**Onco-Logic: Intelligent Machine Oncology**

CST 489/499

Nicholas Marolla, Federico Murrieta, Rahim Siddiq, Harris Popal

Dr. Alireza Abdoli

Summer 2025

# Executive Summary

The goal of this project is to develop a predictive model for breast cancer survival using clinical data from the SEER cancer registry, sourced from Kaggle. The project focuses on equipping participants with skills in data cleaning, feature selection, and survival analysis using machine learning techniques.

Breast cancer is a leading cause of cancer-related mortality, and early identification of high-risk patients is essential for improving outcomes. Unlike traditional classification models that predict binary outcomes, this project focuses on survival analysis, which estimates the time until an event such as death based on clinical features including age, tumor size, lymph node status, and hormone receptor status. This approach provides more detailed and actionable insights for clinical decision-making

The project has two primary objectives: (1) to train reliable survival models using algorithms such as survival XGBoost and survival random forest, and (2) to interpret model outputs using survSHAP to identify the most important prognostic factors. The dataset consists of 4,024 de-identified patient records that include both demographic and tumor-related variables.

Expected outcomes include the development of an accurate and interpretable survival prediction model, along with a deeper understanding of which clinical features most influence patient prognosis. Aligned with the Cedars-Sinai AI Campus mission, this project emphasizes collaborative learning, ethical data use, and practical application of machine learning in healthcare. The results will support personalized treatment planning and improved decision-making in breast cancer care.

# Table of Contents

[**Executive Summary 2**](#_5ny0aqrjd8sj)

[**Table of Contents 3**](#_cak2uq6mr3e4)

[**Introduction/Background 4**](#_z5nbsg5cx6ty)

[Project Name and Description 4](#_aa7m3nvp26yv)

[Problem and/or Issue in Technology 5](#_m5xu0xhzjqa7)

[Solution to the Problem and/or Issue in Technology 5](#_gqwz52tzpcax)

[**Environmental Scan / Literature Review 6**](#_u22sh6pvzaad)

[**Stakeholders 7**](#_yx2ts520k7dk)

[Ethical Considerations 9](#_ukedc83p9pfv)

[Legal Considerations 10](#_40yejz26en20)

[**References 11**](#_foozx2yxrg4e)

# Introduction/Background

Our project is based on *Breast Cancer Prognosis: Leveraging Cancer Registry Data for Survival Prediction*, an official project within the Cedars-Sinai AI Campus initiative. This proposal outlines the project’s goals, the specific clinical and technical problems it aims to address, and the machine learning solutions it proposes. It also includes a review of related research, a discussion of stakeholder roles, and an overview of the ethical and legal considerations relevant to the project’s development and implementation.

## Project Name and Description

**Onco-Logic: Intelligent Machine Oncology**

This project focuses on building a predictive model for breast cancer survival using publicly available clinical data from the SEER cancer registry. By applying machine learning algorithms to estimate patient-specific survival outcomes, the model aims to identify patterns and risk factors associated with reduced survival.

Key clinical features used in the model include age, tumor size, lymph node status, and hormone receptor status. These variables have been shown to play a significant role in determining patient prognosis and treatment response.

Designed for healthcare professionals, data scientists, and interdisciplinary learners, this project provides a practical tool for early risk identification and personalized treatment planning. By improving the accuracy of survival predictions, it supports data-driven decisions that can lead to better outcomes for patients and more efficient use of clinical resources.

## **Problem and/or** I**ssue in** T**echnology**

Breast cancer remains one of the most common and deadly cancers worldwide. While significant progress has been made in treatment, accurately predicting patient survival remains a difficult and unresolved challenge. Traditional models typically produce binary outcomes, indicating whether a patient is likely to survive or not. However, such models overlook the timing of outcomes, which is essential for effective clinical decision-making.

Survival prediction, which estimates the time until an event such as death, offers more detailed prognostic insight but requires advanced analytical techniques that many existing tools lack. Current models often fall short in accuracy, generalizability, and interpretability, limiting their use in personalized treatment planning. As a result, high-risk patients may not be identified early enough to benefit from timely interventions.

There is a clear need for more accurate, data-driven survival prediction models that can account for complex clinical features and generate individualized risk assessments. A machine learning approach, applied to high-quality registry data, offers the potential to close this gap and improve both prognosis and care delivery.

## Solution to the Problem and/or Issue in Technology

To address the limitations of current survival prediction tools, this project proposes the development of a machine learning model that estimates individualized breast cancer survival probabilities using clinical data. The model will be built using algorithms such as survival XGBoost and survival random forest, both of which are well-suited for time-to-event analysis and capable of handling complex, non-linear relationships in the data.

The dataset, sourced from Kaggle and derived from the SEER cancer registry, includes over 4,000 de-identified patient records with features such as age, tumor size, lymph node involvement, hormone receptor status, and other tumor characteristics. These features will be used to train, validate, and test the model to ensure both accuracy and generalizability.

To enhance interpretability, the project will apply survSHAP, a model technique that identifies and ranks the most influential factors driving survival predictions. This interpretability is critical for clinical adoption, allowing healthcare professionals to understand and trust the model's output.

The final outcome will be a robust, explainable tool that provides clinicians with actionable insights to support early risk stratification and informed treatment decisions, ultimately aiming to improve patient outcomes and optimize care delivery.

# Environmental Scan / Literature Review

Breast cancer survival prediction is a well-established area of research, with growing interest in using machine learning to improve prognostic accuracy. Phung et al. (2019) conducted a comprehensive review of breast cancer prediction models and emphasized the importance of incorporating clinical features such as tumor size and hormone receptor status. Their findings highlight the limitations of traditional models and the potential of data-driven approaches.

Efthimiou et al. (2024) offered a practical guide for developing clinical prediction models, outlining best practices for feature selection, validation, and interpretation. These guidelines are directly applicable to this project and support its emphasis on transparency and methodological rigor.

Recent advancements in machine learning have expanded the use of time-to-event models for survival analysis. Krzyziński et al. (2023) demonstrated that survival XGBoost and survival random forests can effectively capture complex patterns in survival data. They also introduced survSHAP, a method for interpreting survival models by identifying the time-dependent influence of individual features. This technique will be applied in the current project to support clinical interpretability.

The SEER cancer registry provides a strong foundation for model development. With over 4,000 de-identified patient records and detailed clinical variables, the dataset is well suited for building and validating survival models. Tools like the PREDICT breast cancer website (Grootes et al., 2024) further illustrate how predictive models can be integrated into clinical workflows and inform real-world treatment decisions.

These studies and resources collectively validate the project’s approach and support the application of machine learning methods to improve breast cancer prognosis.

# Stakeholders

The stakeholders in this project include healthcare professionals, patients, data scientists, project mentors, and potentially healthcare institutions. Healthcare professionals are the primary beneficiaries, as the predictive model will support clinical decision-making by identifying high-risk patients earlier and more accurately. This can lead to more personalized treatment plans and improved allocation of medical resources.

Patients benefit indirectly from these improvements in care. More precise survival predictions can guide earlier interventions, contributing to better treatment outcomes and increased survival rates.

Data scientists and developers involved in the project will gain valuable hands-on experience working with clinical data and applying machine learning techniques in a real-world health context. Through tasks such as data preprocessing, model development, and the use of interpretability tools like survSHAP, they will strengthen their skills in both technical modeling and applied analytics.

Project mentors, including those affiliated with Cedars-Sinai AI Campus, serve as crucial guides throughout the development process. Their support ensures that the project remains methodologically sound, ethically responsible, and aligned with the learning objectives of the AI Campus initiative.

Finally, healthcare institutions and research organizations may benefit from the implementation of the final predictive model. By incorporating the tool into clinical workflows, these organizations could improve the accuracy and efficiency of breast cancer prognosis, leading to enhanced patient care and research outcomes.

## Ethical Considerations

The primary ethical consideration in this project involves the responsible use of patient data. Although the dataset is derived from the SEER cancer registry and accessed through Kaggle in a de-identified format, maintaining privacy and confidentiality remains essential. The project will comply with all relevant data protection guidelines, including principles outlined by HIPAA, even though the dataset does not contain personal health information (PHI). All analyses will be conducted using anonymized data, and care will be taken to ensure that no model outputs can be linked back to individual patients.

Another key ethical consideration is accessibility. The predictive tool should be designed to be usable by healthcare providers with varying levels of technical expertise. Outputs will be presented in a clear, interpretable manner to support real-world clinical decision-making without requiring specialized training in machine learning.

In addition, the project will incorporate safeguards to promote fairness and transparency. Model performance will be regularly audited for potential bias across demographic groups, such as race, age, and gender. Any disparities in model behavior will be investigated, and steps will be taken to mitigate unintended bias to ensure equitable outcomes for all patient populations.

Overall, the project aims to uphold high ethical standards in data use, system design, and outcome fairness throughout the development and evaluation process.

## **Legal Considerations**

The primary legal considerations for this project involve ensuring that all data sources are properly documented and used in accordance with applicable licensing terms. The clinical dataset is publicly available through Kaggle and originates from the SEER cancer registry, which provides de-identified patient data for research purposes. No personally identifiable information (PII) or protected health information (PHI) is included in the dataset, and no additional permissions are required for its use in this context.

Nevertheless, the project will adhere to relevant data protection standards, including those outlined in HIPAA and, where applicable, the General Data Protection Regulation (GDPR). Although the dataset itself is exempt from these regulations due to its public and anonymized nature, the principles of data ethics and responsible handling will still be observed throughout the project.

In addition, any software tools, libraries, or frameworks used in the development of the predictive model will be properly cited and utilized under their respective licenses. Intellectual property created during the project, including model code, notebooks, and documentation, will be clearly attributed and shared under an appropriate open-source license, such as the MIT License, to encourage transparency, collaboration, and reproducibility.

# References

Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., & Salanti, G. (2024). Developing

clinical prediction models: A step-by-step guide. *BMJ*, 386, 1–10.

https://doi.org/10.1136/bmj.2024.065367

Grootes, I., Wishart, G. C., & Pharoah, P. D. (2024). An updated PREDICT breast cancer

prognostic model including the benefits and harms of radiotherapy. *NPJ Breast Cancer, 10*(1), 6. https://doi.org/10.1038/s41523-024-00612-1

Krzyziński, M., Spytek, M., Baniecki, H., & Biecek, P. (2023). SurvSHAP (t): Time-dependent

explanations of machine learning survival models. *Knowledge-Based Systems, 262*, 110234. https://doi.org/10.1016/j.knosys.2023.110234

Phung, M. T., Tin Tin, S., & Elwood, J. M. (2019). Prognostic models for breast cancer: A

systematic review. *BMC Cancer, 19*, 1–18. https://doi.org/10.1186/s12885-019-5894-2