Tab 1

* Problem Definition & Scope
  + Build on your project charter - add more details including hosting organization's mission.
  + Clear problem identification with justification (importance/impact).
  + Defined project scope and measurable objectives.
* Background Research & Context
  + Overview of existing solutions or related work.
* Methodology & Design
  + Description of chosen methodology, design, and overall approach.
* Resource Identification and Management
  + Identification and management plan for necessary resources.
* Progress and Preliminary Results
  + Completed work, achieved milestones, and encountered challenges.
  + Presentation of preliminary findings or data (if applicable).
* Future Work
  + Suggested future steps and overall project plan.

Tab 2

**Midterm Progress Report: Predictive Modeling of Immune-Related Adverse Events in Melanoma Patients Treated with Immune Checkpoint Inhibitors**

**Problem Definition and Scope**

This capstone project is conducted in collaboration with the Lombardi Comprehensive Cancer Center (LCCC), an NCI-designated comprehensive cancer center committed to reducing the burden of cancer through research, clinical care, and translational science. Leveraging the LCCC-curated Immuno-Oncology (IO) Patient Registry, the project aims to address a critical unmet need in immunotherapy: the prediction of immune-related adverse events (irAEs) prior to initiating treatment with immune checkpoint inhibitors (ICIs).

While ICIs have transformed the therapeutic landscape for advanced cancers, their use is frequently complicated by irAEs, which can range in severity from mild to life-threatening. Currently, clinical decision-making lacks robust tools to stratify irAE risk before therapy initiation. As a result, monitoring and management of irAEs are largely reactive, contributing to avoidable hospitalizations, treatment interruptions, and reduced patient safety.

This project aims to close that gap by developing a machine learning classifier to estimate the probability of irAE occurrence before ICI therapy begins. The model is trained on structured clinical and demographic data from the IO Registry. To improve signal quality and clinical relevance, the study population is restricted to melanoma patients receiving ICI monotherapy as a first-line treatment with no prior exposure to ICIs. By narrowing the cohort in this way, the model focuses on the clinical scenario where predictive insights are most actionable: the initial treatment decision.

The scope of this project is limited to structured data and does not include free-text clinical notes or imaging. Modeling efforts prioritize interpretability by employing explainable machine learning techniques such as XGBoost and logistic regression. The primary objective is to achieve strong predictive performance (AUROC ≥ 0.75) while ensuring model transparency to facilitate future clinical integration.

**Background Research and Context**

Immune checkpoint inhibitors, including agents targeting PD-1, PD-L1, and CTLA-4, have substantially improved survival in patients with advanced melanoma. However, these therapies are associated with immune-related adverse events (irAEs), which may involve multiple organ systems and vary widely in severity and onset. Identifying patients at high risk of irAEs prior to treatment initiation remains a major challenge in immuno-oncology.

Despite the clinical importance of irAEs, current guidelines from the National Comprehensive Cancer Network (NCCN) and the American Society of Clinical Oncology (ASCO) focus on post-hoc detection and management. This reactive orientation means clinicians have few tools to proactively tailor surveillance or treatment regimens based on individual risk. As a result, patients with elevated risk may go unidentified, and low-risk patients may be over-monitored, leading to inefficient and potentially harmful care.

Recent studies have attempted to characterize clinical and demographic predictors of irAE risk. Huang et al. (2022) found associations between irAEs and factors such as female sex, normal BMI, and preexisting autoimmune disorders in NSCLC patients receiving PD-1/PD-L1 inhibitors. Altan et al. further suggested that machine learning models incorporating features such as EGFR mutation status and smoking history could improve irAE risk stratification. However, these findings have not yet translated into widely adopted clinical tools.

Gao et al. (2023) demonstrated the feasibility of modeling irAE risk using structured registry data in NSCLC patients, achieving favorable AUROC scores. However, their model lacked stratification by therapy line and regimen modality.

This project builds directly on these prior efforts by focusing on a narrowly defined, clinically homogeneous population—melanoma patients receiving first-line ICI monotherapy. The use of a curated registry with detailed, structured variables enables a more precise and transparent modeling process, increasing the likelihood that findings can inform real-world clinical workflows.

**Methodology and Design**

The modeling approach is grounded in supervised learning, with the goal of predicting the binary outcome of irAE occurrence (io\_toxicity == 'Yes') based solely on pre-treatment structured data. All analyses are conducted on de-identified records from the Immuno-Oncology Patient Registry.

**Cohort Selection**

A strict set of inclusion and exclusion criteria was applied to define the modeling cohort:

* Inclusion: Patients with primary melanoma (currentcancer = 'Melanoma'), no history of ICI therapy (prevrx\_io\_therapy = 'No'), first-line therapy (current\_io\_line == 1), monotherapy only (io\_agent1 not containing '+'), and complete treatment initiation data (io\_dose1date). Patients must not have had an irAE prior to ICI start and must have at least one follow-up at least four weeks after treatment (do\_followup > io\_dose1date + 28 days).
* Exclusion: Duplicate records or retreatment episodes were removed.

This selection strategy ensures a well-defined temporal window, minimizes confounding, and supports confident negative labeling of irAE outcomes.

**Feature Set Design**

The feature set includes demographic variables (age, gender, race), cancer descriptors (stage, metastases), performance status (ECOG), comorbidity indicators (e.g., checkboxes and dates), and laboratory values (e.g., LDH, albumin, BMI). Engineered variables include:

* A "first irAE onset" variable derived from the earliest value among all "Date of Onset" irAE fields
* Categorical therapy indicators derived from the naming structure of io\_agent1
* A Charlson Comorbidity Index (CCI) score computed from relevant comorbidities

Only variables available prior to the treatment start date are retained for model training.

**Modeling Strategy**

Supervised classification models (logistic regression, decision trees, XGBoost) are used, with hyperparameter tuning via cross-validation. Model performance is evaluated using AUROC, and feature importance will be interpreted using SHAP or Gini-based methods to support explainability.

**Resource Identification and Management**

The project draws on several resources:

* **Data**: The IO Registry provides structured REDCap data. Access is governed by IRB approval and data use agreements.
* **Computing**: Google Colab is used for scalable model training and visualization.
* **Software**: Python libraries include pandas, scikit-learn, XGBoost, matplotlib, seaborn.
* **Mentorship**: Clinical and technical guidance is provided by Dr. Yili Zhang, Dr. Neil Shah, and Dr. Adil Alaoui.

Project phases are planned across data cleaning, cohort selection, exploratory analysis, and model development. Key deliverables are managed via modular scripts and version-controlled notebooks.

**Progress and Preliminary Results**

As of the midterm checkpoint, extensive exploratory data analysis (EDA) and variable engineering have been completed. Field labels were aligned to current REDCap dictionaries, and candidate variables were categorized into directly matched, expanded checkboxes, or missing.

Cohort selection was programmatically applied using the inclusion/exclusion criteria. This resulted in a cleaner, clinically meaningful subset of the registry.

Initial EDA insights include:

* **Distribution of melanoma cases**: Melanoma is among the most common primary cancers in the dataset.
* **irAE timing**: CDF plots revealed most irAEs occur within 100 days of treatment, while follow-up durations are more widely distributed. This insight guides labeling confidence for patients with no documented irAEs.
* **Missingness**: Over 30% missingness in several variables was identified. Diagnosis date fields were excluded from modeling due to high sparsity and limited clinical interpretability.
* **Variable utility**: AUROC-based screening showed several variables (e.g., CCI, baseline labs) with scores >0.6, marking them as strong candidates.

These efforts set the stage for model training and evaluation in the next project phase.

**Future Work**

The next steps involve translating exploratory insights into actionable models and integrating clinical domain knowledge. Meetings are planned with Dr. Shah and Dr. Lev-Ari to discuss variable relevance, tolerable levels of missingness, and potential feature exclusions. Their clinical insight will refine the final feature set and guide model interpretation.

In parallel, the dataset will be expanded to incorporate laboratory test data, which may provide more dynamic and mechanistically relevant predictors of irAEs.

Model development will then proceed using cross-validated training, performance evaluation, and interpretation via feature attribution tools. The aim is to produce a high-performing, interpretable risk prediction tool that could be prospectively validated and deployed in clinical risk stratification workflows.

**Citations**

Gao, W., Liu, Q., Zhou, Y., Yang, M., & Yu, Y. (2023). The Predictive Model Construction for Immune-Related Adverse Events in Non-Small Cell Lung Cancer Patients Receiving Immunotherapy. *Technology in Cancer Research & Treatment*, *22*, 15330338231206705.<https://doi.org/10.1177/15330338231206705>

Huang, Y., Soon, Y. Y., Aminkeng, F., Tay, S. H., Ang, Y., Kee, A. C. L., Goh, B. C., Wong, A. S. C., & Soo, R. A. (2022). Risk factors for immune-related adverse events from anti-PD-1 or anti-PD-L1 treatment in an Asian cohort of nonsmall cell lung cancer patients. *International Journal of Cancer*, *150*(4), 636–644.<https://doi.org/10.1002/ijc.33822>