## **Project Charter: "Utilizing Immuno-Oncology Registry Data for Enhanced Immune-related Adverse Events Predictions”**

### **1. Introduction & Background**

Immunotherapy, particularly through Immune Checkpoint Inhibitors (ICIs), has transformed cancer treatment by enabling the immune system to recognize and attack tumor cells. However, this approach is associated with a wide range of immune-related adverse events (irAEs), which can significantly affect patient quality of life and treatment outcomes. Predicting these adverse effects prior to ICI therapy administration remains a major clinical challenge, due to variability in patient responses and the complex interplay of biological, environmental, and social factors.

This project, in collaboration with the Lombardi Comprehensive Cancer Center, seeks to address this gap by applying machine learning (ML) techniques to the Immuno-Oncology Patient Registry. The registry contains a rich set of clinical, demographic, and potentially time-series and social determinants of health (SDOH) data, which will be leveraged to develop a predictive model for irAEs.

### **2. Problem Statement and Justification**

Immune Checkpoint Inhibitors (ICIs) have revolutionized cancer care, offering durable responses in various malignancies. However, their use is frequently complicated by immune-related adverse events (irAEs)—unintended inflammatory reactions that affect healthy tissues. These events are often unpredictable, vary in severity, and may lead to treatment discontinuation, hospitalization, or permanent damage.

Currently, there are no robust tools to predict the likelihood of irAEs in individual patients prior to treatment. Risk factors are heterogeneous and scattered across structured and unstructured clinical data. This lack of personalized risk assessment limits oncologists’ ability to optimize treatment strategies, pre-empt toxicities, or provide truly informed consent.

#### **Business Case and Benefits**

A predictive model for irAEs offers significant value:

* **Clinically**: Enables proactive care, tailored treatment decisions, and earlier intervention to prevent complications.
* **Operationally**: Reduces unplanned hospitalizations and costly emergency care.
* **Scientifically**: Leverages underutilized real-world data, including NLP-extracted adverse events, to generate actionable insights.

### **3. Project Objectives and Goals**

* Develop and evaluate a machine learning pipeline to predict the likelihood of irAEs in cancer patients undergoing ICI therapy.
* Integrate diverse data types including medical history, laboratory tests, and social determinants of health.
* Explore the predictive value of time-series features (e.g., lab trends).
* Enable risk-informed decision-making for oncologists and care teams.

### **4. Project Scope & Boundaries**

This project will focus on the development, evaluation, and documentation of a predictive ML model that forecasts the likelihood of immune-related adverse events (irAEs) in patients undergoing immune checkpoint inhibitor (ICI) therapy. The work will include data preprocessing, feature engineering, model training and evaluation, and preliminary interpretability analysis.

The scope **does not** include model deployment into clinical systems or prospective validation in a clinical trial setting. However, the outputs will be structured in a way that enables future translation or pilot testing.

### **5. In-Scope / Out-of-Scope**

| **In-Scope** | **Out-of-Scope** |
| --- | --- |
| Acquisition and cleaning of Immuno-Oncology Registry data | Real-time integration into EHRs or clinical decision systems |
| Feature engineering from clinical history, lab tests, and SDOH data | Development of a patient-facing application or visualization dashboard |
| Training and comparing multiple ML algorithms for the classification of irAEs | Legal or regulatory approval of the predictive tool |
| Exploratory analysis of time-series data for additional predictive power | Manual chart review or detailed physician annotation beyond what’s in the registry |
| Producing a reproducible Jupyter notebook or Python script with documentation | Integration with insurance or billing systems |

### **6. Deliverables**

* A cleaned and preprocessed version of the Immuno-Oncology Registry dataset tailored for ML model development.
* A trained machine learning classification model capable of predicting irAE occurrence prior to ICI administration.
* Evaluation metrics (e.g., precision, recall, AUC-ROC) and stratified performance reports.
* Documentation of features, model architecture, assumptions, and limitations.
* A final capstone report and presentation summarizing methodology, findings, and future directions.

### **7. Key Stakeholders**

| **Stakeholder** | **Role / Interest** |
| --- | --- |
| **Clinicians and Researchers from the I/O Registry Consortium** | End-users of the predictive model; interested in identifying early warning signals for irAEs to improve treatment planning and patient outcomes. |
| **MedStar Health Network & Hackensack Meridian Health System** | Institutional data providers; interested in ensuring ethical data use, model generalizability, and improved patient care across sites. |
| **Cancer Center Administrators** | Interested in leveraging predictive analytics to reduce treatment complications, optimize resource allocation, and meet value-based care goals. |
| **Patients Undergoing ICI Therapy** | Indirect beneficiaries of improved safety and personalization of treatment. |

### **8. Team Members & Roles**

| **Team Member** | **Contact** | **Role** |
| --- | --- | --- |
| **Austin Cherian** | [Austin Cherian](mailto:azc10@georgetown.edu)  (414)-573-0047 | Principal Investigator / Data Scientist-in-Training. Responsible for end-to-end execution including literature review, data curation, model development, evaluation, and reporting. |
| **Yili Zhang** | [yz932@georgetown.edu](mailto:yz932@georgetown.edu) | HIDS Mentor; provides guidance on modeling strategy, validation methods, and technical writing. |
| **Adil Alaoui** | [Adil Alaoui](mailto:alaoui@georgetown.edu)  2026879108 | Course Director; supports project structuring, milestone tracking, and deliverable review. |
| **Neil Shah** | ShahN6@mskcc.org | Clinical mentor; interprets clinical relevance, identifies meaningful features, and supports clinical validation pathways. |

### **9. Project Approach**

#### **Methodology and Research Design**

### This project will use a retrospective observational design, leveraging de-identified data from the Immuno-Oncology Registry. The central approach involves building a supervised classification model to predict the risk of immune-related adverse events (irAEs) before initiating immune checkpoint inhibitor (ICI) therapy. The methodology includes:

### Descriptive analysis of patient cohorts

### Feature engineering across structured (EHR) and unstructured (NLP-derived) fields

### Training and evaluation of ML models using stratified cross-validation

#### **Data Sources and Collection Methods**

### **Primary source**: Immuno-Oncology Registry data curated by the I/O consortium

### **Data types**: Diagnosis codes (ICD), labs, medication history, radiology/pathology summaries, clinician-entered toxicities, and NLP-extracted irAE labels from clinical notes

### **Collection method**: Multi-hospital ETL pipeline into a centralized research warehouse

#### **System Development Lifecycle (SDLC)**

### **Data Engineering Phase**: Preprocessing, cleaning, normalization

### **Model Development Phase**: Iterative training, evaluation, tuning

### **Interpretability & Documentation**: Explainable ML and final deliverable packaging

### **No deployment or productionization is planned in this phase**

#### **Analysis Techniques and Tools**

### **Languages/Tools**: Python (pandas, scikit-learn, XGBoost, SHAP), Jupyter

### **Techniques**: Feature importance analysis, stratified sampling, time-series feature extraction, class imbalance correction (e.g., SMOTE or class weights)

### **10. Assumptions**

* The Immuno-Oncology Registry data has undergone a rigorous and well-documented ETL (Extract, Transform, Load) process to ensure accurate integration from source systems (EHRs, lab systems, pathology, radiology, and cancer registries).
* Adverse events are captured through both structured data (e.g., ICD codes) and in-house NLP extraction from unstructured EHR notes.
* Clinician-labeled irAE data, including those extracted from unstructured clinical notes using NLP methodologies, is accurate and reflective of true patient outcomes.
* Patient data has been de-identified in compliance with HIPAA and institutional review board (IRB) standards.
* The dataset is of sufficient quality and completeness to support robust machine learning analysis, particularly for modeling irAE occurrence.
* Metadata, including treatment timelines and lab timestamps, is complete and correctly aligned to individual patient records.
* Access to relevant features—including clinical history, lab results, and social determinants of health—will be provided for analysis.
* Stakeholders and mentors will be available for timely feedback on interim findings and model outputs.

**11. Constraints**

| **Constraint** | **Description** |
| --- | --- |
| **Data Access Limitations** | Access to the Immuno-Oncology Registry may be limited to pre-approved extracts or require additional approvals for new data pulls. |
| **Data Quality Variability** | Given the integration of multi-institutional data, there may be inconsistencies or missingness in key variables. |
| **Computational Resources** | Project is limited to academic computing resources; large-scale model training or hyperparameter tuning may be constrained. |
| **Time Constraints** | Must complete all deliverables by the end of the capstone semester (Summer 2025). |
| **Single-Person Project Execution** | Execution relies entirely on one student investigator, limiting redundancy and capacity. |
| **Lack of Prospective Validation** | The model will be evaluated retrospectively and may require further validation before clinical use. |

### 

### **12. Risks**

| **Risk** | **Mitigation Strategy** |
| --- | --- |
| **Incomplete or missing labels for irAEs** | Perform data quality assessment early. If critical samples are missing irAE labels, either (1) manually extract labels from unstructured clinical notes where feasible, or (2) exclude those samples from analysis if the volume of missing data is acceptably low. |
| **Model overfitting due to limited irAE-positive cases** | Employ cross-validation, regularization techniques, and class balancing strategies. |
| **Bias introduced by data imbalance or institution-specific patterns** | Stratify data where possible and evaluate performance across hospital systems. |
| **Difficulty interpreting model predictions in a clinical context** | Include model explainability techniques (e.g., SHAP values) to enhance clinical utility. |
| **Delays in feedback or mentorship due to faculty schedules** | Proactively schedule check-ins and share progress updates regularly to avoid bottlenecks. |

### **13. Success Metrics**

| **Metric** | **Target / Evaluation Criteria** |
| --- | --- |
| **Model performance (e.g., AUC-ROC, F1 Score)** | Achieve clinically meaningful performance (e.g., AUC > 0.75) on hold-out data |
| **Feature relevance and explainability** | Generate interpretable feature importance rankings using SHAP or similar tools |
| **Stakeholder feedback** | Positive evaluation from mentors regarding methodological soundness and relevance |
| **Reproducibility** | Code and documentation allow independent replication of results |
| **Presentation & Final Report** | Meets or exceeds HIDS capstone standards; clearly communicates goals, methods, and implications |

### 

### **14. Timeline (High-Level - Updated)**

| **Phase** | **Date / Week** | **Milestones / Activities** |
| --- | --- | --- |
| **Project Kick-off & Problem Definition** | Week 1 (May 21, 2025) | Onboarding, problem definition, review of potential solutions and project overview. |
| **Planning & Research** | Week 2 (May 26, 2025) | Continue literature review, project charter development. |
| **Charter Finalization** | Week 3 (June 2, 2025) | **Complete Project Charter** (due end of Week 3). |
| **Design Foundation** | Week 4 (June 9, 2025) | Block diagram, functional requirements, solution specifications. |
| **Intermediate Design & Feedback** | Week 5 (June 16, 2025) | Initial implementation of design, receive feedback, iterate. |
| **Comprehensive Design & Finalization** | Week 6 (June 23, 2025) | Finalize solution design and release package. |
| **Progress Checkpoint** | Week 7 (June 30, 2025) | **Midterm Progress Report Due**: Submit progress report on model development and data work. |
| **Final Design (Cont.)** | Week 8 (July 7, 2025) | Continue feature refinement, model tuning, or design upgrades. |
| **Testing & Evaluation** | Week 9 (July 14, 2025) | Perform solution testing, model validation, and performance assessment. |
| **Documentation & Reporting** | Week 10 (July 21, 2025) | Write final report, finalize code notebooks, prepare for submission. |
| **Report Submission & Final Review** | Week 11 (July 28, 2025) | **Final Report Due & Review**: Submit written report and receive feedback. |
| **Presentation** | Week 12 (August 4, 2025) | **Final Project Showcase**: Present project between **3:00 – 6:00 PM**. |

### **15. Budget**

* **Budget:** N/A  
   (No external funding or purchases are expected. Computational and software needs will be met using Georgetown University academic resources.)

### **16. Approval Sign-Offs**

| **Name** | **Role** | **Signature / Date** |
| --- | --- | --- |
| **Austin Cherian** | Student Investigator | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Yili Zhang, PhD** | HIDS Mentor | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Adil Alaoui, PhD** | Capstone Course Director | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Neil Shah, MD** | LCCC Clinical Mentor | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |