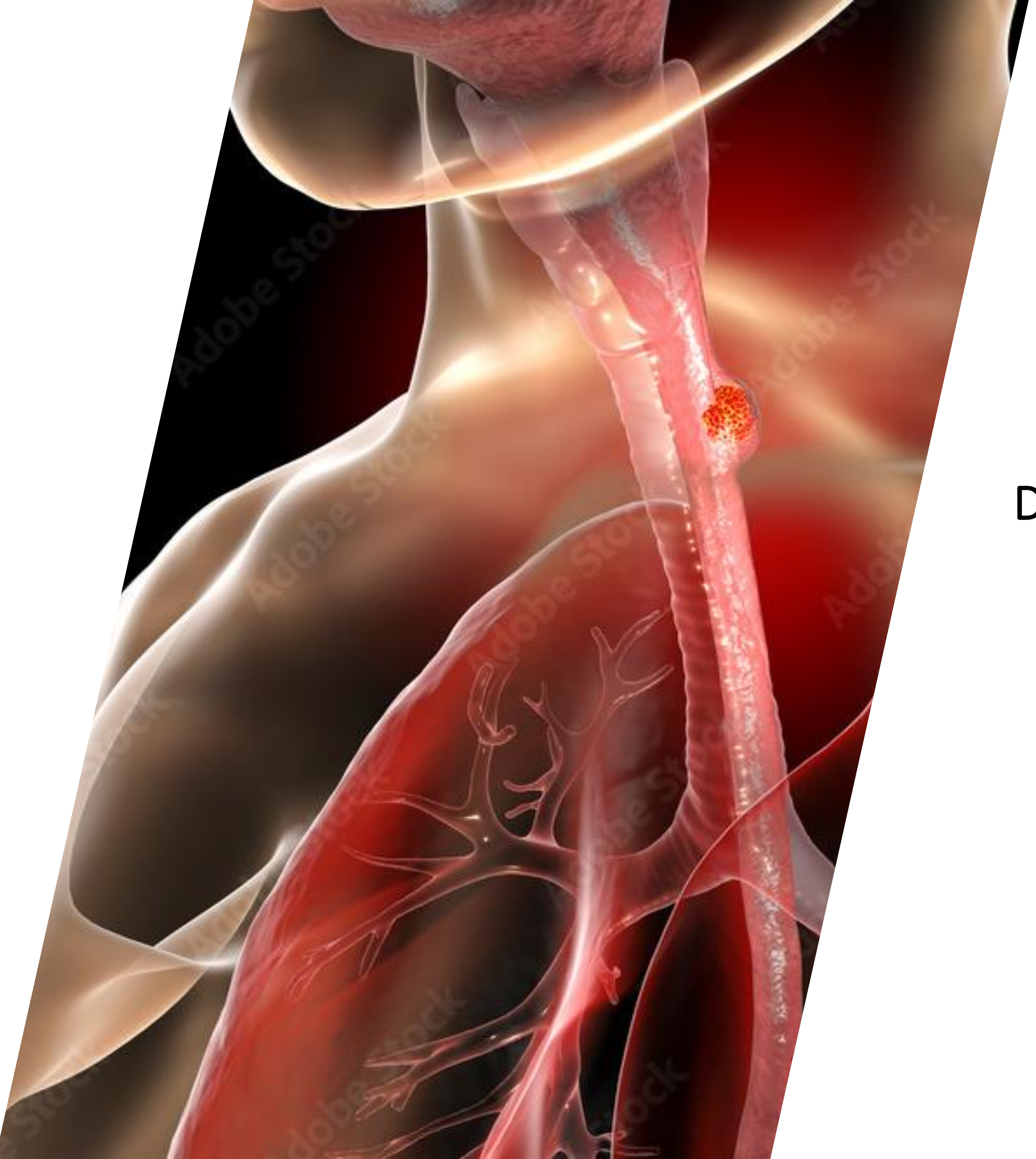
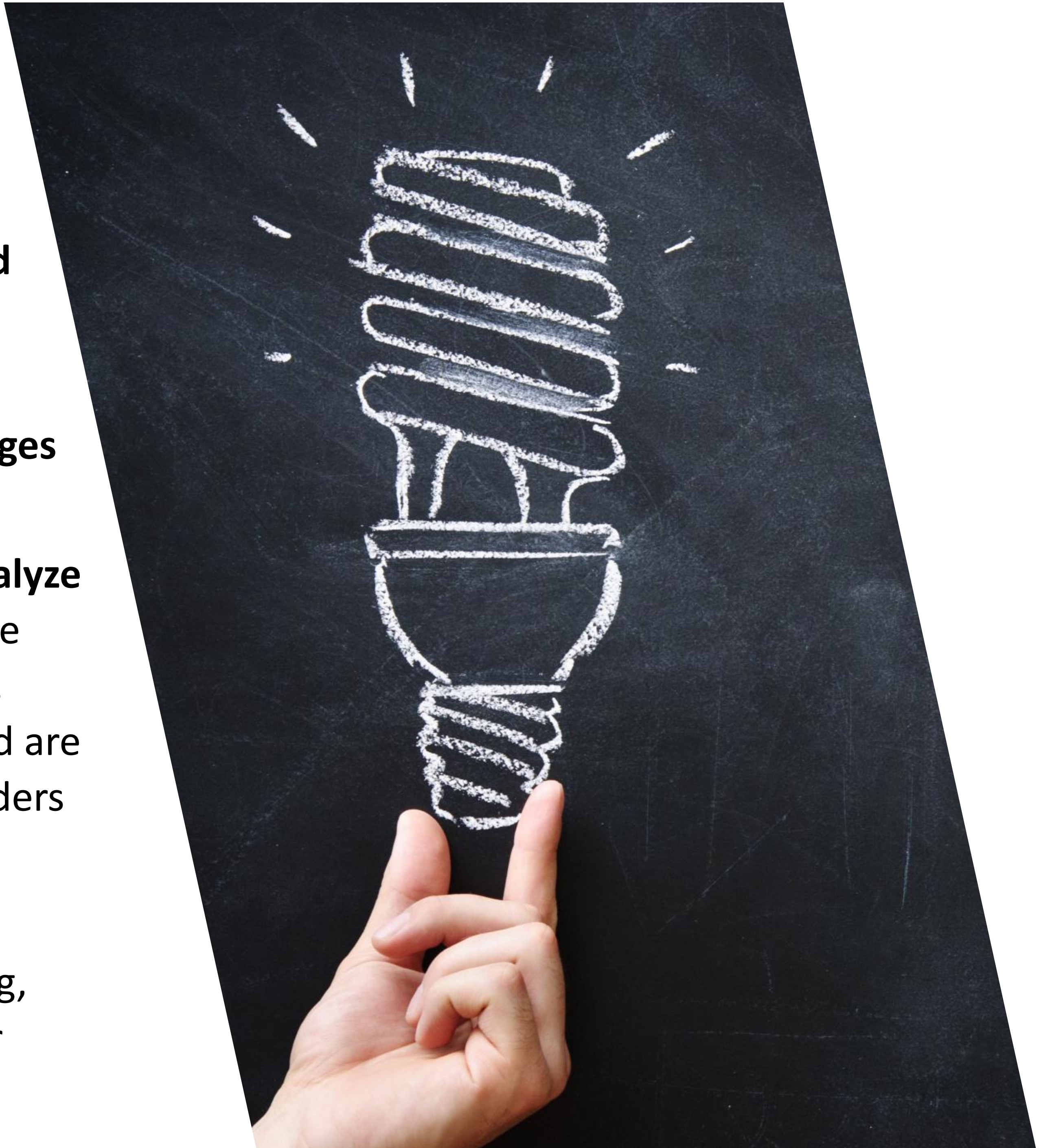


DESIGN AND IMPLEMENTATION OF AN ADVANCED HEALTHCARE DATABASE SYSTEM: OPTIMIZING DATA STORAGE, MANAGEMENT, AND ANALYTICAL INSIGHTS FOR ESOPHAGEAL CANCER TREATMENT OUTCOMES

SAMEER MOHAMMAD

INTRODUCTION

- Esophageal cancer is a leading cause of **cancer-related** deaths, often diagnosed late in its progression.
- The complexity of managing patient records, tumor progression, and treatment outcomes creates **challenges** in delivering efficient, data-driven care.
- Healthcare systems often struggle to **organize** and **analyze** this data in a structured way. A well-designed database system is crucial to managing esophageal cancer data, ensuring that **clinical records** are stored efficiently and are accessible for **analysis**, and enabling healthcare providers to make informed decisions that **improve** patient outcomes.
- By integrating visualizations and geographical mapping, the system will help clinicians **analyze** patient data for improved treatment outcomes.



DATASET

Dataset: https://www.kaggle.com/datasets/abhinaba1biswas/esophageal-cancer-dataset/data

The dataset contains detailed information on esophageal cancer patients, covering **demographic, clinical, and treatment**-specific data points.

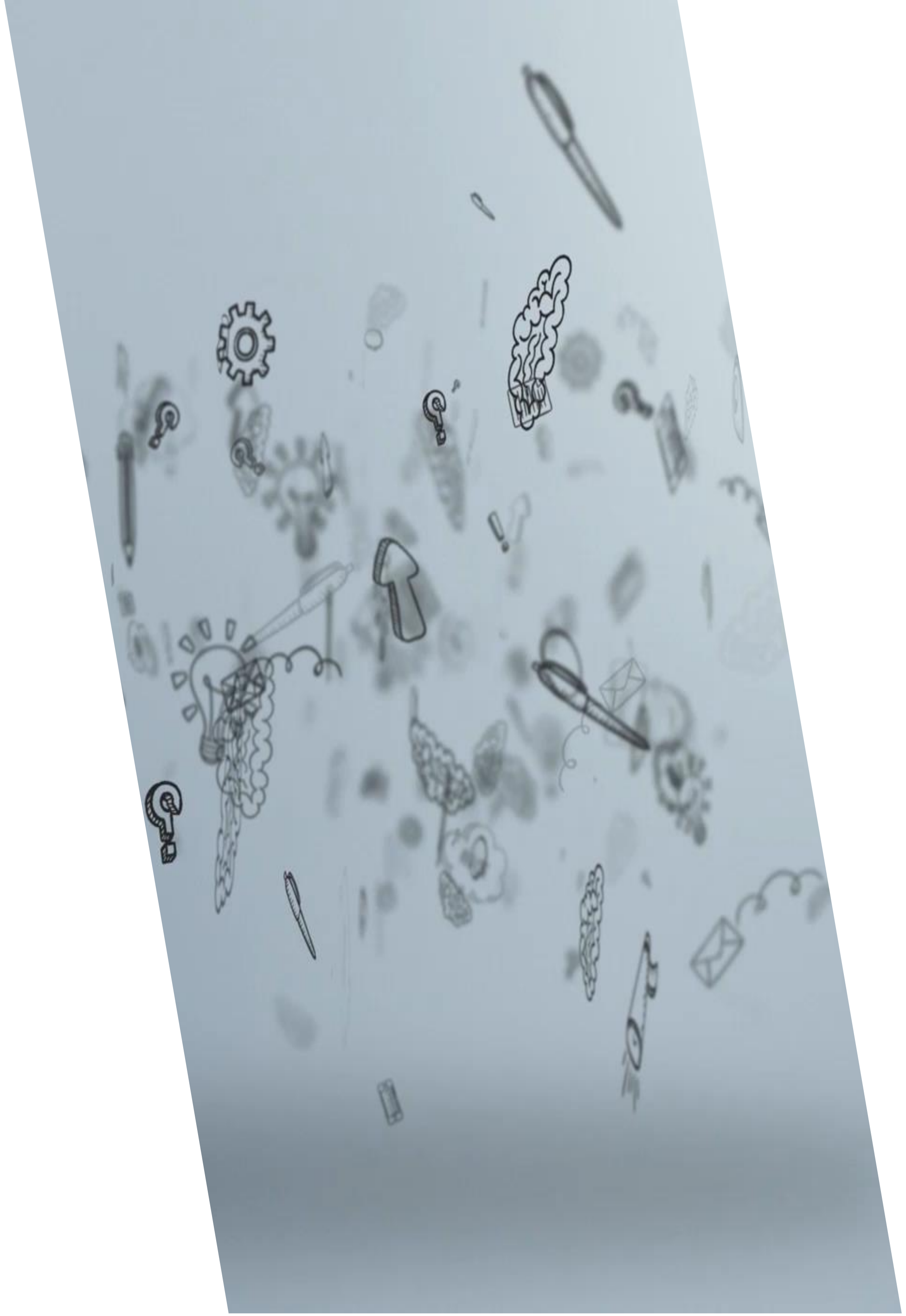
Key fields include:

Patient Demographics: patient_id, gender, country_of_birth, country_of_procurement, race_list.

Tumor Data: icd_o_3_site, icd_o_3_histology, icd_10, primary_pathology_tumor_tissue_site, primary_pathology_histological_type.

Treatment Data: person_neoplasm_cancer_status, vital_status, days_to_last_followup, days_to_death, treatment_prior_to_surgery, radiation_therapy, postoperative_rx_tx.

Outcome Data: karnofsky_performance_score, primary_pathology_residual_tumor, primary_pathology_lymph_node_examined_count, number_of_lymphnodes_positive_by_he.



OBJECTIVE

- This project focuses on designing and implementing a relational database for esophageal cancer data, aiming to enable healthcare providers and researchers to analyze patient demographics, treatment effectiveness, and clinical outcomes.

- The database incorporates advanced features, including data visualization and geographical mapping, highlighting trends and disparities across different regions. These capabilities will provide valuable insights into clinical data, facilitating better understanding and decision-making in cancer treatment and research.

- Additionally, the database is structured for seamless integration with machine learning models to support predictive analytics in the future.

- This design ensures scalability and adaptability, enabling the inclusion of additional datasets and features such as clinical decision support systems (CDSS).

- By prioritizing efficient data storage, retrieval, and advanced analytics, the project establishes a strong foundation for data-driven healthcare improvements and enhanced patient care outcomes.

ROLES AND RESPONSIBILITIES

NAMES	Role	Responsibilities
SAMEER	Database Architect	Created the database design and optimized its structure to ensure efficient data storage and retrieval. Applied normalization techniques and indexing strategies to improve query performance. Refined the Entity-Relationship (ER) diagram to ensure it aligned with project requirements.
BHARGAV	Data Analyst	Preprocessed the dataset using Python to clean and standardize the data for analysis and machine learning modeling. Developed interactive visualizations using Plotly and Matplotlib to showcase trends in patient data, treatment effectiveness, and demographics. Implemented visualizations and geographical mapping to visualize regional trends in esophageal cancer prevalence and treatment outcomes.
VERONICA	Visualization and User Experience Designer	Prepared comprehensive documentation detailing the system’s use, database schema, ER diagram, and machine learning models Worked on preparing the user interface for the database and improving the overall user experience. Enhanced the visualization components, focusing on making the data accessible and actionable through interactive interfaces and graphical representations.

METHODOLOGY

METHODS

Database Design Overview

The relational database is designed using MySQL, structured to store and analyze esophageal cancer data efficiently. It organizes information into core entities such as patients, diagnoses, treatments, visits, outcomes, and country details, ensuring optimized data retrieval and maintaining data integrity..

ER Diagram Details

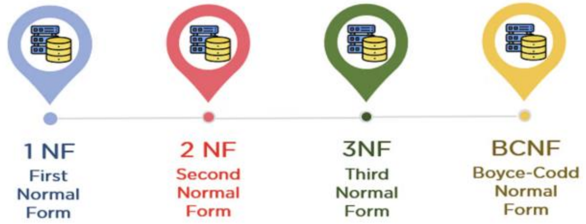
The database comprises the following entities and attributes:

1. **Patient Table**: Includes unique patient identifiers **(patient_id),** demographic details **(e.g., country_of_birth),** and tissue-related information.
2. **Country Table**: Normalizes geographical data with attributes like **country_of_birth, country_of_procurement, and city/state details**.
3. **Visit Table**: This table tracks **patient visits** with **visit dates** and **visit-related IDs**, linking patients to their clinical timelines.
4. **Diagnosis Table**: Records **initial diagnoses**, **cancer status, tumor characteristics, and pathology results**.
5. **Treatment Table**: Captures details of treatments provided, such as types of **therapies** and **residual tumor statuses**.
6. **Outcome Table**: Documents patient outcomes, including performance **scores** (e.g., Karnofsky score), **ECOG status,** and **vital status.**

Schema Design

Each table is defined with a **primary key** (e.g., patient_id, diagnosis_id) and linked via **foreign keys** to ensure referential integrity. The design supports normalization (3NF), reducing redundancy and optimizing esophageal cancer data analysis storage and querying. This structure facilitates efficient data visualization, geographical trend mapping, and predictive analytics integration..

NORMALIZATION



1NF

All tables (patient_table, visit_table, treatment_table, diagnosis_table, country_table, outcome_table) meet 1NF criteria as they have atomic columns and no repeating groups.

2NF

All tables meet **2NF** Criteria :patient_table: No partial dependencies, as all attributes depend on **patient_id.** visit_table: All attributes depend on **visit_id**.treatment_table, diagnosis_table, country_table, outcome_table: All attributes depend on their respective primary keys.

3NF

All tables meet **3NF** Criteria: No transitive dependencies exist. Attributes in each table are entirely functionally dependent on their respective primary keys.

BCNF

The schema adheres to BCNF as all functional dependencies are resolved, and there are no violations.

RELATIONSHIPS

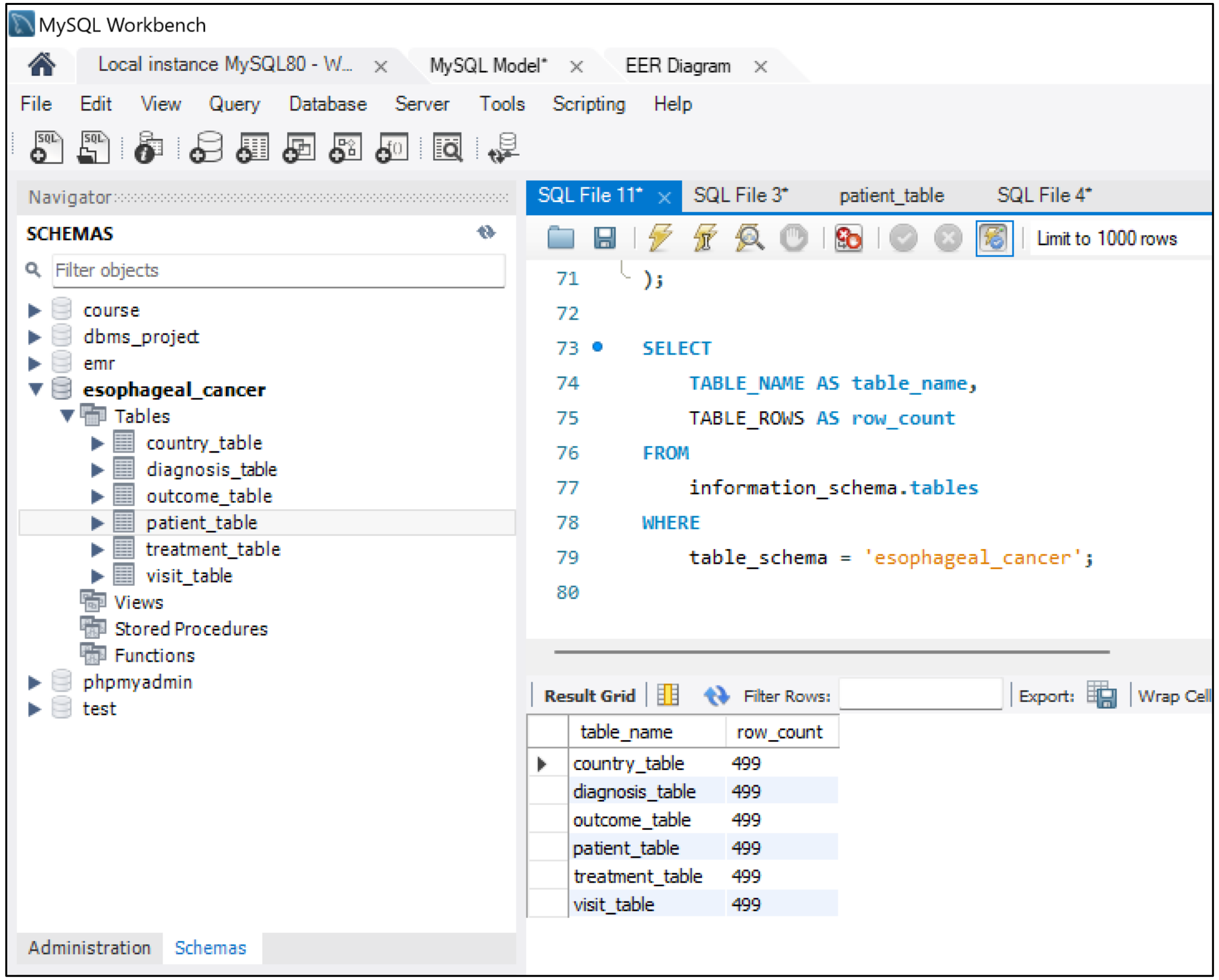
One-to-Many Relationships:

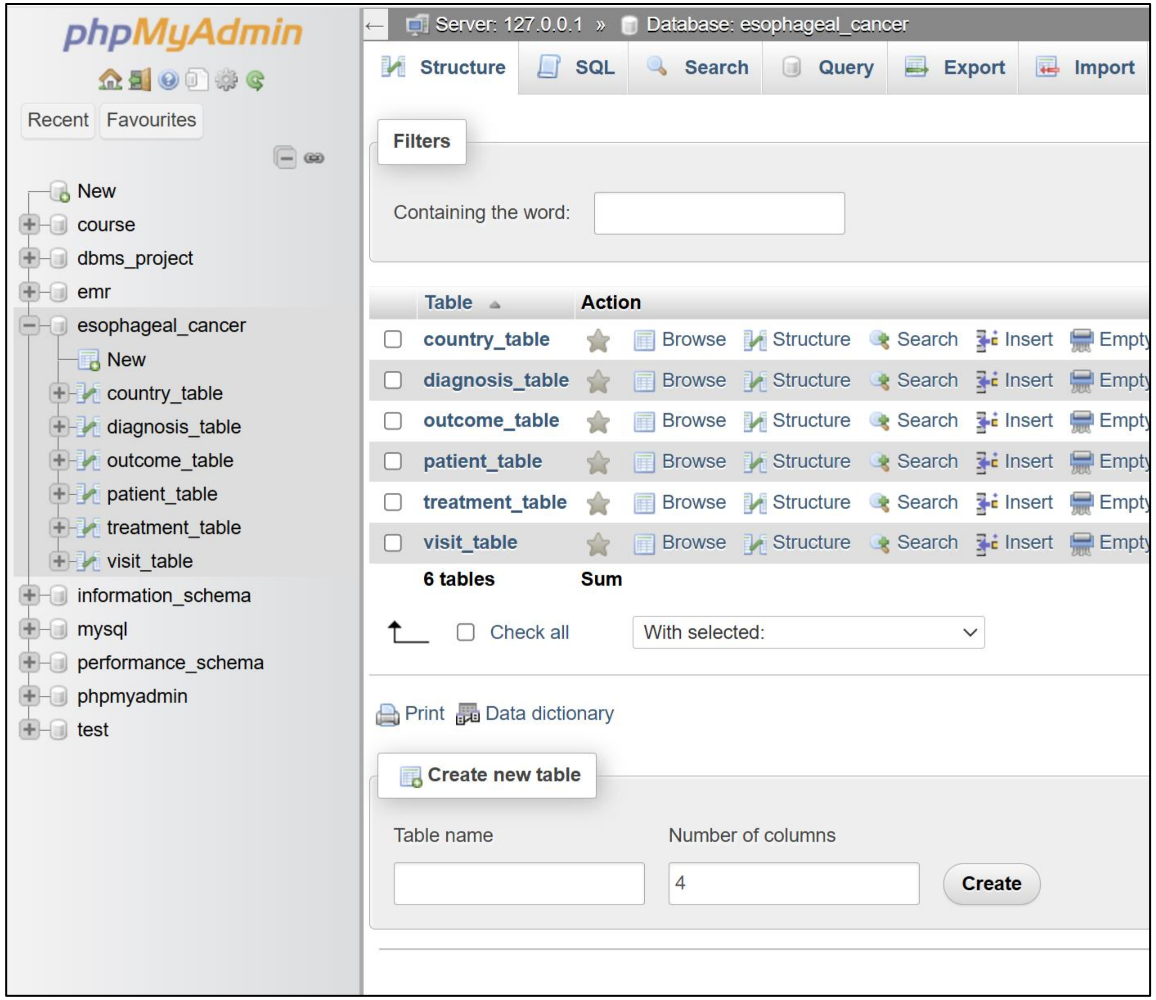
1. **Patient Table to Visit Table**
One patient can have multiple visits, but each visit is associated with only one patient.
2. **Visit Table to Diagnosis Table**
One visit can have multiple diagnoses, but each diagnosis is linked to only one visit.
3. **Visit Table to Outcome Table**
One visit can result in multiple outcomes, but each outcome is associated with only one visit.
4. **Visit Table to Treatment Table**
One visit can be associated with multiple treatments, but each treatment is linked to only one visit.

Many-to-One Relationships:

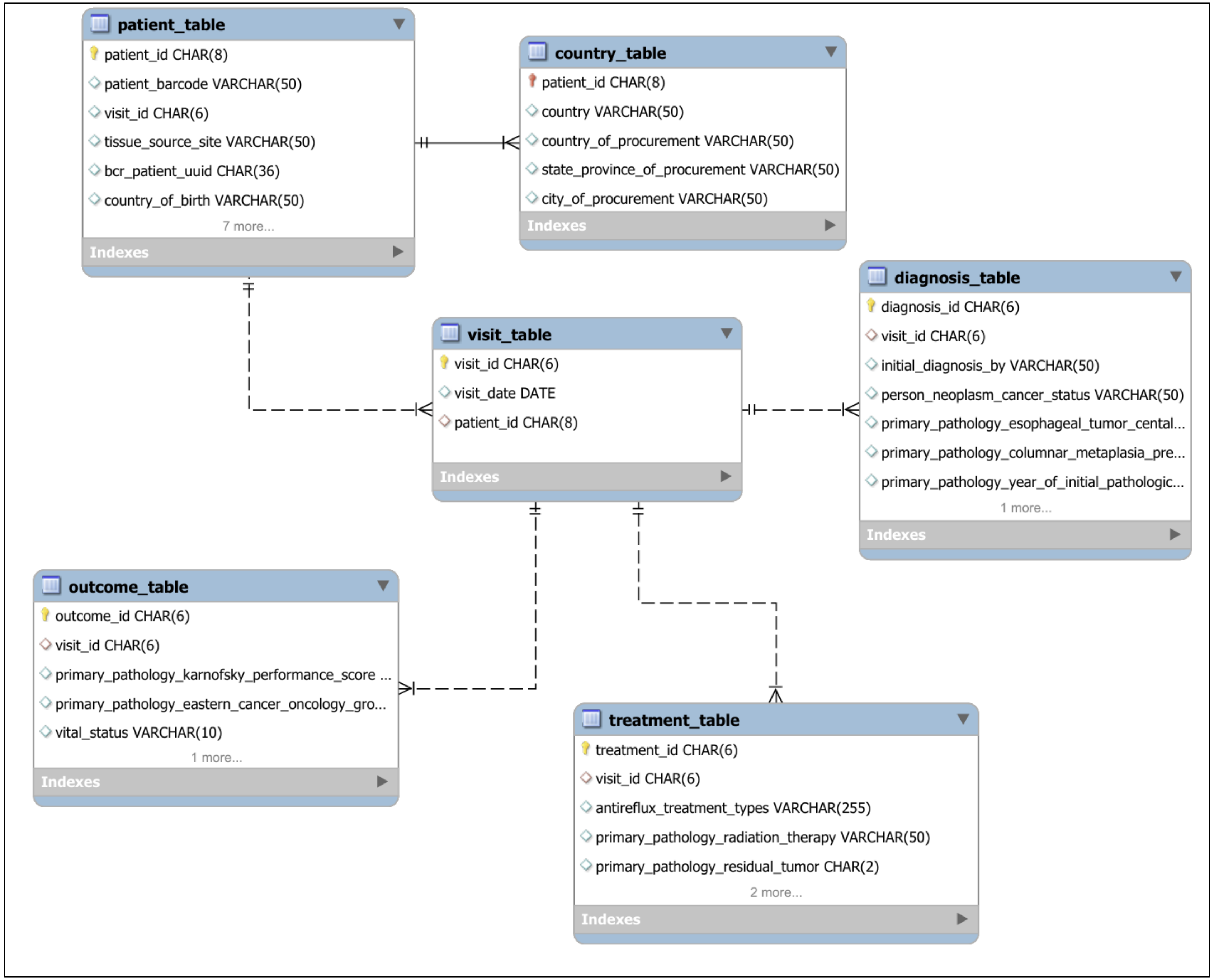
1. **Patient Table to Country Table**
Each patient is associated with only one country (for birth and procurement), but a country can have many patients.
2. **Visit Table to Country Table**
Each visit is associated with only one country (through the country of procurement), but a country can have multiple visits

MY SQL WORKBENCH AND PHP MY ADMIN



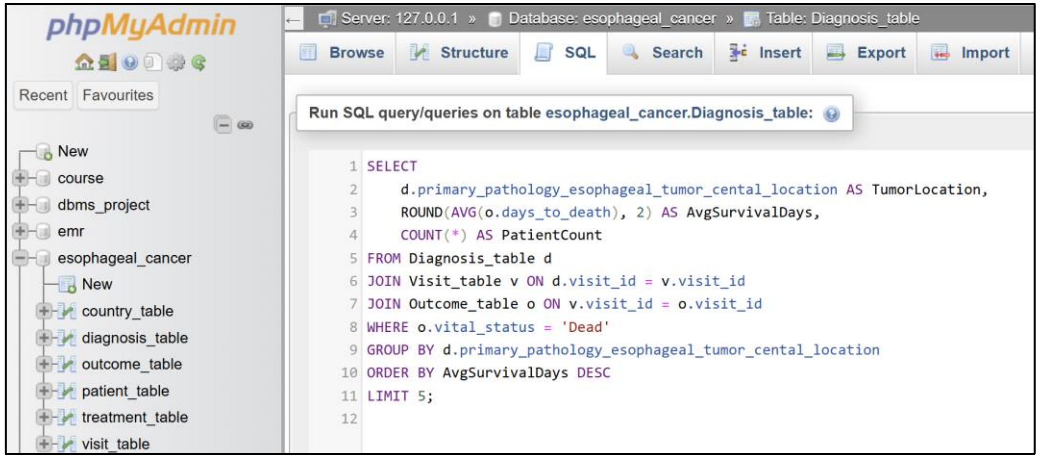


ENTITY RELATIONSHIP DIAGRAM FOR ESOPHAGEAL CANCER DATABASE - SQL SCHEMA



1. PATIENTS WITH THE LONGEST SURVIVAL TIMES BY TUMOR LOCATION

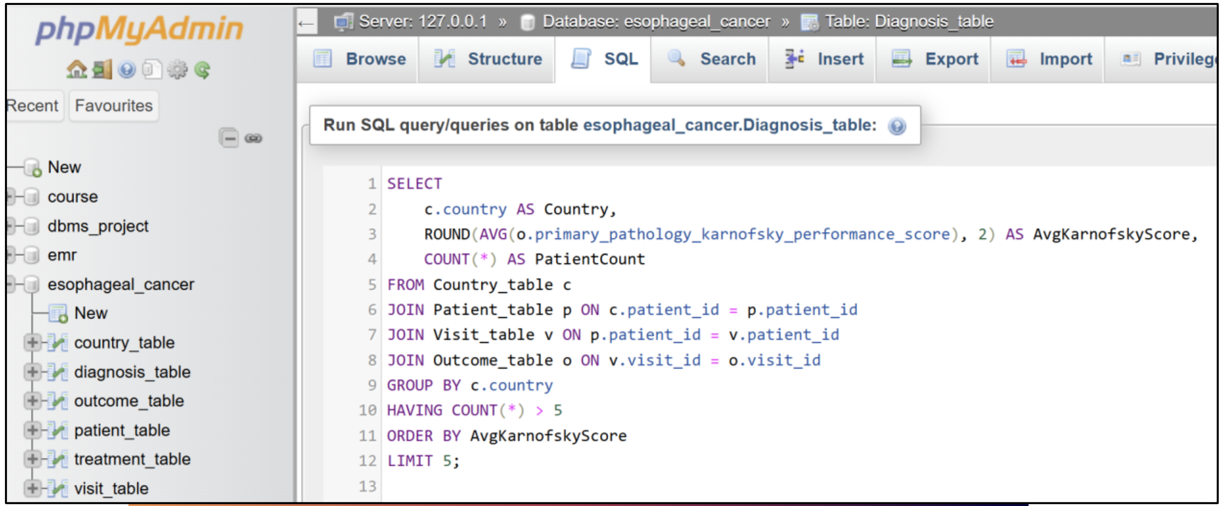
IDENTIFIES THE TUMOR LOCATIONS WITH THE LONGEST AVERAGE SURVIVAL TIMES AND LISTS THE TOP 5 LOCATIONS.



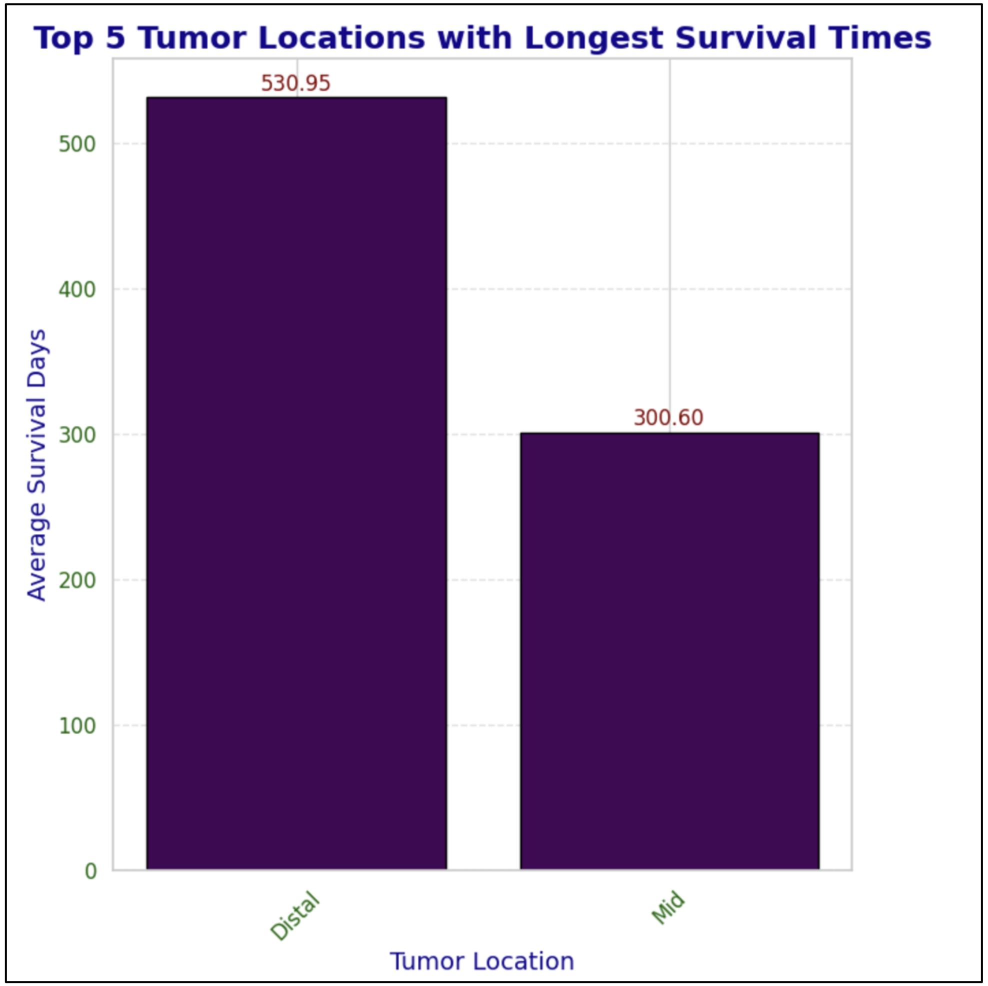
TumorLocation	AvgSurvivalDays 1	PatientCount
Distal	530.95	128
Mid	300.60	25

2. HIGH-RISK COUNTRIES BASED ON KARNOFSKY PERFORMANCE SCORES

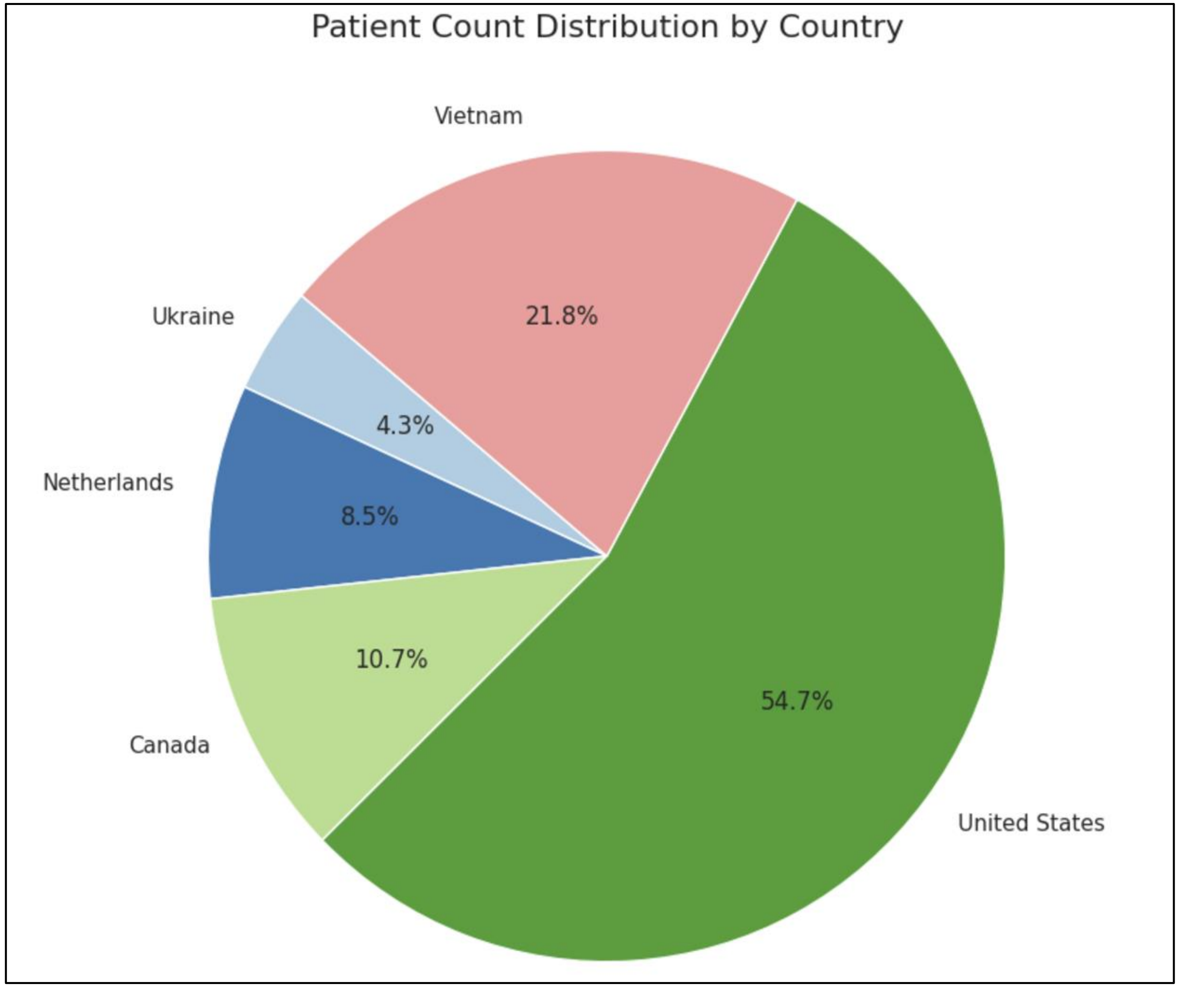
IDENTIFIES THE COUNTRIES WITH THE LOWEST AVERAGE KARNOFSKY PERFORMANCE SCORES, INDICATING POOR PATIENT CONDITIONS.



Country	AvgKarnofskyScore 1	PatientCount
Ukraine	0.00	18
Netherlands	0.00	36
Canada	0.00	45
United States	5.15	231
Vietnam	43.26	92



Tumors in the "Distal" location have the highest survival time (530.95 days), followed by "Mid" tumors (300.60 days).



Distribution of patients by country with the United States contributing the largest share (54.7%) and Vietnam (21.8%) and Canada (10.7%) show smaller proportions.

SQL QUERY

3. Analyzing Treatment Frequency and Ranking by Antireflux Treatment Types

Identify and ranking antireflux treatment types by their frequency in the treatment_table, sorted by rank in ascending order.

Run SQL query/queries on table esophageal_cancer.treatment_table:

```sql
SELECT
    t.antireflux_treatment_types,
    COUNT(t.treatment_id) AS treatment_frequency,
    RANK() OVER (ORDER BY COUNT(t.treatment_id) DESC) AS treatment_rank
FROM treatment_table t
GROUP BY t.antireflux_treatment_types
ORDER BY treatment_rank ASC;
```

antireflux_treatment_types	treatment_frequency	treatment_rank
0	369	1
Medically Treated	104	2
No Treatment	22	3
Medically TreatedSurgically Treated	4	4

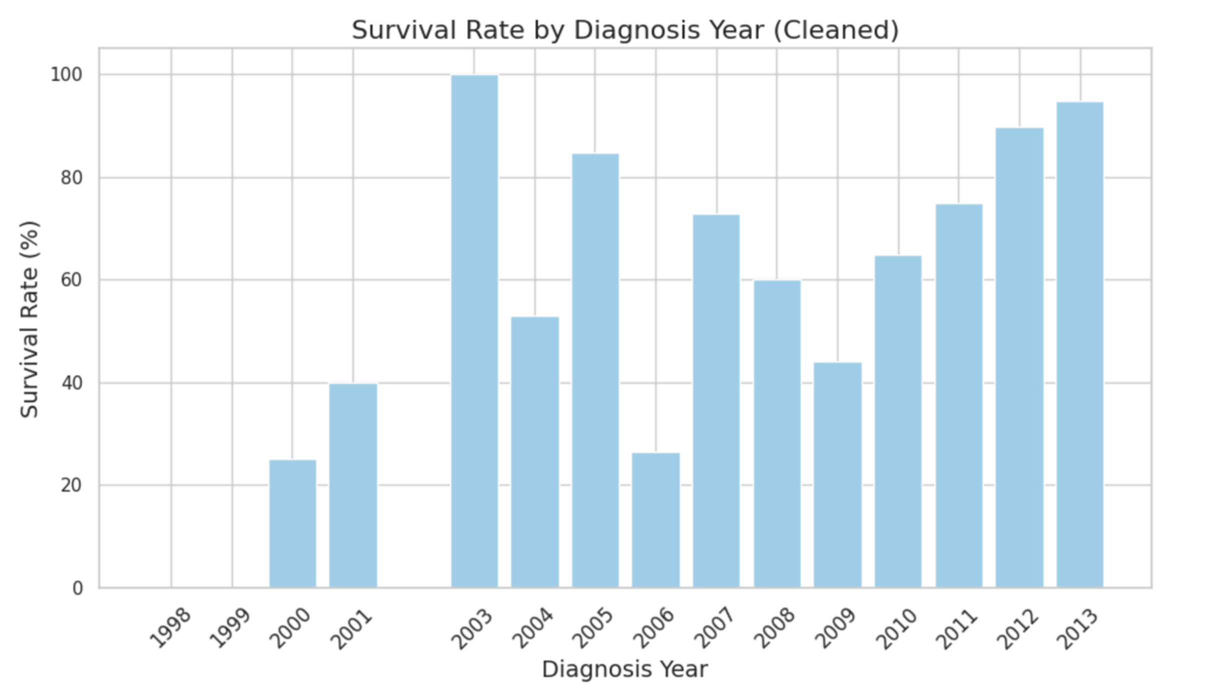
4. Calculating Patient Survival Rates by Primary Pathology Diagnosis Year

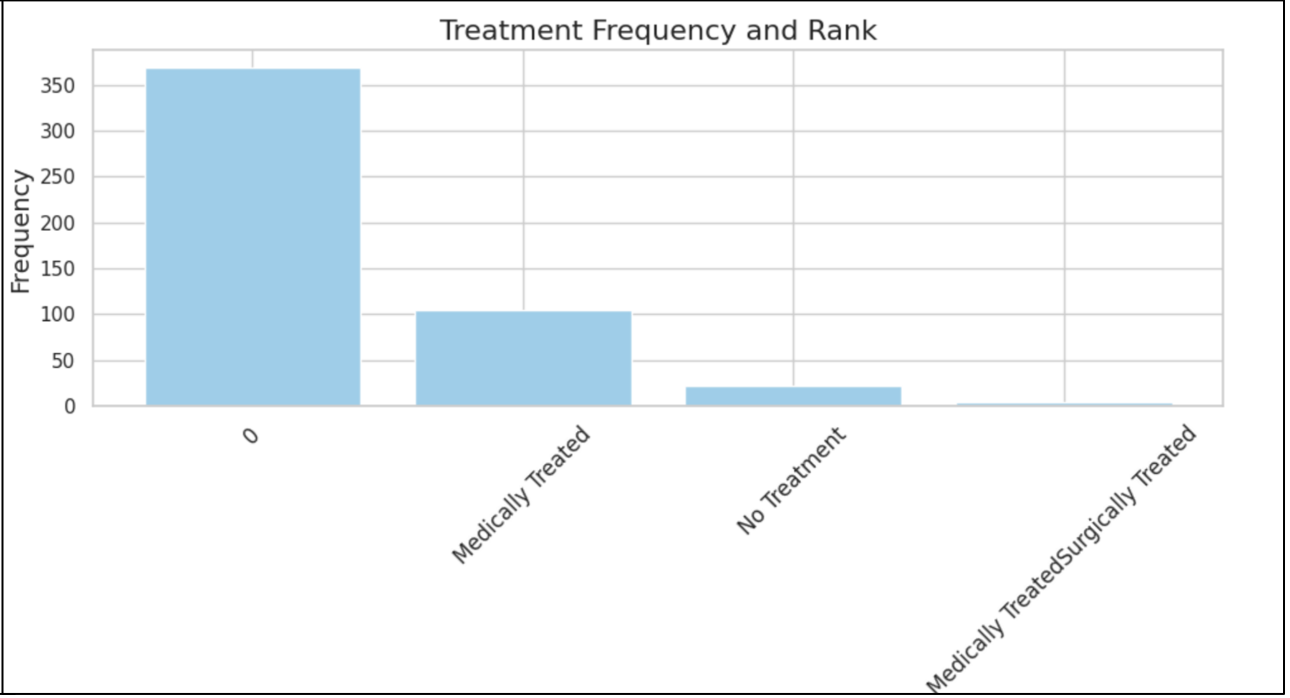
Write an SQL query to calculate the survival rate of patients based on their primary pathology diagnosis year, including the total number of patients, the number of alive patients, and the survival rate percentage, grouped by diagnosis year and sorted in ascending order.

Run SQL query/queries on table esophageal_cancer.treatment_table:

```sql
SELECT
    d.primary_pathology_year_of_initial_pathologic_diagnosis AS diagnosis_year,
    COUNT(DISTINCT CASE WHEN o.vital_status = 'Alive' THEN p.patient_id END) AS alive_patients,
    COUNT(DISTINCT p.patient_id) AS total_patients,
    ROUND(
        (COUNT(DISTINCT CASE WHEN o.vital_status = 'Alive' THEN p.patient_id END) / COUNT(DISTINCT p.patient_id)) * 100,
        2
    ) AS survival_rate
FROM diagnosis_table d
JOIN visit_table v ON d.visit_id = v.visit_id
JOIN outcome_table o ON v.visit_id = o.visit_id
JOIN patient_table p ON v.patient_id = p.patient_id
GROUP BY d.primary_pathology_year_of_initial_pathologic_diagnosis
ORDER BY diagnosis_year;
```

diagnosis_year	alive_patients	total_patients	survival_rate
0	10	14	71.43
1998	0	3	0.00
1999	0	9	0.00
2000	6	24	25.00
2001	16	40	40.00
2003	3	3	100.00
2004	9	17	52.94
2005	11	13	84.62
2006	5	19	26.32
2007	8	11	72.73
2008	3	5	60.00
2009	11	25	44.00
2010	35	54	64.81
2011	51	68	75.00
2012	104	116	89.66
2013	74	78	94.87





Survival rates by diagnosis year show a peak in 2003 and an upward trend, indicating improved treatments and patient care over time.

Most patients received no treatment, followed by medical treatment. Combined medical-surgical treatments were the least frequent

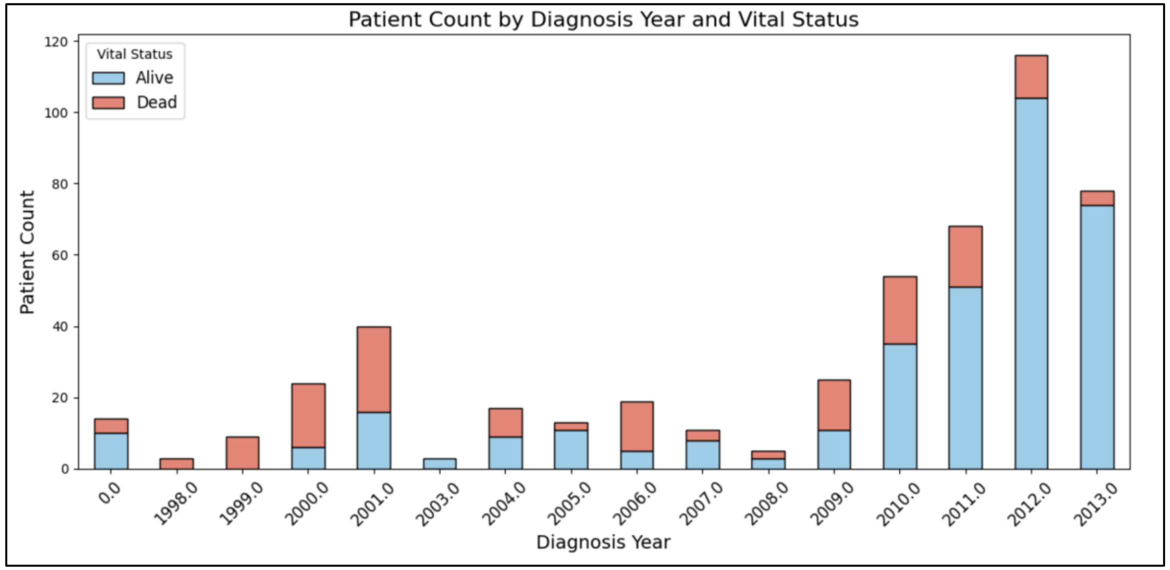
SQL QUERY

5. Analyzing Patient Counts by Diagnosis Year and Vital Status

Write an SQL query to determine the count of unique patients grouped by their primary pathology diagnosis year and vital status, sorted by diagnosis year and patient count in descending order.

Showing rows 0 - 24 (29 total, Query took 0.0069 seconds.)

```sql
SELECT
    d.primary_pathology_year_of_initial_pathologic_diagnosis AS diagnosis_year,
    o.vital_status,
    COUNT(DISTINCT p.patient_id) AS patient_count
FROM diagnosis_table d
JOIN visit_table v ON d.visit_id = v.visit_id
JOIN outcome_table o ON v.visit_id = o.visit_id
JOIN patient_table p ON v.patient_id = p.patient_id
GROUP BY d.primary_pathology_year_of_initial_pathologic_diagnosis, o.vital_status
ORDER BY diagnosis_year, patient_count DESC;
```

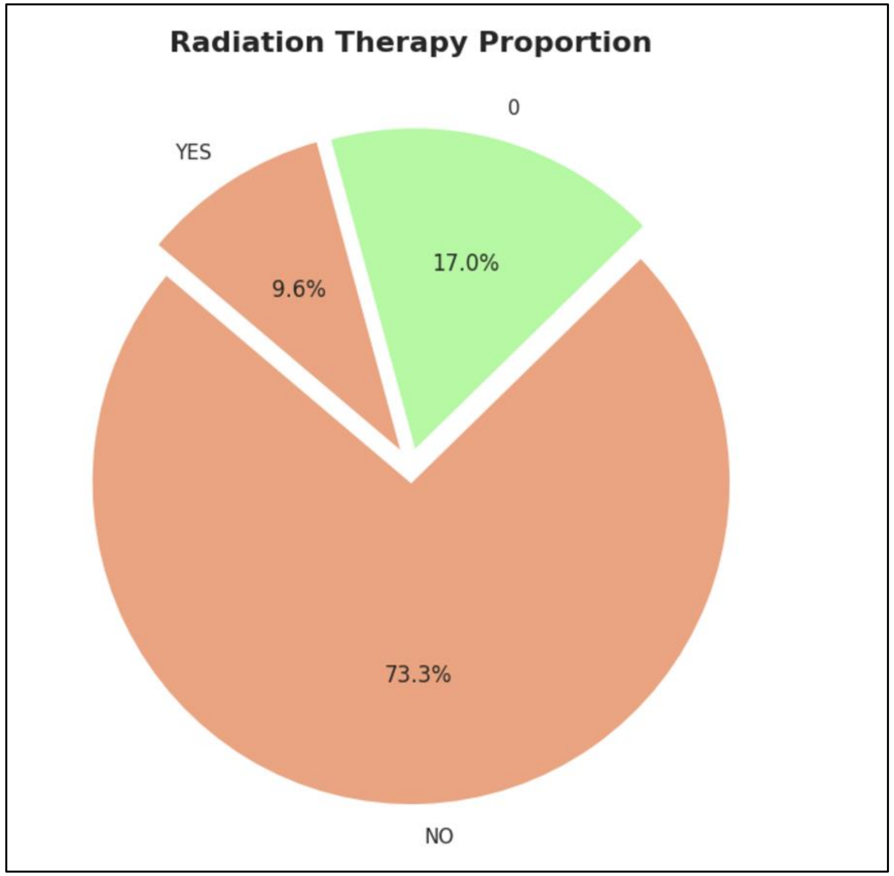


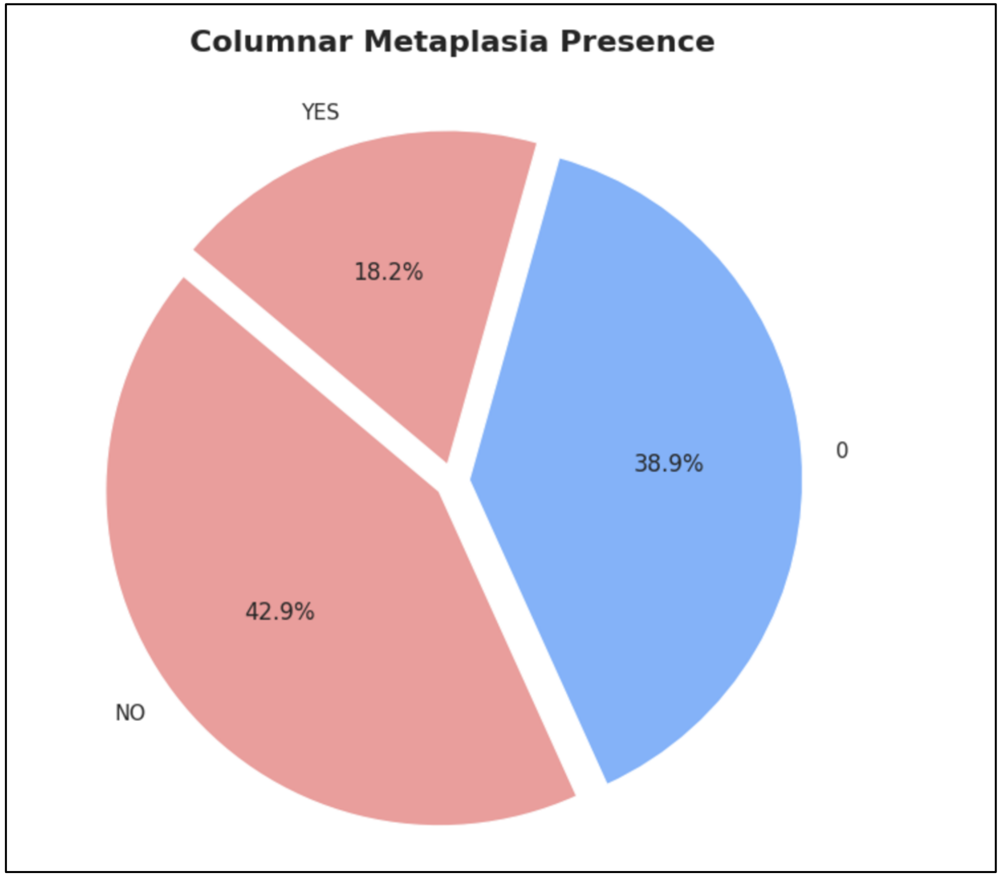
diagnosis_year 1	vital_status	patient_count 2
0	Alive	10
0	Dead	4
1998	Dead	3
1999	Dead	9
2000	Dead	18
2000	Alive	6
2001	Dead	24
2001	Alive	16
2003	Alive	3
2004	Alive	9
2004	Dead	8
2005	Alive	11
2005	Dead	2
2006	Dead	14
2006	Alive	5
2007	Alive	8
2007	Dead	3
2008	Alive	3
2008	Dead	2
2009	Dead	14
2009	Alive	11
2010	Alive	35
2010	Dead	19
2011	Alive	51
2011	Dead	17

1 > >> Show all | Number of rows: 25

Console

VISUALIZATION

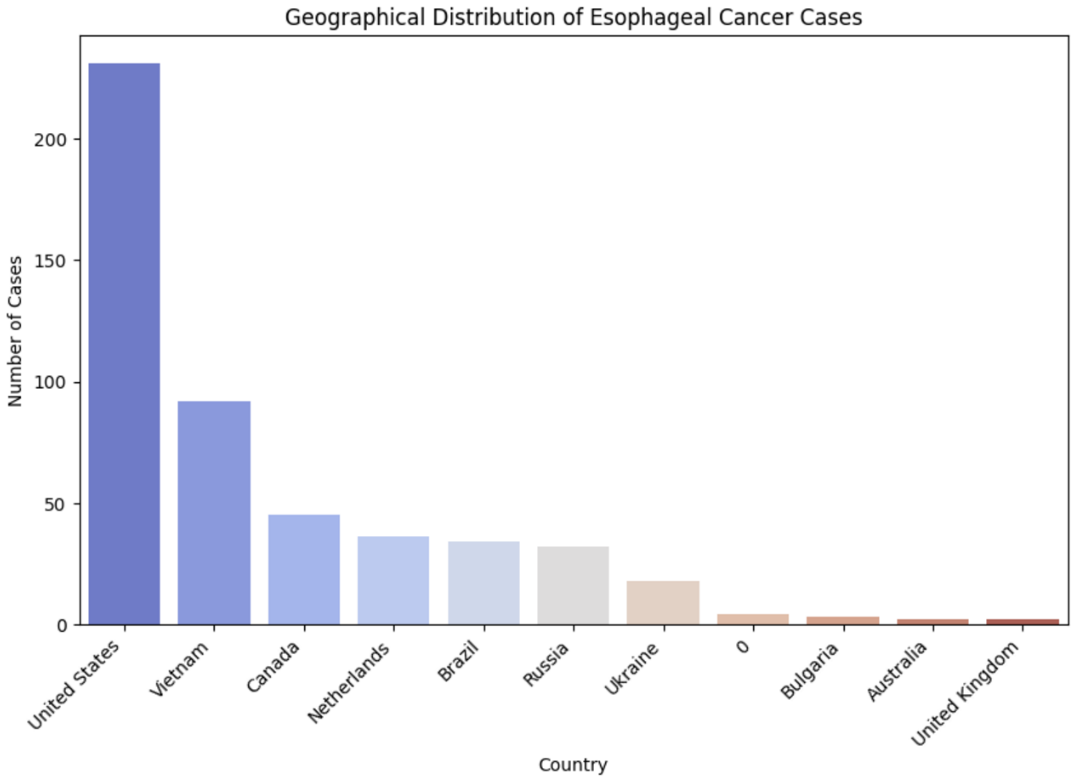




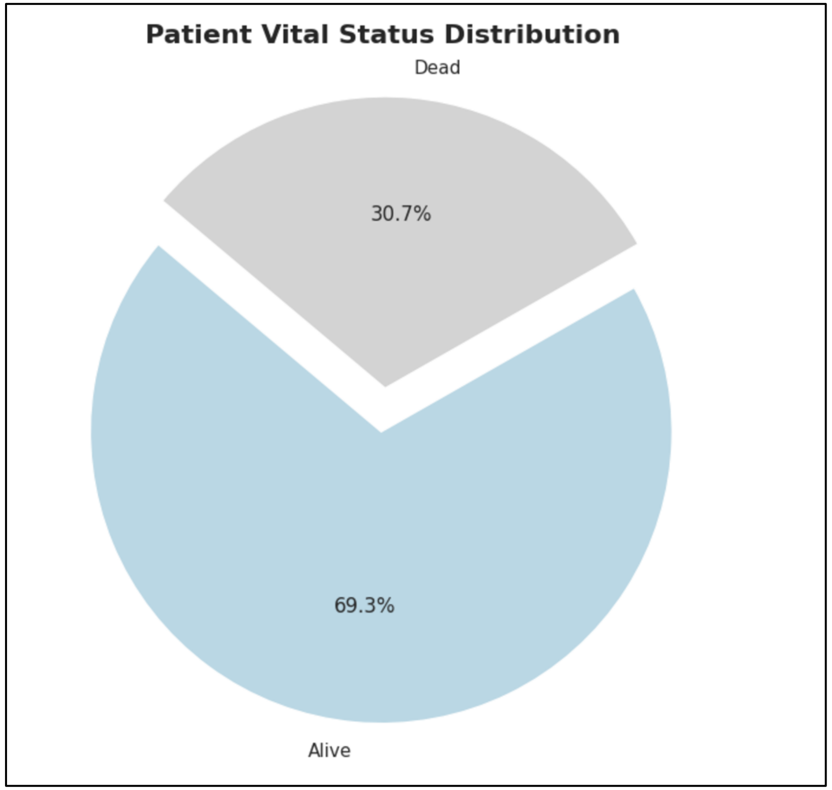
73.3% of patients did not undergo radiation therapy, 17% have unknown therapy status, and 9.6% underwent radiation therapy.

42.9% of patients tested negative for columnar metaplasia, 18.2% tested positive, and 38.9% have unknown results.

VISUALIZATION



The United States has the highest number of esophageal cancer cases, followed by Vietnam and Canada. Other countries contribute significantly fewer cases.



Patient vital status distribution shows 69.3% of patients are alive, while 30.7% have succumbed to the disease.

DISTRIBUTION OF ESOPHAGEAL CANCER CASES BASED ON PATIENT DATA: A GEOGRAPHIC VISUALIZATION

ANALYSIS

Model Evaluation and Feature Importance

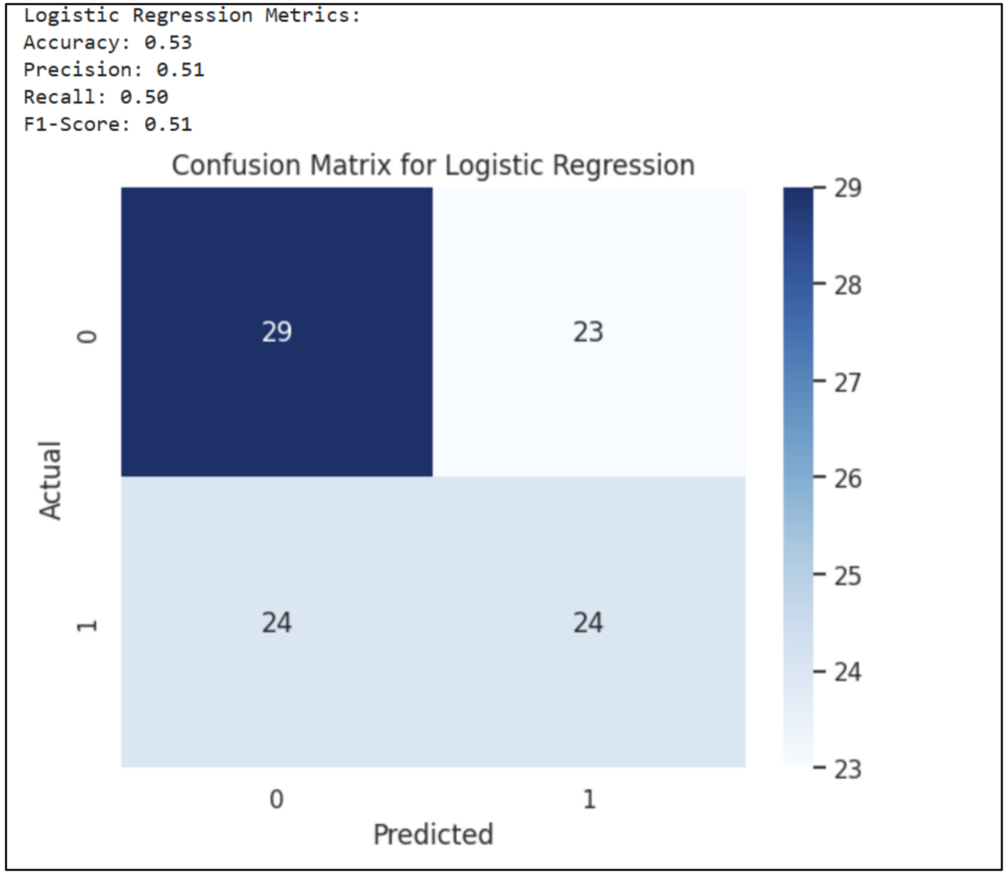
The results from the **Logistic Regression** and **Random Forest** models show that both models perform moderately, with accuracy values of 0.53 and 0.49, respectively.

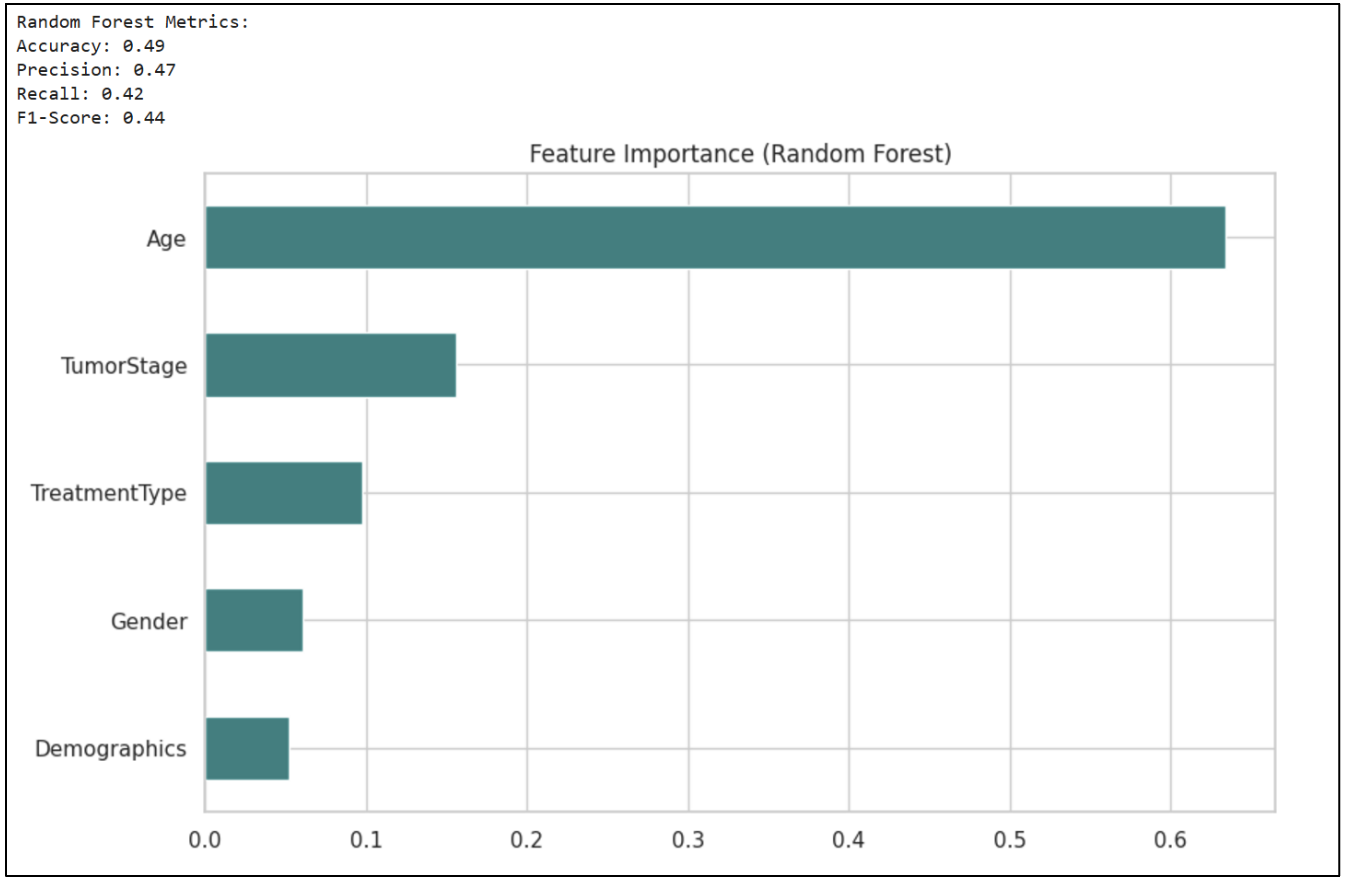
The **confusion matrix for Logistic Regression** reveals a relatively balanced prediction outcome, with 29 true positives and 23 false positives, indicating room for improvement in predictive accuracy.

On the other hand, the Random Forest model provides similar metrics with an accuracy of 0.49, precision of 0.47, and recall of 0.42. These values suggest that further tuning or feature engineering may be necessary for optimal performance.

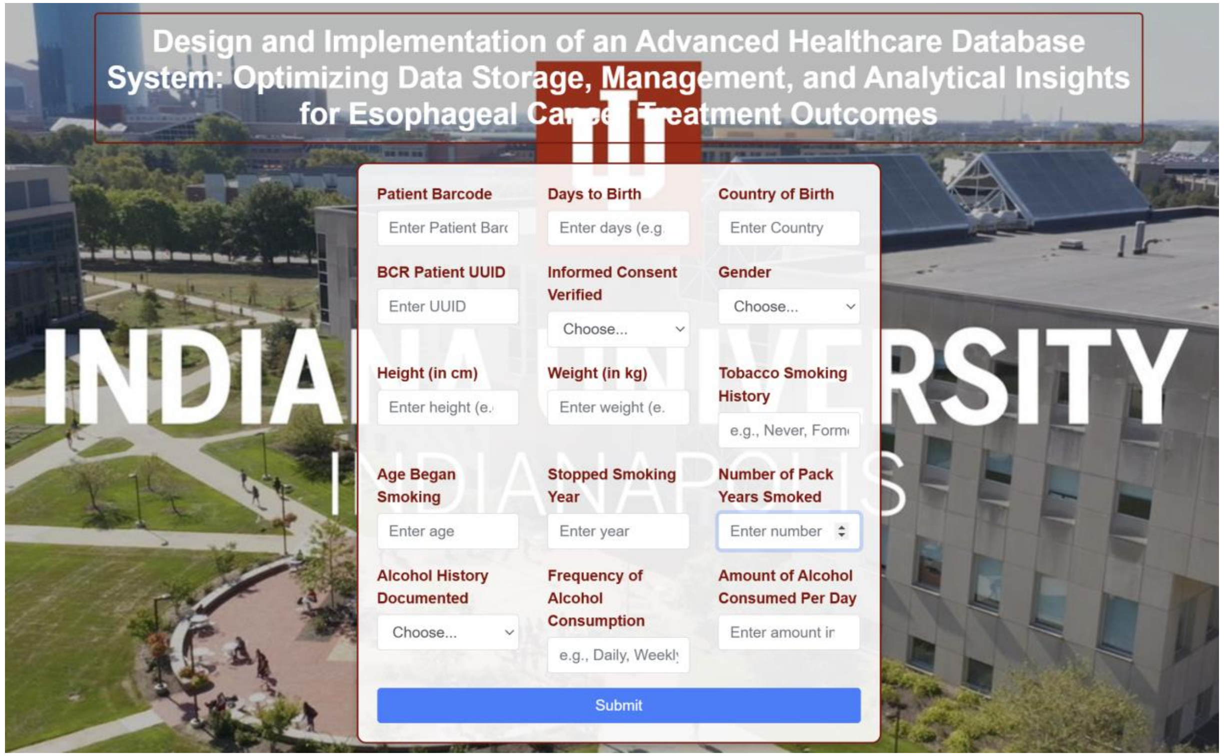
Feature Importance Analysis
The **Random Forest feature importance chart** highlights the significant predictors of treatment outcomes. **Age** emerges as the most important feature, followed by **Tumor Stage, Treatment Type**, and **Gender**. This information can guide clinicians in identifying critical factors influencing treatment effectiveness and patient survival.

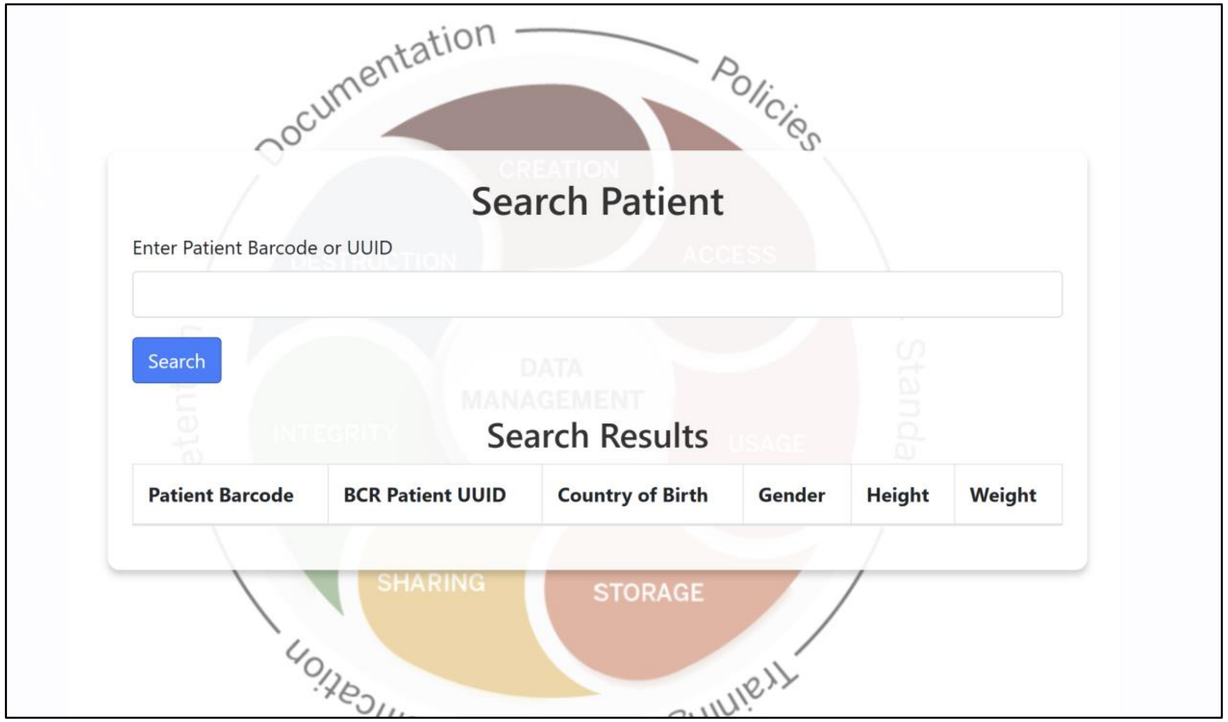
LOGISTIC REGRESSION & RANDOM FOREST





WEB INTERFACE





CONCLUSION

- Designed a relational database for esophageal cancer data, integrating patient demographics, tumor details, treatment outcomes, and geographical information to support efficient data management and analysis.
- Normalized data into a structured schema, enabling advanced analytics, real-time data querying, and predictive modeling for improved clinical decision-making and research insights.
- Incorporated visualizations and geographical mapping, offering a comprehensive view of patient distributions, treatment trends, and regional disparities to aid healthcare interventions.
- Identified critical insights, such as longer survival rates for distal tumors and high-risk regions like Ukraine, helping prioritize clinical and policy efforts.
- Applied machine learning models, including logistic regression and random forest, to predict patient outcomes, optimize treatment strategies, and trigger early alerts for interventions.
- Analyzed survival trends and treatment frequencies, visualizing key findings like increased antireflux treatments and improving survival rates by diagnosis year.
- Established a future-ready system, supporting the inclusion of additional datasets and integration with clinical decision support systems (CDSS) for broader applications.

THANK YOU