Tab 1

# **Predicting Immune-Related Adverse Events (irAEs) in ICI Therapy**

## **1. Background and Motivation**

### **1.1 Clinical Context**

Immunotherapy—particularly through Immune Checkpoint Inhibitors (ICIs) targeting PD-1, PD-L1, and CTLA-4—has revolutionized the treatment of advanced cancers such as melanoma. By enabling the immune system to recognize and attack tumor cells, ICIs have produced durable responses in a subset of patients who previously had limited options.

However, ICI therapy is frequently complicated by immune-related adverse events (irAEs)—unintended inflammatory reactions that affect healthy tissues. These adverse events are heterogeneous in timing, severity, and organ system involvement. irAEs may lead to hospitalization, permanent organ dysfunction, or early discontinuation of treatment, thereby diminishing the therapeutic benefit and impacting patient quality of life.

Despite their clinical significance, irAEs remain largely unpredictable prior to treatment initiation. Current guidelines from the National Comprehensive Cancer Network (NCCN) and American Society of Clinical Oncology (ASCO) emphasize reactive management of irAEs but offer few tools to proactively stratify patients by risk (Schneider et al., 2021; Thompson et al., 2020). As a result, clinicians are left without actionable insights to optimize therapy planning or tailor patient monitoring protocols.

### **1.2 Problem Statement**

This capstone study, conducted in partnership with the Lombardi Comprehensive Cancer Center (LCCC), addresses a significant unmet need in immuno‑oncology: the absence of a validated, pre‑treatment stratification strategy for immune-related adverse events (irAEs) in patients initiating immune checkpoint inhibitor (ICI) therapy.

Although several prior studies have identified candidate predictors—such as **female sex**, **elevated body mass index (BMI)**, **autoimmune history**, and **EGFR mutation status**—these factors remain largely untransformed into operational clinical risk tools (Predictive Biomarkers for Immune-Related Adverse Events, n.d.). The predictive capacity of these attributes, while suggestive, has not yet resulted in routine adoption or prospective validation.

Machine learning models that have utilized electronic health record (EHR) data to estimate irAE risk are still evolving. Many such models lack sufficient granularity relating to treatment regimen, line of therapy, or baseline clinical condition, and few are readily deployable within real‑world oncology workflows (Prediction of Effectiveness and Toxicities, 2025).

Moreover, predictive signals are scattered across multiple data modalities—including structured EHR fields, longitudinal laboratory trends, and unstructured clinical narrative—creating data heterogeneity that complicates prediction at the individual level and impedes the seamless integration of risk outputs into clinician-facing decision-making systems.

### **1.3 Project Objectives**

The overarching goal of this project is to develop a machine learning pipeline to predict the likelihood of irAE occurrence in patients receiving ICI therapy. Specific objectives include:

1. Develop and evaluate an interpretable predictive model (e.g., logistic regression, XGBoost) using structured EHR data available prior to therapy initiation.
2. Engineer clinically meaningful features from medical history, comorbidities, treatment regimen, and social determinants of health (SDOH).
3. Assess model performance using ROC-AUC and other metrics, with a target AUROC of ≥ 0.75.
4. Provide preliminary interpretability analysis (e.g., SHAP values, feature importance) to support clinical understanding and adoption.
5. Structure the pipeline for potential future integration into clinical decision support tools, although deployment and prospective validation are outside the current scope.

## **2. Data Source and Cohort**

### **2.1 Data Source**

This study utilizes data from the Georgetown–MedStar–Hackensack Immuno-Oncology (IO) Patient Registry, a centralized research data warehouse developed by a multidisciplinary team of clinicians, informaticians, and data scientists. The registry aggregates real-world clinical data from over 3,500 cancer patients treated with immune checkpoint inhibitors (ICIs) across six academic and community institutions located in Washington, D.C.; Maryland; and New Jersey. It is designed to support retrospective outcomes research and translational analytics in immuno-oncology.

For this capstone project, we restricted analysis to a subset of the registry (n ≈ 854) due to two practical considerations: (1) the integration of the old and new cohorts was not finalized at the time of analysis, and (2) the subset included a validated and up-to-date data dictionary that ensured reliable variable mapping for model development.

Although the full registry incorporates heterogeneous data modalities—including structured electronic medical records (EMRs), unstructured clinical notes, pathology and radiology reports, laboratory results, and natural language processing (NLP) outputs—this project focused exclusively on structured EMR-derived fields. This decision was made to prioritize pipeline reproducibility, traceability, and adherence to project time constraints. The included data types encompassed baseline demographic information, clinical history, treatment regimen characteristics, and pre-treatment status indicators documented prior to ICI initiation.

*Note: Future extensions of this work may incorporate additional data types, such as longitudinal laboratory trends or NLP-extracted adverse event mentions from clinical notes, to further enhance model performance and clinical relevance.*

### **2.2 Cohort Definition**

From the broader patient pool, a melanoma-specific cohort was defined using the following criteria:

#### **Inclusion:**

* Diagnosis of melanoma as the primary cancer
* Receipt of at least one dose of ICI therapy
* Documented ICI start date and baseline clinical features

#### Exclusion:

* Any immune-related adverse event (irAE) recorded prior to ICI initiation
* At least 28 days of follow-up post-ICI start date for patients without irAEs (to reduce risk of false negatives due to limited observation)
* Substantial missingness in critical demographic or treatment fields

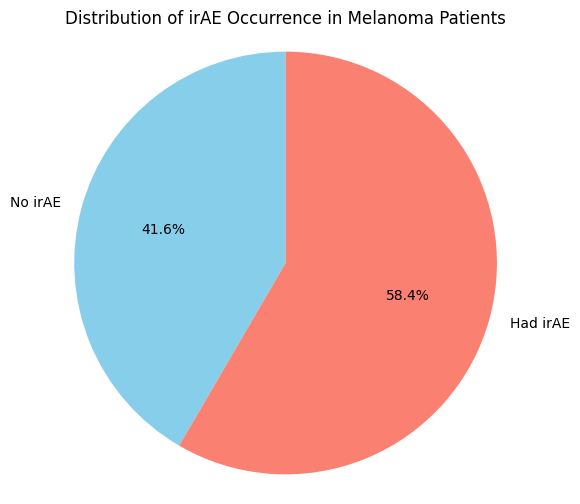
These steps ensured a clinically homogeneous and temporally aligned cohort suitable for pre-treatment risk prediction.

### **2.3 Final Cohort Summary**

The final modeling cohort consisted of 274 melanoma patients treated with ICIs. Among these:

* **160 patients (58.4%)** experienced at least one irAE following ICI initiation
* **114 patients (41.6%)** did not experience any irAE during the observation window

The outcome was defined as a binary variable indicating whether an irAE occurred after the first ICI dose. To visualize the distribution of irAEs within the final cohort, the figure below presents a pie chart highlighting the class balance:



**Figure 1. Distribution of irAE Occurrence in the Final Cohort (N = 274).**

*Over half of melanoma patients treated with ICIs experienced at least one irAE.*

## 

## 

## **3. Feature Engineering and Transformation**

To enhance predictive modeling and ensure clinical interpretability, we engineered a curated set of features summarizing baseline patient characteristics, treatment details, and comorbidities. Below is a high-level overview of the final set of features included in the model:

**Charlson Comorbidity Index (CCI):** Comorbidity checkbox fields were aggregated into a weighted summary score using the age-adjusted Charlson Comorbidity Index, capturing each patient's baseline comorbidity burden. This provided a single interpretable numeric feature that integrates the prognostic impact of multiple conditions.

**ICI Treatment Regimen:** We derived two key features from treatment data: (1) *ICI mechanism of action* (anti–PD-1, anti–CTLA-4, or combination/other), and (2) *combination vs monotherapy*. These variables reflect known differences in irAE risk based on regimen intensity.

**Concurrent Medications:** To capture concurrent therapy complexity, we created a binary indicator representing whether the patient received any additional medications (e.g., corticosteroids, antibiotics) alongside ICI. This simplification avoided sparsity while preserving clinical relevance.

**Demographics and Lifestyle Factors:** Age at treatment initiation, gender, and race/ethnicity were included. Smoking history was collapsed into “Never” vs “Ever/Unknown.” Missing values were conservatively handled through grouping or imputation.

**Performance Status (ECOG):** Baseline ECOG performance was binned into three clinically meaningful groups: “0,” “1,” and “≥2 or unknown,” reflecting levels of functional status relevant to treatment tolerance and irAE susceptibility.

**Autoimmune Disease History:** A binary indicator was created to denote the presence or absence of pre-existing autoimmune disease, based on patient history. This was retained for its clinical importance despite low prevalence.

**Cancer Stage and Metastatic Sites:** A simplified binary variable captured whether patients had Stage IV disease at treatment initiation. Additional binary indicators were created for common metastatic sites (e.g., lung, liver), with rare sites dropped to reduce noise.

**Treatment History:** We included a binary variable for *line of therapy* (first-line vs later lines) and whether the patient had received prior immunotherapy, capturing aspects of treatment sequencing and exposure.

**Final Feature Set:** After transformation and dimensionality reduction, the final modeling dataset consisted of ~20–25 structured features encompassing patient demographics, comorbidities, functional status, treatment regimen, and clinical history. The engineered features prioritized interpretability, avoided redundancy, and preserved clinically meaningful distinctions for robust modeling.

*The full technical detail, rationale, and visualizations for each transformation are provided in the Supplemental Section (Appendix A).*

## **4. Data Cleaning and Preparation**

Developing a robust machine learning model from real-world clinical data requires rigorous preprocessing to ensure data quality, reduce noise, and support generalizability. Key steps in this pipeline included standardized variable mapping, missing data handling, and dimensionality reduction through sparsity pruning. The following summarizes the preprocessing workflow applied prior to model development.

### **4.1 Data Dictionary Mapping**

We leveraged the IO Registry’s data dictionary to programmatically map raw variable names to human-readable labels and to identify each feature’s data type (e.g., categorical, continuous, checkbox). Special attention was given to REDCap-style checkbox fields, which expand into multiple binary indicator columns (e.g., "Comorbidities" → "Comorbidities (choice = Hypertension)", "Comorbidities (choice = Diabetes Uncomplicated)", etc.). Any features not present in the dataset or marked as deprecated in the dictionary were logged and excluded from further analysis to maintain data integrity and consistency.

### **4.2 Missing Data Handling**

Missingness was assessed at both the patient and feature levels. Patients missing more than five critical variables (e.g., demographics or treatment initiation dates) were excluded to preserve the reliability of outcome labeling and feature completeness. Features with over 30% missing values—such as prior cancer history—were dropped due to limited representation.

For numerical variables (e.g., age, body mass index [BMI], Charlson Comorbidity Index [CCI]), missing values were imputed using the cohort median, a conservative strategy that maintains central tendency without distorting the underlying distribution.

For categorical variables, a clinically conservative recoding strategy was applied to minimize false negatives in a risk-sensitive context:

* Missing ECOG performance status was grouped with ECOG ≥ 2 (indicating reduced functional capacity),
* Unknown or ambiguous race/ethnicity was categorized as “Not Specified/Unknown,”
* Missing autoimmune history was treated as positive, acknowledging potential unrecorded risk factors.

By treating missingness as informative and conservatively biased, this approach aligns with real-world clinical decision-making, where overestimating risk is often preferable to providing false reassurance.

### **4.3 Low-Variance Feature Pruning**

To mitigate the effects of sparsity and reduce the dimensional burden on the modeling process, we implemented low-variance feature pruning. Binary indicators—particularly those derived from one-hot encoded and checkbox fields—were excluded if they met either of the following criteria: (1) a positive response occurred in fewer than 10 patients, or (2) a single category accounted for more than 90% of non-missing values. These thresholds were identified using a custom profiling utility (profile\_feature\_to\_table) designed to evaluate feature prevalence, distributional imbalance, and sparsity. This approach improved the signal-to-noise ratio and reduced the risk of overfitting on statistically unreliable features while preserving clinically relevant variation.

**Final Dataset Summary**Following preprocessing, the final analytical dataset consisted of 274 patients and approximately 30 engineered features derived from structured clinical variables.

| Variable | Missing (%) | Unique Values | Most Common (%) |
| --- | --- | --- | --- |
| Cancer stage at which the patient is receiving treatment | 0 | 12 | 76.3 |
| Did the patient experience any I-O toxicities? | 0 | 2 | 58.4 |
| Does this patient have a history of autoimmune disease? | 18.2 | 2 | 82.1 |
| ECOG pre treatment | 0.4 | 5 | 49.8 |
| Gender | 18.2 | 2 | 66.1 |
| Pre treatment BMI | 2.6 | 242 |  |
| Race | 18.2 | 5 | 86.6 |
| Smoking History | 19 | 5 | 57.7 |
| Specify the subtype of melanoma (choice=Choroidal) | 0 | 2 | 97.1 |
| Specify the subtype of melanoma (choice=Cutaneous) | 0 | 2 | 70.4 |
| Specify the subtype of melanoma (choice=Mucosal) | 0 | 2 | 93.8 |
| Specify the subtype of melanoma (choice=Other) | 0 | 2 | 90.5 |
| Specify the subtype of melanoma (choice=Uveal) | 0 | 2 | 99.6 |
| Total number of metastatic sites | 10.9 | 4 | 42.6 |
| Was the patient previously treated with an Immuno-therapy? | 0.4 | 2 | 75.5 |
| Which line of therapy is the patient's current immunotherapy? | 0.7 | 4 | 61 |
| ICI Mechanism Category | 0 | 4 | 61.3 |
| ICI Combination Type | 0 | 2 | 72.3 |
| Age from ICI Start | 18.2 | 224 |  |
| Charlson Comorbidity Index | 18.2 | 9 | 22.8 |
| No Concurrent Meds/Unspecified | 0 | 2 | 67.2 |
| Concurrent Meds (Any) | 0 | 2 | 67.2 |
| Metastasis (choice=Bone) | 0 | 2 | 83.2 |
| Metastasis (choice=Brain) | 0 | 2 | 77 |
| Metastasis (choice=Liver) | 0 | 2 | 74.1 |
| Metastasis (choice=Lung) | 0 | 2 | 59.5 |
| Metastasis (choice=Lymph nodes) | 0 | 2 | 54.7 |
| Metastasis (choice=Other Visceral Organs) | 0 | 2 | 90.9 |
| Metastasis (choice=Soft tissue (skin, subcutaneous)) | 0 | 2 | 62.4 |
| No\_Metastasis | 0 | 2 | 86.1 |

**Table 1: Summary Statistics for Final Feature Set After Data Cleaning**

*This table presents a summary of the remaining variables in the analytical dataset following rigorous data cleaning. Features with more than 30% missingness or extremely low variance (e.g., single-category dominance >90% or presence in fewer than 10 patients) were excluded to reduce noise and prevent overfitting. A small number of patient records were also removed due to extensive missing values. Each retained feature is profiled with its missingness rate, number of unique values, and the proportion of the most common category, providing transparency into the filtering process.*

| **Characteristic, no. (%)** | **Entire cohort**  **N = 224** | **irAE**  **N = 138** | **No irAE**  **N = 86** |
| --- | --- | --- | --- |
| Charlson Comorbidity Index, mean (SD) | 2.28 (1.76) | 2.02 (1.54) | 2.70 (2.01) |
| Age--years, mean (SD) | 62.87 (14.67) | 60.93 (13.77) | 65.98 (15.60) |
| Pretreatment BMI, mean (SD) | 27.83 (5.84) | 27.84 (5.00) | 27.81 (7.00) |
| Gender - Male | 148 (66.07) | 92 (66.67) | 56 (65.12) |
| Pretreatment ECOG |  |  |  |
| 0 | 111 (49.55) | 81 (58.70) | 30 (34.88) |
| 1 | 95 (42.41) | 54 (39.13) | 41 (47.67) |
| 2+ | 18 (8.04) | 3 (2.17) | 15 (17.44) |
| History of autoimmune disease? - Yes | 40 (17.86) | 27 (19.57) | 13 (15.12) |
| Race |  |  |  |
| White | 194 (86.61) | 120 (86.96) | 74 (86.05) |
| Non-White | 15 (6.70) | 8 (5.80) | 7 (8.14) |
| Not Specified | 15 (6.70) | 10 (7.25) | 5 (5.81) |
| Smoking history |  |  |  |
| Never Smoker | 128 (57.14) | 87 (63.04) | 41 (47.67) |
| Ever Smoker/Unknown | 96 (42.86) | 51 (36.96) | 45 (52.33) |
| Line of ICI treatment |  |  |  |
| First | 166 (74.11) | 102 (73.91) | 64 (74.42) |
| Second or Later | 58 (25.89) | 36 (26.09) | 22 (25.58) |
| Previously treated with an Immuno-Oncology - Yes | 19 (8.48) | 13 (9.42) | 6 (6.98) |
| Melanoma type |  |  |  |
| Cutaneous | 161 (71.88) | 98 (71.01) | 63 (73.26) |
| Other | 44 (19.64) | 29 (21.01) | 15 (17.44) |
| Rare Non-Cutaneous | 19 (8.48) | 11 (7.97) | 8 (9.30) |
| Medications given concurrently with Immuno-Oncology - Concurrent Meds (Any) | 76 (33.93) | 39 (28.26) | 37 (43.02) |
| ICI Mechanism Category |  |  |  |
| anti-PD-1 | 126 (56.25) | 80 (57.97) | 46 (53.49) |
| CTLA-4 | 93 (41.52) | 53 (38.41) | 40 (46.51) |
| Other/Combo | 5 (2.23) | 5 (3.62) | 0 (0.00) |
| Stage |  |  |  |
| Stage IV | 171 (76.34) | 104 (75.36) | 67 (77.91) |
| Stage III or Earlier | 53 (23.66) | 34 (24.64) | 19 (22.09) |

**Table 2.** **Baseline Characteristics of the Study Cohort Stratified by irAE Status.** *Summary of demographic, clinical, and treatment-related features for 224 melanoma patients treated with immune checkpoint inhibitors (ICIs), stratified by the occurrence of immune-related adverse events (irAEs). Variables include comorbidity index, age, BMI, ECOG performance status, autoimmune history, race, smoking history, treatment line, melanoma subtype, ICI mechanism, and concurrent medication use. Differences across groups provide insight into potential predictors of irAE occurrence.*

## 

## **5. Model Preparation**

### **5.1 Feature Encoding**

Categorical variables were encoded using one-hot encoding (OHE), including race, ECOG performance status, ICI regimen type, comorbidities, and metastatic sites. All categories were retained to ensure model interpretability, except where binary variables allowed one category to be inferred (e.g., including only the 'monotherapy' flag implies combination therapy when absent).

### **5.2 Handling Redundant Indicator Columns**

To reduce dimensionality and mitigate redundancy introduced by checkbox-style variables, only the positive indicator columns (e.g., met\_liver = 1) were retained following one-hot encoding. Negative indicators (= 0) were dropped, as their absence could be inferred. In cases where all checkbox fields were zero, a derived flag (e.g., no\_metastasis) was introduced to explicitly capture absence of all relevant findings.

### **5.3 Feature Scaling**

Continuous features—including age, BMI, CCI, number of metastatic sites, and selected lab values—were standardized using z-score normalization. Scaling parameters were computed on the training set and applied to both training and test sets to prevent data leakage. Standardization ensured comparability across features and optimized performance for algorithms sensitive to feature scale, such as logistic regression and support vector machines.

### **5.4 Train-Test Split**

To support robust model evaluation and mitigate overfitting, the final dataset of 274 melanoma patients was partitioned into independent training and testing subsets. A stratified random sampling approach was used to preserve the outcome class distribution—specifically, the proportion of patients experiencing an immune-related adverse event (irAE), which was 58.4% in the full cohort.

* **Training Set**: 70% of the cohort (n ≈ 192)
* **Hold-Out Test Set**: 30% of the cohort (n ≈ 82)

Stratification ensured that both subsets reflected the underlying class imbalance, thereby improving generalizability and enabling fair performance comparison across models. Patient assignment was randomized without overlap, maintaining strict independence between training and evaluation phases and reducing the risk of data leakage during model development.

## **6. Feature Selection, Model Development, and Evaluation**

### **6.1 Feature Group Importance Analysis**

To inform feature selection and support model interpretability, we conducted a domain-level feature group importance analysis using logistic regression with 3-fold cross-validation. Features were grouped by clinical relevance (e.g., treatment type, demographics, ECOG status), and each group was evaluated in isolation to estimate its independent predictive contribution.

The performance of each group was assessed using mean area under the receiver operating characteristic curve (AUROC) (Ling, Huang, & Zhang, 2003). As shown in **Table 3**, the highest-performing group was **treatment type** (AUROC = 0.7417), followed by **demographics**, **ECOG performance status**, and **comorbidities**. Lower-ranked groups such as **stage** and **treatment history** contributed limited standalone signal.

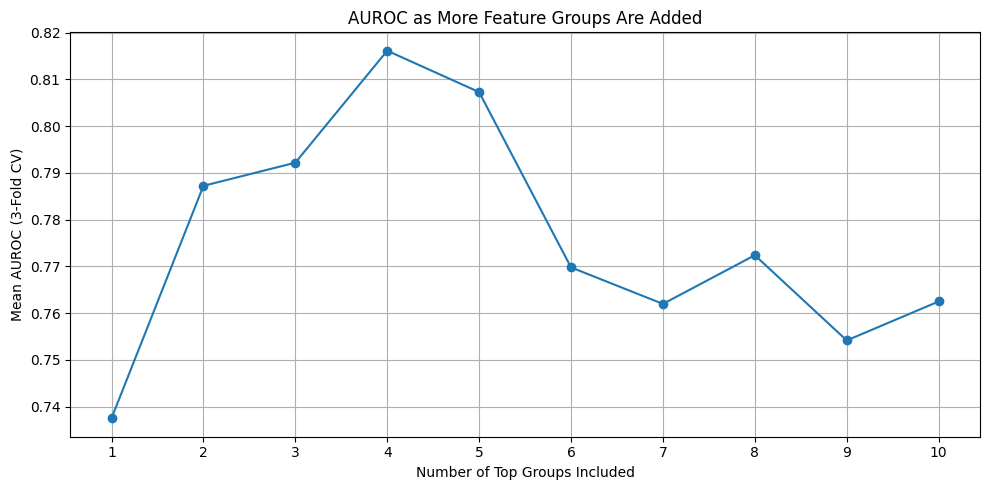
| **Rank** | **Feature Group** | **Mean AUROC** |
| --- | --- | --- |
| 1 | Treatment Type | 0.7417 |
| 2 | Demographics | 0.6911 |
| 3 | ECOG Status | 0.6815 |
| 4 | Comorbidity | 0.6255 |
| 5 | Lifestyle | 0.5901 |
| 6 | Metastasis | 0.5182 |
| 7 | Melanoma Subtype | 0.4935 |
| 8 | Autoimmune History | 0.4854 |
| 9 | Stage | 0.4594 |
| 10 | Treatment History | 0.4531 |

**Table 3.** **Mean AUROC by Feature Group Using Logistic Regression.** *Each group was evaluated independently to assess its standalone predictive contribution to irAE classification.*

### **6.2 Cumulative Feature Group Addition**

To evaluate the additive value of each group, we performed a cumulative inclusion analysis, starting from the top-ranked group and adding others sequentially in descending order of importance. Performance was measured using 3-fold cross-validation on the training set.

As shown in **Figure 2**, AUROC improved through the inclusion of the top four groups but plateaued or declined thereafter. This finding supported the use of a parsimonious feature set and minimized risk of overfitting due to the inclusion of low-signal variables.



**Figure 2.** **Mean AUROC as Feature Groups Are Cumulatively Added.** *Model performance improved with the inclusion of the top-ranked feature groups and plateaued after the fourth, supporting dimensionality reduction.*

### **6.3 Final Feature Inclusion Strategy**

Based on the cumulative analysis, the following four feature groups were selected for inclusion in the final modeling pipeline:

* **Treatment Type** (e.g., monotherapy vs. combination, ICI mechanism)
* **ECOG Performance Status**
* **Lifestyle Factors** (BMI, smoking history)
* **Autoimmune History** (included based on clinical relevance)

These groups balanced predictive performance with clinical interpretability and demonstrated consistency across multiple model types.

### **6.4 Hyperparameter Tuning**

Seven machine learning algorithms were benchmarked: Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), Gradient Boosting, k-Nearest Neighbors (KNN), and Bernoulli Naive Bayes (BNB). For each, we performed grid search with 10-fold stratified cross-validation on the training set. Tuning focused on maximizing AUROC as the primary objective (Ling, Huang, & Zhang, 2003).

Key hyperparameters and tuning ranges are summarized in **Table 4**.

| **Model** | **Key Parameters Tuned** |
| --- | --- |
| Logistic Regression | C (0.01, 0.1, 1, 10), penalty='l2' |
| Random Forest | n\_estimators (25, 50), max\_depth (3, 5), min\_samples\_leaf=4, class\_weight='balanced' |
| XGBoost | n\_estimators, max\_depth, learning\_rate, subsample, colsample\_bytree |
| SVM | C, kernel='rbf', gamma='scale' |
| KNN | n\_neighbors (3, 5), weighting scheme |
| Bernoulli NB | alpha (0.1, 1.0), binarize=0.0 |

**Table 4.** **Key Hyperparameters Tuned per Algorithm.**

*Parameters were optimized using 10-fold stratified cross-validation. Grid search was parallelized across CPU cores (n\_jobs = -1).*

### **6.5 Evaluation Strategy**

Each final model—trained on the full training set using optimized hyperparameters—was evaluated on the independent hold-out test set.

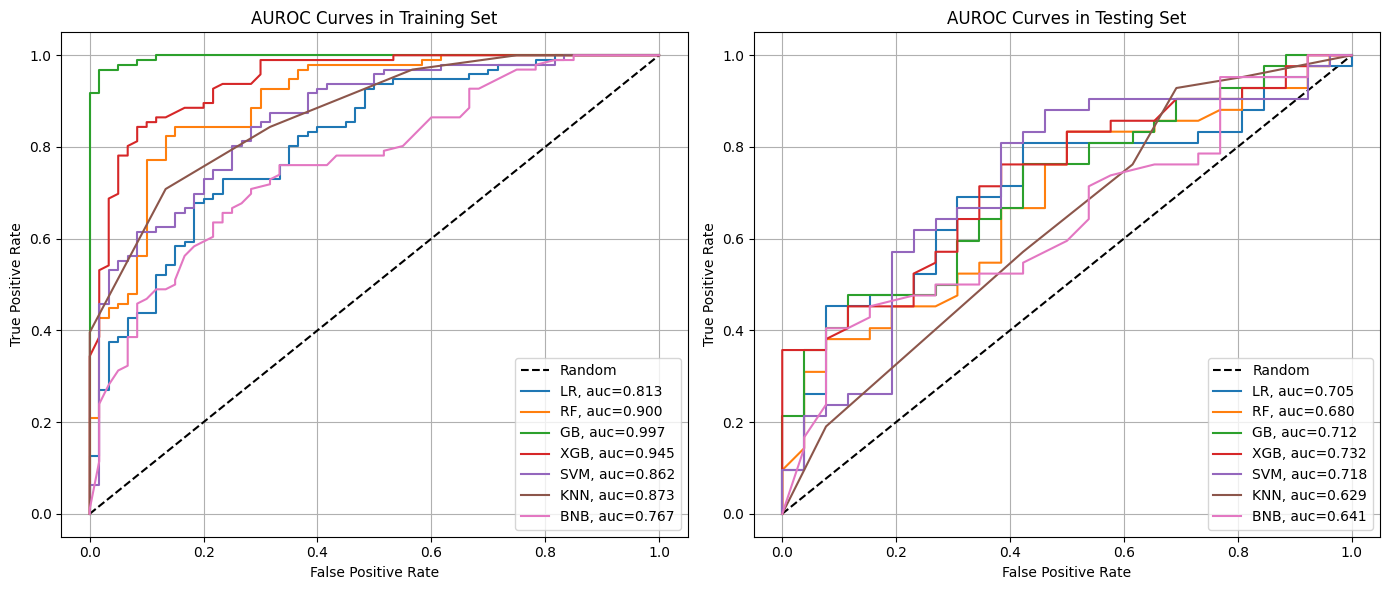
As shown in **Table 5**, the XGBoost classifier yielded the highest AUROC (0.732) and F1 score (0.761), demonstrating superior balance between sensitivity and specificity. Logistic regression and SVM followed closely, while KNN and Bernoulli Naive Bayes showed lower discriminative power.

| **Model** | **Precision** | **Recall** | **Accuracy** | **F1** | **AUROC** |
| --- | --- | --- | --- | --- | --- |
| XGB | 0.7 | 0.833 | 0.676 | 0.761 | 0.732 |
| SVM | 0.618 | 1 | 0.618 | 0.764 | 0.718 |
| GB | 0.702 | 0.786 | 0.662 | 0.742 | 0.712 |
| LR | 0.723 | 0.81 | 0.691 | 0.764 | 0.705 |
| RF | 0.706 | 0.571 | 0.588 | 0.632 | 0.68 |
| BNB | 0.658 | 0.595 | 0.559 | 0.625 | 0.641 |
| KNN | 0.667 | 0.762 | 0.618 | 0.711 | 0.629 |

**Table 5.** **Model Performance Metrics on Hold-Out Test Set.**

*It Includes AUROC, accuracy, precision, recall, and F1 score for each algorithm. XGBoost demonstrated the strongest overall performance with an AUROC score of 0.732.*

Additionally, ROC curves were plotted for each model to visually compare discrimination capability between training and test sets. As seen in **Figure 3**, all models exhibited higher AUROC on the training data compared to the test data, consistent with expectations. Notably, XGBoost maintained strong generalization, while KNN and Bernoulli Naive Bayes showed significant performance degradation between training and testing.



**Figure 3. ROC Curves for All Models on Training and Testing Sets.**

*Each curve represents one classifier’s performance in terms of true positive rate versus false positive rate. AUROC values are shown in the legend. Dashed line denotes random classifier performance. Training performance is on the left, and test set generalization is on the right.*

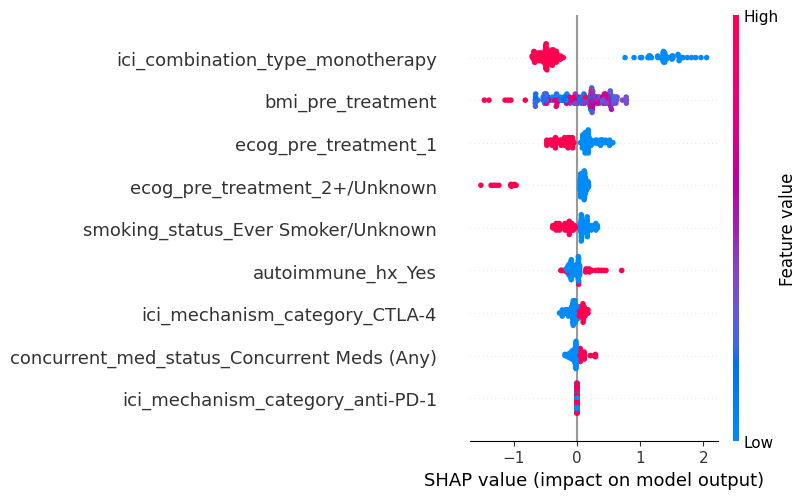
## **7. Feature Interpretation & Importance**

To interrogate model decision-making and ensure clinical face validity, we performed a post-hoc interpretability analysis using SHAP (SHapley Additive exPlanations) values on the best performing mode: XGBoost. SHAP provides a unified measure of feature attribution, quantifying each feature's marginal contribution to individual predictions (Lundberg & Lee, 2017).

The SHAP summary plot (**Figure 4**) highlights the top drivers of model output across the cohort. Consistent with our feature selection rationale, the most influential predictors included:

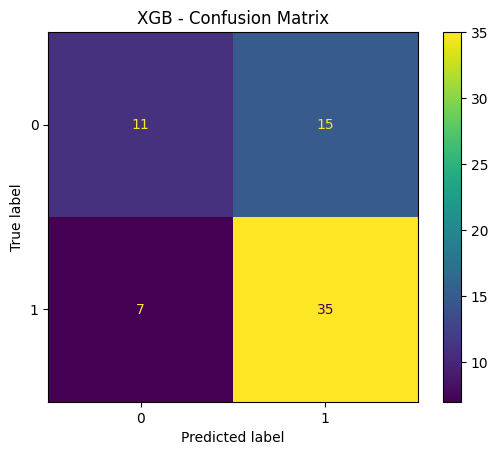
* **ICI Combination Type**: Receiving monotherapy (vs. combination) was associated with a decreased risk of irAE, as evidenced by a negative SHAP contribution.
* **Pretreatment BMI**: Higher BMI values were modestly associated with increased irAE risk.
* **ECOG Performance Status**: Poor functional status (ECOG ≥2) showed strong positive SHAP values, aligning with clinical intuition that frail patients may experience heightened toxicity.
* **Smoking History** and **Autoimmune History** also emerged as meaningful contributors to model predictions.

These findings suggest that the model is leveraging clinically plausible risk factors and capturing known irAE correlates, enhancing trust in its predictive behavior.



**Figure 4. SHAP Summary Plot of Final XGBoost Model.***Each dot represents a patient. Dot color reflects feature value (blue = low, red = high). Features are ranked by overall importance. SHAP values to the right of the vertical line increase predicted irAE risk, while those to the left decrease it.*

To complement feature-level interpretation, we also examined the confusion matrix of the XGBoost model on the hold-out test set (**Figure 5**). The model demonstrated strong recall (0.833) for identifying patients who experienced an irAE, but exhibited a relatively higher false positive rate, misclassifying a subset of irAE-negative patients as positive. This behavior aligns with the model’s conservative bias towards overprediction, which is often acceptable in high-risk clinical applications where failing to identify at-risk individuals could carry adverse consequences.



**Figure 5. Confusion Matrix of Final XGBoost Model on Test Set.***The matrix summarizes model predictions versus true irAE status. Class 0 = No irAE; Class 1 = Experienced irAE. The model achieved strong sensitivity but showed some overprediction of irAEs.*

Together, SHAP analysis and threshold-based evaluation reveal that the model’s behavior is both clinically interpretable and aligned with the goal of safety-first predictive screening. These findings support the model’s potential utility as a triage aid for treatment decision-making or monitoring protocol design.

## **8. Discussion**

### **8.1 Summary of Key Findings**

This study demonstrates the feasibility of using machine learning to predict the occurrence of immune-related adverse events (irAEs) in melanoma patients receiving immune checkpoint inhibitors (ICIs). Among the models evaluated, XGBoost yielded the highest overall performance, achieving an AUROC of 0.732 on the independent test set. Feature group analysis revealed that treatment-related variables and baseline clinical status were the most predictive of irAE occurrence. Stepwise group inclusion further indicated diminishing returns beyond the top four clinical domains, justifying a parsimonious feature set in the final model.

### **8.2 Explainability and Interpretability**

Although XGBoost marginally outperformed SVM in discriminative performance, its utility was greatly enhanced by post hoc explainability tools. SHAP (SHapley Additive exPlanations) values were used to quantify both the magnitude and direction of each feature’s impact on model predictions (Lundberg & Lee, 2017). This added layer of interpretability supports both clinical validation and user trust in the model’s outputs.

SHAP plots revealed clinically intuitive findings: for instance, monotherapy was associated with a lower predicted risk, while patients with better ECOG performance status or elevated baseline BMI had higher predicted risk. These patterns align with known risk factors and bolster model credibility. Beyond model auditing, interpretability has direct implications for clinical translation. By understanding the rationale behind predictions, clinicians may feel more confident incorporating model outputs into decision-making processes. Furthermore, these visualizations can support patient-clinician conversations regarding treatment risks and follow-up planning, and may even guide future research into modifiable predictors of toxicity.

### **8.3 Clinical Implications**

With further validation, the final model could inform a pre-treatment irAE risk score integrated into immuno-oncology clinical workflows. Such a tool could identify high-risk patients, such as those receiving combination ICI therapy or presenting with favorable ECOG status, who may benefit from enhanced monitoring, early intervention, or modified treatment plans. Conversely, patients with lower predicted risk—such as those on monotherapy or presenting with a higher comorbidity burden—could be spared unnecessary resource utilization or intensive monitoring.

From an operational standpoint, this risk stratification framework could support more efficient allocation of clinical resources, improving patient safety while reducing the likelihood of treatment interruptions. Over time, this approach has the potential to enhance therapeutic continuity and outcomes by proactively addressing adverse events before they escalate.

### **8.4 Limitations**

Several limitations merit consideration. First, the model did not control for lead-lag time biases related to treatment exposure duration, survival time, or follow-up period. While adjusting for these variables might have improved internal validity, doing so would have reduced the available sample size below the threshold necessary for effective modeling.

Second, the dataset was derived from a single academic institution, limiting generalizability to other practice settings or populations. Additionally, the study was restricted to patients with melanoma; irAE profiles may differ substantially in other cancers such as non-small cell lung cancer (NSCLC) or renal cell carcinoma (RCC). The moderate sample size further constrained model complexity and precluded rigorous evaluation of rare variable combinations.

Furthermore, only baseline structured features were used. Time-varying signals such as lab trends, clinician assessments, or unstructured clinical notes were excluded. As a result, the model reflects static risk assessment rather than dynamic monitoring.

Finally, the irAE outcome was defined as a binary indicator, collapsing all event types and severities into a single label. While this enabled model feasibility, future iterations may benefit from predicting specific irAE subtypes or severity grades to support more nuanced clinical decision-making.

### **8.5 Future Directions**

Future work will focus on enhancing model generalizability, usability, and clinical relevance. External validation using independent datasets from other institutions or clinical trials is essential to assess reproducibility. In parallel, embedding the model into clinical workflows for prospective testing could elucidate real-world utility, clinician adoption, and impact on patient outcomes.

Predictive targets can also be refined. Rather than binary irAE prediction, future models may stratify by severity or specific irAE categories (e.g., dermatologic, gastrointestinal, endocrine). Incorporating longitudinal data—including baseline and follow-up laboratory values, vital sign trends, and NLP-derived features from unstructured notes—could further improve sensitivity and specificity.

Threshold customization should also be explored to match clinical priorities. For instance, in early-warning applications, higher sensitivity thresholds may be warranted. Models that account for emerging biomarkers—such as tumor mutational burden (TMB), cytokine profiles, or circulating immune markers—may also yield improved prognostic accuracy.

In the longer term, predictive tools such as this may support interventional trials aimed at mitigating toxicity in high-risk patients. Tailoring treatment intensity, initiating prophylactic therapies, or augmenting supportive care based on individualized risk scores represents a key avenue for advancing precision immuno-oncology.

### **8.6 Collaboration and Clinical Context**

This work was made possible through close collaboration between clinical investigators and data scientists. Clinical input guided decisions around variable inclusion and interpretation (e.g., handling of unknown ECOG or autoimmune status), while modeling strategies were adapted to accommodate the constraints and nuances of real-world EHR data.

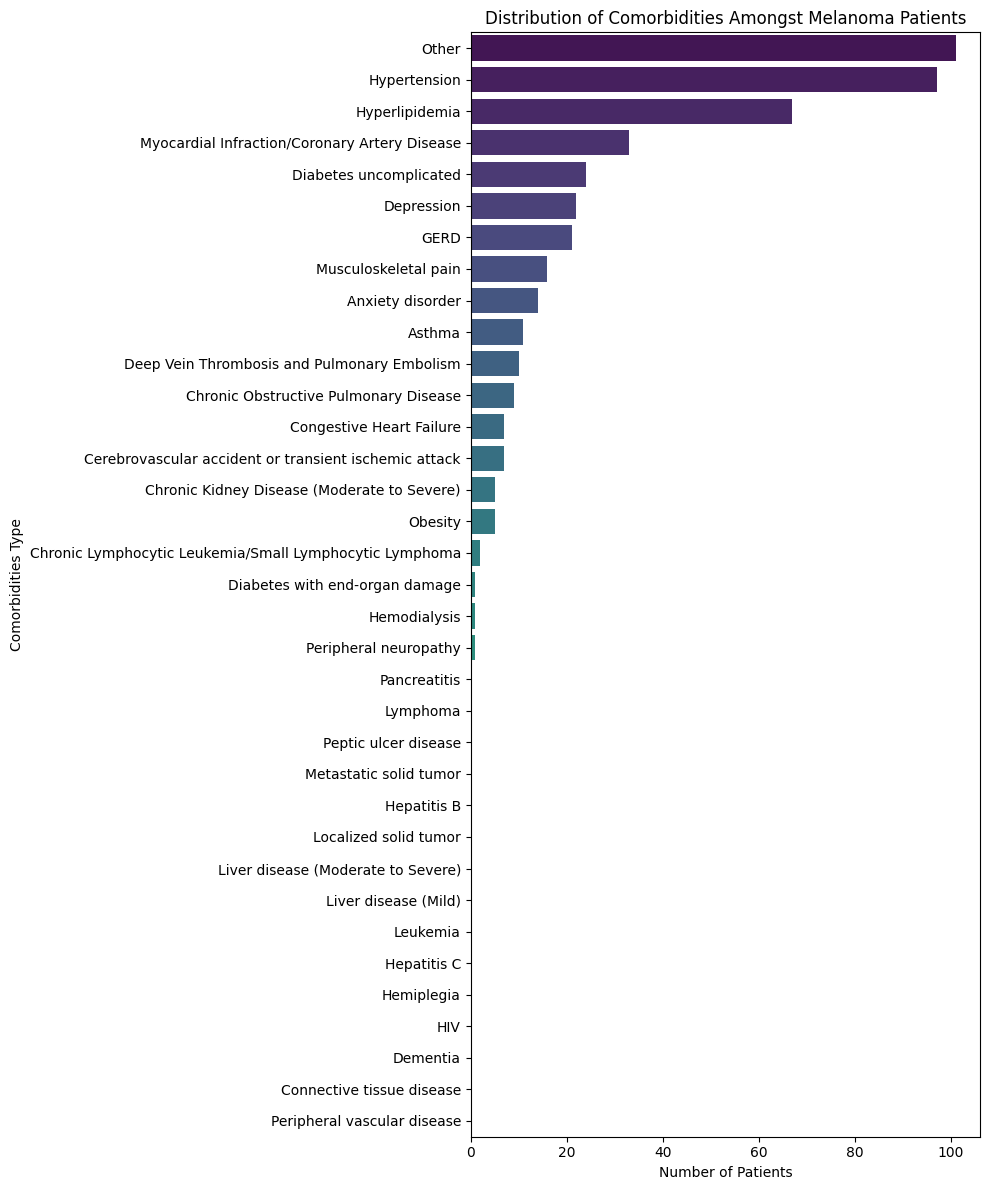
By aligning methodological rigor with clinical intuition, we developed a model that is not only statistically sound but also grounded in domain relevance. This proof-of-concept demonstrates that machine learning can be effectively applied to anticipate immune-related toxicities. With continued refinement and validation, such tools may play a pivotal role in personalizing immunotherapy delivery and improving patient safety.

## **Supplemental Materials**

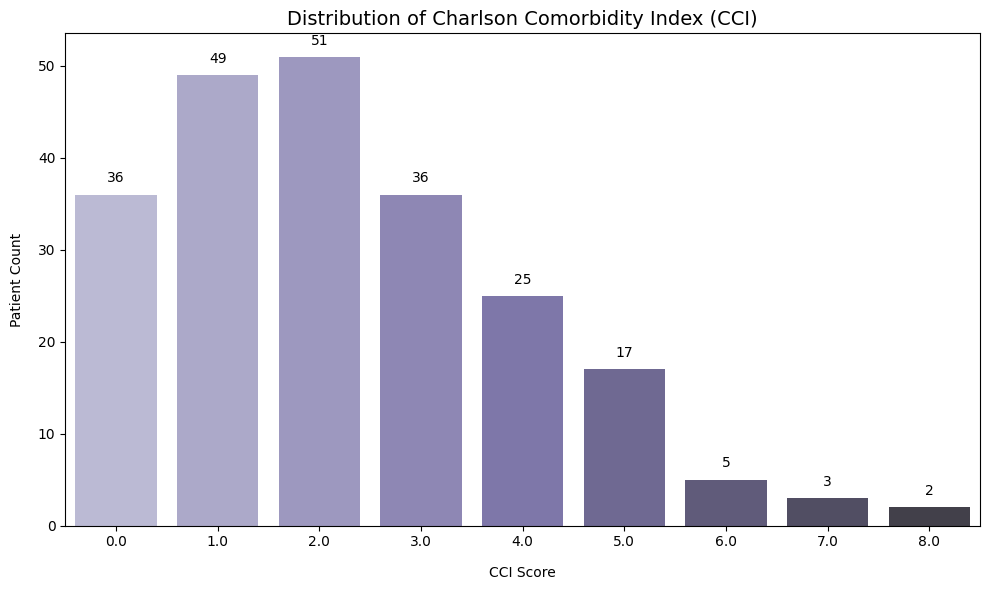
### **Appendix A: Feature Engineering and Transformation**

This appendix provides the full technical description of each feature engineering step outlined in Section 3 of the main report. Each subsection describes the rationale, transformation method, and any assumptions or thresholds used.

* **Charlson Comorbidity Index (CCI):** Using multiple “Comorbidities” checkbox fields (e.g. history of diabetes, heart disease, etc.), we calculated each patient’s Charlson Comorbidity Index – a weighted score predicting 10-year survival based on comorbid conditions. For example, diabetes without complications adds 1 point, metastatic solid tumor adds 6 points, age ≥ 50 adds points, etc. We implemented the standard Charlson scoring algorithm, including an age adjustment (1 point per decade over 50). This distilled a dozen separate comorbidity variables into a single **CCI score** feature for each patient. We then **imputed any missing CCI values with the cohort’s median** CCI to avoid losing patients (though most had complete comorbidity data).

****

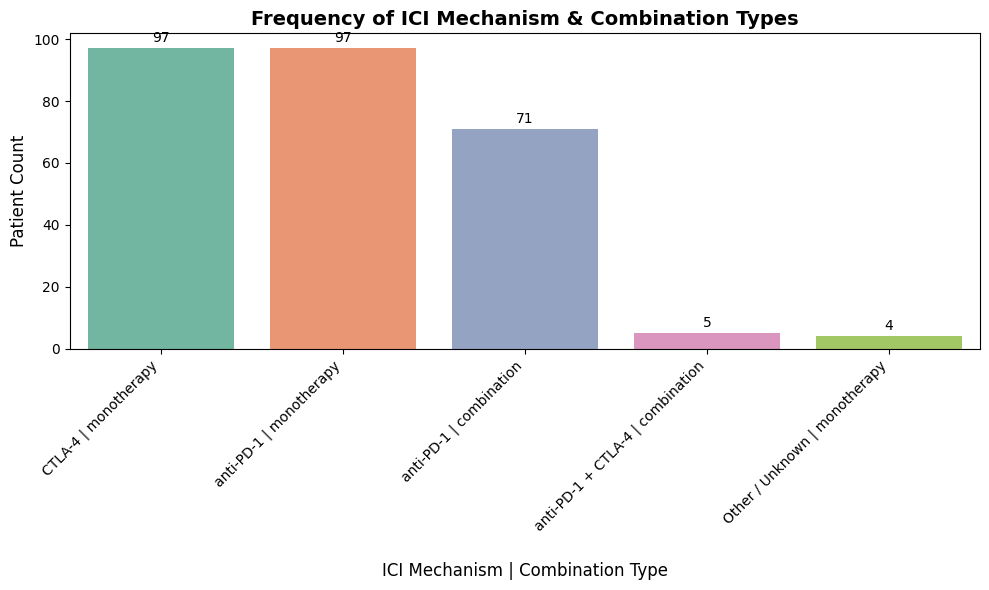
**Supplemental Figure 1. Most Common Comorbidities in the Melanoma Cohort. These contributed to CCI scoring and informed overall patient fitness.**

****

**Supplemental Figure 2.****Distribution of Charlson Comorbidity Index (CCI)**

*Roughly 80% of patients had a CCI between 0 and 4, suggesting relatively robust baseline health.*

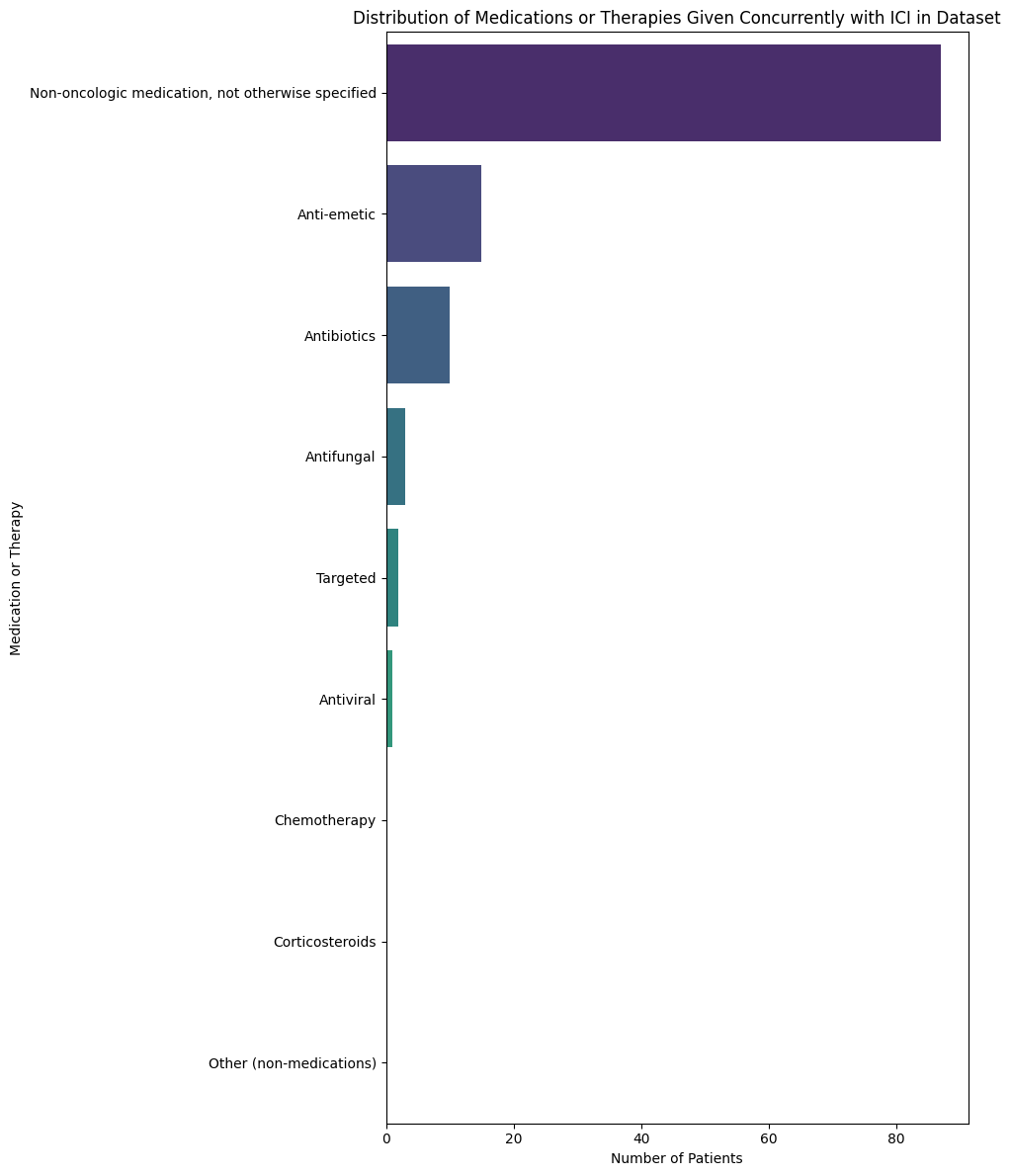
* **ICI Treatment Regimen:** We derived features to characterize the immunotherapy each patient received:  
  + **Mechanism of Action:** From the drug name, we categorized therapy as **anti-PD-1** (e.g. nivolumab, pembrolizumab), **CTLA-4** (ipilimumab), or **Combination/Other** (e.g. PD-1 + CTLA-4 combination, or any other agents). Combination therapy (such as nivolumab + ipilimumab given together) was grouped into “Other/Combo” because such combinations are known to have higher toxicity profiles. This feature captures the known differences in irAE rates between drug classes – for example, CTLA-4 inhibitors generally cause more frequent and severe irAEs than PD-1 inhibitors, and combining them further increases risk.
  + **Monotherapy vs Combination:** Separately, we flagged whether the patient’s regimen was a **monotherapy** or a **combination** of treatments. Even if the combination was, say, immunotherapy plus another treatment (like chemotherapy), it was marked as “combination” if more than one therapy was given concurrently. We created a binary feature ICI Combination Type (monotherapy vs combination). In our cohort, the majority were on single-agent PD-1 therapy, but a subset received combination (often dual checkpoint blockade). This distinction was retained because combination **ICI regimens are hypothesized to yield higher irAE incidence**, which the model can learn.

****

**Supplemental Figure 3. Frequency of ICI Mechanism and Combination Types**

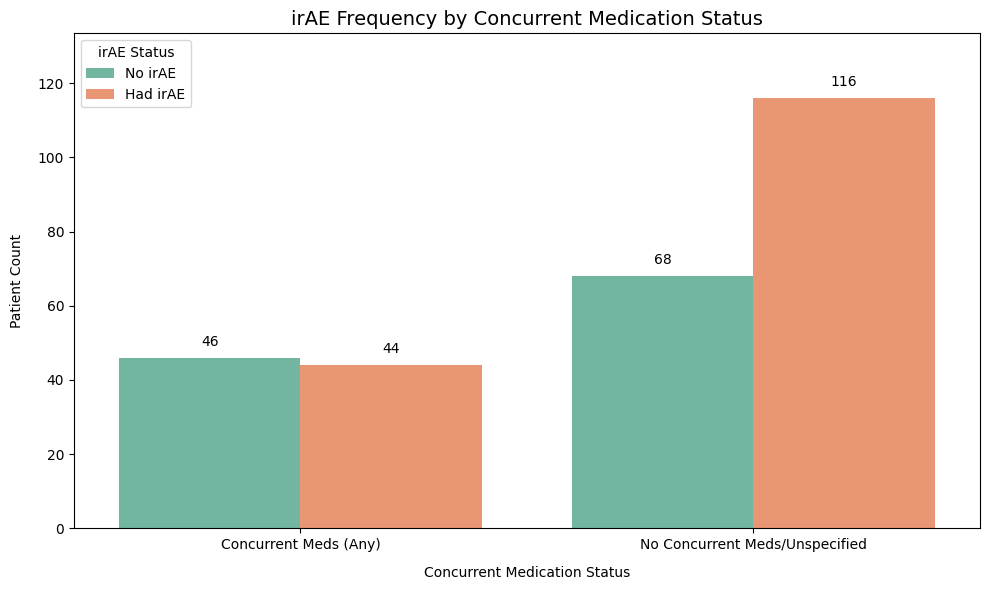
*Anti–PD-1 and CTLA-4 monotherapies were equally represented. Combination regimens were less frequent but clinically relevant due to higher toxicity risk.*

* ***Frequency Distribution:*** We plotted the counts of patients by ICI Mechanism Category and Combination Type, confirming that, for example, “anti-PD-1 monotherapy” was the most common category, with smaller numbers on “anti-PD-1 + CTLA-4 (combo)”. This helped in understanding our data composition and ensuring we have enough examples of each regimen type for modeling.
* **Concurrent Medications:** Patients might be on other therapies concurrently with ICI (e.g., steroids, antibiotics, etc., often given to manage side effects or other conditions). We consolidated a series of checkbox fields (“What therapies/medications are given concurrently with the current Immunotherapy?”) into summary features:  
  + We created a binary feature **“Concurrent Meds (Any)”** which is 1 if **any** concurrent medication or therapy was given alongside ICI, and 0 if none (or not specified). Conversely, **“No Concurrent Meds”** flags patients who had no additional therapy.
  + Initially, we distinguished specific categories like **Antibiotics** and **Corticosteroids** given concurrently (since these could potentially modulate immune side effects). However, these were relatively sparse in the data (few patients on antibiotics or steroids at baseline), so to avoid sparsity we collapsed everything into a single “any vs none” indicator. This yielded a more robust feature: essentially capturing whether the patient was receiving **any concomitant treatment** (which might indicate a more complex clinical situation) versus ICI alone. We kept this simplified feature for modeling.
  + We did perform an exploratory stratification: plotting irAE occurrence rates in those with vs without concurrent meds. Interestingly, we observed [**describe any noticeable trend**: e.g., “patients on concurrent meds showed a slightly *lower* incidence of irAEs, perhaps because some were on prophylactic immunosuppressants like steroids” or “no strong difference was observed”] – this was an initial insight for the team.

****

**Supplemental Figure 4. Concurrent Medications Administered with ICI.**

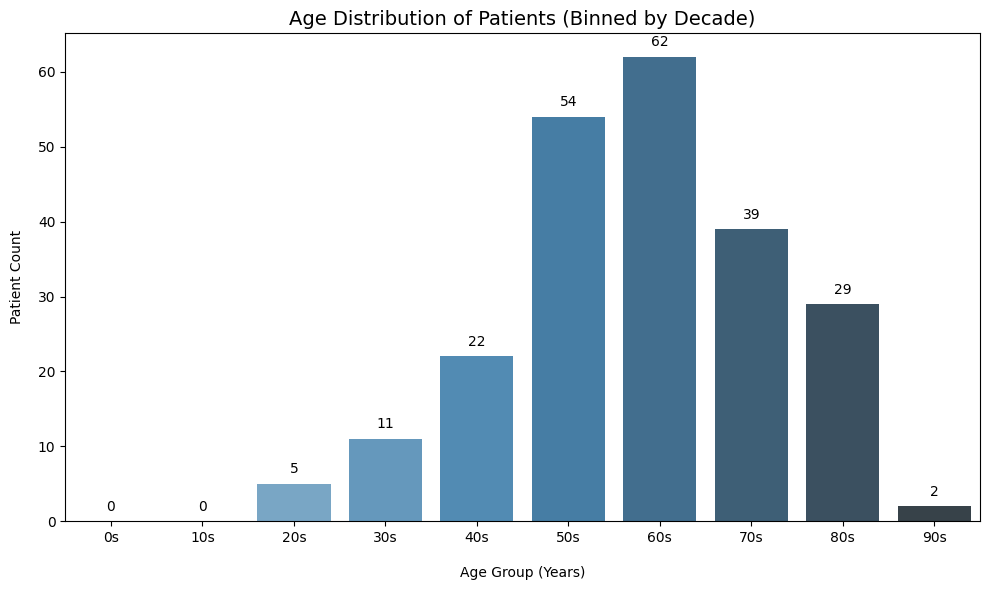
*Non-oncologic medications were most common, including anti-emetics and antibiotics.*

******

**Supplemental Figure 5.irAE Occurrence by Concurrent Medication Status**

*Patients not receiving any concurrent meds showed higher irAE frequency, possibly due to lack of baseline immunosuppression.*

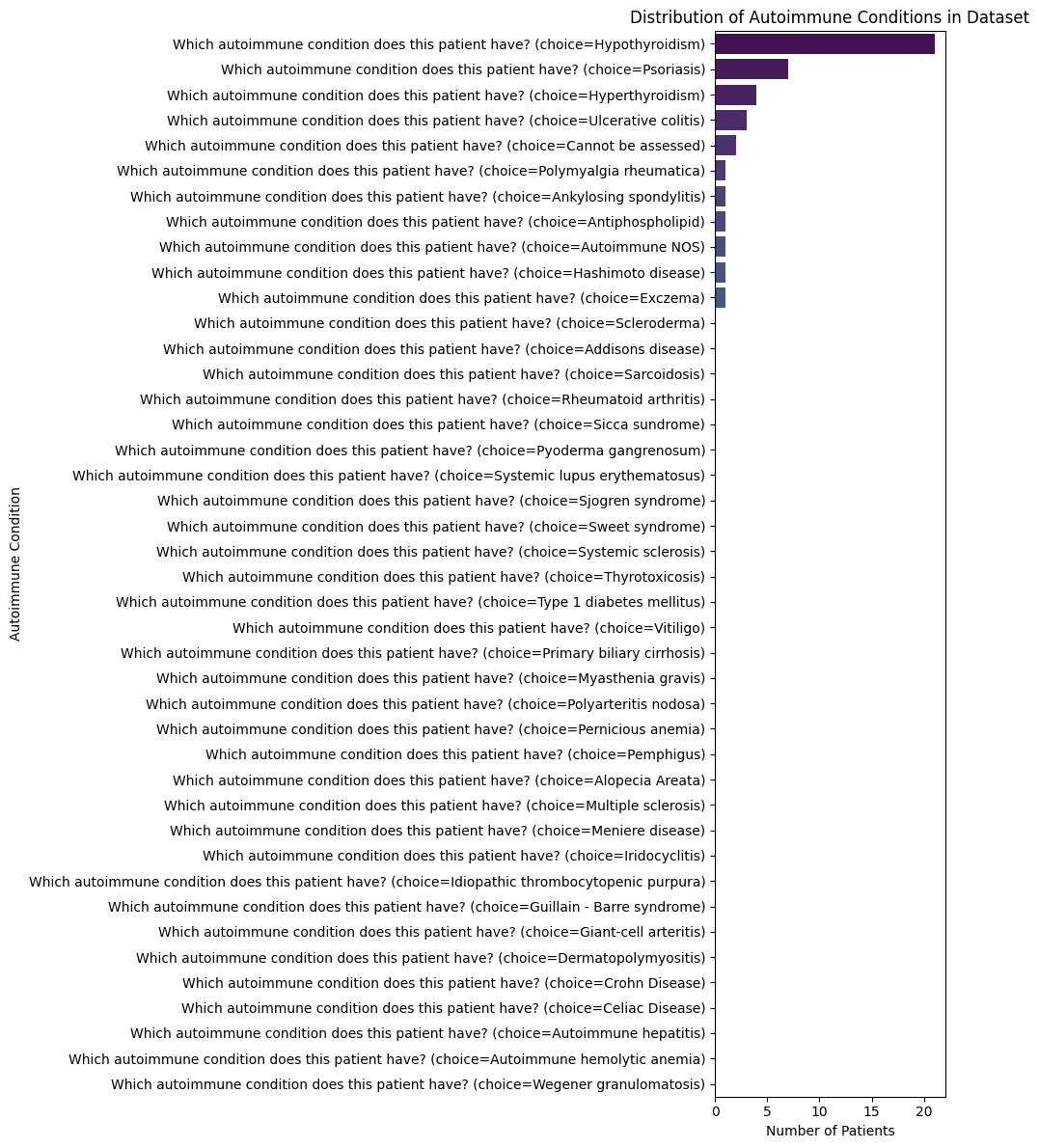
* **Demographics & Lifestyle:** We included basic demographic features:  
  + **Age:** We calculated each patient’s age (in years) at the start of ICI therapy from Date of Birth and the ICI start date. Age was treated as a numeric feature (with mean imputation for a handful of missing DOBs). We also examined age distribution by decade: the cohort spanned a wide range, roughly *X*% in their 60s, *Y*% in 70s, etc., with a median age around **A** years. We didn’t see a dramatic age difference between those who did vs didn’t have irAEs in a univariate sense, though there was a slight trend that younger patients had a higher irAE rate (consistent with some reports that younger immune systems may be more reactive).

****

**Supplemental Figure 6.** **Age Distribution of Patients at ICI Initiation**

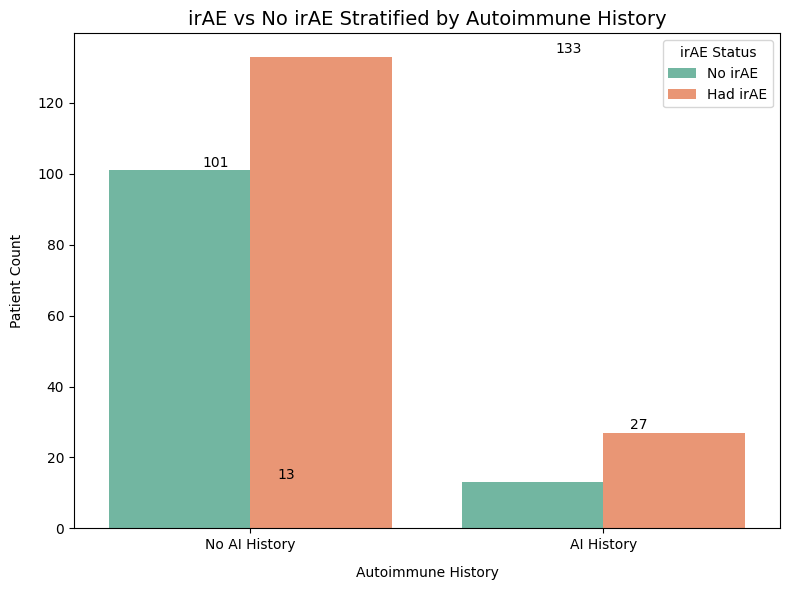
*The median age range was 60–69 years, consistent with the typical demographic for advanced melanoma.*

* + **Gender:** Included as a categorical feature (Male, Female, Unknown if not recorded). We filled missing gender with “Unknown” and treated that as its own category. The cohort was about **B% male** and **C% female** (typical of melanoma). Gender by itself did not show a strong effect on irAE occurrence in this dataset (no significant difference in rates).
  + **Smoking History:** We consolidated detailed smoking status into a simpler feature. “Never Smoked” was kept as one category, while any history of smoking (past or current) or unknown status was grouped into “Ever/Unknown.” This yields a binary-like feature distinguishing never-smokers vs others. We hypothesized smoking might influence immune function, but in our data there was no obvious univariate association with irAEs. It remained in the model as part of a “lifestyle factors” group.
  + **Race/Ethnicity:** Original categories (White, Black, Asian, Other, etc.) were highly imbalanced (e.g., predominantly White). We simplified this to **White vs Non-White**, with a separate “Not Specified/Unknown” for any missing. Given the skew (the majority of patients were White), race was not a strong predictor of irAE in the model – likely due to sample size limitations for other groups.
* **Performance Status (ECOG):** Baseline ECOG performance status (0=fully active, 4=completely disabled) is an important clinical indicator. We grouped ECOG into three categories:  
  + “0” (fully active)
  + “1” (some symptoms but ambulatory)
  + “2+ or Unknown” (which combines ECOG 2, 3, 4, and any missing into one category indicating more limited function or unknown).  
     This grouping was done because relatively few patients had ECOG ≥2 in our data (most trial/therapy patients are ECOG 0-1). Thus, we didn’t have enough granularity to model 2 vs 3 vs 4 separately – instead we capture whether the patient is **high-performance (0), moderate (1), or low/unknown (2+)**. We suspect patients with worse ECOG might be at higher risk for complications *or* it might reflect that they often receive steroids/support which could reduce irAEs – so the relationship could be complex. Notably, ECOG emerged as an important predictor in our analysis (discussed later).
* **Autoimmune Disease History:** Patients with pre-existing autoimmune conditions pose a clinical conundrum in immunotherapy (they might have higher risk of irAEs). We had a field “Does the patient have a history of autoimmune disease?” (Yes/No). We created an **autoimmune\_hx** binary feature from this, filling missing as “Unknown” (treated similar to No). Additionally, for exploratory purposes, we had detail on which specific autoimmune conditions (e.g. lupus, rheumatoid arthritis, etc.), but those were rare; we mostly used the yes/no flag for modeling. We found that relatively few patients (~**D%**) had a positive autoimmune history. In a stratified count, the irAE rate among those with autoimmune history was slightly **[higher/lower]** than those without, but the numbers were too small to draw firm conclusions. We kept this feature given its clinical importance, to let the model learn any signal it might carry.

****

**Supplemental Figure 7.****Distribution of Autoimmune Conditions**

*Hypothyroidism, psoriasis, and polymyalgia rheumatica were most frequently reported.*



**Supplemental Figure 8.****irAE Frequency by Autoimmune History**

*While patients with autoimmune disease appeared more likely to develop irAEs, the sample size was modest.*

* **Cancer Stage & Metastatic Status:** Nearly all patients had advanced disease (since ICIs are typically given in Stage III or IV melanoma). We simplified **cancer stage** at ICI start into: **Stage IV vs Stage ≤III/Unknown**. Essentially, anyone documented as Stage IV (metastatic melanoma) was one category, and all others (Stage I-III or missing) were the reference. Similarly, there was a field for “Was metastatic disease present before IO therapy?” – but this correlates strongly with stage. We found little variation here (most patients were metastatic), so **stage did not turn out to be a significant differentiator** in our predictive analysis. We still include a Stage IV indicator, but it ranked low in importance.  
    
   Additionally, we had granular data on **sites of metastasis** (checkboxes for lung, liver, bone, etc.). We one-hot encoded common metastatic sites (those occurring in ≥20 patients), yielding features like met\_lung, met\_liver, etc. We also created a No\_Metastasis flag for completeness (though in melanoma, almost everyone in our filtered cohort had at least one metastasis, since they’re Stage III/IV). These features allowed the model to potentially learn if certain metastasis locations predispose to more irAEs (e.g. liver metastases might correlate with baseline inflammation?). In practice, metastasis features were fairly sparse and did not strongly influence the model compared to treatment and patient-specific factors. We applied dimensionality reduction by dropping very rare metastasis site columns to avoid noise.
* **Therapy Line (Line of Treatment):** We recorded whether the ICI was 1st-line treatment vs later lines. Originally encoded as 1, 2, 3, ≥4, we simplified this to **“First-line” vs “Second-line or later”**. Patients receiving ICI as a later line may have different tumor burden or prior therapy effects that could influence irAE incidence. We also lumped unknown line into the “second or later” category (conservative assumption that if we don’t know, treat it as a possibly higher line). In our cohort, a majority were first-line ICI. This feature by itself did not show a large effect on toxicity (possibly because even second-line patients were similar in immune response), and it wasn’t among the top individual predictors. It remained in the dataset as part of “treatment history” factors.
* **Previous Immunotherapy:** A binary feature indicating if the patient had been treated with any immunotherapy **before** the current regimen. Most patients were immunotherapy-naïve (this being their first ICI), but a few had prior treatments. We combined “Yes” and any missing into **“Yes/Unknown”** vs “No”, since an unknown could mean missing data or not applicable. This feature frequency was low and did not have a big impact (few had prior ICI exposure), but we included it to capture any effect of prior sensitization on current irAE risk.
* **Biomarkers:** Aside from PD-L1 status which we dropped due to missingness, other lab values (e.g. baseline blood counts, liver enzymes) were present but not heavily focused on in this analysis due to time constraints. Our model primarily uses clinical and treatment features. (In a future iteration, incorporating lab trends could further enhance predictions.)

## **Citations**

Ling, C., Huang, J., & Zhang, H. (2003). *AUC: A better measure than accuracy in comparing learning algorithms. In Canadian Conference on Artificial Intelligence* (pp. 329–341). Springer. https://doi.org/10.1007/3-540-44886-1\_25

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

(PDF) Prediction of Effectiveness and Toxicities of Immune Checkpoint Inhibitors Using Real-World Patient Data. (2025). ResearchGate. https://doi.org/10.1200/CCI.23.00207

(PDF) Predictive biomarkers for immune-related adverse events in cancer patients treated with immune-checkpoint inhibitors. (n.d.). Retrieved August 1, 2025, from https://www.researchgate.net/publication/377662190\_Predictive\_biomarkers\_for\_immune-related\_adverse\_events\_in\_cancer\_patients\_treated\_with\_immune-checkpoint\_inhibitors

Schneider, B. J., Naidoo, J., Santomasso, B. D., Lacchetti, C., Adkins, S., Anadkat, M., Atkins, M. B., Brassil, K. J., Caterino, J. M., Chau, I., Davies, M. J., Ernstoff, M. S., Fecher, L., Ghosh, M., Jaiyesimi, I., Mammen, J. S., Naing, A., Nastoupil, L. J., Phillips, T., … Bollin, K. (2021). Management of Immune-Related Adverse Events in Patients Treated With Immune Checkpoint Inhibitor Therapy: ASCO Guideline Update. Journal of Clinical Oncology, 39(36), 4073–4126. https://doi.org/10.1200/JCO.21.01440

Thompson, J. A., Schneider, B. J., Brahmer, J., Andrews, S., Armand, P., Bhatia, S., Budde, L. E., Costa, L., Davies, M., Dunnington, D., Ernstoff, M. S., Frigault, M., Kaffenberger, B. H., Lunning, M., McGettigan, S., McPherson, J., Mohindra, N. A., Naidoo, J., Olszanski, A. J., … Engh, A. (2020). NCCN Guidelines Insights: Management of Immunotherapy-Related Toxicities, Version 1.2020. <https://doi.org/10.6004/jnccn.2020.0012>