Utilizing Electronic Health Records (EHR) and Tumor Panel Sequencing to Demystify Prognosis of Cancer of Unknown Primary (CUP) patients

Intae Moon[1,3,*], **Sylvan C. Baca**[2, 5], **Kenneth L. Kehl**[3], **and Alexander Gusev**[3, 4, 5]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
[3]Division of Population Sciences, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
[4]Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[5]The Broad Institute of MIT & Harvard, Cambridge, MA, USA
[*]itmoon@mit.edu

May 19, 2022

Background

When a standardized diagnostic test fails to locate the primary site of a metastatic cancer, it is diagnosed as a cancer of unknown primary (CUP). This type of cancer represents about 3-5% of all cancers[1]. Due to the hidden nature of primary sites for CUP tumors, oncologists are often forced to resort to empiric treatments and patients with CUPs typically have very poor outcomes[2]. Therefore, there is great need for an accurate method to identify the primary site of CUP and empower more informed clinical decision making.

Methods

We used the next generation sequencing (NGS) data collected at three different cancer centers in routine clinical care as part of the AACR project GENIE[3]: Dana-Farber Cancer Institute (DFCI, n=18,013), Memorial Sloan Kettering Cancer (MSK, n=16,294) center, and Vanderbilt-Ingram Cancer Center (VICC, n=1,335). We trained an XGBoost[4] model to classify 22 OncoTree-based cancer types (training tumor samples: n = 28,353, 10-fold cross-validation; test tumor samples: n = 7,289), using molecular features abstracted from the NGS data. To identify important features for each predicted cancer types, we used the recently proposed feature interpretation tool for tree-based models, called TreeExplainer[5]. We evaluated median survival of patients with CUP (n = 838) as well as metastatic Cancer with Known Primary (CKP) counterparts (n = 8,373) using the Kaplan-Meier method. Finally, among 131 patients with CUP who received first-line palliative treatment at DFCI, we evaluated whether the concordance between predicted cancer type and disease center was associated with improved survival using a multivariable Cox Proportional Hazard regression.

Result

Our classifier achieved high performance on held-out test data consisting of 7,289 primary and metastatic tumor samples from 22 known cancer types (weighted F1 : 0.789) as shown in Fig. 1a, 1b. The classifier performance was robust to factors including cancer center, tumor sample type, sequencing panel version, and patient ethnicity (Fig. 1c). We identified shared genetic features between CUP tumor samples and their CKP counterparts across predicted primary cancer types, using the recently proposed model interpretation method SHAP (Fig. 2a, 2b). Additionally, we provided a local explanation for a primary site prediction of each CUP tumor sample (Fig. 2c). Applying our predictive algorithm to 838 patients with CUP, we identified subtypes with significant prognostic differences (Fig. 3a), which were also consistent with relative median survival in their corresponding CKPs (Spearman's rho 0.810, p-value : 0.015; Fig. 3b). Importantly, patients with CUP that were assigned to disease centers concordant with their predicted primary sites showed better 5-year survival than those assigned to discordant disease centers (H.R. 0.63, 95% C.I. 0.45 - 0.83, p-value : 0.007) (See Fig. 3c).

Conclusion

We demonstrated accurate primary site classification from routinely collected, multi-center NGS panels. Classification predictions stratified patients with CUP into significantly different survival groups, and disease center assignments consistent with the predictions were predictive of improved survival. Our classifier offers interpretable predictions and the potential for meaningful clinical decision support for managing patients with CUP.

References

1. Qaseem, A., Usman, N., Jayaraj, J. S., Janapala, R. N. & Kashif, T. Cancer of Unknown Primary: A Review on Clinical Guidelines in the Development and Targeted Management of Patients with the Unknown Primary Site. *Cureus* **11**, e5552 (2019).

2. Bochtler, T. & Krämer, A. Does cancer of unknown primary (cup) truly exist as a distinct cancer entity? *Front. Oncol.* **9**, 402, DOI: 10.3389/fonc.2019.00402 (2019).

3. Aacr project genie: Powering precision medicine through an international consortium. *Cancer Discov.* **7**, 818–831, DOI: 10.1158/2159-8290.CD-17-0151 (2017). https://cancerdiscovery.aacrjournals.org/content/7/8/818.full.pdf.

4. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794, DOI: 10.1145/2939672.2939785 (ACM, New York, NY, USA, 2016).

5. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67, DOI: 10.1038/s42256-019-0138-9 (2020).

Figures

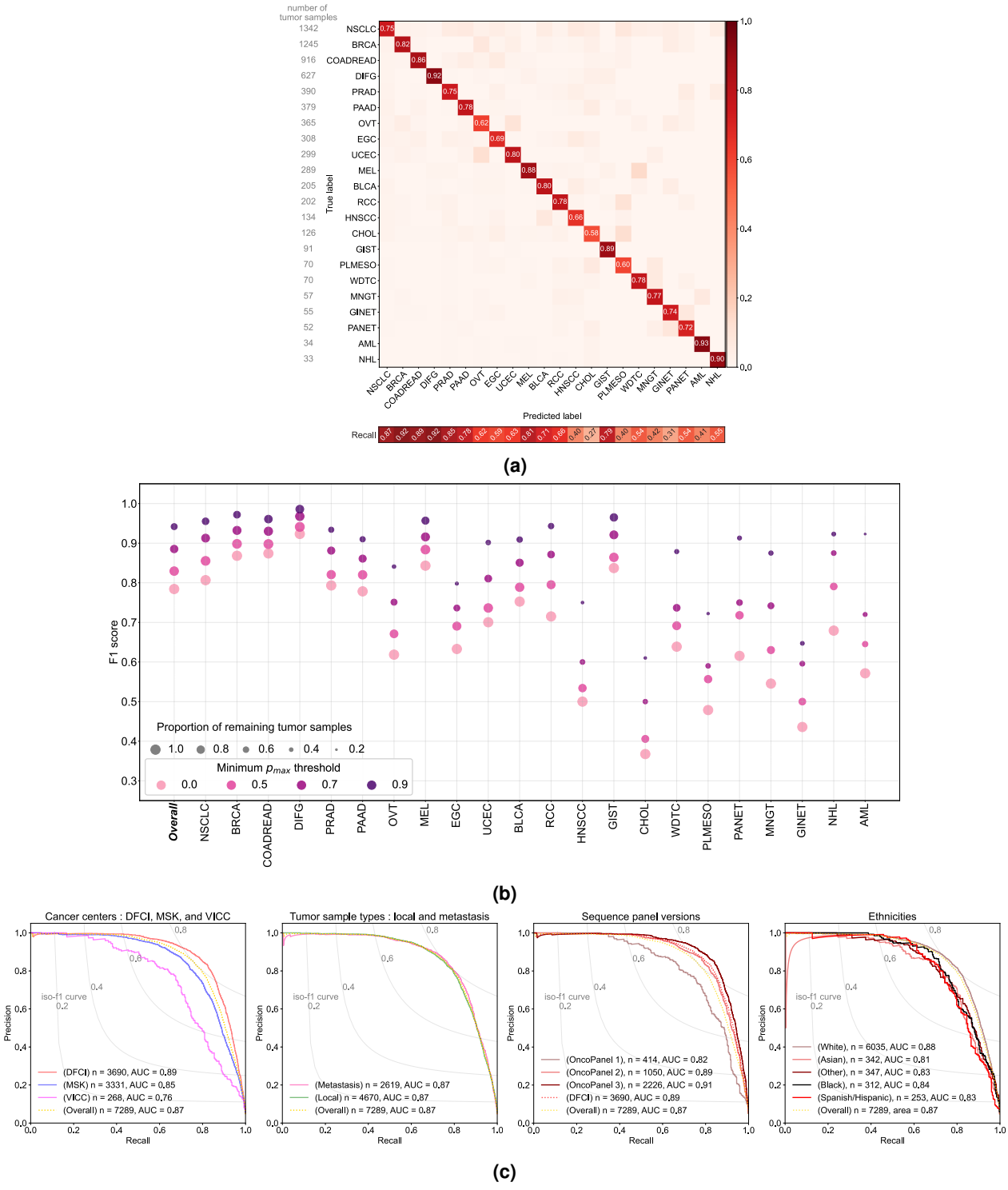
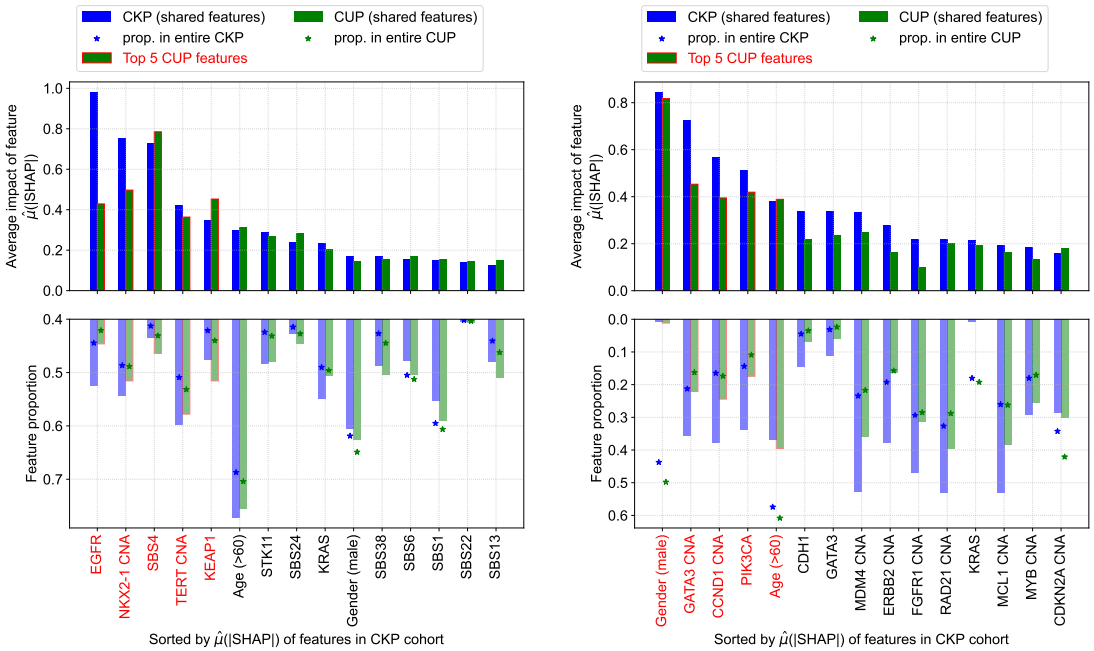
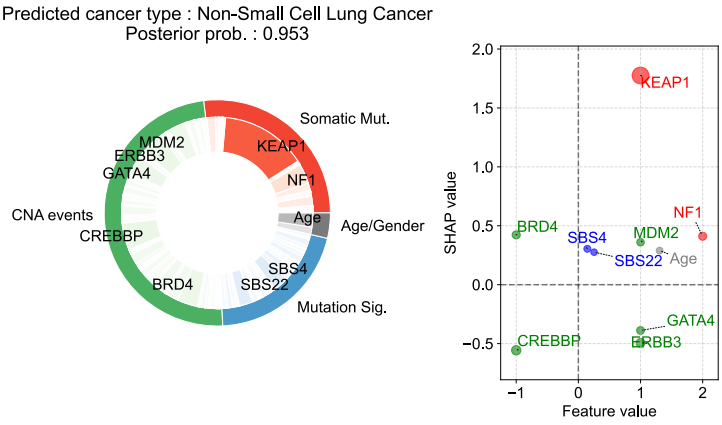


Figure 1. **(a)** The normalized confusion matrix of the classifier model performance on the test set (n = 7,289) across 22 different cancer types. Each diagonal entry of the matrices indicated the precision for each cancer type prediction. A recall for each cancer type is shown under the confusion matrix. Finally, the number of tumor samples in each label is shown next to the label. **(b)** The performance of the classifier model on the test set across cancer types at different minimum p_{max} thresholds (0.0, 0.5, 0.7, and 0.9). The quality of performance is captured in F1 score along the y-axis, which shows the overall model performance as well as label-specific performance along the x-axis. p_{max} is the maximum posterior probability for each cancer type prediction and reflects the confidence of the classification. Finally, each dot size is scaled by the proportion of remaining tumor samples in each label after applying p_{max} threshold. **(c)** Precision-recall curves showing the performance of the classifier model across the subgroups formed by different factors including cancer center, tumor sample type, sequence panel version, and ethnicity. The yellow dotted curve corresponds to the baseline performance of the model on the overall test set.



(a) Non-Small Cell Lung Cancer (NSCLC)

(b) Invasive Breast Carcinoma (BRCA)



(c)

Figure 2. Top 15 most important input features sorted by their $\hat{\mu}(|SHAP|)$ in CUP-CKP cohorts across **(a)** Non-Small Cell Lung Cancer (NSCLC) and **(b)** Invasive Breast Carcinoma (BRCA). Blue bars correspond to $\hat{\mu}(|SHAP|)$ of those features in the CKP cohort, and green bars correspond to their $\hat{\mu}(|SHAP|)$ in the predicted CUP cohort. Top 5 most important features in a CUP cohort of each predicted cancer type are shown in red. Blue and green bars going downwards show the proportion of each feature in CUP and CKP cohorts, respectively. Finally, the blue and green star-shaped dots show the proportions of each feature across entire CKP tumor samples and CUP tumor samples, respectively. **(c)** Local explanation for a tumor sample collected from a 78 year-old, female patient diagnosed with CUP. (Left) The two-level pie chart where combined impact of features in each feature type is shown in the outer ring and impact of each feature in each feature type is shown in the inner ring. Features with higher impacts, hence larger sections of the pie, have a darker color. Finally, we label top 10 highest impact features based on $|SHAP|$. (Right) The scatter plot shows each of the important features as a dot where its size is scaled by its impact on the model outcome (i.e. $|SHAP|$) and its color is chosen based on a type of the feature.

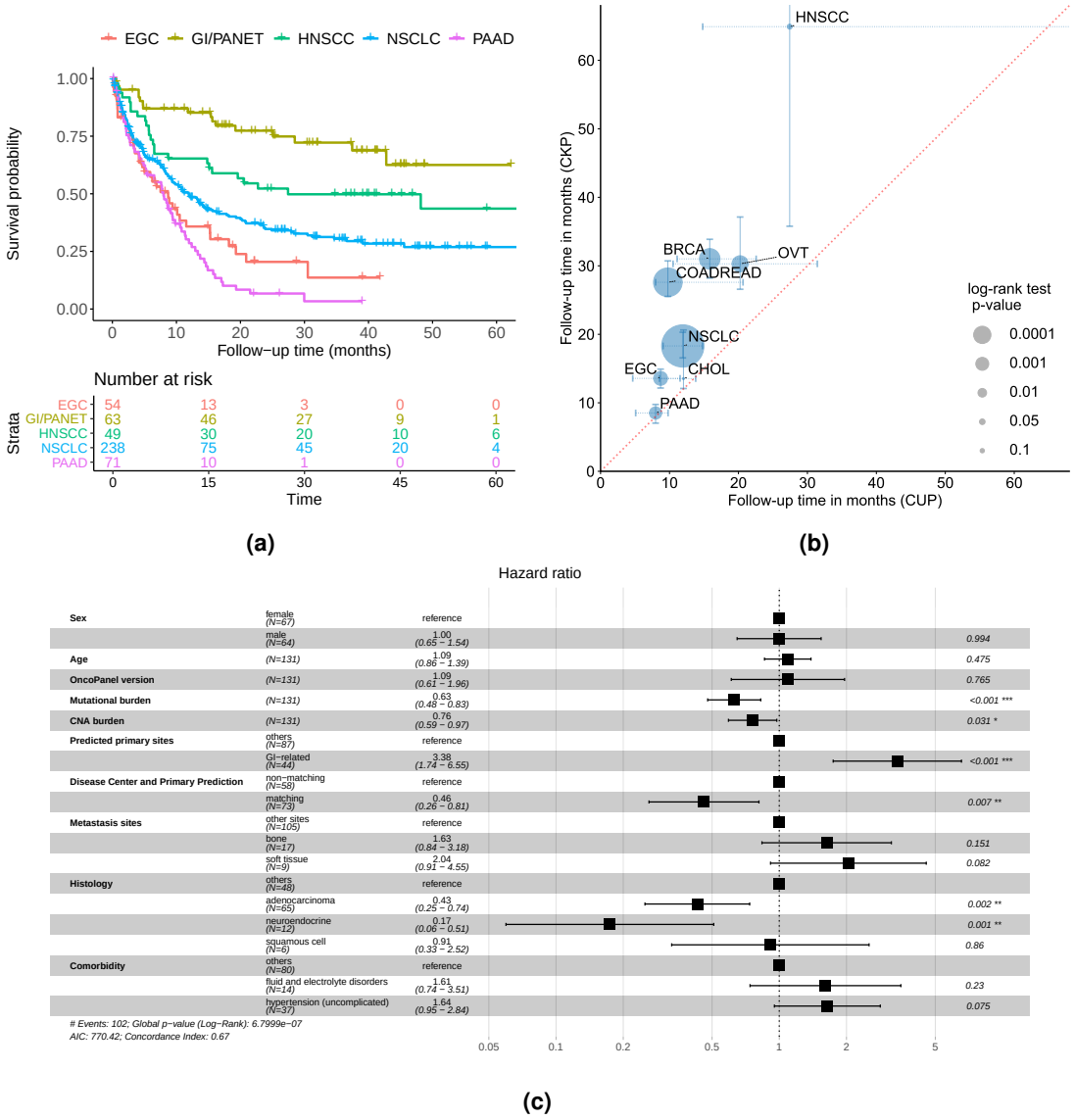


Figure 3. **(a)** Risk stratification of patients with CUP based on their predicted cancer types. We used the Kaplan-Meier method to estimate survival function for each predicted cancer type over the follow-up time of 60 months from OncoPanel sequence date. **(b)** The plot shows correlation between median survival time (in months) of CUP predicted cancer types (x-axis) and those of CKP cancer types (y-axis): Spearman's rho 0.810 (p-value : 0.0149). Note that cancer types with at least 30 CUP tumor samples were chosen. The size of each dot is adjusted by a p-value of the log-rank test which quantifies how significantly different survival functions are between CUP-CKP pairs. **(c)** The summary of multivariable Cox Regression on patients in the CUP cohort with first-line palliative treatment records available (n = 131). The binary indicator, Disease Center and Primary Prediction, of whether an assigned disease center and predicted primary cancer type are concordant, is significantly associated with a 5-year mortality of patients in the cohort (H.R. 0.46, 95% C.I. 0.28 - 0.81, p-value : 0.007).