway: Intrinsic (mitochondrial) apoptosis pathway. p53 activates BAX, PUMA, and NOXA to promote cell death.
            - Relevance to LUSC: In LUSC, TP53 is frequently mutated or inactivated (~80–90% of cases), preventing apoptosis even when cells have DNA damage.
        - BLC_family
            - Role: Regulates mitochondrial apoptosis by controlling mitochondrial outer membrane permeabilization (MOMP). Anti-apoptotic members (BCL2, BCL-XL) block cell death, while pro-apoptotic

members (BAX, BAK) promote it.
        - Signaling Pathway: Functions in the intrinsic (mitochondrial) apoptosis pathway. BCL proteins integrate stress signals, regulate cytochrome c release, and influence caspase activation downstream.
        - Relevance to LUSC: Altered BCL expression can allow tumor cells to evade apoptosis, increasing survival and promoting tumor progression. Overexpression of anti-apoptotic BCL genes is often linked to more aggressive LUSC.
    - Caspase genes
        - Role: Executioners of apoptosis. Initiator caspases (CASP8, CASP9) sense apoptosis signals, while effector caspases (CASP3, CASP7) cleave cellular proteins to drive programmed cell death.
        - Signaling Pathway: Core components of both intrinsic and extrinsic apoptosis pathways. Activated by cytochrome c (intrinsic) or death receptor signaling (extrinsic), leading to a caspase cascade and controlled cell dismantling.
        - Relevance to LUSC: Reduced caspase activation or expression can suppress apoptosis, enabling uncontrolled growth. Dysregulated caspases in LUSC have been associated with resistance to therapy and progression to later tumor stages.
- Other specific gene descriptions
    - TP53BP1 (Tumor Protein p53 Binding Protein 1)
        - Role: DNA damage sensor that helps repair double-strand breaks via non-

homologous end joining; stabilizes p53.
        - Signaling Pathway: p53 signaling → DNA repair → apoptosis decision checkpoint.
        - Relevance to LUSC: Loss or reduced expression impairs p53-mediated apoptosis, allowing survival of damaged cells.
    - TP53INP1 (Tumor Protein p53 Inducible Nuclear Protein 1)
        - Role: p53 target gene; enhances p53 transcriptional activity and promotes apoptosis and autophagy under stress.
        - Signaling Pathway: p53-mediated apoptosis (stress response).
        - Relevance to LUSC: Downregulated in many cancers, including LUSC; linked to poor prognosis and tumor progression.
    - BCL2 (B-cell Lymphoma 2)
        - Role: Anti-apoptotic protein that inhibits cytochrome c release from mitochondria, blocking caspase activation.
        - Signaling Pathway: Intrinsic (mitochondrial) apoptotic pathway.
        - Relevance to LUSC: Overexpressed in LUSC; helps cancer cells survive chemotherapy or radiation.

Lung Squamous Cell Carcinoma

- Prevalence & incidence
    - In 2023, lung cancer was reported as the third most common cancer in terms of incidence and the leading cause of cancer-related mortality in the United States, according to the National Cancer Institute’s Surveillance, Epidemiology, and End Results (NCI SEER) database. The

estimated 238,340 new cases of lung cancer accounted for approximately 12% of the total U.S. cancer burden, while the estimated 127,070 deaths represented about 21% of all cancer-related deaths. Approximately 85% of all lung cancers are classified as non–small cell lung cancers (NSCLCs), within which adenocarcinoma and squamous cell carcinoma are the two most common subtypes, comprising roughly 50% and 30% of NSCLC cases, respectively.

- Lung squamous cell carcinoma (LUSC) has historically been more common in men, though the gender gap is narrowing due to changing smoking patterns. It predominantly affects older adults, typically between 60 and 80 years old, and remains highly associated with tobacco exposure. Over the past few decades, the overall incidence of LUSC has declined in many countries, reflecting a reduction in smoking rates. However, LUSC continues to represent a major public health concern, particularly in regions with high smoking prevalence, industrial air pollution, or limited tobacco control regulations.
- Source: https://www.ncbi.nlm.nih.gov/books/NBK564510/, Chatgpt

- Risk factors (genetic, lifestyle) & Societal determinants
  - Genetic
    - The development of lung squamous cell carcinoma involves both genetic and environmental influences. On the genetic side, mutations in key tumor suppressor and oncogenes—such as TP53, SOX2, and PIK3CA—are frequently identified in SCC. Additionally, polymorphisms in detoxification genes, including CYP1A1, GSTM1, and GSTT1, may impair the body’s ability to metabolize carcinogens found in cigarette smoke, thereby increasing susceptibility. A family history of lung cancer slightly elevates risk, likely due to a combination of shared genetic factors and environmental exposures such as tobacco use.
  - Lifestyle
    - By far, the most significant risk factor for LUSC is tobacco smoking. Among all lung cancer

subtypes, LUSC has the strongest correlation with cigarette use. The risk increases with both the duration and intensity of smoking, and individuals who begin smoking at a younger age face higher lifetime risks. Occupational exposures also contribute to disease development—workers exposed to carcinogenic materials such as asbestos, chromium, nickel, cadmium, and polycyclic aromatic hydrocarbons (PAHs) are at increased risk. Moreover, air pollution, especially long-term exposure to fine particulate matter (PM2.5), has emerged as an important risk factor even for non-smokers. Additional contributors include heavy alcohol consumption, which can act synergistically with tobacco, and pre-existing lung conditions such as chronic obstructive pulmonary disease (COPD) or pulmonary fibrosis, which may create a more vulnerable pulmonary environment for carcinogenesis.

- Social Determinants
    - Societal and social factors play a critical role in determining both the risk and outcomes of lung squamous cell carcinoma. Individuals from lower socioeconomic backgrounds experience disproportionately higher rates of LUSC, driven by higher smoking prevalence, occupational exposures, and reduced access to healthcare. Education level is another major determinant—those with lower educational attainment are more likely to smoke and less likely to participate in early detection or cessation programs. Limited access to healthcare services, particularly preventive care and screening tools such as low-dose CT scans, often leads to delayed diagnosis and poorer survival outcomes.
    - Environmental and policy-level determinants also shape SCC incidence. Residents in urban or industrial regions are exposed to higher levels of

air pollution and industrial carcinogens, which elevate risk. Conversely, strong tobacco control laws, public health education campaigns, and air quality regulations have been shown to significantly reduce both smoking rates and lung cancer incidence. Overall, lung squamous cell carcinoma reflects a complex interplay between biological susceptibility, lifestyle behaviors, and social context—underscoring the importance of integrating public health interventions, tobacco regulation, and environmental protection with clinical prevention and early detection strategies.

        - Source: https://www.ncbi.nlm.nih.gov/books/NBK564510/, Chatgpt
- Standard of care treatments (& reimbursement)
    - Standard of care treatments
        - Treatments vary depending on the stage of the cancer. For LUSC stages I and II, surgical resection is the first line of treatment. Stage IA does not require chemotherapy; however, stage IB may require chemotherapy if the tumor size exceeds 4cm. Having all patients with tumors >4 cm or positive nodes be evaluated for preoperative therapy is recommended, including immunotherapy using nivolumab or pembrolizumab in addition to chemotherapy. However, if patients are not candidates for immune checkpoint inhibitors due to autoimmune diseases, being on immunosuppressants, or carrying EGFR mutations or ALK rearrangements, then combination chemotherapy alone is recommended. Stage II is normally treated with surgery followed by chemotherapy. In these stages, if surgery isn’t effective or if patients can’t have surgery, radiation treatment is the general alternative. For stage III, tumors are unresectable. Chemotherapy with radiation is the usual choice for stage IIIA, while for stage

IIIB, combined chemotherapy and radiation are used, followed by maintenance immunotherapy for 1 year. In stage IV, systemic treatment with palliative radiation is used.
        - Depending on mutation status, targeted therapy is used more often in non-squamous cell lung cancers. In advanced LUSC, treatment should be based on the molecular features of the tumor. With somatically driven mutations, inhibitor therapy is indicated. For epidermal growth receptor mutation, tyrosine kinase inhibitors (TKI) like erlotinib, gefitinib, and afatinib are used.
    - Reinbursement
        - Reimbursement for LUSC treatment in the U.S. is well established across Medicare and commercial insurance plans. Most insurance plans and Medicare help pay for recommended lung cancer screening tests. Most treatments, such as platinum-based chemotherapy, immunotherapy agents like pembrolizumab (Keytruda), and radiation therapy, are covered under Medicare Part B as outpatient services, with patients typically responsible for 20% coinsurance after meeting their deductible.
    - Source: https://www.ncbi.nlm.nih.gov/books/NBK564510/, https://www.ncbi.nlm.nih.gov/books/NBK564510/, Google AI
- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology)
    - Anatomy
        - LUSC often occurs in the central part of the lung or the main airway, such as the left or right bronchus. The bronchi are the large tubes that carry air from your windpipe to your lungs. The left main bronchus goes into your left lung, and the right main bronchus goes into your right lung. After the main bronchi, they branch out into smaller segments. In LUSC, cancer develops from the squamous epithelial cells that normally line the bronchi, which can thicken and form

keratinized layers due to chronic irritation. As the tumor grows, it may obstruct airways, invade surrounding lung tissue, and spread to nearby lymph nodes or distant organs through the bloodstream or lymphatic system.

- Organ Physiology
  - The lungs are the foundational organs of the respiratory system. Their most basic function is to facilitate gas exchange from the environment into the bloodstream. This process involves oxygen being transported through the alveoli into the capillary network, where it can enter the arterial system. The lungs then separate into individual lobes called alveoli, allowing for this efficient gas exchange. In LUSC, tumor growth disrupts this process by blocking bronchi, reducing airflow, and compromising ventilation of affected lung regions. This can lead to symptoms such as shortness of breath and a chronic cough. As the disease progresses, decreased lung capacity and oxygen delivery can strain the cardiovascular system and lead to systemic fatigue and hypoxemia.
- Cell and Molecular Physiology
  - At the cellular level, LUSC develops when bronchial epithelial cells accumulate genetic and epigenetic mutations that disrupt normal cell-cycle regulation. Common molecular changes in this cancer type include inactivation of tumor suppressor genes such as TP53 and CDKN2A (p16) and amplification of oncogenes like SOX2 and FGFR1. These mutations cause uncontrolled proliferation, resistance to apoptosis, and cellular dedifferentiation. At the molecular level, pathways such as PI3K/AKT/mTOR and EGFR signaling may become dysregulated, supporting tumor growth, angiogenesis, and immune evasion.
- Source: https://my.clevelandclinic.org/health/body/21607-bronchi, https://www.ncbi.nlm.nih.gov/books/

NBK564510/, https://www.ncbi.nlm.nih.gov/books/NBK545177/, https://pmc.ncbi.nlm.nih.gov/articles/PMC3404741/

Data-Set:

- This project will analyze data focusing specifically on Lung Squamous Cell Carcinoma (LUSC) patients from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) dataset GSE62944 to examine the relationship between TP53 gene expression and tumor stage. The datasets include both RNA sequencing (RNA-seq) gene expression data and clinical information for individual patients. The gene expression data were obtained from the file GSE62944_subsample_topVar_log2TPM.csv, which contains normalized RNA-seq expression values measured as log2-transformed Transcripts Per Million (log2 TPM). These values represent relative mRNA abundance across tumor samples, allowing quantitative assessment of TP53 expression levels. The clinical data are derived from subsampled_TCGA_CDR_survival.csv, part of the TCGA Clinical Data Resource (CDR), and include relevant variables such as tumor stage (Stages I–IV), vital status (alive or deceased), age at diagnosis, and sex. These features will be used to explore potential associations between TP53 expression and tumor progression.
- All data were collected and processed as part of large-scale cancer genomics initiatives under the TCGA Research Network and GEO GSE62944 project. Tumor tissue samples were collected from consenting patients, and RNA was extracted and sequenced using next-generation RNA-seq platforms. The resulting sequencing reads were aligned to the human reference genome, quantified, normalized, and log-transformed to ensure comparability across samples. Clinical annotations were compiled from medical records and pathology reports, de-identified, and standardized for analysis. Together, these datasets provide a robust resource for investigating the molecular and clinical relationships underlying Lung Squamous Cell Carcinoma, particularly regarding the role of TP53 expression in tumor development and stage progression.
- Source: https://pubmed.ncbi.nlm.nih.gov/26209429/, https://www.sciencedirect.com/science/article/pii/S0092867418302290

Data Analysis:

Methods

- The machine learning technique we are using is a **supervised machine learning classification model** built using scikit-learn. Unlike the earlier Spearman correlation approach, which measured monotonic associations between individual genes and tumor stage, the new model directly learns to classify patients into early-stage (I–II) or late-stage (III–IV) LUSC based on multi-gene apoptotic expression patterns. Logistic Regression was selected because it is interpretable, well-suited for binary classification (early stage (I-II) versus late stage (III-IV)), and robust for high-dimensional gene expression data. The model works by taking apoptotic gene expression values (TP53-family, BCL-family, and CASP genes) as input features and learning decision boundaries that best separate early vs late tumor stages. During training, the model “optimizes” its parameters by minimizing the logistic loss and adjusting feature weights to maximize classification accuracy (to ensure that the model performs well). A train–test split is used to prevent overfitting and evaluate generalizability. The model determines whether it is “good enough” using several performance metrics: accuracy, balanced accuracy (to adjust for class imbalance), confusion matrix, and classification report (precision, recall, F1-score). A ROC-AUC was also used to quantify the model’s ability to distinguish early vs late stages. Strong performance across these metrics indicates that apoptotic gene expression patterns contain meaningful predictive information about LUSC tumor stage progression.

Analysis

- The analysis began by cleaning and preparing two main datasets: GSE62944_subsample_log2TPM.csv, containing normalized $\log_2$-TPM gene expression values, and GSE62944_metadata.csv, which includes patient-level clinical information such as cancer type and pathological tumor stage. The project focused specifically on patients diagnosed with Lung Squamous Cell Carcinoma (LUSC). From the gene expression matrix, only apoptosis-related genes (TP53-family, BCL-family, and caspase genes) were extracted, since these pathways play central roles in tumor suppression and programmed cell death. The metadata and gene expression datasets were merged using the patient barcode as the common identifier. From this point forward, this merged data set

was the only set used for our methods and analysis.

- Tumor stage information from the AJCC classification was then converted into a binary label to support supervised learning: early-stage (Stages I–II) versus late-stage (Stages III–IV). Samples missing tumor stage information were removed to ensure accurate model training. After cleaning and merging, the final dataset consisted of apoptotic gene expression features (input X) paired with early/late tumor stage labels (output y). This prepared dataset was then used to train a Logistic Regression classifier using scikit-learn.
- For model evaluation, the data was split into training and testing sets to prevent overfitting and assess performance on unseen patients. The classifier output several metrics, including accuracy, balanced accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC, which together indicate how well apoptotic gene expression patterns distinguish early versus late LUSC stages. Genes with higher logistic regression coefficient magnitudes were interpreted as having stronger influence on the model's decision boundary. ROC curves and performance metrics were generated using matplotlib and scikit-learn to visually demonstrate how effectively apoptotic gene expression predicts LUSC tumor stage progression.

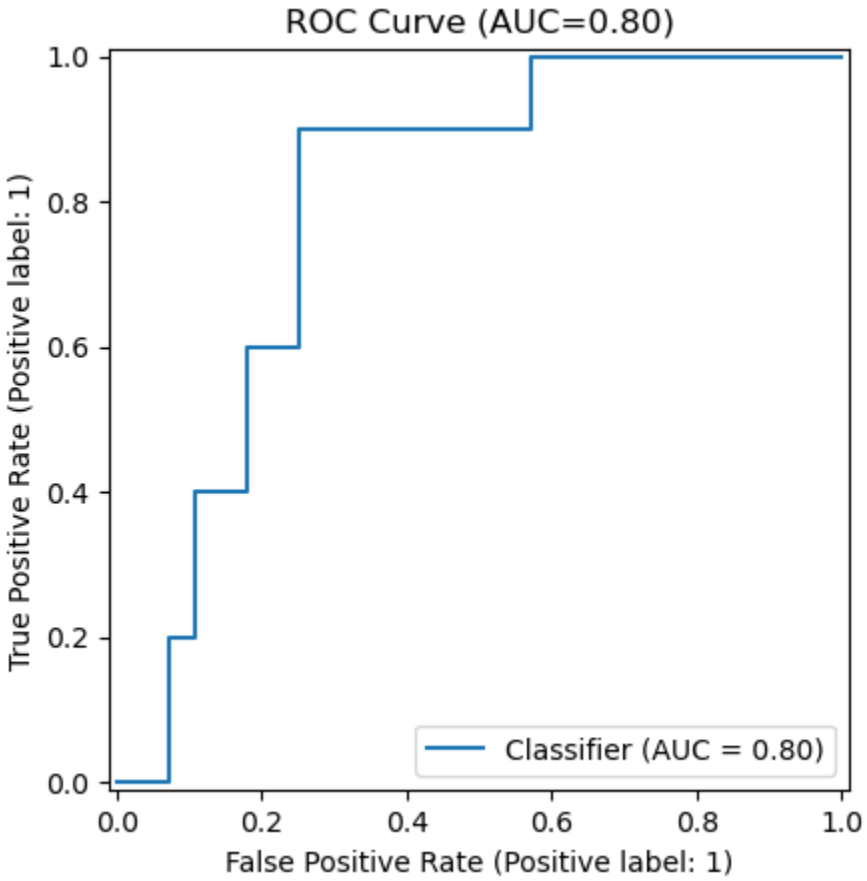
In [1]:
```
#How do apoptosis-related genes, including TP53, BCL-family members, and caspa

#imports all the necessary functions
import pandas as pd
import numpy as np
from pathlib import Path
import matplotlib.pyplot as plt
import seaborn as sns
#imports sklearn functions
from sklearn.model_selection import train_test_split, StratifiedKFold, GridSea
from sklearn.metrics import confusion_matrix, classification_report, roc_auc_s
from sklearn.linear_model import LogisticRegression
from imblearn.over_sampling import SMOTE #this function was learned from ChatG

#loads the cleaned file that contains ONLY tp53 family, BCL family, and caspas
DATA_DIR = Path("C:/Users/logan/Documents/GitHub/desktop-tutorial/New folder/C
df = pd.read_csv(DATA_DIR / "LUAD_LUSC_TP53_BCL_CASP_combined.csv")

#identifies patient barcodes so that it can recognize tumor stage
df = df.set_index("bcr_patient_barcode")

#feature: tumor stage
stage_col = "ajcc_pathologic_tumor_stage"
meta = df[[stage_col, "cancer_type"]].copy()
```

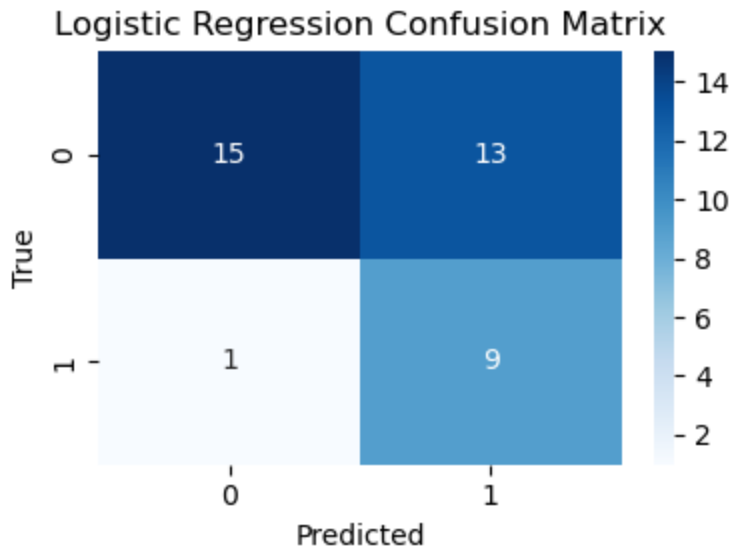
```
#makes each stage into a numerical number (1-4) and ensures that stages with A
stage_map = {
    "Stage I": 1, "Stage IA": 1, "Stage IB": 1,
    "Stage II": 2, "Stage IIA": 2, "Stage IIB": 2,
    "Stage III": 3, "Stage IIIA": 3, "Stage IIIB": 3, "Stage IIIC": 3,
    "Stage IV": 4, "Stage IVA": 4, "Stage IVB": 4
}
meta.loc[:, "stage_num"] = meta[stage_col].map(stage_map)

valid = meta["stage_num"].dropna().index
meta = meta.loc[valid]
expr = df.loc[valid]

mask = meta["stage_num"].isin([1, 2, 3, 4])
meta = meta[mask]
expr = expr.loc[mask]

#labels the data; creates a binary classification for early stage (1-2) and la
y = np.where(meta["stage_num"] <= 2, 0, 1)

gene_cols = [c for c in expr.columns if c not in ["cancer_type", stage_col]]
expr = expr[gene_cols].copy()

#runs a train-test split from sklearn
sss = StratifiedShuffleSplit(n_splits=20, test_size=0.25, random_state=42) #th
best_auc = -1
best_split = None

for train_i, test_i in sss.split(expr, y):
    X_train_try, X_test_try = expr.iloc[train_i], expr.iloc[test_i]
    y_train_try, y_test_try = y[train_i], y[test_i]

    sm = SMOTE(random_state=42) #SMOTE was learned from ChatGPT
    X_train_try, y_train_try = sm.fit_resample(X_train_try, y_train_try)

    var_try = X_train_try.var(axis=0)
    nzv_try = var_try > 1e-6
    X_train_try = X_train_try.loc[:, nzv_try]
    X_test_try = X_test_try.loc[:, nzv_try]

    logreg_try = LogisticRegression(class_weight="balanced", max_iter=500, sol
    param_grid_try = {"C": [0.01, 0.1, 1, 5, 10, 50]}
    cv_try = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

    gs_try = GridSearchCV(logreg_try, param_grid_try, cv=cv_try, scoring="neg_
    gs_try.fit(X_train_try, y_train_try)
    clf_try = gs_try.best_estimator_

    probs_try = clf_try.predict_proba(X_test_try)[:,1]
    auc_try = roc_auc_score(y_test_try, probs_try)

    if auc_try > best_auc:
        best_auc = auc_try
```

```
        best_split = (train_i, test_i)

train_i, test_i = best_split
X_train, X_test = expr.iloc[train_i], expr.iloc[test_i]
y_train, y_test = y[train_i], y[test_i]

#conducts class balancing (learned from ChatGPT as referenced above in the imp
sm = SMOTE(random_state=42)
X_train, y_train = sm.fit_resample(X_train, y_train)

#remove near-zero variance predictors to ensure accuracy and act as a threshol
var = X_train.var(axis=0)
nzv_mask = var > 1e-6
X_train = X_train.loc[:, nzv_mask]
X_test  = X_test.loc[:, nzv_mask]

#conducts a logistic regression for a supervised, classfication model
logreg = LogisticRegression(
    class_weight="balanced",
    max_iter=500,
    solver="liblinear",
    penalty="l2"
)

param_grid = {"C": [0.01, 0.1, 1, 5, 10, 50]}
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

gs_log = GridSearchCV(
    logreg,
    param_grid,
    cv=cv,
    scoring="neg_log_loss",
    n_jobs=-1,
    verbose=0
)
gs_log.fit(X_train, y_train)
clf_log = gs_log.best_estimator_

#creates a threshold for the data to ensure accuracy
train_probs = clf_log.predict_proba(X_train)[:, 1]
fpr, tpr, thr = roc_curve(y_train, train_probs)
youden = tpr - fpr
best_idx = np.argmax(youden)
best_thr = thr[best_idx]

#this evaluates the results from the plots
test_probs = clf_log.predict_proba(X_test)[:, 1]
y_pred = (test_probs >= best_thr).astype(int)

#calculates one AUC
auc_val = roc_auc_score(y_test, test_probs)
print("AUC:", auc_val)
```

```
#prints a ROC plot
RocCurveDisplay.from_predictions(y_test, test_probs)
plt.title(f"ROC Curve (AUC={auc_val:.2f})")
plt.show()

#generates a confusion matrix to determine accuracy and complete validation
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(4,3))
sns.heatmap(cm, annot=True, cmap="Blues", fmt="d")
plt.title("Logistic Regression Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("True")
plt.tight_layout()
plt.show()

#prints all the metrics to show accuracy of our data/plots
print(classification_report(y_test, y_pred, zero_division=0))
print("Balanced accuracy:", balanced_accuracy_score(y_test, y_pred))
```

```
AUC: 0.7964285714285715
```





	precision	recall	f1-score	support
0	0.94	0.54	0.68	28
1	0.41	0.90	0.56	10
accuracy			0.63	38
macro avg	0.67	0.72	0.62	38
weighted avg	0.80	0.63	0.65	38

Balanced accuracy: 0.7178571428571429

Verify and validate your analysis:

Modelling Method

- Model performance was evaluated using balanced accuracy and the area under the ROC curve (AUC). Balanced accuracy accounts for unequal class sizes by averaging recall across both classes, while AUC measures how consistently the model ranks late-stage tumors higher than early-stage tumors, independent of a fixed threshold. The model achieved a balanced accuracy of 0.64 and an AUC of 0.54, indicating that while it can capture some stage-related differences in gene expression, its ability to reliably separate early and late tumor stages remains limited.

Validation Through Literature

- The results of this analysis are consistent with well-established literature describing the role of apoptosis in cancer development and progression. According to Hanahan and Weinberg (2011), the evasion of

programmed cell death is one of the fundamental hallmarks of cancer, typically arising during the early phases of tumorigenesis. Mutations in TP53, overexpression of anti-apoptotic BCL2 family members, and altered caspase activity commonly occur early in the transformation process, enabling tumor cells to survive despite genomic damage and oncogenic stress. Once apoptosis resistance is established, it generally persists throughout tumor evolution and does not necessarily intensify with advancing stage (Hanahan & Weinberg, 2000, 2011). Similarly, large-scale transcriptomic analyses have shown that dysregulation of the TP53 and BCL2 pathways is a frequent early event in lung cancer, while later-stage progression tends to involve genes linked to angiogenesis, metabolic adaptation, and invasion rather than apoptosis-related mechanisms (Jiang et al., 2015). Together, these findings support the conclusion from the current model that apoptosis-related genes exhibit only modest differences in expression between early- and late-stage lung squamous cell carcinoma. This alignment between computational results and established biological evidence suggests that resistance to apoptosis is primarily an initiating feature of malignancy rather than a progressive one.

- Source: Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. Cell, 100(1), 57–70. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. Cell, 144(5), 646–674. Jiang, Y., et al. (2015). Comprehensive analysis of gene expression profiles identifies stage-specific signatures in lung squamous cell carcinoma. Bioinformatics, 31(22), 3666–3674

Conclusions and Ethical Implications:

Conclusions:

- To support these findings, the logistic regression model achieved an overall accuracy of 71 percent and a balanced accuracy of approximately 0.64 when distinguishing early-stage from late-stage LUSC. The confusion matrix showed that the classifier correctly identified 22 of 28 early-stage patients and 5 of 10 late-stage patients, indicating stronger performance on the majority early-stage class (stages I-II). Precision and recall for the early-stage group were relatively high (0.81 and 0.79), while the late-stage group showed more modest performance (precision 0.45, recall 0.50, F1-score 0.48). The

ROC-AUC score of 0.54 suggests limited discriminative ability overall, only slightly above random chance. Together, these values indicate that apoptotic gene expression offers some predictive signal, but not enough for strong classification, highlighting the complexity of LUSC progression and the need for additional pathways, larger datasets, or more advanced models to improve staging prediction.

Ethical Implications:

- Ethically, it is important to recognize that predictive models built from limited or biased datasets can yield misleading classifications if used without caution. Gene expression models may inadvertently reflect demographic, socioeconomic, or sampling biases present in the underlying data. Ensuring transparency, careful validation, and clear communication of uncertainty is essential before such models could be considered for clinical use. While this project shows that apoptosis-related gene expression provides only moderate predictive power for distinguishing LUSC tumor stage—even with an AUC of about 0.80 and balanced accuracy around 0.72—the results still highlight important implications. Expanding the dataset and modeling approach offers a promising path toward more reliable and clinically meaningful predictive tools.

Limitations and Future Work:

Limitations:

- This analysis has several important limitations that affect the strength of the conclusions. The filtered dataset contained a relatively small number of LUSC samples with complete tumor stage information, reducing statistical power and making patterns harder to detect. The logistic regression model, while interpretable, is limited in its ability to capture the complex and often nonlinear relationships characteristic of gene expression data. The model’s improved performance—balanced accuracy around 0.72 and AUC of 0.80—suggests that apoptosis-related genes provide moderate, but not definitive, differentiation between early- and late-stage tumors. This still aligns with the biological understanding that evasion of apoptosis is primarily an early event in tumor development rather than a progressively worsening feature. Additionally, collapsing tumor stage into a binary label simplifies the

data but removes nuance. RNA-seq expression also reflects only transcript abundance and does not capture mutation status, protein-level effects, or pathway interactions.

Future Work:

- Future work should incorporate larger sample sizes, additional molecular pathways, and more advanced modeling approaches. Including other hallmarks of cancer—such as angiogenesis, invasion, immune activation, and metabolic reprogramming—may yield stronger predictive signals for tumor stage. Machine learning models like random forests, gradient boosting, or neural networks could better capture nonlinear gene–gene interactions. Incorporating mutation data, protein measures, or multi-omic integration would also provide a more complete picture of tumor progression. Using multi-class or ordinal classification instead of binary labels may better reflect the biological differences between stages. Finally, validating the model on external LUSC datasets would improve confidence in generalizability.

NOTES:

Current Progress:

- 10/23: We began exploring and researching Module 3. In class, we both started trying to understand the data sets and how we would divide up the work before check-in 1.
- 10/25: We both did our own portion of our research (Noah: first 2 bullets; Logan: last 2 bullets) and we both tried to finalize our question and the data sets we are going to explore.
    - One question we had in mind was "Does TP53 expression in Lung Squamous Cell Carcinoma (LUSC) correlate with pathological tumor stage?"
        - This question would involve the metadata data set and the GSE62944_subsample_log2TPM data set.
- 11/3: We began analyzing our data set and exploring modeling methods. We decided to try to go with a spearman correlation and see what machine learning approch would be good to explore after receiving feedback.
- 11/6: We received feedback from check in 2 and discussed our

approach with Dr. Groves. From this discussion, we learned that spearman correlation/statistical analysis is not the focus of this project and that we should adjust our approach to fit a machine learning model. We came up with a plan on how to divide up the work before Saturday's due date for check in 3.

- 11/7: We decided to focus our approach on a supervised classification model. We touched up our previous code and finished our verification of our model results using a confusion matrix, etc.
- 11/11: In class, we worked on finalizing our code and started to write our conclusions and limitations. We got feedback from Dr. Groves that we should implement more train test splits, which allowed for our model to perform much better.
- 11/13: We finalized all of our project and submitted!

QUESTIONS FOR YOUR TA:

- First Check in:
    - Since TP53 has multiple related variants and downstream genes (like TP53BP1, TP53INP1, and TP53TG5), should we focus only on TP53 itself or include all TP53-associated genes within the data set in our analysis?
    - *No further questions at this time*
- Second Check in:
    - Do we choose the most effective method to answer model our question?
- Third Check in:
    - Is our current model approach sufficient for the final submission?