Title: *Modeling Cancer Treatment Response Over Time Using Longitudinal Clinical Data*

1. Problem Statement

Cancer treatment response is inherently temporal. A patient's condition evolves across multiple follow-up visits, and changes in disease status, treatment, and outcomes depend on what happened previously. Traditional snapshot-based models struggle to capture these time-dependent patterns.

The goal of this project is to model cancer treatment response as a **longitudinal sequence prediction problem**, where each patient's clinical history is represented as a timeline of events. By learning patterns across these sequences, the project aims to predict disease progression and treatment response more accurately than static models.

2. Context

Cancer care is increasingly driven by data, yet much of the information available in electronic health records is complex and longitudinal. Healthcare providers, hospitals, and research institutions want to better understand how patients respond to treatment so they can intervene earlier and personalize care.

The client for this project can be viewed as a cancer research or healthcare organization (such as a hospital system or research institute) that wants to use historical patient data to improve understanding of treatment outcomes. The results of this analysis could help clinicians monitor patients more closely, prioritize follow-up care, or identify groups of patients with higher risk of relapse.

3. Data Sources

The primary dataset for this project is **The Cancer Genome Atlas (TCGA)**, a publicly available cancer dataset provided by the U.S. National Cancer Institute.

TCGA includes:

- Patient demographics
- Cancer type and diagnosis information
- Follow-up visits
- Disease status (e.g., tumor progression or stable disease)
- Survival and outcome variables
- Treatment indicators

TCGA data can be accessed through the NIH Genomic Data Commons (GDC) Portal:

- https://portal.gdc.cancer.gov/
- https://registry.opendata.aws/tcga/
- https://www.cancer.gov/ccg/research/genome-sequencing/tcga

This dataset supports longitudinal analysis because it contains time-based follow-up information for each patient, allowing patient histories to be represented as sequences over time.

4. Criteria for Success

The project will be considered successful if it can:

- Accurately model how cancer patients transition between disease states over time
- Demonstrate that sequence-based models outperform or add value over non-temporal baselines
- Identify meaningful temporal patterns associated with disease progression or stability

Performance will be evaluated using predictive metrics (e.g., accuracy, AUC, or time-to-event error) as well as by visualizing predicted vs. observed patient trajectories.

5. Scope of the Solution Space

The project will focus on one or a small number of cancer types (such as breast or lung cancer) to keep the scope manageable. The analysis will be limited to clinical and follow-up data and will not require advanced genomic interpretation, although genomic features may be added later if feasible.

The solution will focus on modeling how patient outcomes evolve over time rather than making one-time predictions.

6. Constraints

- TCGA data requires some preprocessing and filtering to construct longitudinal patient timelines.
- Follow-up data may be irregularly spaced in time, which can complicate modeling.
- The dataset is de-identified and observational, so causal conclusions cannot be made.
- Compute and time limitations will require selecting a manageable subset of the data.

7. Stakeholders

The main stakeholders for this project include:

- Cancer researchers
- Clinicians and oncologists
- Healthcare systems and hospital administrators
- Patients, who benefit from improved understanding of disease progression

These stakeholders care about improving patient outcomes and using data to support better treatment decisions.

8. Proposed Approach

This project will treat cancer treatment response as a **time-series prediction problem**. Each patient in the TCGA dataset will be represented as a sequence of clinical observations collected across follow-up visits, including disease status, treatment indicators, and outcome variables.

After constructing these patient timelines, two modeling approaches will be explored:

1. **Baseline models**, which ignore temporal ordering and use aggregated patient features
2. **Deep learning sequence models**, which explicitly model how patient states evolve over time

Deep learning models such as **recurrent neural networks (LSTM or GRU)** or **temporal transformers** will be used to process each patient's sequence of visits. These models are designed to learn long-term dependencies, such as how early treatment response influences later outcomes.

The models will be trained to predict clinically relevant outcomes such as:

- Whether a patient's disease will progress
- Whether the patient remains stable or improves
- Time-dependent survival or relapse risk

By comparing deep learning models with simpler baselines, the project will demonstrate the value of sequence-based modeling for cancer outcome prediction.

9. Why Deep Learning Is Essential for This Project

Cancer progression is not determined by a single data point, but by the **entire trajectory** of a patient's disease. Deep learning is well-suited for this problem because:

- It can model **long-term dependencies** across visits
- It handles **irregularly spaced clinical events**
- It learns **nonlinear relationships** between treatments and outcomes
- It supports **multitask learning** (e.g., predicting progression and survival simultaneously)

This makes deep learning a natural and justified choice for modeling cancer treatment response over time.

10. Deliverables

The final deliverables for this project will include:

- A GitHub repository containing all code and documentation
- A written report describing the data, analysis, and results
- Visualizations showing patient trajectories and model outputs
- A final PDF of this project proposal and results