Robust and fair time-to-event framework for predicting cancer-associated Venous Thromboembolism (VTE) using routinely-collected clinical and panel-sequencing data

Intae Moon[1], Hyewon Jeong[1], Alexander Gusev[2,3], and Marzyeh Ghassemi[1]

[1]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Division of Population Sciences, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
[3]The Broad Institute of MIT and Harvard, Cambridge, MA, USA

November 18, 2023

Venous Thromboembolism (VTE) is a common and potentially fatal complication in cancer patients, leading to increased mortality, reduced quality of life, and interruptions in treatment. The widely-used Khorana score assesses VTE risk based on five factors including cancer type and lab results but often fails to reflect the diverse risk profiles due to variations in tumor types, cancer stages, and patient ethnicities. In our work, we aim to improve VTE prediction for a broad and diverse range of cancer patients. To this end, we are integrating a comprehensive set of data, including personalized patient data and genetic panel sequencing data routinely collected at Dana-Farber Cancer Institute (DFCI). We aim to utilize a robust optimization framework that prioritizes minority subgroups, reducing performance disparities. Our integrative and equitable approach seeks to establish new standards and best practices in VTE risk assessment, contributing to the advancement of personalized and precision-oriented cancer care for a wider demographic.

Patient Analysis Cohort

Our study utilized a large cohort from the Dana-Farber Cancer Institute (DFCI), encompassing 16,833 ambulatory cancer patients aged 18-80. These individuals, who received treatment and follow-up at DFCI starting from June 1, 2015, were specifically chosen because they had not experienced any acute VTE episodes in the six months before their treatment. This criterion ensured that we could analyze VTE risk in patients who began their chemotherapy without recent VTE events, which may not typically be classified as high risk by clinical standards.

Heterogeneity of Cancer-Associated VTE Incidence Across Diverse Patient Subgroups

Identifying patients at high risk for cancer-associated VTE is challenging due to the heterogenous incidence rates among different patient subpopulations. To quantify this heterogeneity, we applied the Aalen–Johansen estimator to calculate the Cumulative Incidence Function (CIF) for VTE events, taking into account all-cause mortality as a competing risk, from the commencement of the first treatment regimen. Our findings, illustrated in Figure 1, indicate a striking variance in VTE incidence, with certain cancer types such as pancreatic, cancer of unknown primary (CUP), stomach, and bowel showing notably higher risks. Similarly, black patients and those aged between 70-80 also presented increased VTE risks. These disparities underscore the necessity for personalized risk assessment models that can adapt to the unique profiles of each patient subgroup to improve prevention and treatment strategies.

Association of Panel Sequencing and Clinical Features with cancer-associated VTE

Improving the prediction of VTE across a diverse range of cancer patients necessitates integrating a comprehensive set of data, including detailed clinical information and genetic panel sequencing data. This holistic approach allows for a nuanced understanding of how various factors contribute to VTE risk. In our detailed analysis, we utilized univariate cause-specific Cox Proportional Hazard regression for panel sequencing somatic variant features and clinical features. For the initial treatment regimens received by patients in the cohort, we conducted multivariable analysis, adjusting for known cancer groups and features of the Khorana score, to account for cancer-specific effects and established clinical knowledge of VTE risk. Through iteration of the Cox model, we calculated hazard ratio—a metric that quantifies the influence of each feature on the risk of developing VTE over time—and p-values for each variable. As illustrated in Figure 2, our analysis showed statistically significant associations between VTE incidence and somatic mutations in genes such as TP53 (Hazard ratio: 0.212, p-value: 6.86×10^{-13}) and KRAS (Hazard ratio: 0.342, p-value: 5.18×10^{-11}), along with Copy Number deletions in NEIL2 (Hazard ratio: 0.298, p-value: 5.50×10^{-7}) and GATA4 (Hazard ratio: 0.263, p-value: 8.02×10^{-7}). Clinically, being in the pancreatic cancer group and having elevated levels of alkaline phosphatase (LAB_ALKP, Hazard ratio: 0.191, p-value: 9.18×10^{-19}), absolute neutrophil count (LAB_ANEU, Hazard ratio: 0.183, p-value: 5.41×10^{-15}), and platelets (LAB_PLT, Hazard ratio: 0.194, p-value: 1.06×10^{-14}) were identified as notable risk factors, while a higher albumin level (LAB_ALB, Hazard ratio: -0.293, p-value: 4.38×10^{-30}) offered a protective effect. Moreover, treatment regimens that include fluorouracil, irinotecan, and leucovorin (FLUORO/IRINOT/LEUCOV) — a combination typically employed in colorectal cancer chemotherapy — as well as bevacizumab-based therapies — frequently used to treat cancers such as colorectal, lung, and kidney cancers — were associated with an increased risk of VTE (Hazard ratio: 0.669, p-value: 3.87×10^{-9}, and Hazard ratio: 0.716, p-value: 1.18×10^{-10}, respectively). These significant findings inform the development of a predictive model inclusive of a broad range of genetic and clinical parameters, marking a significant step toward refining VTE risk stratification and enhancing personalized treatment strategies in oncology.

Prediction of Time-to-Cancer Associated VTE

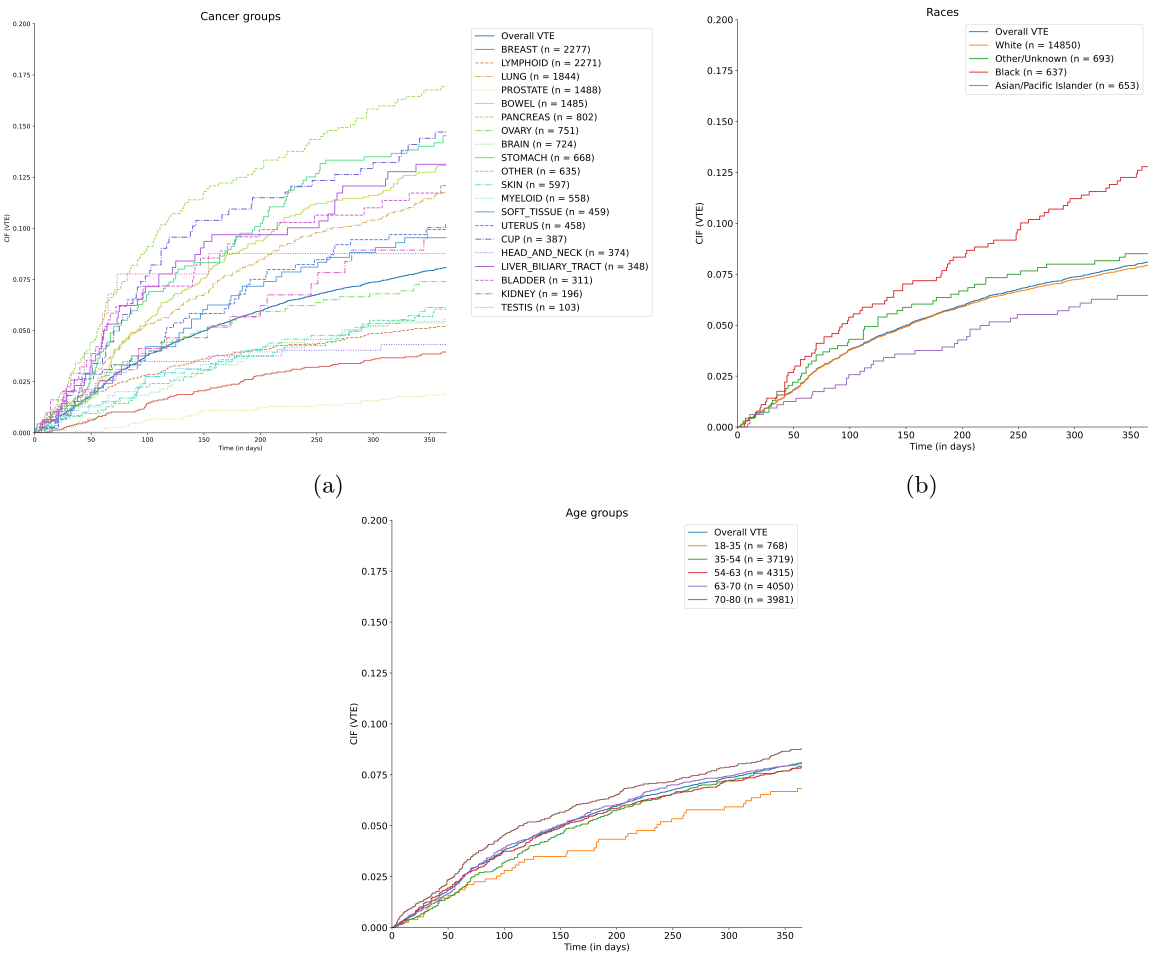
In efforts to improve predictive modeling for VTE, the heterogeneity across patient demographics requires moving beyond a one-size-fits-all approach. Therefore, we investigated a variety of model configurations. Our analysis utilized time-to-event models, including Cox Proportional Hazards (CPH) and DeepSurv—a nonlinear version of the Cox model designed to capture complex interactions among features. These models were evaluated against the performance of the Khorana score, the current clinical standard for predicting VTE risk. Our investigation highlighted the differences between a generic feature set and a personalized set that incorporated all clinical and treatment data, including factors like cancer type, age, ethnicity, and sex. Although the personalized set improved overall model performance—reflected in a higher mean Area Under the Curve (AUC), a measure of model accuracy—it did not consistently benefit all patient subgroups, as evidenced in Figure 3. Specifically, the personalized set resulted in a reduction in prediction accuracy for groups such as those with pancreatic and soft tissue cancers, and among black patients in both CPH and DeepSurv models. These findings are now steering us towards the development of nuanced modeling strategies. We aim to construct a multi-faceted framework that addresses the limitations inherent in a single model configuration, ensuring more accurate and equitable VTE risk predictions across the diverse patient population.

Ongoing and Future Work

Building upon our ongoing research, we are actively integrating state-of-the-art optimization and genetic risk assessment strategies to enhance the accuracy of VTE risk prediction models. To mitigate performance disparities, we utilize frameworks like Distributionally Robust Optimization (DRO), designed to prioritize minority subgroups who often suffer from such disparities. Concurrently, we are incorporating Polygenic Risk Scores (PRS) to account for germline contributions to VTE risk, providing potentially unique predictive insights. This is particularly beneficial for demographics such as younger patients at elevated VTE risk, who traditionally are not well-represented in existing models.

With the understanding that no single model can adequately serve all, we are currently developing a framework that aggregates models based on specific subgroup performances. This endeavor not only strives to advance beyond existing VTE risk prediction methods including Khorana Score but also to better meet the varied needs of a diverse cancer patient population, reinforcing our commitment to inclusivity and precision in oncological care.

Figures

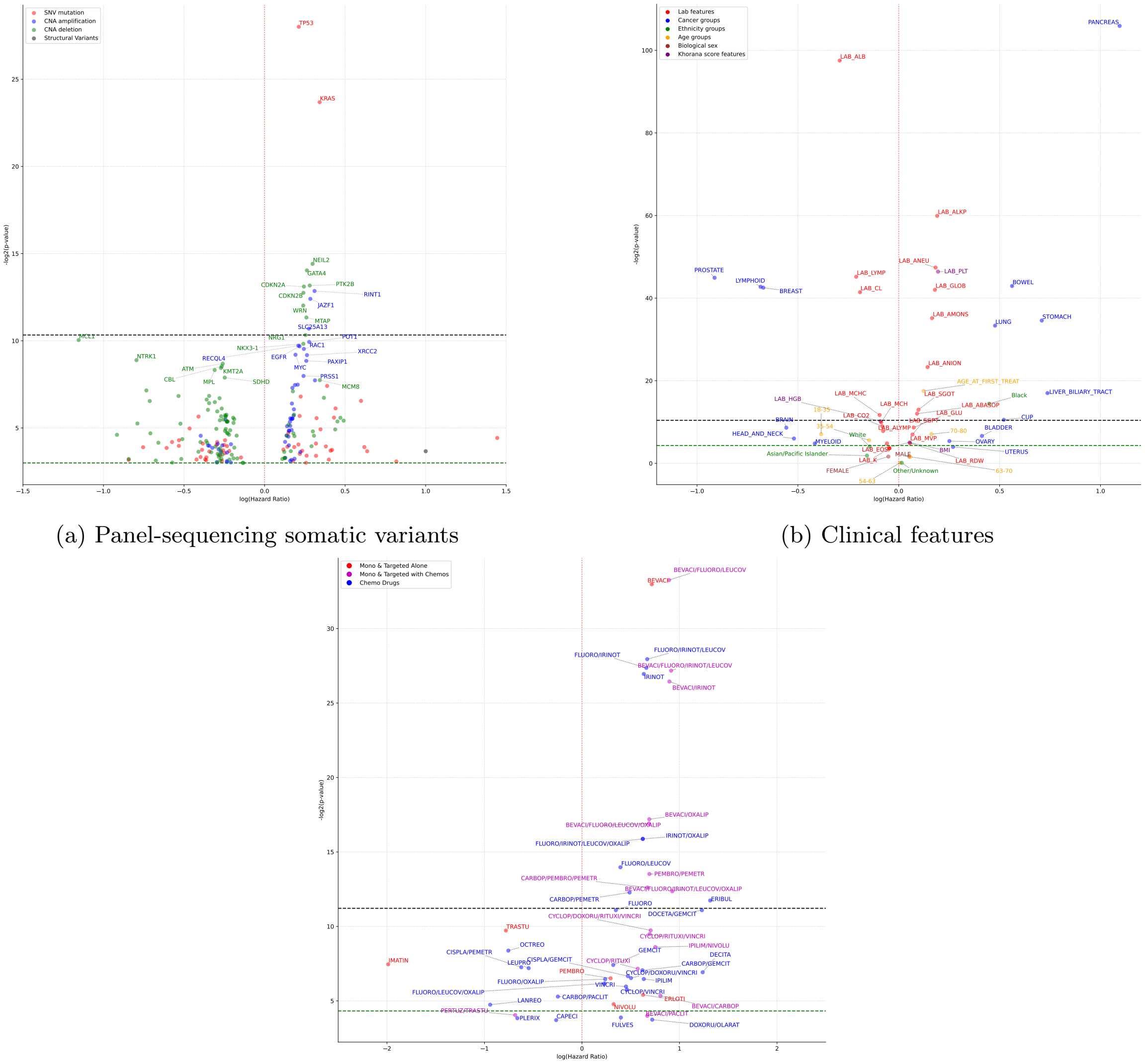


(a)

(b)

(c)

Figure 1: Heterogeneity of Venous Thromboembolism (VTE) Incidence in Cancer Patients. (a) Cumulative incidence of VTE by cancer type. Each line represents a different type of cancer, with the number of patients in each group provided in parentheses. The incidence of VTE varies significantly by cancer type, with pancreatic, cancer of unknown primary (CUP), stomach, and bowel cancers showing the highest cumulative incidence over time. (b) Cumulative incidence of VTE by race. The graph illustrates a higher incidence in black patients compared to other racial groups. (c) Cumulative incidence of VTE by age group. It shows that patients aged between 70-80 have a higher cumulative incidence.

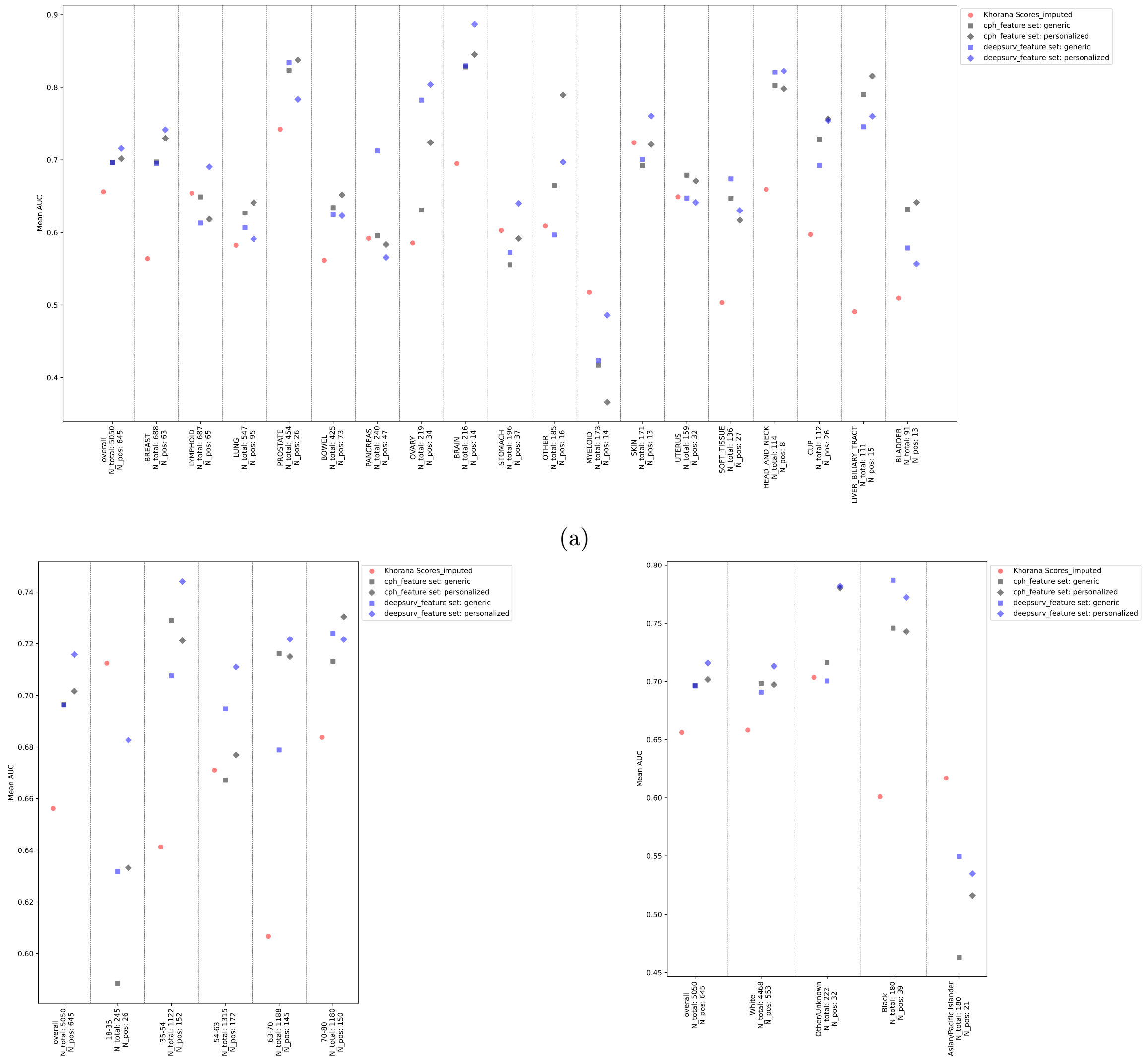


(a) Panel-sequencing somatic variants

(b) Clinical features

(c) First treatment regimens

Figure 2: Associations of Somatic Variants (a), Clinical Features (b), and Treatment Regimens (c) with VTE Incidence. The dotted green line signifies nominal significance (p-value < 0.05), and the black dotted line indicates the Bonferroni-corrected significance level. The x-axis represents the log(hazard ratio), and the y-axis corresponds to -log(p-value), highlighting the strength and significance of each association.



(a)

(b)

(c)

Figure 3: Comparison of Model Performance Across Patient Subgroups. (a) Model accuracy for different cancer types. This dot plot compares the mean Area Under the Curve (AUC) for various cancer types using different feature sets in predictive models. (b) Model accuracy by race. The plot illustrates the AUC for racial groups, comparing the performance of generic versus personalized feature sets in prediction models. (c) Model accuracy by age group.