Can we Predict Rectal Cancer Outcomes using Clinical Data? A Comparative Analysis of Different Techniques.

Karina Krishnan Grade 11, Beachwood High School | *Beachwood, Ohio, 44122*

Advisors: Professor Dr. Satish Viswanath, Mr. Thomas DeSilvio, Dr. Charlems Alvarez Jimenez (Case Western Reserve University)

Background

Rectal cancer is a subtype of colorectal cancer (CRC), the third most common cancer and the second leading cause of cancer-related deaths globally. Treatment differs from colon cancer due to the rectum's proximity to other organs, making surgical planning complex. Advances in MRI imaging have improved treatment decisions and outcome predictions.

Rectal cancer is staged using the TNM system, which assesses tumor size (T-stage), lymph node involvement (N-stage), and metastasis (M-stage).

Objective

The study aims to identify pre-surgery and MRI variables that significantly predict rectal cancer outcomes, measured by pathologic TNM staging and recurrence. This information is crucial for prognostic assessment, surgical planning, and treatment evaluation.

Data

The small sample but high-dimensional data used in my analysis is collected from Case Western Reserve University's Department of Biomedical Engineering, from 55 patients treated at University Hospitals Cleveland Medical Center.

Variables analyzed:

- Pre-surgery (control) variables: Sex, BMI, race, days from diagnosis to surgery, initial cancer staging, and tumor marker levels.
- MRI variables: Mucin production, tumor margins, lymph node involvement, and invasion of nearby structures.

There are 4 outcome variables:

- Pathologic T-Stage (*path_t_stage*): Measures tumor size and invasion, ranging from 0 (no tumor) to 4 (tumor spreading to nearby organs and lymph nodes)
- Pathologic N-Stage (*path_n_stage*): Assesses lymph node involvement, with 0 indicating no spread, 1 indicating limited metastasis, and 2 indicating extensive lymph node involvement.
- Pathologic M-Stage (*path_m_stage*): Evaluates distant metastasis, where 0 means no spread beyond nearby lymph nodes, 1 indicates distant metastasis, and 2 signifies more extensive spread.
- *recurrence*: A binary measure where 0 indicates no cancer recurrence after treatment, and 1 signifies recurrence within the follow-up period.

Hypothesis

Among the various pre-surgery and MRI variables available for colorectal cancer patients, Initial staging, or Clinical TNM among the pre-surgery variables, and the extent of lymph node involvement by cancer cells, in imaging variables are significantly associated with pathologic TNM and recurrence, consistently across all the different regression methods used.

Methodology, Data, & Results

All the continuous explanatory variables are first standardized into the z scores by subtracting the sample mean and dividing by sample standard deviation, to remove the dimensionality for the data but preserve the variability.

Then three different regression techniques are utilized to determine whether any variable of the variables is consistently and significantly associated with outcomes. Using Stata and Python programming, the regression results are examined with (Panel A of each table) only the imaging variables and (Panel B of each table) with both imaging variables and pre-surgery variables (or control variables).

Method 1: Tobit and Logit Regression

Since the dependent variables are not continuous variables, the Tobit regression method is utilized for *path_t_stage, path_n_stage,* and *path_m_stage* which can take ordered values, and the Logit regression method is used for *recurrence* that is 0 or 1 indicator variable.

Results

Panel A:

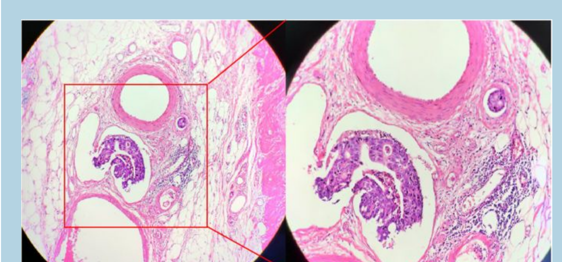
Significant Coefficients for Path T Stage	*Tobit*
mucin_present	**2.879**

Significant Coefficients for Path M Stage	*Tobit*
number_of_positive_lymph_n	**0.274**
lymphovascular_invasion	**3.314**

Panel B:

Significant Coefficients for Path M Stage	*Tobit*
number_of_positive_lymph_n	**0.572**
lymphovascular_invasion	**0.393**
sex	**-4.353**

number_of_positive_lymph_n and *lymphovascular_invasion* imaging variables are significantly associated with the path M Stage outcome in both Panels A and B.



presence of positive lymphovascular invasion

Method 3: Ridge & ElasticNet Regression

In Ridge regression, overfitting deals with multicollinearity problems by imposing penalties on the regression coefficients. ElasticNet is also chosen for its ability to combine the advantages of LASSO and Ridge regression, providing a robust approach to handling high-dimensional data and multicollinearity.

Results

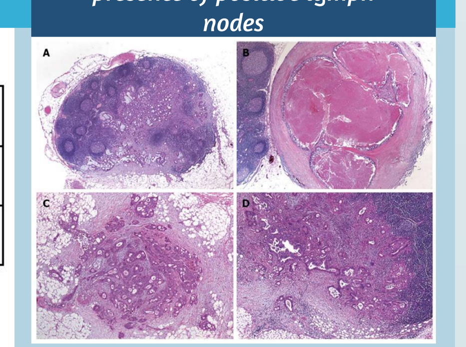
Panel A:

Significant Coefficients for Path N Stage	*Ridge*	*ElasticNet*
number_of_positive_lymph_n	**0.691**	**0.656**
lymphovascular_invasion	0.147	

Panel B:

Significant Coefficients for Path N Stage	*Ridge*	*ElasticNet*
init_clinical_staging_m	0.177	
number_of_positive_lymph_n	**0.514**	**0.562**
large_vessel_invasion	-0.139	

presence of positive lymph nodes



Imaging variable, *number_of_positive_lymph_n*, is significantly associated with path N Stage in both Panels A and B.

Method 2: Adaptive LASSO, SCAD, & MCP Regression

LASSO (Least Absolute Shrinkage and Selection Operator) performs variable selection and regularization, effectively selecting the variables that are most important to the response variable. Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP) improve model performance.

Results

Panel A:

Non-Zero Coefficients for Path T Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.073**	**0.036**	
distance_to_proximal_margin		-0.031	
number_of_lymph_nodes_exam		0.044	
Non-Zero Coefficients for Path N Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.214**	**0.156**	**0.090**

Non-Zero Coefficients for Path M Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	0.044		
distance_to_proximal_margin		-0.034	
distance_to_distal_margin		-0.022	
number_of_lymph_nodes_exam		-0.045	

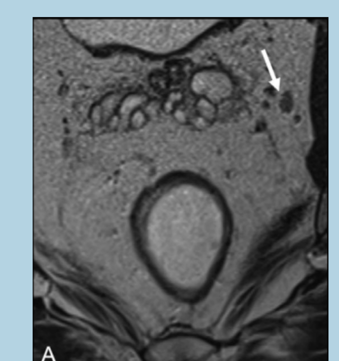
Panel B:

Non-Zero Coefficients for Path T Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.073**	**0.054**	
Sex		**-0.367**	**-0.340**
init_clinical_staging_m		**-0.174**	**-0.217**
Bmi		**0.158**	**0.058**
days_from_diagnosis_to_surgery		**-0.190**	**-0.132**
distance_to_distal_margin		**0.079**	**0.026**
number_of_lymph_nodes_exam		0.024	

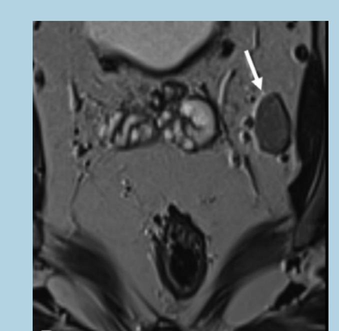
Non-Zero Coefficients for Path N Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.214**	**0.144**	**0.060**
Race		**0.177**	**0.297**
init_clinical_staging_m		**0.217**	**0.346**

Non-Zero Coefficients for Path M Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	0.044		
Race		**0.209**	**0.200**
init_clinical_staging_m		**0.148**	**0.206**
Bmi		-0.046	
days_from_neo_xrt_to_surgery		0.047	
distance_to_proximal_margin		-0.018	
distance_to_distal_margin		-0.044	
number_of_lymph_nodes_exam		**-0.035**	**-0.030**

Non-Zero Coefficients for Recurrence	*Adaptive Lasso*	*SCAD*	*MCP*
init_clinical_staging_m		-0.024	



no significant positive lymph nodes



presence of significant positive lymph nodes

Imaging variable, *number_of_positive_lymph_n*, is significantly associated with path T Stage and path N Stage outcomes in both Panels A and B. Among the pre-surgery or control variables, *init_clinical_staging_m* appears to be a significant predictor of path T Stage, path N Stage and path M Stage outcomes, and *race* appears to be a significant predictor of path N Stage and path M Stage outcomes in Panel B.

Discussion

In general, across almost all methods I used, the *number_of_positive_lymph_n* imaging variable is significant and positively associated with *path_t_stage, path_m_stage,* and *path_n_stage* outcomes. The *number_of_positive_lymph_n* refers to the number of lymph nodes to which cancer has spread, also known as the n-stage. Clinically, this aligns with the understanding that lymph node involvement worsens outcomes, as cancer spreads through the lymphatic system, increasing the risk of metastasis. Studies (Kroon et al., 2022; Sluckin et al., 2022) highlight the impact of lateral lymph node metastasis, particularly in locally advanced rectal cancer, on recurrence and survival rates.

Among control or pre-surgery variables, *init_clinical_staging_m* and *race* are generally significantly associated with *path_t_stage, path_m_stage,* and *path_n_stage outcomes. init_clinical_staging_m* refers to clinical M, or metastatic, stage, determined at diagnosis prior to any treatment. This emphasizes how racial disparities exist in rectal cancer survival, with black patients showing worse survival rates than white patients, even with similar treatments, likely due to biological and systemic factors.

Conclusion

The study supports the hypothesis that *number_of_positive_lymph_n* (imaging variable) and *init_clinical_staging_m* and *race* (pre-surgery variables) are key predictors of rectal cancer outcomes.

Despite the small sample size (55 cases), the study provides multidimensional insights into rectal cancer prognosis. Future work should expand the dataset and test findings over a longer period to improve reliability. As imaging technology advances, its role in predicting and improving patient outcomes will likely become even more critical.

Acknowledgements

I thank Professor Dr. Satish Viswanath, Mr. Thomas DeSilvio, and Dr. Charlems Alvarez Jimenez of Case Western Reserve University's Department of Biomedical Engineering, for the data and regular guidance in this project.

References

Krishnan, K. (2025). Can We Predict Rectal Cancer Outcomes Using Clinical Data? A Comparative Analysis of Different Techniques. Beachwood High School.

Adam Wetzel, Satish Viswanath, Emre Gorgun, Ilker Ozgur, Daniela Allende, David Liska, Andrei S Purysko, Staging and Restaging of Rectal Cancer with MRI: A Pictorial Review, Seminars in Ultrasound, CT and MRI, Volume 43, Issue 6, 2022, Pages 441-454, ISSN 0887-2171, https://doi.org/10.1053/j.sult.2022.06.003 A

Iessandra Borgheresi, Federica De Muzio, Andrea Agostini,Letizia Ottaviani,Alessandra Bruno, Vincenza Granata, Roberta Fusco, Ginevra Danti, Federica Flammia, Roberta Grassi, Francesca Grassi, Federico Bruno, Pierpaolo Palumbo, Antonio Barile, Vittorio Miele, and Andrea Giovagnoni,"Lymph Nodes Evaluation in Rectal Cancer: Where Do We Stand and Future Perspective." Journal of Clinical Medicine. 2022 May; 11(9), 2599,

Required Photographic/Graphics Source Identification

Photographs taken by: Nagtegaal et al, Smith et al, Wetzel et al

Graphics from outside sources are from: World Journal of Surgical Oncology (2021), World Journal of Gastroenterology (2013), Wetzel et al., Seminars in Ultrasound, CT, and MRI (2022)

Photographic permissions were obtained and are located: Open-Access Sources and Dr. Viswanath gave permission to use photographs from *Staging and Restaging of Rectal Cancer with MRI: A Pictorial Review* (Wetzel et al., 2022).

Required Photographic/Graphics Source Identification

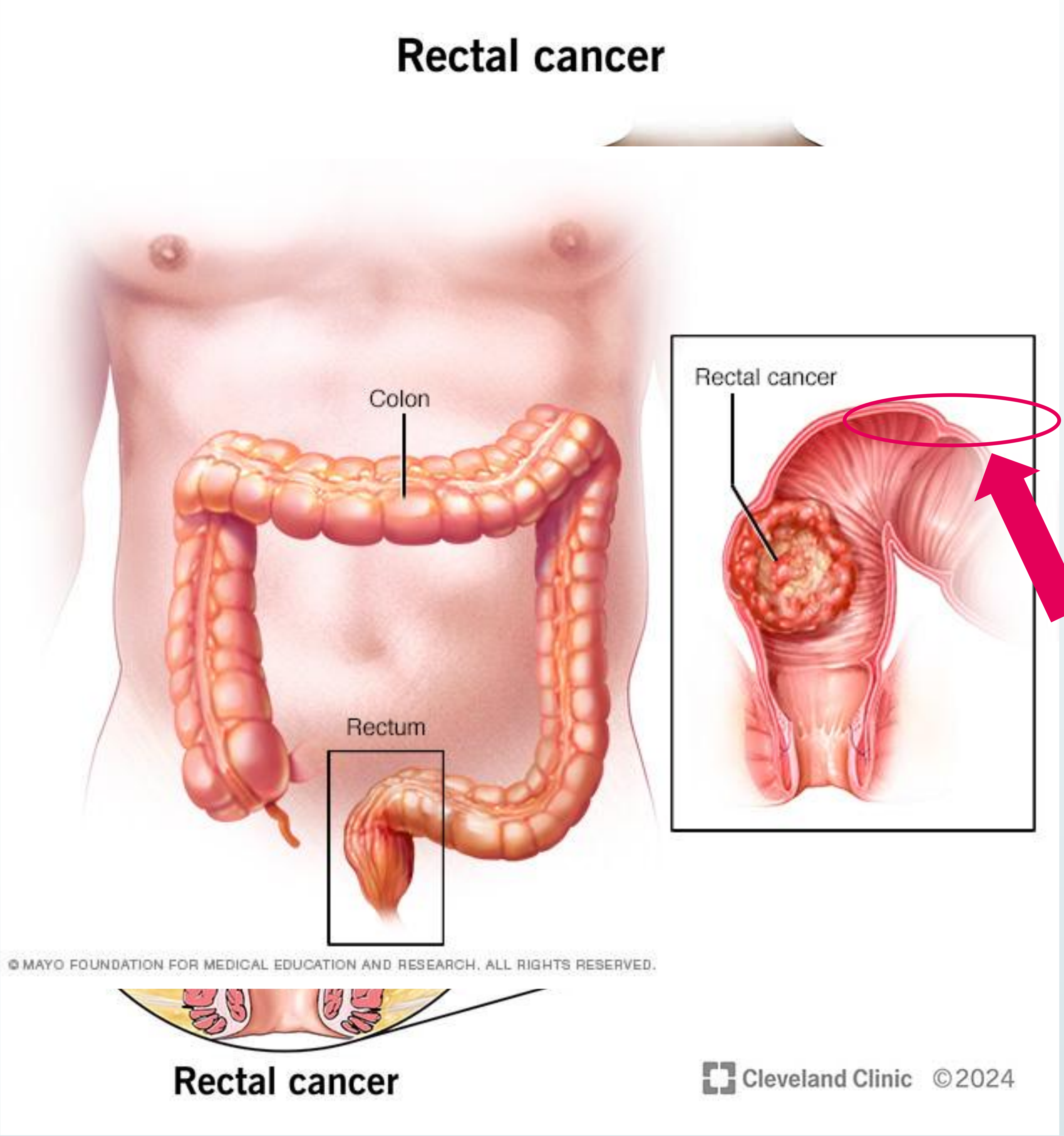
Graphics from outside sources are from: Cleveland Clinic et al, IHME Global Burden of Disease (2024) et al, Mayo Clinic et al,

Photographic permissions were obtained and are located: These are Open Access Sources

Background

Rectal cancer is a subtype of colorectal cancer (CRC), the third most common cancer and the second leading cause of cancer-related deaths globally. Treatment differs from colon cancer due to the rectum's proximity to other organs, making surgical planning complex. Advances in MRI imaging have improved treatment decisions and outcome predictions.

Rectal cancer is staged using the TNM system, which assesses tumor size (T-stage), lymph node involvement (N-stage), and metastasis (M-stage).



Objective

The study aims to identify pre-surgery and MRI variables that significantly predict rectal cancer outcomes, measured by pathologic TNM staging and recurrence. This information is crucial for prognostic assessment, surgical planning, and treatment evaluation.

Stata & Python Code:

Stata code:

```
tobit depvar [indepvars], ll[(#)] ul[(#)] [options]

  options          Description

  ------------------------------------------------------------------------------------------

  Model

    noconstant        suppress constant term

  * ll[(#)]          left-censoring limit

  * ul[(#)]          right-censoring limit

logit depvar [indepvars][, options]

  options              Description

  ------------------------------------------------------------------------------------------

  Model

    noconstant             suppress constant term
```

Python code:

- **Lasso**

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LassoCV, ElasticNetCV
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
from pyglmnet import GLM
import matplotlib.pyplot as plt

data = pd.read_excel('z score both.xlsx')

# Check for any missing values and drop rows with NaNs for LASSO regression
data = data.dropna()

# Separate independent and dependent variables
X = data[['sex', 'race', 'init_clinical_staging_t', 'init_clinical_staging_n', 'init_clinical_staging_m',
     'bmi', 'days_from_diagnosis_to_surgery', 'initial_cea', 'days_from_neo_xrt_to_surgery',
     'mucin_present', 'signet_ring_features', 'number_of_positive_lymph_n',
     'lymphovascular_invasion', 'perineural_invasion', 'peritumor_lymphocytic_resp',
     'large_vessel_invasion', 'ulceration_present', 'distance_to_proximal_margin',
     'distance_to_distal_margin', 'number_of_lymph_nodes_exam']]

y_path_t_stage = data['path_t_stage']
y_path_n_stage = data['path_n_stage']
y_path_m_stage = data['path_m_stage']
y_recurrence = data['recurrence']

# Adaptive Lasso implementation
def adaptive_lasso(X, y, cv=5):
    initial_lasso = make_pipeline(StandardScaler(), LassoCV(cv=cv)).fit(X, y)
    initial_coefs = np.abs(initial_lasso.named_steps['lassocv'].coef_)
    weights = 1 / (initial_coefs + 1e-2)
    adaptive_lasso = make_pipeline(StandardScaler(), LassoCV(cv=cv))
    adaptive_lasso.set_params(lassocv__alphas=np.linspace(0.1, 1, 10))
    adaptive_lasso.named_steps['lassocv'].fit(X * weights, y)
    return adaptive_lasso.named_steps['lassocv']

# Function to instantiate models afresh each time
def get_models(X, y):
    return {
        'Lasso': make_pipeline(StandardScaler(), LassoCV(cv=5)),
        'Adaptive Lasso': adaptive_lasso(X, y),
        'ElasticNet': make_pipeline(StandardScaler(), ElasticNetCV(cv=5)),
        'SCAD': GLM(distr='gaussian', reg_lambda=0.1, alpha=0.5, solver='cdfast', learning_rate=1e-3),
        'MCP': GLM(distr='gaussian', reg_lambda=0.1, alpha=1.0, solver='cdfast', learning_rate=1e-3)
    }

# Function to fit and plot regressions for a given target variable
def fit_and_plot_regressions(X, y, target_name):
    coefficients = {}
    models = get_models(X, y)
    for name, model in models.items():
        if name in ['SCAD', 'MCP']:
            model.fit(X.values, y.values)
            coefs = model.beta_.flatten()
        else:
            model.fit(X, y)
            if hasattr(model, 'named_steps'):
                coefs = model.named_steps[f'{name.lower()}cv'].coef_
            else:
                coefs = model.coef_
        coefficients[name] = pd.Series(coefs, index=X.columns)
        plt.figure(figsize=(10, 6))
        coefficients[name].plot(kind='bar')
        plt.title(f'{name} Coefficients for {target_name}')
        plt.show()
    return coefficients

# Fit and plot regressions for each dependent variable
coefficients_path_t_stage = fit_and_plot_regressions(X, y_path_t_stage, 'Path T Stage')
coefficients_path_n_stage = fit_and_plot_regressions(X, y_path_n_stage, 'Path N Stage')
coefficients_path_m_stage = fit_and_plot_regressions(X, y_path_m_stage, 'Path M Stage')
coefficients_recurrence = fit_and_plot_regressions(X, y_recurrence, 'Recurrence')

# Function to extract non-zero coefficients
def extract_non_zero_coefficients(coefficients):
    non_zero_coefs = {}
    for model, coef_series in coefficients.items():
        non_zero_coefs[model] = coef_series[coef_series != 0]
    return non_zero_coefs

# Extract non-zero coefficients
non_zero_coefficients_path_t_stage = extract_non_zero_coefficients(coefficients_path_t_stage)
non_zero_coefficients_path_n_stage = extract_non_zero_coefficients(coefficients_path_n_stage)
non_zero_coefficients_path_m_stage = extract_non_zero_coefficients(coefficients_path_m_stage)
non_zero_coefficients_recurrence = extract_non_zero_coefficients(coefficients_recurrence)
# Function to display non-zero coefficients in a table format
def display_non_zero_coefficients(non_zero_coefficients, target_name):
    print(f"Non-Zero Coefficients for {target_name}:")
    for model, coefs in non_zero_coefficients.items():
        print(f"\n{model}:")
        print(coefs)

# Display non-zero coefficients
display_non_zero_coefficients(non_zero_coefficients_path_t_stage, 'Path T Stage')
display_non_zero_coefficients(non_zero_coefficients_path_n_stage, 'Path N Stage')
display_non_zero_coefficients(non_zero_coefficients_path_m_stage, 'Path M Stage')
display_non_zero_coefficients(non_zero_coefficients_recurrence, 'Recurrence')
```

- **Ridge**

```
import pandas as pd
import numpy as np
from sklearn.linear_model import RidgeCV, Ridge, ElasticNetCV
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import cross_val_score, KFold
import matplotlib.pyplot as plt

# Load your new data
data = pd.read_excel('z score both.xlsx')

# Check for any missing values and drop rows with NaNs for Ridge regression
data = data.dropna()

# Separate independent and dependent variables
X = data[['sex', 'race', 'init_clinical_staging_t', 'init_clinical_staging_n', 'init_clinical_staging_m',
      'bmi', 'days_from_diagnosis_to_surgery', 'initial_cea', 'days_from_neo_xrt_to_surgery',
      'mucin_present', 'signet_ring_features', 'number_of_positive_lymph_n',
      'lymphovascular_invasion', 'perineural_invasion', 'peritumor_lymphocytic_resp',
      'large_vessel_invasion', 'ulceration_present', 'distance_to_proximal_margin',
      'distance_to_distal_margin', 'number_of_lymph_nodes_exam']]

y_path_t_stage = data['path_t_stage']
y_path_n_stage = data['path_n_stage']
y_path_m_stage = data['path_m_stage']
y_recurrence = data['recurrence']

# Define a threshold for significant coefficients (based on Gelman and Hill, 2007)
threshold = 0.1

# Function to fit and plot Ridge regression for dependent variables
def fit_and_plot_ridge(X, y, target_name):
    alphas = np.logspace(-6, 6, 13)
    ridge = make_pipeline(StandardScaler(), RidgeCV(alphas=alphas, cv=5))
    ridge.fit(X, y)
    coefficients = pd.Series(ridge.named_steps['ridgecv'].coef_, index=X.columns)

    # Plot coefficients
    plt.figure(figsize=(10, 6))
    coefficients.plot(kind='bar')
    plt.title(f'Ridge Coefficients for {target_name}')
    plt.show()

    # Report only significant coefficients
    significant_coefficients = coefficients[coefficients.abs() > threshold]

    print(f"Significant Ridge Regression Coefficients for {target_name}:")
    print(significant_coefficients)

    return significant_coefficients

# Function to fit and plot Multiple Penalty Ridge regression
def fit_and_plot_multiple_penalty_ridge(X, y, target_name):
    alphas = np.logspace(-6, 6, 13)
    ridge_scores = []
    best_alpha = 0
    best_score = -np.inf

    kf = KFold(n_splits=5, shuffle=True, random_state=1)

    for alpha in alphas:
        ridge = make_pipeline(StandardScaler(), Ridge(alpha=alpha))
        scores = cross_val_score(ridge, X, y, cv=kf, scoring='neg_mean_squared_error')
        mean_score = np.mean(scores)
        ridge_scores.append(mean_score)

        if mean_score > best_score:
            best_score = mean_score
            best_alpha = alpha

    # Fit the model with the best alpha
    best_ridge = make_pipeline(StandardScaler(), Ridge(alpha=best_alpha))
    best_ridge.fit(X, y)
    coefficients = pd.Series(best_ridge.named_steps['ridge'].coef_, index=X.columns)

    # Plot coefficients
    plt.figure(figsize=(10, 6))
    coefficients.plot(kind='bar')
    plt.title(f'Multiple Penalty Ridge Coefficients for {target_name}')
    plt.show()

    # Report only significant coefficients
    significant_coefficients = coefficients[coefficients.abs() > threshold]
    print(f"Significant Multiple Penalty Ridge Regression Coefficients for {target_name}:")
    print(significant_coefficients)

    return significant_coefficients

# Function to fit and plot Elastic Net regression for dependent variables
def fit_and_plot_elastic_net(X, y, target_name):
    l1_ratios = np.linspace(0.01, 1, 10)
    elastic_net = make_pipeline(StandardScaler(), ElasticNetCV(l1_ratio=l1_ratios, cv=5))
    elastic_net.fit(X, y)
    coefficients = pd.Series(elastic_net.named_steps['elasticnetcv'].coef_, index=X.columns)

    # Plot coefficients
    plt.figure(figsize=(10, 6))
    coefficients.plot(kind='bar')
    plt.title(f'Elastic Net Coefficients for {target_name}')
    plt.show()

    # Report only significant coefficients
    significant_coefficients = coefficients[coefficients.abs() > threshold]
    print(f"Significant Elastic Net Regression Coefficients for {target_name}:")
    print(significant_coefficients)

    return significant_coefficients

# Fit and plot Ridge regression for different dependent variables
coefficients_path_t_stage = fit_and_plot_ridge(X, y_path_t_stage, 'Path T Stage')
coefficients_path_n_stage = fit_and_plot_ridge(X, y_path_n_stage, 'Path N Stage')
coefficients_path_m_stage = fit_and_plot_ridge(X, y_path_m_stage, 'Path M Stage')
coefficients_recurrence = fit_and_plot_ridge(X, y_recurrence, 'Recurrence')

# Fit and plot Multiple Penalty Ridge regression for different dependent variables
coefficients_path_t_stage_multi = fit_and_plot_multiple_penalty_ridge(X, y_path_t_stage, 'Path T
Stage')
coefficients_path_n_stage_multi = fit_and_plot_multiple_penalty_ridge(X, y_path_n_stage, 'Path N
Stage')
coefficients_path_m_stage_multi = fit_and_plot_multiple_penalty_ridge(X, y_path_m_stage, 'Path M
Stage')
coefficients_recurrence_multi = fit_and_plot_multiple_penalty_ridge(X, y_recurrence, 'Recurrence')

# Fit and plot Elastic Net regression for different dependent variables
coefficients_path_t_stage_en = fit_and_plot_elastic_net(X, y_path_t_stage, 'Path T Stage')
coefficients_path_n_stage_en = fit_and_plot_elastic_net(X, y_path_n_stage, 'Path N Stage')
coefficients_path_m_stage_en = fit_and_plot_elastic_net(X, y_path_m_stage, 'Path M Stage')
coefficients_recurrence_en = fit_and_plot_elastic_net(X, y_recurrence, 'Recurrence')
```


PYTHON & STATA REFERENCES

Python References:

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, 105-142.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference.*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Greene, W. H. (2012). *Econometric analysis* (7th ed.). Pearson.

Stata References:

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics, 24*(1-2), 3-61.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies, 81*(2), 608-650.

Nichols, A., & Schaffer, M. (2007). Practice: Tobit models. *Stata Journal, 7*(2), 167-182.

StataCorp. (2021). *LASSO reference manual*. Stata Press.

StataCorp. (2021). *Stata 17 base reference manual*. Stata Press.

Wooldridge, J. M. (2010). *Econometric analysis of cross-section and panel data*. MIT Press.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics, 12*(1), 69-82.

Hypothesis

Among the various pre-surgery and MRI variables available for colorectal cancer patients, Initial staging, or Clinical TNM among the pre-surgery variables, and the extent of lymph node involvement by cancer cells, in imaging variables are significantly associated with pathologic TNM and recurrence, consistently across all the different regression methods used.

Methodology, Data, & Results

All the continuous explanatory variables are first standardized into the z scores by subtracting the sample mean and dividing by sample standard deviation, to remove the dimensionality for the data but preserve the variability.

Then three different regression techniques are utilized to determine whether any variable of the variables is consistently and significantly associated with outcomes. Using Stata and Python programming, the regression results are examined with (Panel A of each table) only the imaging variables and (Panel B of each table) with both imaging variables and pre-surgery variables (or control variables).

Method 1: Tobit and Logit Regression

Since the dependent variables are not continuous variables, the Tobit regression method is utilized for *path_t_stage, path_n_stage,* and *path_m_stage* which can take ordered values, and the Logit regression method is used for *recurrence* that is 0 or 1 indicator variable.

Results

Panel A:

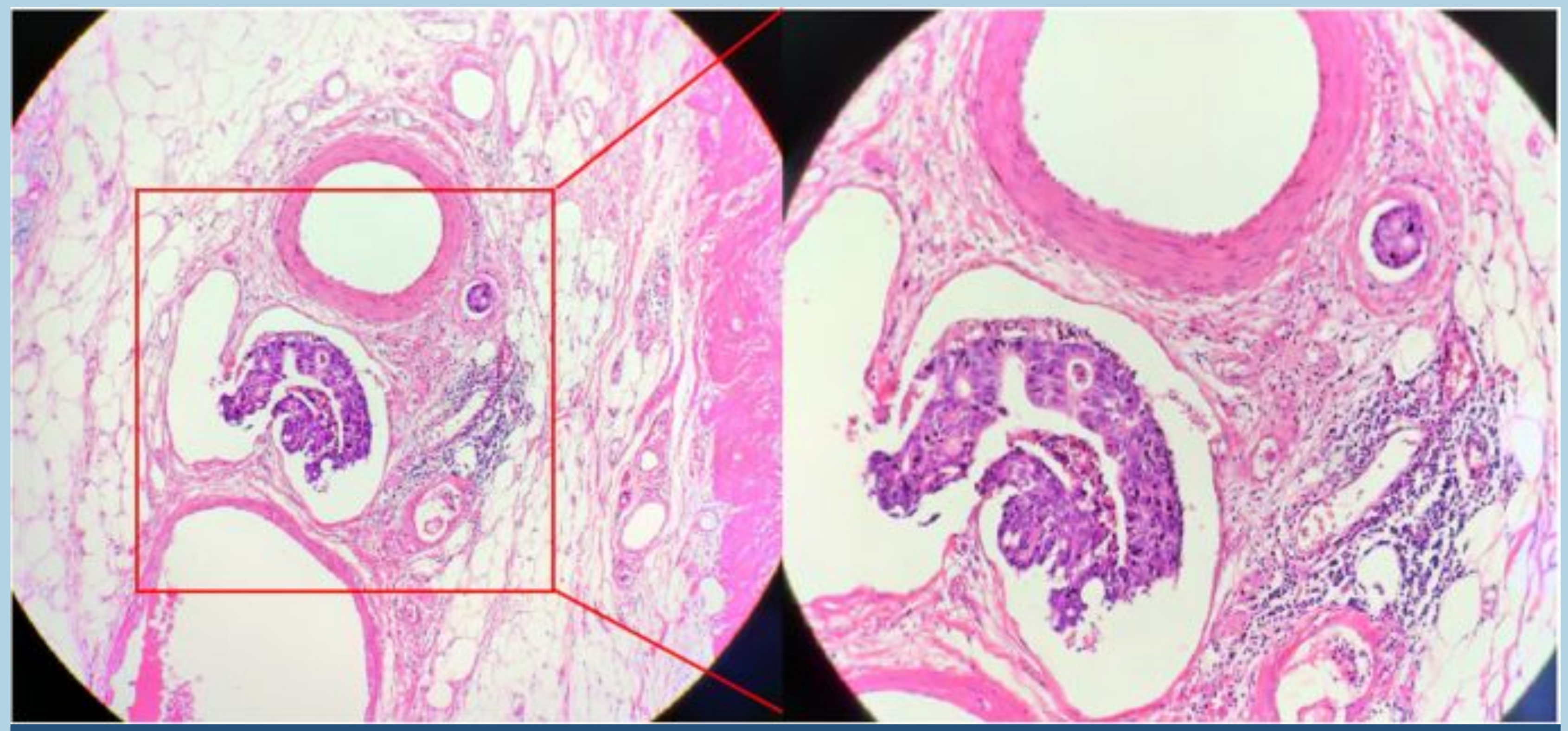
Significant Coefficients for Path T Stage	*Tobit*
mucin_present	**2.879**

Significant Coefficients for Path M Stage	*Tobit*
number_of_positive_lymph_n	**0.274**
lymphovascular_invasion	**3.314**

Panel B:

Significant Coefficients for Path M Stage	*Tobit*
number_of_positive_lymph_n	**0.572**
lymphovascular_invasion	**0.393**
sex	**-4.353**

number_of_positive_lymph_n and *lymphovascular_invasion* imaging variables are significantly associated with the path M Stage outcome in both Panels A and B.



presence of positive lymphovascular invasion

Can we Predict Rectal Cancer Outcomes using Clinical Data? A Comparative Analysis of Different Techniques.

Karina Krishnan Grade 11, Beachwood High School | *Beachwood, Ohio, 44122*

Advisors: Professor Dr. Satish Viswanath, Mr. Thomas DeSilvio, Dr. Charlems Alvarez Jimenez (Case Western Reserve University)

Background

Rectal cancer is a subtype of colorectal cancer (CRC), the third most common cancer and the second leading cause of cancer-related deaths globally. Treatment differs from colon cancer due to the rectum's proximity to other organs, making surgical planning complex. Advances in MRI imaging have improved treatment decisions and outcome predictions.

Rectal cancer is staged using the TNM system, which assesses tumor size (T-stage), lymph node involvement (N-stage), and metastasis (M-stage).

Objective

The study aims to identify pre-surgery and MRI variables that significantly predict rectal cancer outcomes, measured by pathologic TNM staging and recurrence. This information is crucial for prognostic assessment, surgical planning, and treatment evaluation.

Data

The small sample but high-dimensional data used in my analysis is collected from Case Western Reserve University's Department of Biomedical Engineering, from 55 patients treated at University Hospitals Cleveland Medical Center.

Variables analyzed:

- Pre-surgery (control) variables: Sex, BMI, race, days from diagnosis to surgery, initial cancer staging, and tumor marker levels.
- MRI variables: Mucin production, tumor margins, lymph node involvement, and invasion of nearby structures.

There are 4 outcome variables:

- Pathologic T-Stage (*path_t_stage*): Measures tumor size and invasion, ranging from 0 (no tumor) to 4 (tumor spreading to nearby organs and lymph nodes)
- Pathologic N-Stage (*path_n_stage*): Assesses lymph node involvement, with 0 indicating no spread, 1 indicating limited metastasis, and 2 indicating extensive lymph node involvement.
- Pathologic M-Stage (*path_m_stage*): Evaluates distant metastasis, where 0 means no spread beyond nearby lymph nodes, 1 indicates distant metastasis, and 2 signifies more extensive spread.
- *recurrence*: A binary measure where 0 indicates no cancer recurrence after treatment, and 1 signifies recurrence within the follow-up period.

Hypothesis

Among the various pre-surgery and MRI variables available for colorectal cancer patients, Initial staging, or Clinical TNM among the pre-surgery variables, and the extent of lymph node involvement by cancer cells, in imaging variables are significantly associated with pathologic TNM and recurrence, consistently across all the different regression methods used.

Methodology, Data, & Results

All the continuous explanatory variables are first standardized into the z scores by subtracting the sample mean and dividing by sample standard deviation, to remove the dimensionality for the data but preserve the variability.

Then three different regression techniques are utilized to determine whether any variable of the variables is consistently and significantly associated with outcomes. Using Stata and Python programming, the regression results are examined with (Panel A of each table) only the imaging variables and (Panel B of each table) with both imaging variables and pre-surgery variables (or control variables).

Method 1: Tobit and Logit Regression

Since the dependent variables are not continuous variables, the Tobit regression method is utilized for *path_t_stage, path_n_stage,* and *path_m_stage* which can take ordered values, and the Logit regression method is used for *recurrence* that is 0 or 1 indicator variable.

Results

Panel A:

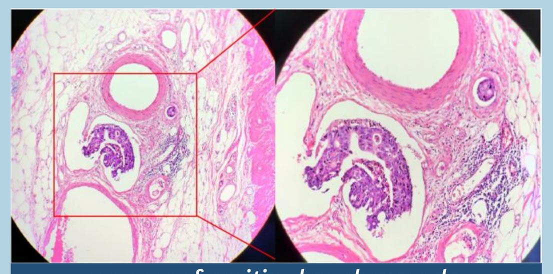
Significant Coefficients for Path T Stage	*Tobit*
mucin_present	**2.879**

Significant Coefficients for Path M Stage	*Tobit*
number_of_positive_lymph_n	**0.274**
lymphovascular_invasion	**3.314**

Panel B:

Significant Coefficients for Path M Stage	*Tobit*
number_of_positive_lymph_n	**0.572**
lymphovascular_invasion	**0.393**
sex	**-4.353**

number_of_positive_lymph_n and *lymphovascular_invasion* imaging variables are significantly associated with the path M Stage outcome in both Panels A and B.



presence of positive lymphovascular invasion

Method 3: Ridge & ElasticNet Regression

In Ridge regression, overfitting deals with multicollinearity problems by imposing penalties on the regression coefficients. ElasticNet is also chosen for its ability to combine the advantages of LASSO and Ridge regression, providing a robust approach to handling high-dimensional data and multicollinearity.

Results

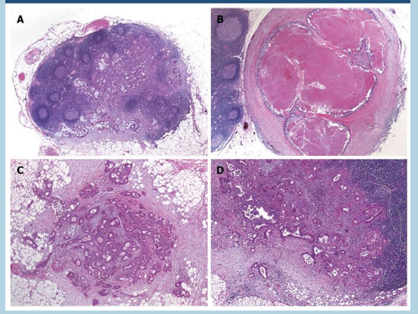
Panel A:

Significant Coefficients for Path N Stage	*Ridge*	*ElasticNet*
number_of_positive_lymph_n	**0.691**	**0.656**
lymphovascular_invasion	0.147	

Panel B:

Significant Coefficients for Path N Stage	*Ridge*	*ElasticNet*
init_clinical_staging_m	0.177	
number_of_positive_lymph_n	**0.514**	**0.562**
large_vessel_invasion	-0.139	

presence of positive lymph nodes



Imaging variable, *number_of_positive_lymph_n*, is significantly associated with path N Stage in both Panels A and B.

Method 2: Adaptive LASSO, SCAD, & MCP Regression

LASSO (Least Absolute Shrinkage and Selection Operator) performs variable selection and regularization, effectively selecting the variables that are most important to the response variable. Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP) improve model performance.

Results

Panel A:

Non-Zero Coefficients for Path T Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.073**	**0.036**	
distance_to_proximal_margin		-0.031	
number_of_lymph_nodes_exam		0.044	
Non-Zero Coefficients for Path N Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.214**	**0.156**	**0.090**

Non-Zero Coefficients for Path M Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	0.044		
distance_to_proximal_margin		-0.034	
distance_to_distal_margin		-0.022	
number_of_lymph_nodes_exam		-0.045	

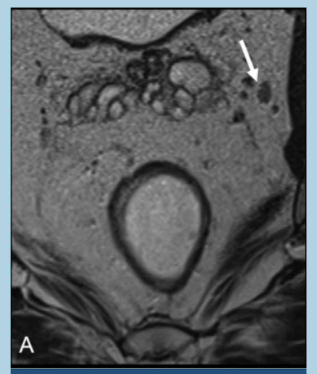
Panel B:

Non-Zero Coefficients for Path T Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.073**	**0.054**	
Sex		**-0.367**	**-0.340**
init_clinical_staging_m		**-0.174**	**-0.217**
Bmi		**0.158**	**0.058**
days_from_diagnosis_to_surgery		**-0.190**	**-0.132**
distance_to_distal_margin		**0.079**	**0.026**
number_of_lymph_nodes_exam		0.024	

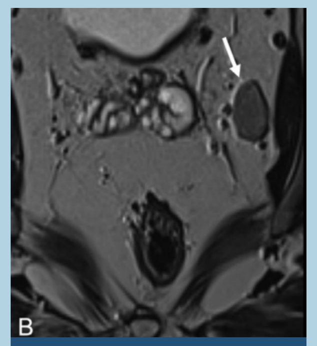
Non-Zero Coefficients for Path N Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.214**	**0.144**	**0.060**
Race		**0.177**	**0.297**
init_clinical_staging_m		**0.217**	**0.346**

Non-Zero Coefficients for Path M Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	0.044		
Race		**0.209**	**0.200**
init_clinical_staging_m		**0.148**	**0.206**
Bmi		-0.046	
days_from_neo_xrt_to_surgery		0.047	
distance_to_proximal_margin		-0.018	
distance_to_distal_margin		-0.044	
number_of_lymph_nodes_exam		**-0.035**	**-0.030**

Non-Zero Coefficients for Recurrence	*Adaptive Lasso*	*SCAD*	*MCP*
init_clinical_staging_m		-0.024	



no significant positive lymph nodes



presence of significant positive lymph nodes

Imaging variable, *number_of_positive_lymph_n*, is significantly associated with path T Stage and path N Stage outcomes in both Panels A and B. Among the pre-surgery or control variables, *init_clinical_staging_m* appears to be a significant predictor of path T Stage, path N Stage and path M Stage outcomes, and *race* appears to be a significant predictor of path N Stage and path M Stage outcomes in Panel B.

Discussion

In general, across almost all methods I used, the *number_of_positive_lymph_n* imaging variable is significant and positively associated with *path_t_stage, path_m_stage,* and *path_n_stage* outcomes. The *number_of_positive_lymph_n* refers to the number of lymph nodes to which cancer has spread, also known as the n-stage. Clinically, this aligns with the understanding that lymph node involvement worsens outcomes, as cancer spreads through the lymphatic system, increasing the risk of metastasis. Studies (Kroon et al., 2022; Sluckin et al., 2022) highlight the impact of lateral lymph node metastasis, particularly in locally advanced rectal cancer, on recurrence and survival rates.

Among control or pre-surgery variables, *init_clinical_staging_m* and *race* are generally significantly associated with *path_t_stage, path_m_stage,* and *path_n_stage outcomes. init_clinical_staging_m* refers to clinical M, or metastatic, stage, determined at diagnosis prior to any treatment. This emphasizes how racial disparities exist in rectal cancer survival, with black patients showing worse survival rates than white patients, even with similar treatments, likely due to biological and systemic factors.

Conclusion

The study supports the hypothesis that *number_of_positive_lymph_n* (imaging variable) and *init_clinical_staging_m* and *race* (pre-surgery variables) are key predictors of rectal cancer outcomes.

Despite the small sample size (55 cases), the study provides multidimensional insights into rectal cancer prognosis. Future work should expand the dataset and test findings over a longer period to improve reliability. As imaging technology advances, its role in predicting and improving patient outcomes will likely become even more critical.

Acknowledgements

I thank Professor Dr. Satish Viswanath, Mr. Thomas DeSilvio, and Dr. Charlems Alvarez Jimenez of Case Western Reserve University's Department of Biomedical Engineering, for the data and regular guidance in this project.

References

Krishnan, K. (2025). Can We Predict Rectal Cancer Outcomes Using Clinical Data? A Comparative Analysis of Different Techniques. Beachwood High School.

Adam Wetzel, Satish Viswanath, Emre Gorgun, Ilker Ozgur, Daniela Allende, David Liska, Andrei S Purysko, Staging and Restaging of Rectal Cancer with MRI: A Pictorial Review, Seminars in Ultrasound, CT and MRI, Volume 43, Issue 6, 2022, Pages 441-454, ISSN 0887-2171, https://doi.org/10.1053/j.sult.2022.06.003 A

lessandra Borgheresi, Federica De Muzio, Andrea Agostini,Letizia Ottaviani,Alessandra Bruno, Vincenza Granata, Roberta Fusco, Ginevra Danti, Federica Flammia, Roberta Grassi, Francesca Grassi, Federico Bruno, Pierpaolo Palumbo, Antonio Barile, Vittorio Miele, and Andrea Giovagnoni,"Lymph Nodes Evaluation in Rectal Cancer: Where Do We Stand and Future Perspective." Journal of Clinical Medicine. 2022 May; 11(9), 2599,

Required Photographic/Graphics Source Identification

Photographs taken by: Nagtegaal et al, Smith et al, Wetzel et al

Graphics from outside sources are from: World Journal of Surgical Oncology (2021), World Journal of Gastroenterology (2013), Wetzel et al., Seminars in Ultrasound, CT, and MRI (2022)

Photographic permissions were obtained and are located: Open-Access Sources and Dr. Viswanath gave permission to use photographs from *Staging and Restaging of Rectal Cancer with MRI: A Pictorial Review* (Wetzel et al., 2022).

Method 2: Adaptive LASSO, SCAD, & MCP Regression

LASSO (Least Absolute Shrinkage and Selection Operator) performs variable selection and regularization, effectively selecting the variables that are most important to the response variable. Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP) improve model performance.

Results

Panel A:

Non-Zero Coefficients for Path T Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.073**	**0.036**	
distance_to_proximal_margin		-0.031	
number_of_lymph_nodes_exam		0.044	

Non-Zero Coefficients for Path N Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.214**	**0.156**	**0.090**

Non-Zero Coefficients for Path M Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	0.044		
distance_to_proximal_margin		-0.034	
distance_to_distal_margin		-0.022	
number_of_lymph_nodes_exam		-0.045	

Panel B:

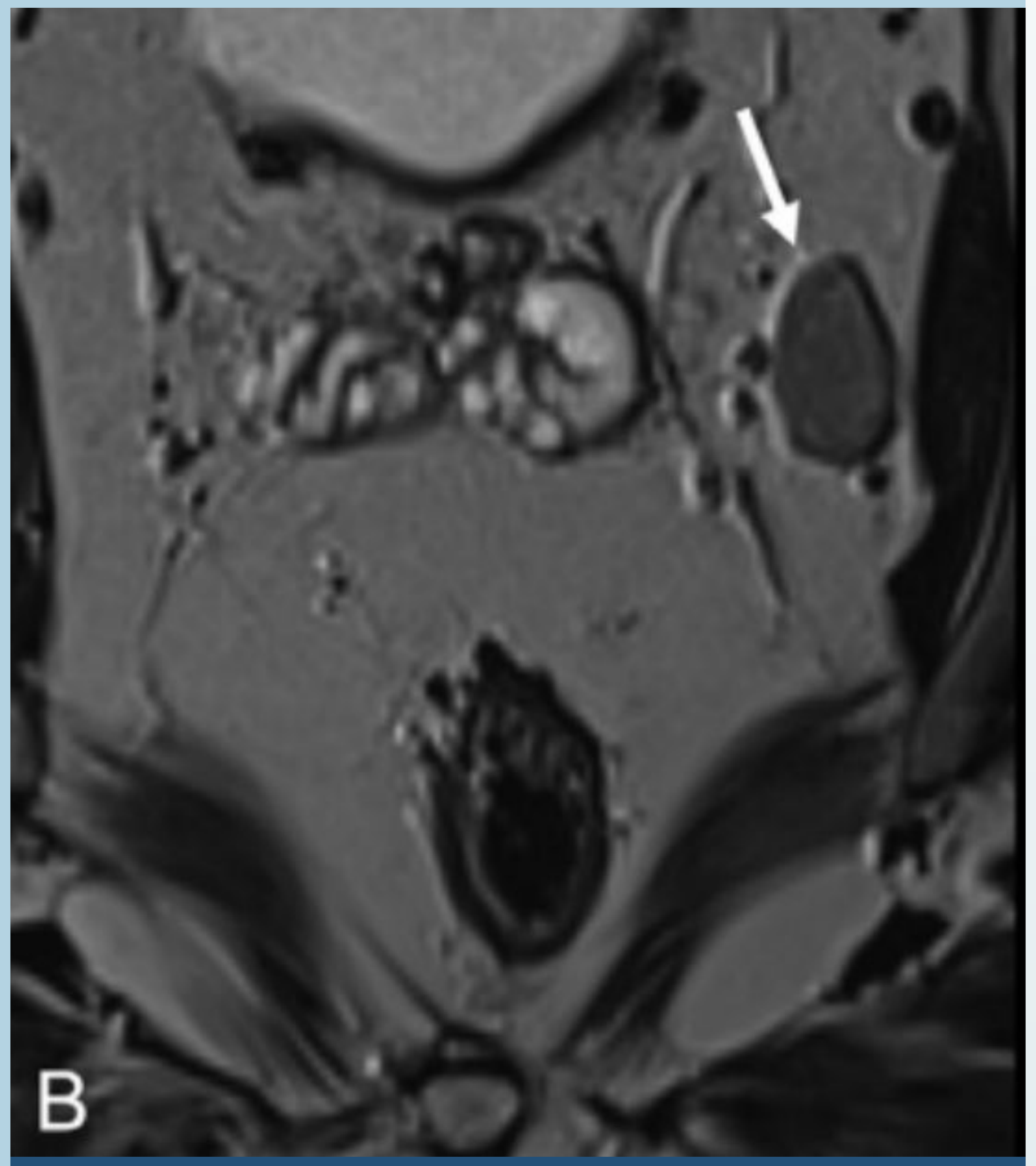
Non-Zero Coefficients for Path T Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.073**	**0.054**	
Sex		**-0.367**	**-0.340**
init_clinical_staging_m		**-0.174**	**-0.217**
Bmi		**0.158**	**0.058**
days_from_diagnosis_to_surgery		**-0.190**	**-0.132**
distance_to_distal_margin		**0.079**	**0.026**
number_of_lymph_nodes_exam		0.024	

Non-Zero Coefficients for Path N Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.214**	**0.144**	**0.060**
Race		**0.177**	**0.297**
init_clinical_staging_m		**0.217**	**0.346**

Non-Zero Coefficients for Path M Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	0.044		
Race		**0.209**	**0.200**
init_clinical_staging_m		**0.148**	**0.206**
Bmi		-0.046	
days_from_neo_xrt_to_surgery		0.047	
distance_to_proximal_margin		-0.018	
distance_to_distal_margin		-0.044	
number_of_lymph_nodes_exam		**-0.035**	**-0.030**

Non-Zero Coefficients for Recurrence	*Adaptive Lasso*	*SCAD*	*MCP*
init_clinical_staging_m		-0.024	

no significant positive lymph nodes



presence of significant positive lymph nodes

Imaging variable, *number_of_positive_lymph_n*, is significantly associated with path T Stage and path N Stage outcomes in both Panels A and B. Among the pre-surgery or control variables, *init_clinical_staging_m* appears to be a significant predictor of path T Stage, path N Stage and path M Stage outcomes, and *race* appears to be a significant predictor of path N Stage and path M Stage outcomes in Panel B.

Method 3: Ridge & ElasticNet Regression

In Ridge regression, overfitting deals with multicollinearity problems by imposing penalties on the regression coefficients. ElasticNet is also chosen for its ability to combine the advantages of LASSO and Ridge regression, providing a robust approach to handling high-dimensional data and multicollinearity.

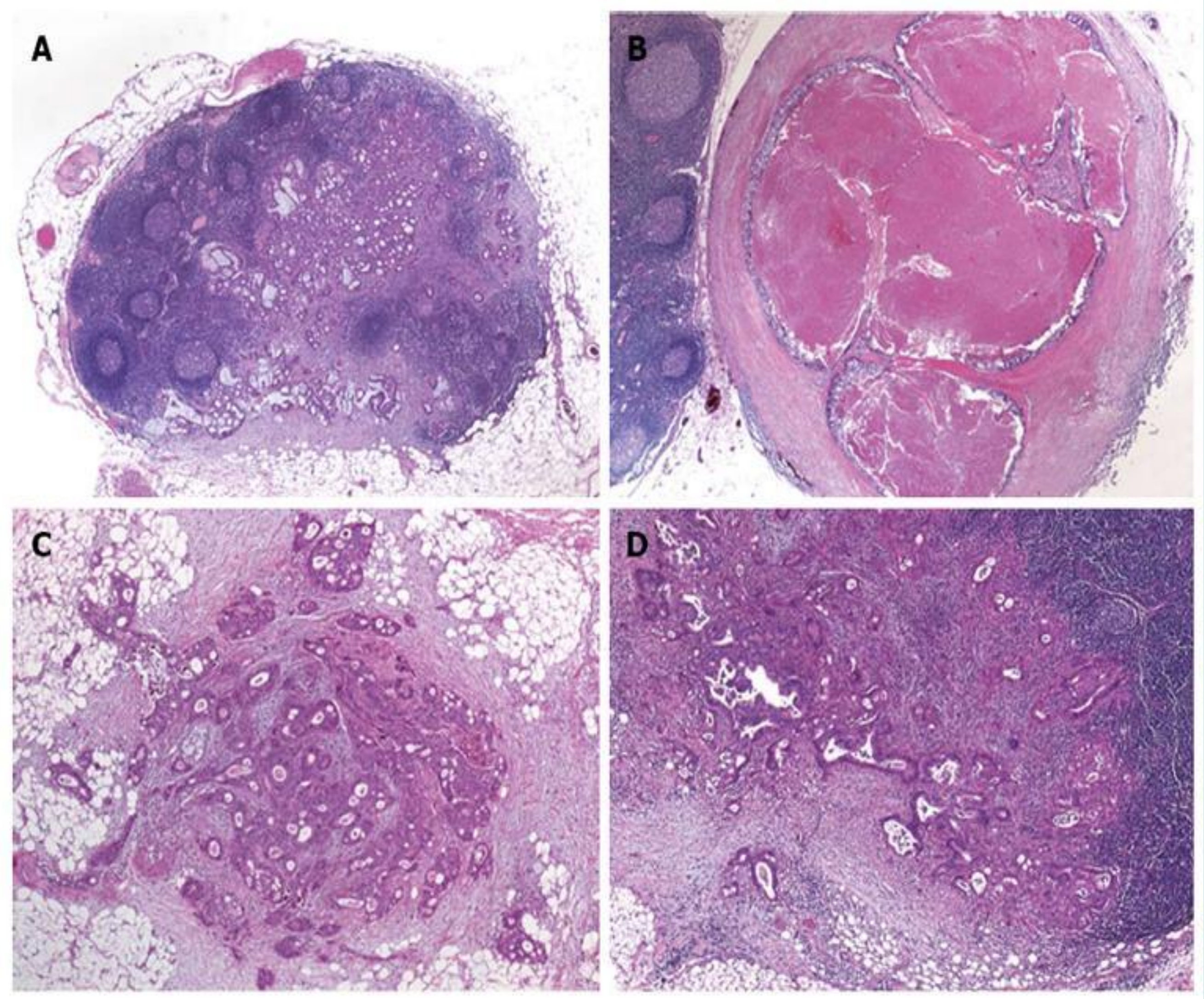
Results

Panel A:

Significant Coefficients for Path N Stage	*Ridge*	*ElasticNet*
number_of_positive_lymph_n	**0.691**	**0.656**
lymphovascular_invasion	0.147	

Panel B:

Significant Coefficients for Path N Stage	*Ridge*	*ElasticNet*
init_clinical_staging_m	0.177	
number_of_positive_lymph_n	**0.514**	**0.562**
large_vessel_invasion	-0.139	



presence of positive lymph nodes

Imaging variable, *number_of_positive_lymph_n*, is significantly associated with path N Stage in both Panels A and B.

Required Photographic/Graphics Source Identification

Photographs taken by: Doubeni et al, Nagtegaal et al, Smith et al, Wetzel et al

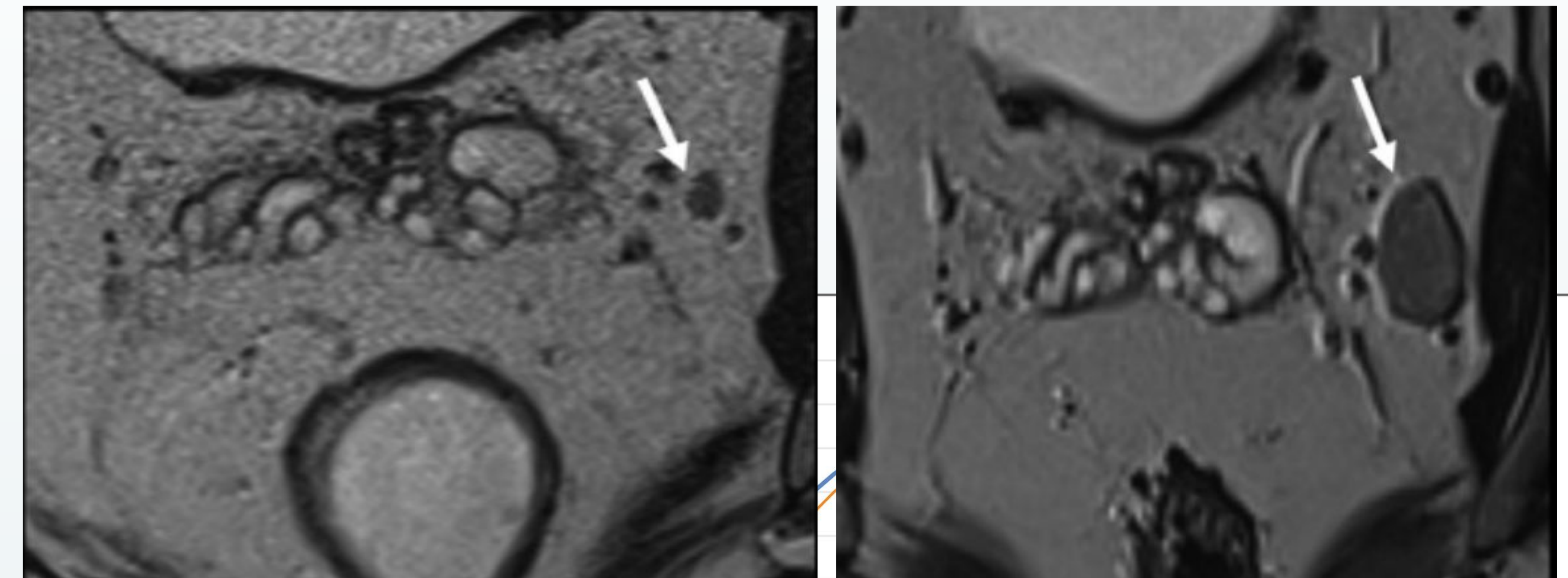
Graphics from outside sources are from: World Journal of Surgical Oncology (2021), World Journal of Gastroenterology (2013), Wetzel et al., Seminars in Ultrasound, CT, and MRI (2022), The New England Journal of Medicine (2022)

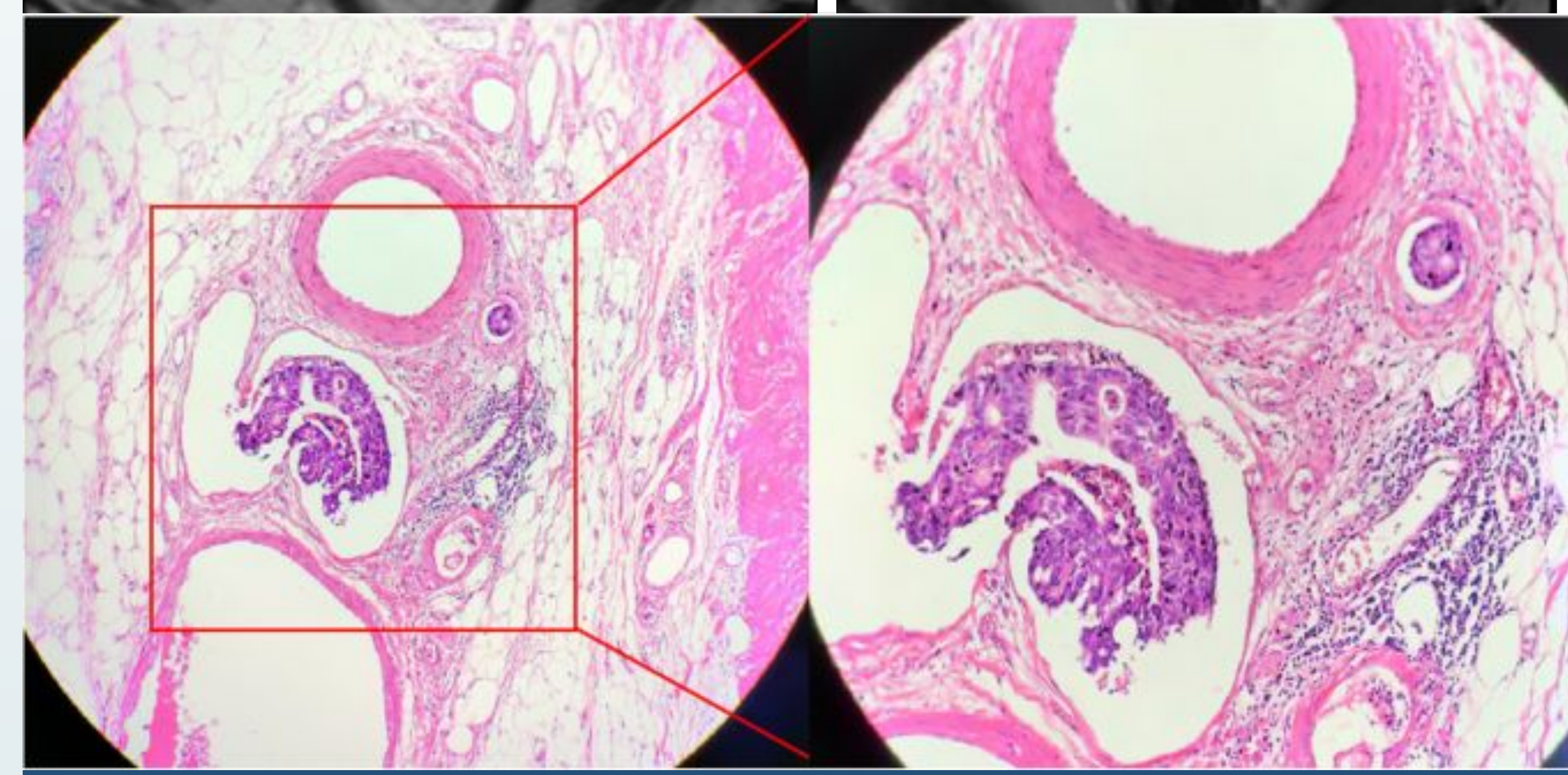
Photographic permissions were obtained and are located: Open-Access Sources and Dr. Viswanath gave permission to use photographs from *Staging and Restaging of Rectal Cancer with MRI: A Pictorial Review* (Wetzel et al., 2022).

Discussion

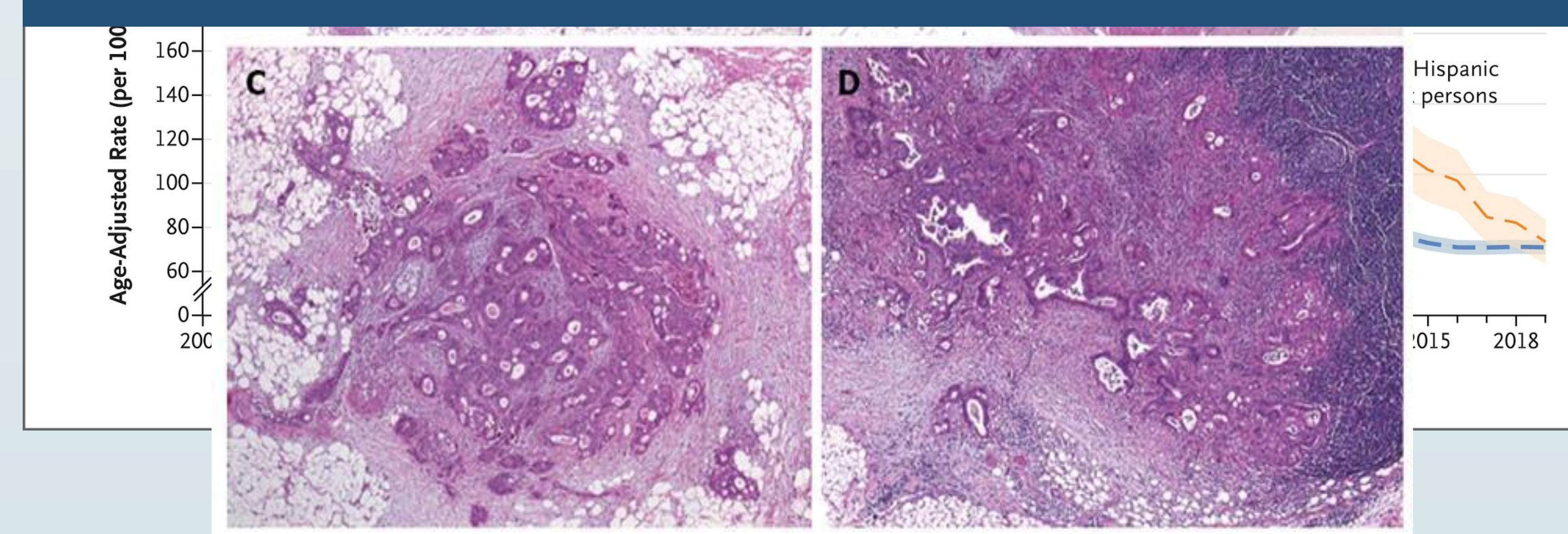
In general, across almost all methods I used, the *number_of_positive_lymph_n* imaging variable is significant and positively associated with *path_t_stage, path_m_stage,* and *path_n_stage* outcomes. The *number_of_positive_lymph_n* refers to the number of lymph nodes to which cancer has spread, also known as the n-stage. Clinically, this aligns with the understanding that lymph node involvement worsens outcomes, as cancer spreads through the lymphatic system, increasing the risk of metastasis. Studies (Kroon et al., 2022; Sluckin et al., 2022) highlight the impact of lateral lymph node metastasis, particularly in locally advanced rectal cancer, on recurrence and survival rates.

Among control or pre-surgery variables, *init_clinical_staging_m* and *race* are generally significantly associated with *path_t_stage, path_m_stage,* and *path_n_stage outcomes*. *init_clinical_staging_m* refers to clinical M, or metastatic, stage, determined at diagnosis prior to any treatment. This emphasizes how racial disparities exist in rectal cancer survival, with black patients showing worse survival rates than white patients, even with similar treatments, likely due to biological and systemic factors.

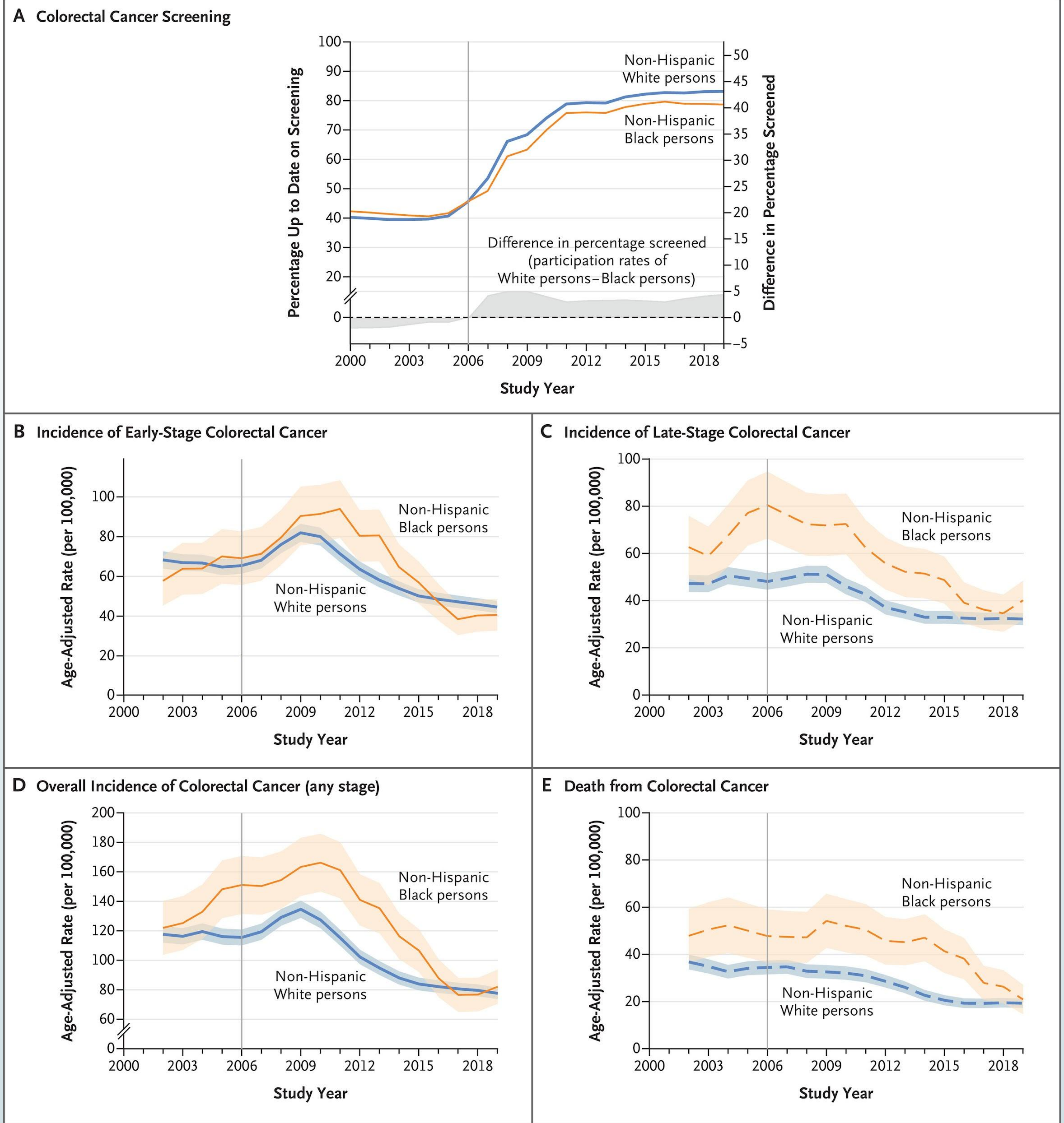




presence of positive lymphovascular invasion



presence of positive lymph nodes


A Colorectal Cancer Screening
Percentage Up to Date on Screening
Difference in Percentage Screened
Non-Hispanic White persons
Non-Hispanic Black persons
Difference in percentage screened (participation rates of White persons−Black persons)
Study Year
B Incidence of Early-Stage Colorectal Cancer
C Incidence of Late-Stage Colorectal Cancer
D Overall Incidence of Colorectal Cancer (any stage)
E Death from Colorectal Cancer
Age-Adjusted Rate (per 100,000)
Non-Hispanic Black persons
Non-Hispanic White persons
Study Year

Conclusion

The study supports the hypothesis that *number_of_positive_lymph_n* (imaging variable) and *init_clinical_staging_m* and *race* (pre-surgery variables) are key predictors of rectal cancer outcomes.

Despite the small sample size (55 cases), the study provides multidimensional insights into rectal cancer prognosis. Future work should expand the dataset and test findings over a longer period to improve reliability. As imaging technology advances, its role in predicting and improving patient outcomes will likely become even more critical.

Can we Predict Rectal Cancer Outcomes using Clinical Data? A Comparative Analysis of Different Techniques.

Karina Krishnan Grade 11, Beachwood High School | *Beachwood, Ohio, 44122*

Advisors: Professor Dr. Satish Viswanath, Mr. Thomas DeSilvio, Dr. Charlems Alvarez Jimenez (Case Western Reserve University)

Background

Rectal cancer is a subtype of colorectal cancer (CRC), the third most common cancer and the second leading cause of cancer-related deaths globally. Treatment differs from colon cancer due to the rectum's proximity to other organs, making surgical planning complex. Advances in MRI imaging have improved treatment decisions and outcome predictions.

Rectal cancer is staged using the TNM system, which assesses tumor size (T-stage), lymph node involvement (N-stage), and metastasis (M-stage).

Objective

The study aims to identify pre-surgery and MRI variables that significantly predict rectal cancer outcomes, measured by pathologic TNM staging and recurrence. This information is crucial for prognostic assessment, surgical planning, and treatment evaluation.

Data

The small sample but high-dimensional data used in my analysis is collected from Case Western Reserve University's Department of Biomedical Engineering, from 55 patients treated at University Hospitals Cleveland Medical Center.

Variables analyzed:

- Pre-surgery (control) variables: Sex, BMI, race, days from diagnosis to surgery, initial cancer staging, and tumor marker levels.
- MRI variables: Mucin production, tumor margins, lymph node involvement, and invasion of nearby structures.

There are 4 outcome variables:

- Pathologic T-Stage (*path_t_stage*): Measures tumor size and invasion, ranging from 0 (no tumor) to 4 (tumor spreading to nearby organs and lymph nodes)
- Pathologic N-Stage (*path_n_stage*): Assesses lymph node involvement, with 0 indicating no spread, 1 indicating limited metastasis, and 2 indicating extensive lymph node involvement.
- Pathologic M-Stage (*path_m_stage*): Evaluates distant metastasis, where 0 means no spread beyond nearby lymph nodes, 1 indicates distant metastasis, and 2 signifies more extensive spread.
- *recurrence*: A binary measure where 0 indicates no cancer recurrence after treatment, and 1 signifies recurrence within the follow-up period.

Hypothesis

Among the various pre-surgery and MRI variables available for colorectal cancer patients, Initial staging, or Clinical TNM among the pre-surgery variables, and the extent of lymph node involvement by cancer cells, in imaging variables are significantly associated with pathologic TNM and recurrence, consistently across all the different regression methods used.

Methodology, Data, & Results

All the continuous explanatory variables are first standardized into the z scores by subtracting the sample mean and dividing by sample standard deviation, to remove the dimensionality for the data but preserve the variability.

Then three different regression techniques are utilized to determine whether any variable of the variables is consistently and significantly associated with outcomes. Using Stata and Python programming, the regression results are examined with (Panel A of each table) only the imaging variables and (Panel B of each table) with both imaging variables and pre-surgery variables (or control variables).

Method 1: Tobit and Logit Regression

Since the dependent variables are not continuous variables, the Tobit regression method is utilized for *path_t_stage, path_n_stage,* and *path_m_stage* which can take ordered values, and the Logit regression method is used for *recurrence* that is 0 or 1 indicator variable.

Results

Panel A:

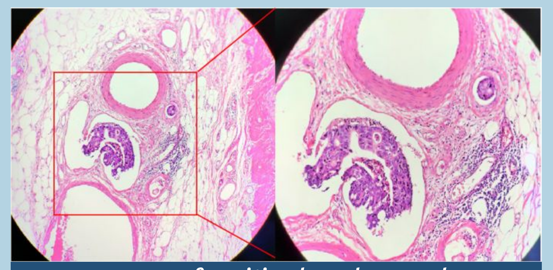
Significant Coefficients for Path T Stage	*Tobit*
mucin_present	**2.879**

Significant Coefficients for Path M Stage	*Tobit*
number_of_positive_lymph_n	**0.274**
lymphovascular_invasion	**3.314**

Panel B:

Significant Coefficients for Path M Stage	*Tobit*
number_of_positive_lymph_n	**0.572**
lymphovascular_invasion	**0.393**
sex	**-4.353**

number_of_positive_lymph_n and *lymphovascular_invasion* imaging variables are significantly associated with the path M Stage outcome in both Panels A and B.



presence of positive lymphovascular invasion

Method 2: Adaptive LASSO, SCAD, & MCP Regression

LASSO (Least Absolute Shrinkage and Selection Operator) performs variable selection and regularization, effectively selecting the variables that are most important to the response variable. Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP) improve model performance.

Results

Panel A:

Non-Zero Coefficients for Path T Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.073**	**0.036**	
distance_to_proximal_margin		-0.031	
number_of_lymph_nodes_exam		0.044	
Non-Zero Coefficients for Path N Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.214**	**0.156**	**0.090**

Non-Zero Coefficients for Path M Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	0.044		
distance_to_proximal_margin		-0.034	
distance_to_distal_margin		-0.022	
number_of_lymph_nodes_exam		-0.045	

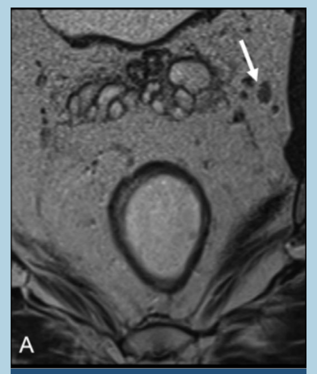
Panel B:

Non-Zero Coefficients for Path T Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.073**	**0.054**	
Sex		**-0.367**	**-0.340**
init_clinical_staging_m		**-0.174**	**-0.217**
Bmi		**0.158**	**0.058**
days_from_diagnosis_to_surgery		**-0.190**	**-0.132**
distance_to_distal_margin		**0.079**	**0.026**
number_of_lymph_nodes_exam		0.024	

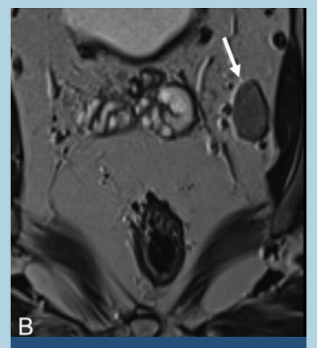
Non-Zero Coefficients for Path N Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	**0.214**	**0.144**	**0.060**
Race		**0.177**	**0.297**
init_clinical_staging_m		**0.217**	**0.346**

Non-Zero Coefficients for Path M Stage	*Adaptive Lasso*	*SCAD*	*MCP*
number_of_positive_lymph_n	0.044		
Race		**0.209**	**0.200**
init_clinical_staging_m		**0.148**	**0.206**
Bmi		-0.046	
days_from_neo_xrt_to_surgery		0.047	
distance_to_proximal_margin		-0.018	
distance_to_distal_margin		-0.044	
number_of_lymph_nodes_exam		**-0.035**	**-0.030**

Non-Zero Coefficients for Recurrence	*Adaptive Lasso*	*SCAD*	*MCP*
init_clinical_staging_m		-0.024	



no significant positive lymph nodes



presence of significant positive lymph nodes

Imaging variable, *number_of_positive_lymph_n*, is significantly associated with path T Stage and path N Stage outcomes in both Panels A and B. Among the pre-surgery or control variables, *init_clinical_staging_m* appears to be a significant predictor of path T Stage, path N Stage and path M Stage outcomes, and *race* appears to be a significant predictor of path N Stage and path M Stage outcomes in Panel B.

Method 3: Ridge & ElasticNet Regression

In Ridge regression, overfitting deals with multicollinearity problems by imposing penalties on the regression coefficients. ElasticNet is also chosen for its ability to combine the advantages of LASSO and Ridge regression, providing a robust approach to handling high-dimensional data and multicollinearity.

Results

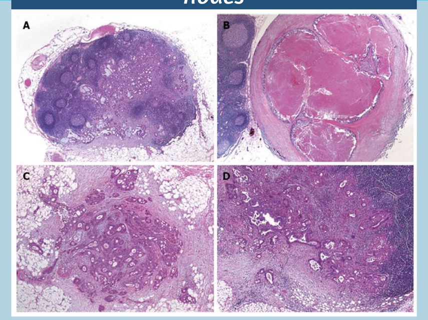
Panel A:

Significant Coefficients for Path N Stage	*Ridge*	*ElasticNet*
number_of_positive_lymph_n	**0.691**	**0.656**
lymphovascular_invasion	0.147	

Panel B:

Significant Coefficients for Path N Stage	*Ridge*	*ElasticNet*
init_clinical_staging_m	0.177	
number_of_positive_lymph_n	**0.514**	**0.562**
large_vessel_invasion	-0.139	

presence of positive lymph nodes



Imaging variable, *number_of_positive_lymph_n*, is significantly associated with path N Stage in both Panels A and B.

Discussion

In general, across almost all methods I used, the *number_of_positive_lymph_n* imaging variable is significant and positively associated with *path_t_stage, path_m_stage,* and *path_n_stage* outcomes. The *number_of_positive_lymph_n* refers to the number of lymph nodes to which cancer has spread, also known as the n-stage. Clinically, this aligns with the understanding that lymph node involvement worsens outcomes, as cancer spreads through the lymphatic system, increasing the risk of metastasis. Studies (Kroon et al., 2022; Sluckin et al., 2022) highlight the impact of lateral lymph node metastasis, particularly in locally advanced rectal cancer, on recurrence and survival rates.

Among control or pre-surgery variables, *init_clinical_staging_m* and *race* are generally significantly associated with *path_t_stage, path_m_stage,* and *path_n_stage outcomes. init_clinical_staging_m* refers to clinical M, or metastatic, stage, determined at diagnosis prior to any treatment. This emphasizes how racial disparities exist in rectal cancer survival, with black patients showing worse survival rates than white patients, even with similar treatments, likely due to biological and systemic factors.

Conclusion

The study supports the hypothesis that *number_of_positive_lymph_n* (imaging variable) and *init_clinical_staging_m* and *race* (pre-surgery variables) are key predictors of rectal cancer outcomes.

Despite the small sample size (55 cases), the study provides multidimensional insights into rectal cancer prognosis. Future work should expand the dataset and test findings over a longer period to improve reliability. As imaging technology advances, its role in predicting and improving patient outcomes will likely become even more critical.

Acknowledgements

I thank Professor Dr. Satish Viswanath, Mr. Thomas DeSilvio, and Dr. Charlems Alvarez Jimenez of Case Western Reserve University's Department of Biomedical Engineering, for the data and regular guidance in this project.

References

Krishnan, K. (2025). Can We Predict Rectal Cancer Outcomes Using Clinical Data? A Comparative Analysis of Different Techniques. Beachwood High School.

Adam Wetzel, Satish Viswanath, Emre Gorgun, Ilker Ozgur, Daniela Allende, David Liska, Andrei S Purysko, Staging and Restaging of Rectal Cancer with MRI: A Pictorial Review, Seminars in Ultrasound, CT and MRI, Volume 43, Issue 6, 2022, Pages 441-454, ISSN 0887-2171, https://doi.org/10.1053/j.sult.2022.06.003 A

Iessandra Borgheresi, Federica De Muzio, Andrea Agostini,Letizia Ottaviani,Alessandra Bruno, Vincenza Granata, Roberta Fusco, Ginevra Danti, Federica Flammia, Roberta Grassi, Francesca Grassi, Federico Bruno, Pierpaolo Palumbo, Antonio Barile, Vittorio Miele, and Andrea Giovagnoni,"Lymph Nodes Evaluation in Rectal Cancer: Where Do We Stand and Future Perspective." Journal of Clinical Medicine. 2022 May; 11(9), 2599,

Required Photographic/Graphics Source Identification

Photographs taken by: Nagtegaal et al, Smith et al, Wetzel et al

Graphics from outside sources are from: World Journal of Surgical Oncology (2021), World Journal of Gastroenterology (2013), Wetzel et al., Seminars in Ultrasound, CT, and MRI (2022)

Photographic permissions were obtained and are located: Open-Access Sources and Dr. Viswanath gave permission to use photographs from *Staging and Restaging of Rectal Cancer with MRI: A Pictorial Review* (Wetzel et al., 2022).