Computational and Structural Biotechnology Journal 21 (2023) 3865–3874

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Primary tumor type prediction based on US nationwide genomic profiling data in 13,522 patients

Yunru Huang [1], Shannon M. Pfeiffer [1,2], Qing Zhang [*]

Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, United States

ARTICLE INFO

Keywords:
Primary tumor type
Cancer diagnosis
Predictive model
Machine learning
Real-world data
Clinico-genomic database

ABSTRACT

Timely and accurate primary tumor diagnosis is critical, and misdiagnoses and delays may cause undue health and economic burden. To predict primary tumor types based on genomics data from a de-identified US nationwide clinico-genomic database (CGDB), the XGBoost-based Clinico-Genomic Machine Learning Model (XC-GeM) was developed to predict 13 primary tumor types based on data from 12,060 patients in the CGDB, derived from routine clinical comprehensive genomic profiling (CGP) testing and chart-confirmed electronic health records (EHRs). The SHapley Additive exPlanations method was used to interpret model predictions. XC-GeM reached an outstanding area under the curve (AUC) of 0.965 and Matthew's correlation coefficient (MCC) of 0.742 in the holdout validation dataset. In the independent validation cohort of 955 patients, XC-GeM reached 0.954 AUC and 0.733 MCC and made correct predictions in 77% of non-small cell lung cancer (NSCLC), 86% of colorectal cancer, and 84% of breast cancer patients. Top predictors for the overall model (e.g. tumor mutational burden (TMB), gender, and *KRAS* alteration), and for specific tumor types (e.g., TMB and *EGFR* alteration for NSCLC) were supported by published studies. XC-GeM also achieved an excellent AUC of 0.880 and positive MCC of 0.540 in 507 patients with missing primary diagnosis. XC-GeM is the first algorithm to predict primary tumor type using US nationwide data from routine CGP testing and chart-confirmed EHRs, showing promising performance. It may enhance the accuracy and efficiency of cancer diagnoses, enabling more timely treatment choices and potentially leading to better outcomes.

1. Introduction

Cancer is the leading cause of mortality worldwide, accounting for approximately 10 million deaths in 2020 [1]. While many landmark discoveries and efforts have been made against cancers with tumor-agnostic approaches, such as entrectinib, larotrectinib and pembrolizumab, the majority of treatment options, including targeted therapies, remain approved in specific tumor types. Oncologist training is also closely associated with the tissue of origin. However, locating the origin of a tumor can still be time-consuming and challenging, despite applying comprehensive clinical and diagnostic work-ups. The median duration of the diagnostic interval for different tumor types varies from 1 to 8 weeks through traditional diagnostic analyses [2], such as histological criteria and immunohistochemical stains, and the diagnostic error rate is between 2.4% and 22.5% across different tumor types [3]. Meanwhile, some patients may even be unable to obtain diagnoses. Inefficiency, misdiagnoses, or missed diagnoses may negatively impact therapy decisions, costs, and outcomes [4,5].

Abbreviations: AUC, area under the curve; BC, breast cancer; CGDB, clinico-genomic database; CGP, comprehensive genomic profiling; CI, confidence interval; CN, copy number; CM, Clinical Modification; CRC, colorectal cancer; CUP, cancer of unknown primary; DNA, deoxyribonucleic acid; EHR, electronic health records; FH, Flatiron Health; FMI, Foundation Medicine, Inc.; GC, gastric and esophageal cancer; HCC, hepatocellular carcinoma; HN, head and neck cancer; HR, hazard ratio; ICD, International Classification of Diseases; MCC, Matthew's correlation coefficient; ML, advanced melanoma; MPC, metastatic prostate cancer; MSI, microsatellite instability; MSK-IMPACT, Memorial Sloan Kettering–Integrated Mutation Profiling of Actionable Cancer Targets; MSS, microsatellite stable; NSCLC, non-small cell lung cancer; OC, ovarian cancer; OS, Overall Survival; PANC, pancreatic cancer; RCC, renal cell carcinoma; SCLC, small cell lung cancer; SHAP, SHapley Additive exPlanations; SoC, standard of care; SV, short variant; TMB, tumor mutational burden; UC, advanced urothelial cancer; XC-GeM, XGBoost-based Clinico-Genomic Machine Learning Model.

* Corresponding author.
E-mail address: zhangq47@gene.com (Q. Zhang).
[1] These authors contributed equally to this work.
[2] Currently at the University of California, Irvine

https://doi.org/10.1016/j.csbj.2023.07.036
Received 26 July 2022; Received in revised form 16 July 2023; Accepted 25 July 2023
Available online 26 July 2023

Apart from enhancing conventional cancer diagnostic methods, efforts have been made to utilize molecular alterations that are indicative of a tumor's origin to improve diagnostic accuracy and efficiency. Many known alterations are more prevalent in specific tissues than others [6]. Tissue-specific genomic alterations are generally maintained after metastasis, making it possible to predict the primary tumor type from a metastatic sample based on genomic information [7]. Publicly available or research-use cancer genomics and transcriptomics data have been used to identify primary tumor type [8]. However, their clinical usage remains limited because transcriptome or whole-genome sequencing is not routinely performed in clinical care. An algorithm was published in 2020 using a clinical sequencing test, the Memorial Sloan Kettering–Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) test [9], but the test is only available at Memorial Sloan Kettering Cancer Center. No algorithms based on a US nationwide comprehensive genomic profiling (CGP) test performed in routine clinical practice have been developed yet.

In this study, we used a US nationwide de-identified clinico-genomic database (CGDB), derived from Foundation Medicine, Inc. (FMI) CGP testing linked with Flatiron Health (FH) electronic health records (EHRs), to develop a novel XGBoost-based Clinico-Genomic Machine Learning Model (XC-GeM) to predict the primary site of tumors. Model performance was also assessed in patients who initially had missing primary tumor diagnoses at the time of CGP, evaluating the utility of XC-GeM in patients with unknown diagnoses at some point. Since CGP testing is increasingly performed in clinical practice according to the standard of care (SoC), XC-GeM may also be used as part of routine care without requiring additional tissue specimens or testing. If clinically validated, application of this algorithm may act as an assistive tool for effective and efficient diagnoses, potentially contributing to better treatment choices and outcomes, especially in patients with tumors whose primary origin is challenging to diagnose using traditional methods.

2. Methods and theory

2.1. Data source

The population used in this study was obtained from the FH-FMI CGDB, a US nationwide de-identified retrospective longitudinal cancer database [10]. Clinical data were curated from multiple sources of EHRs across over 280 US-based cancer centers in the FH network. Structured data (e.g., demographic and diagnosis codes) were harmonized to a standard data model. Specific information from unstructured data (e.g., free text in the physician's note) was collected and curated by technology-enabled abstraction and trained human chart reviewers. For genomic data, 322 clinically relevant cancer genes were included from the validated FoundationOne assay, a tissue-based CGP service for all solid tumors increasingly performed in clinical practice according to SoC [11]. The CGDB is updated every three months. Institutional review board approval of the study protocol was obtained before study conduct and included a waiver of informed consent.

2.2. Study population

The study population comprised patients in the CGDB aged 18 years or older who were diagnosed with only one tumor type after January 1, 2011, limited to patients with solid tumors whose samples were tested with the FoundationOne assay and passed a quality control check performed by FMI.

All patients in the CGDB have primary tumor types abstracted and curated from FH EHRs. However, for some cases, at the time of the CGP test, limited context is available when FMI receives a specimen. Therefore, some patients could have an unknown histopathology at the time of CGP testing. The primary cohort was derived from the CGDB with a data cut from September 30, 2019 and contained patients with known primary tumor types at the time of the CGP test ("patients with known primary"). New patients with known primary from a more recent data cut of the CGDB (June 30, 2020) were included as an independent validation cohort to evaluate model generalization. We further assessed the model's performance in another cohort of patients with missing primary tumor diagnoses at the time of the CGP test ("patients with missing primary"). Patients in this cohort are not considered as patients with cancer of unknown primary (CUP) and eventually obtained primary tumor diagnoses based on the FH EHRs (Supplemental Fig. 1).

2.3. Main outcomes

The main outcomes were 13 primary solid tumor types from the FH EHRs with disease-specific databases in CGDB: advanced urothelial cancer (UC), breast cancer (BC), colorectal cancer (CRC), gastric and esophageal cancer (GC), hepatocellular carcinoma (HCC), head and neck cancer (HN), advanced melanoma (ML), metastatic prostate cancer (MPC), non-small cell lung cancer (NSCLC), ovarian cancer (OC), pancreatic cancer (PANC), renal cell carcinoma (RCC), and small cell lung cancer (SCLC). Pleural mesothelioma also has a disease-specific database in the CGDB; however, it was not included in the study because it did not have any cases with missing primary diagnoses at the time of CGP testing. International Classification of Diseases (ICD)−9-Clinical Modification (CM) or ICD-10-CM were collected by FH to categorize the primary tumor diagnosis, which was further confirmed through chart review of clinical and pathology notes.

2.4. Predictors

Candidate clinical and genomic predictors were selected based on prior research and investigations into their contribution to differentiating tumor types (Supplemental Table 7). Clinical variables were age at diagnosis (years), sex (categorized as either "female" or "male"), race (categorized as either "white" or "non-white"), stage at initial diagnosis (categorized as either "0-II", "III", "IV", or "occult and unknown") and body mass index (BMI, categorized as either "underweight" ($<$18 kg/m^2), "normal" (18–25 kg/m^2), "overweight" (25–30 kg/m^2), or "obese" ($>$30 kg/m^2)). For genomic predicators, the altered gene (e.g., *BRAF)*, variant type (i.e., short variant (SV), copy number (CN), rearrangement (RE), and non-human (NH)), and variant class (i.e., point, truncation, amplification, deletion, rearrangement, and non-human) information were aggregated (e.g., *BRAF*_SV_point, *BRAF*_CN_amplification). Other genetic variables were complex genomic signatures, including tumor mutational burden (TMB) (mutations/Mb) and microsatellite instability (MSI) (categorized as either "high", "intermediate", microsatellite stable ("MSS"), or "unknown" (specimens with low coverage: median $<$250X)) to help inform immunotherapy decisions. We limited genomic alterations to those whose functional consequence in cancer was either "known" (i.e., reported as a confirmed somatic mutation in the Catalog Of Somatic Mutations In Cancer or other literature sources, or so assessed by FMI scientists) or "likely" (i.e., not confirmed as somatic in the literature but occurring in or near a known oncogene mutational hotspot or truncating a known tumor suppressor gene). For patients with multiple CGP tests, alterations from the sample corresponding to the earliest specimen collection date were selected in order to indicate the disease state nearest to their initial diagnosis. Any non-numeric predictors were transformed into binary variables by one-hot encoding, and those present in $<$ 10 patients were removed to reduce overfitting. A total of 586 predictors were included for model development.

2.5. Model development and statistical analyses

Descriptive and univariate analyses were conducted to summarize clinical and genomic characteristics in patients with known primary, patients in the independent validation cohort, and patients with missing primary. Missing data in BMI and race were imputed using multivariate

imputation by chained equations (MICE) [12].

2.5.1. Overall survival (OS) analysis

A descriptive comparison of OS between patients with known primary and missing primary was performed using Kaplan-Meier curves [13] and Cox regression analyses [14]. We restricted cohorts to patients with advanced or metastatic diagnoses, who were more homogenous and thus comparable in terms of life expectancy. We also carried out risk set adjustment to handle left truncation bias introduced by CGP tests received after the advanced or metastatic diagnosis date [15]. Confounders included age at diagnosis, sex, race, stage, TMB (log--transformed, mutations/Mb), BMI and the 13 primary tumor types (only included in the pan-tumor analysis).

2.5.2. XC-GeM

To predict the primary tumor type, we used the XGBoost algorithm [16], an implementation of gradient-boosted decision trees. This method offers several attractive properties, including 1) handling predictor collinearities internally by choosing one collinear variable from a group for each decision tree iteration; 2) offering more accurate prediction and effectively preventing overfitting by optimizing both the training loss and regularization of the model; 3) showing superior performance compared with other classification algorithms, including support vector machines, C4.5 decision trees, logistic regression, and random forest [17,18].

To form the training dataset, we randomly selected 80% of patients from the cohort with known tumor types. The remaining 20% were used for holdout validation, which had a similar distribution to the training set (Supplemental Fig. 1**)**. Within the training set, parameter optimization under repeated 5-fold cross-validation grid search was performed to maximize the area under the receiver operating characteristic curve (AUC) [19]. Since imbalanced data might significantly decrease the classification performance in multiclass machine learning algorithms [20], class weights were also incorporated into the model, imposing a heavier cost when errors are made in smaller classes. In the 5-fold cross-validation, the training data was split into 5 equal-sized subgroups: 4 were used for inner training and the rest for inner validation. This process was repeated 5 times by permuting the data blocks. The optimal model of the multiclass XC-GeM had parameters including: learning rate (eta) = 0.04, depth of the tree (max_depth) = 15, gamma = 0.01, fraction of columns to be randomly sampled for each tree (colsample_bytree) = 0.8, minimum number of instances required in a child node (min_child_weight) = 0.2, subsample ratio of the training instances (subsample) = 1, and maximum number of iterations (nrounds) = 1400.

To evaluate the predictive power of XC-GeM, we calculated a multiclass Hand and Till's AUC (outstanding: >0.9, 0.8–0.9: excellent, 0.7–0.8: acceptable, 0.5: no discrimination), multiclass Matthew's correlation coefficient (MCC, very high positive (negative) correlation: 0.9–1.0 (−0.9 to −1.0), high positive (negative) correlation: 0.7–0.9 (−0.7 to −0.9), moderate positive (negative) correlation: 0.5–0.7 (−0.5 to −0.7), low positive (negative) correlation: 0.3–0.5 (−0.3 to −0.5), negligible correlation: 0–0.3 (0 to −0.3)) [21], confusion matrix, micro-averaged F1 score, micro-averaged precision, and micro-averaged recall. These metrics denote the capability of distinguishing among different imbalanced misclassification distributions. Since class weight was adjusted to overcome imbalanced data, we also scored accuracy, the most popular metric to evaluate classification. The model was fitted on the training set, and the prediction was evaluated on the holdout validation dataset. Moreover, we also assessed the model performance on new patients with known primary, as additional validation, and in patients with missing primary. Overall variables of importance were also evaluated using Gain. Additionally, to assess the contributions of genomic factors to the model, we performed XGBoost with only non-genomic predictors (i.e., age, sex, BMI, race, and stage at initial diagnosis) and compared its performance to that of XC-GeM.

2.5.3. SHapley Additive exPlanations (SHAP) technique

To interpret and understand the mechanisms of XC-GeM, the SHAP technique [22] was incorporated in our study. Thirteen SHAP summary plots were generated, one for each tumor type, to visualize each predictor's importance and influence on classifying the specific tumor type. A SHAP value larger than zero suggested a higher probability of being classified as the specific cancer type (i.e., positive prediction), and a SHAP value below zero suggested a lower probability (i.e., negative prediction).

All analyses were carried out using R 4.0.0.

3. Results

3.1. Baseline characteristics and OS in patients with missing primary and known primary

About 4% of patients had a missing primary diagnosis at the time of CGP testing and varied across different tumor types (Supplemental Table 1**)**. Compared to those with known primary, patients with missing primary were more likely to be older, male, have later initial diagnosis stage and higher TMB. They were also less likely to have *APC* short variant truncation, *EGFR* short variant point alteration and more likely to have *RB1* short variant truncation. Interestingly, *VHL* and *CDH1* short variant truncations were not present in any patients with missing primary. Baseline characteristics are presented in Table 1**.**

A descriptive comparison of OS between advanced/metastatic patients with known primary and missing primary showed median OS of patients with missing primary (7.14 months, 95% confidence interval (CI): 5.86–9.01) was shorter than those with known primary (10.16 months, 95% CI: 9.57–10.69). The difference was not statistically significant after adjusting for confounders (hazard ratio (HR):1.10, 95% CI: 0.99–1.22, $P = 0.08$) (Fig. 1A). For NSCLC patients, the individual tumor type cohort with the largest sample size, median OS of patients with missing primary (6.38 months, 95% CI: 3.82–7.89) was shorter than those with known primary (8.12 months, 95% CI: 7.14–8.85). After adjusting for confounders, NSCLC patients with missing primary were significantly associated with an increased risk of death (HR: 1.42, 95% CI: 1.20–1.68, $P < 0.001$) (Fig. 1B). Relevant patient characteristics are shown in Supplemental Table 2**.**

3.2. Performance of XC-GeM in patients with known primary

XC-GeM achieved compelling performance in both the holdout and independent validation datasets (Table 2), with multiclass AUC of 0.965 and 0.954 and multiclass MCC of 0.742 and 0.733, respectively. The overall accuracy was 78.2% for the holdout validation and 77.4% for the independent validation datasets.

The performance of XC-GeM varies across different tumor types. The confusion matrices in Figs. 2A and 2B show that six tumor types were well-predicted in both validation cohorts (accuracy range: 77–89%), including BC, CRC, ML, MPC, NSCLC, and PANC. However, misclassification also occurred; 21% of OC patients (38 out of 183) in the holdout validation and 16% of OC patients (13 out of 80) in the independent validation cohort were misclassified as BC, whereas 8% of SCLC patients (3 out of 40) in the holdout validation and 40% of SCLC patients (4 out of 10) in the independent validation cohort were misclassified as NSCLC. Additional performance metrics for each indication, such as F1, recall, and precision, are shown in Supplemental Table 4 and 5**.**

3.3. XC-GeM assessment on patients with missing primary

We further assessed XC-GeM's performance in patients with missing primary and observed promising results. The model achieved 0.880 multiclass AUC, 0.540 multiclass MCC, and 62.3% overall accuracy (Table 2). Similar to the holdout validation and independent validation datasets, the model predicted well in MPC (accuracy: 71%), NSCLC

Table 1
Baseline Demographic and Genomic Characteristics for the Study Population.

	Known Primary (N = 12,060)	Independent Validation (N = 955)		Missing Primary (N = 507)	
		Value	***P-value***	**Value**	***P-value***
Age at diagnosis, years, median (IQR)	63 (54–71)	60 (51–68)	< 0.001	64 (57–71.5)	0.006
Sex, n (%)			0.31		0.04
Female	6500 (53.9)	532 (55.7)		250 (49.3)	
Male	5560 (46.1)	423 (44.3)		257 (50.7)	
Race, n (%)			0.39		0.09
Non-white	2760 (22.9)	231 (24.2)		99 (19.5)	
White	9300 (77.1)	724(75.8)		408 (80.5)	
TMB, median (IQR)	4.4 (2.6 −7.8)	3.5 (1.7–7.0)	< 0.001	4.4 (2.6–10.4)	0.02
MSI, n (%)			0.006		< 0.001
MSI high or intermediate	185 (1.5)	10 (1.0)		5 (1.0)	
MSS	8692 (72.1)	650 (68.1)		313 (61.7)	
Unknown	3183 (26.4)	295 (30.9)		189 (37.3)	
Stage at initial diagnosis			< 0.001		< 0.001
0 - II	2268 (18.8)	189 (19.8)		38 (7.5)	
III	2859 (23.7)	207 (21.7)		73 (14.4)	
IV	5963 (49.4)	433 (45.3)		345 (68.0)	
Unknown or Occult	970 (8.0)	126 (13.2)		51 (10.1)	
BMI			0.83		0.65
Underweight	458 (3.8)	39 (4.1)		17 (3.3)	
Normal	4081 (33.8)	320 (33.5)		166 (32.7)	
Overweight	4091 (33.9)	314 (32.9)		167 (32.9)	
Obese	3430 (28.4)	282 (29.5)		157(31.0)	
Alterations					
***APC* short variant truncation, n (%)**			0.03		< 0.001
Yes	22 (19.1)	210 (22.0)		41 (8.1)	
No	9761 (80.9)	745 (78.0)		466 (91.9)	
***KRAS* short variant point, n (%)**			0.82		0.62
Yes	3175 (26.3)	248 (26.0)		128 (25.2)	
No	8885 (73.7)	707 (74.0)		379 (74.8)	
***TERT* short variant point, n (%)**			0.92		1
Yes	799 (6.6)	62 (6.5)		34 (6.7)	
No	11261 (93.4)	893 (93.5)		473 (93.3)	
***EGFR* short variant point, n (%)**			0.47		0.003
Yes	560 (4.6)	39 (4.1)		9 (1.8)	
No	11500 (95.4)	916 (95.9)		498 (98.2)	
***BRAF* short variant point, n (%)**			0.55		0.14
Yes	589 (4.9)	42 (4.4)		17 (3.4)	
No	11471 (95.1)	913 (95.6)		490 (96.6)	
***VHL* short variant point, n (%)**			0.90		-
Yes	87 (0.7)	6 (0.6)		n < 4	
No	11973 (99.3)	949 (99.4)		n > 503	
***VHL* short variant truncation, n (%)**			-		–
Yes	117 (1.0)	n < 4		n < 4	
No	11943 (99.0)	n > 951		n > 503	
***RB1* short variant truncation, n (%)**			0.10		< 0.001
Yes	608 (5.0)	36 (3.8)		44 (8.7)	
No	11452 (95.0)	919 (96.2)		463 (91.3)	
***TP53* short variant point, n (%)**			0.20		0.44
Yes	5430 (45.0)	409 (42.8)		219 (43.2)	
No	6630 (55.0)	546 (57.1)		288 (56.8)	
***HPV16*, n (%)**			0.30		0.003
Yes	84 (0.7)	10 (1.0)		10 (2.0)	
No	11976 (99.3)	945 (99.0)		497 (98.0)	

The P-value for categorical variables were derived from chi-square tests comparing primary tumor types in patients in the independent validation cohort or patients with missing primary (separately) to the patients with known primary. The P-value for continuous variables were derived from Mann-Whitney tests comparing primary tumor types in patients in the independent validation cohort or patients with missing primary (separately) to the patients with known primary.

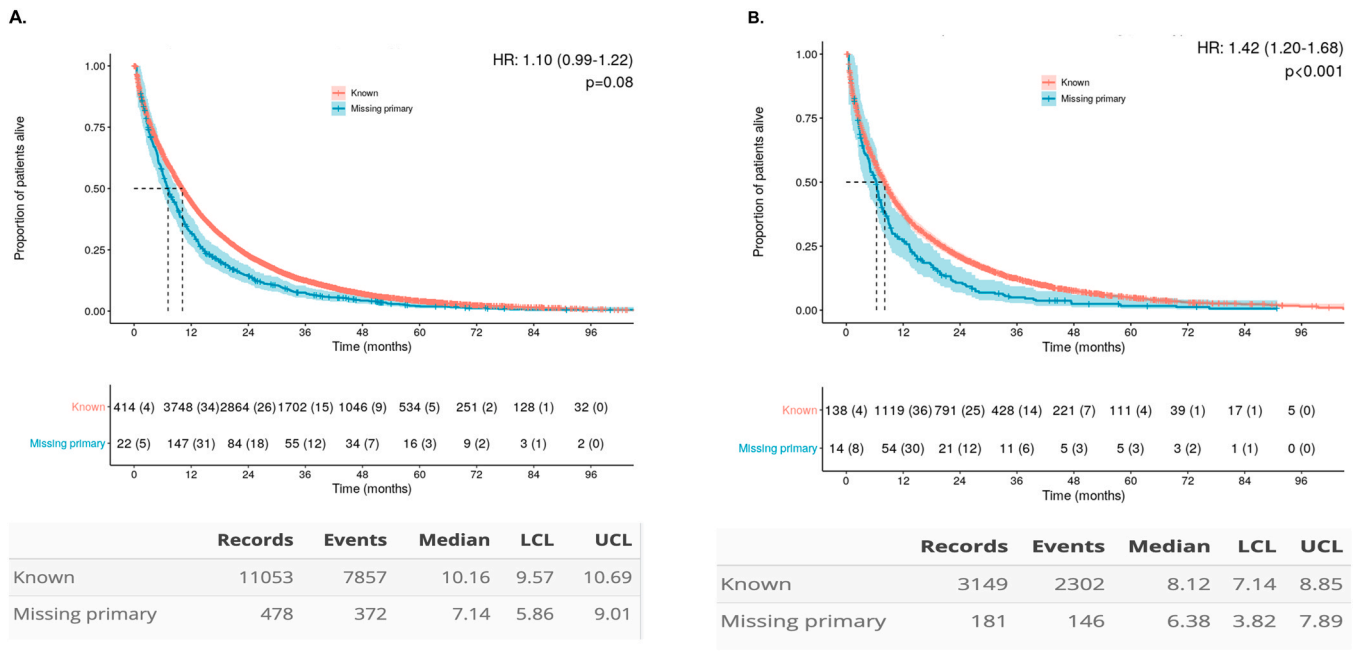
APC, adenomatous polyposis coli; BMI, body mass index; BRAF, proto-oncogene B-Raf; EGFR, epidermal growth factor receptor; HPV, human papillomavirus; IQR, interquartile range; KRAS, Kirsten rat sarcoma viral oncogene homolog; MSI, microsatellite instability; MSS, microsatellite stability; RB1, retinoblastoma protein; TERT, telomerase reverse transcriptase; TMB, tumor mutational burden; TP53, tumor protein P53; VHL, von Hippel-Lindau tumor suppressor;

(accuracy: 73%), and PANC (accuracy: 79%). Misclassification also occurred for some tumor types. For instance, 20% of patients with OC (10 out of 50) were misclassified as BC and 11.1% patients with SCLC (3 out of 27) were misclassified as NSCLC (Fig. 2**C**). Additional performance metrics for each indication, such as F1, recall, and precision, are shown in Supplemental Table 6.

3.4. Important predictors

The Top 10 overall important predictors were TMB score (Gain = 0.10), sex (Gain = 0.07), *KRAS* short variant point alteration (Gain = 0.07), *TERT* short variant point alteration (Gain = 0.06), body mass index (BMI) (Gain = 0.06), *APC* short variant truncation (Gain = 0.06), age at diagnosis (Gain = 0.05), *RB1* short variant point alteration (Gain = 0.04), *BRAF* short variant point alteration (Gain = 0.02), and *HPV-16* non-human alteration (Gain = 0.02) (Table 3 and Supplemental Fig. 2).

Consistent with the heterogeneity of tissue origins for various tumor types, significant variability in the most important predictors was observed across cancer types (Table 3 and Supplemental Figure 3). For example, patients with higher TMB scores were more likely to be predicted as NSCLC or ML. As a critical overall genomic alteration for the multiclass model, the *KRAS* short variant point alteration occurred at high frequency in multiple cancer types, including CRC, NSCLC, and PANC. However, its absence also contributed to the prediction of BC, HN, MPC and SCLC. In addition, some predictors are critical for detecting one particular tumor type, including the presence of *EGFR*



	Records	Events	Median	LCL	UCL
Known	11053	7857	10.16	9.57	10.69
Missing primary	478	372	7.14	5.86	9.01

	Records	Events	Median	LCL	UCL
Known	3149	2302	8.12	7.14	8.85
Missing primary	181	146	6.38	3.82	7.89

Fig. 1. (A) OS of advanced or metastatic patients with known primary tumor site and patients with missing primary tumor site at the time of CGP testing. (A) Patients with known primary vs. patients with missing primary; (B) NSCLC Patients with known primary vs. NSCLC patients with missing primary. NSCLC, non-small cell lung cancer; LCL, lower control limit; OS, overall survival; UCL, upper control limit.

Table 2
Classification Performance of XC-GeM to predict cancer types.

	Holdout validation	Independent validation	Missing Primary
Hand and Till's multiclass AUC	0.965	0.954	0.880
Multiclass MCC	0.742	0.733	0.540
Accuracy	78.2%	77.4%	62.3%
Micro-averaged F1 score	0.786	0.778	0.631
Micro-averaged Precision	0.787	0.778	0.635
Micro-averaged Recall	0.785	0.778	0.628

AUC: area under the receiver operating characteristic curve; MCC: Matthew's correlation coefficient

short variant point alteration for NSCLC and *CTNNB1* short variant point alteration for HCC. Among clinical predictors for the model, sex is the most predictive. Not surprisingly, males were more likely to be classified as MPC, while females were more likely to be classified as BC or OC.

To evaluate the role of genomic predictors in XC-GeM, we developed another XGBoost model with clinical predictors only. This non-genomic model had worse performance than XC-GeM in both the holdout validation (AUC = 0.699, MCC = 0.157, accuracy = 25.7%) and independent validation datasets (AUC = 0.693, MCC = 0.186, accuracy = 28.6%), indicating the importance of genomic predictors in the model (Supplemental Table 3).

4. Discussion

In this study, a novel XC-GeM algorithm was developed to predict 13 primary tumor types based on a US nationwide routine clinical CGP test, achieving outstanding performance in both the holdout validation dataset (AUC = 0.965, MCC = 0.742) and the independent validation dataset (AUC = 0.954, MCC = 0.733). The top predictors associated with specific tumor types were supported by previous literature, suggesting the capability of XC-GeM to accurately capture key biological predictors specific to primary tumor tissue origin. The performance in patients with missing primary was also promising, with an excellent AUC of 0.880 and a moderately positive MCC of 0.540, suggesting potential clinical usages in this population. The XC-GeM might act as an assistive tool for faster and more accurate cancer diagnoses, especially for patients with missing primary at a certain point of time and patients with CUP.

To our knowledge, XC-GeM is the first primary tumor type predictive model based on a US nationwide CGP test performed in routine clinical practice and chart-confirmed EHRs. There are over 200 types of cancer, and timely and accurate diagnosis is critical for subsequent care and survival outcomes [4,5]. A similar algorithm was developed using a single hospital system's data derived from the routine clinical MSK-IMPACT CGP test [9], with 74% accuracy in the independent validation dataset. This model used similar top predictors to XC-GeM for identifying different cancer types, including sex for BC and MPC, *APC* alteration for CRC, and absence of *KRAS* alteration for BC. However, the test is only approved for use in one hospital system, restricting its benefits. The majority of other previous genomic-based primary tumor type predictive algorithms were developed using data from whole-genome or exome, gene expression [8], miRNA [23], and deoxyribonucleic acid (DNA) methylation [24] sequencing techniques, which are not part of routine clinical practice. XC-GeM incorporated data from a US nationwide CGP test used in routine clinical practice according to SoC to guide molecularly targeted therapies. It may improve timely and accurate cancer diagnosis, and has the potential to be integrated as part of routine care without requiring another test or extra tissue samples.

XC-GeM is also the first genomic-based primary tumor type predictive model to assess its performance in patients with missing primary diagnosis at the time of CGP. The accuracy of 62.3% is similar to the 61% accuracy generated from the first and only deep-learning algorithm based on histology slides in patients who had a missing primary tumor diagnosis at some point during their diagnosis process [25]. In our cohort, over 4% of patients were missing primary tumor diagnoses, likely due to the inability to obtain a definitive diagnosis at the time of CGP testing. These patients may encounter delayed or restricted treatment options, especially targeted therapies, because current treatment guidelines and approval of targeted therapies largely remain tissue-based. For example, many physicians use patients' FMI reports to guide the usage of targeted therapies. However, without primary tumor

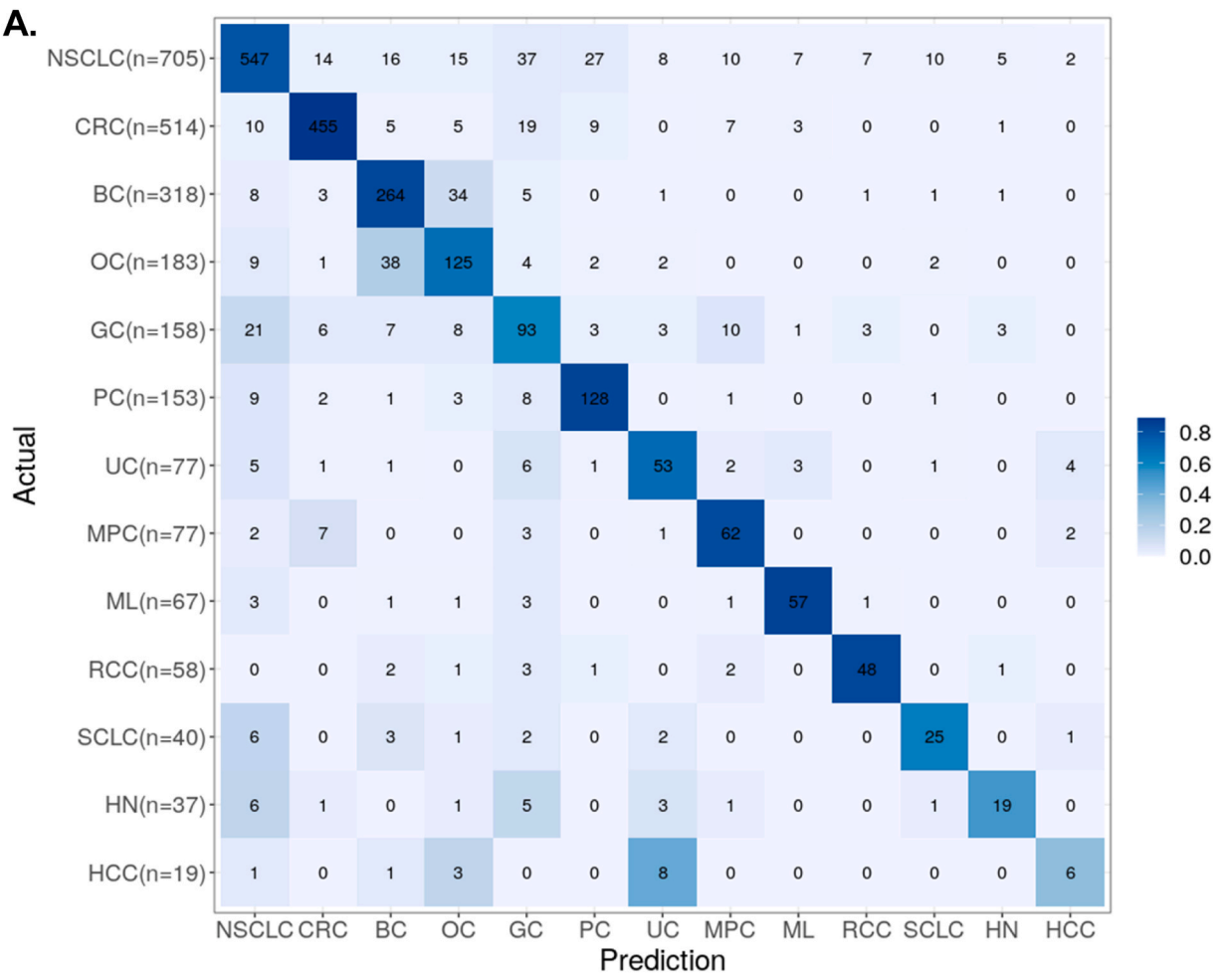


Fig. 2. Confusion Matrix (A) holdout validation, (B) independent validation, and (C) patients with missing primary. Each row corresponds to the actual primary tumor type; Columns correspond to the class prediction emitted by the XC-GeM. Cells are labeled with numbers of tumors of a particular type that were classified by the XC-GeM and colored based on the percentages of correctly predicted tumors of a particular type from dark (high percentages) to light blue (low percentage). UC, advanced urothelial cancer; BC, breast cancer; CRC, colorectal cancer; CUP, cancer of unknown primary; GC, gastroesophageal cancer; HCC, hepatocellular carcinoma; HN, head and neck cancer; ML, advanced melanoma; MPC, metastatic prostate cancer; NSCLC, non-small cell lung cancer; OC, ovarian cancer; PANC, pancreatic cancer; RCC, renal cell carcinoma; SCLC, small cell lung cancer.

diagnosis, approved targeted therapies in specific "other" tumor types will be documented in the latter pages. Therefore, physicians may not initially consider those treatment options, and access to such targeted therapies for patients with missing primary might be challenging and delayed. This primary tumor type predictive model may also benefit patients with CUP [26]. 2.3–5% of cancer patients receive a diagnosis of CUP, commonly experiencing treatment delays and disproportionately ranking third to fourth in cancer-related deaths with a median survival of less than a year [27]. While the value of site-specific treatment in CUP is still unclear [28], a recent randomized study [29] demonstrated that site-specific treatment, including molecularly targeted therapy based on profiling of gene expression and gene alterations by next-generation sequencing, can contribute to improved outcomes in patients with the unfavorable subset of CUP, the major clinicopathologic subtypes. Interestingly, the predictive model based on the MSK-IMPACT assay [9] was able to classify likely tissues of origin among 95 of 141 (67.3%) CUP patients using the 50% probability threshold. XC-GeM's performance in CUP was not assessed because this population was not included in the current CGDB as an individual tumor type. In addition, since the true primary tumor origin is unknown for CUP, the performance of primary tumor type predictive models cannot truly be evaluated for this disease.

The high performance of XC-GeM is amplified by its accurate capture of key biological predictors specific to primary tumor tissue origin, as evidenced by the observation that predictors and their negative/positive contributions to predicting each specific cancer type are consistent with prior literature. For example, similar to previous research that TMB varies markedly in tumor types [30], XC-GeM used TMB score as the most vital overall predictor. Contributions of *KRAS* to positive predictions of PANC, CRC, and NSCLC in XC-GeM are supported by previous research that KRAS alterations predominantly occur in those cancer types [31]. Top predictors used by XC-GeM in identifying NSCLC were also reported in other studies, including higher TMB [30], *EGFR* alterations [32,33], *STK11* alterations [34], *KRAS* alterations [35], and absence of *TERT* alterations [36]. Unsurprisingly, cancers with smaller sample sizes, such as HCC, were not classified well. However, XC-GeM still captured predictors known to be enriched in HCC, including *CTNNB1* alterations [37], *TERT* alterations [38], lower TMB [39], and the absence of *CDKN2A* alterations [40]. Some of these correlations are well known, and an oncologist or geneticist could guess the primary location from such data. However, as the number of entries rises, manual evaluation may no longer be feasible, and a machine learning approach like XC-GeM, when integrated with traditional methods, may be better than human experts.

The performance of XC-GeM in the cohort with missing primary was

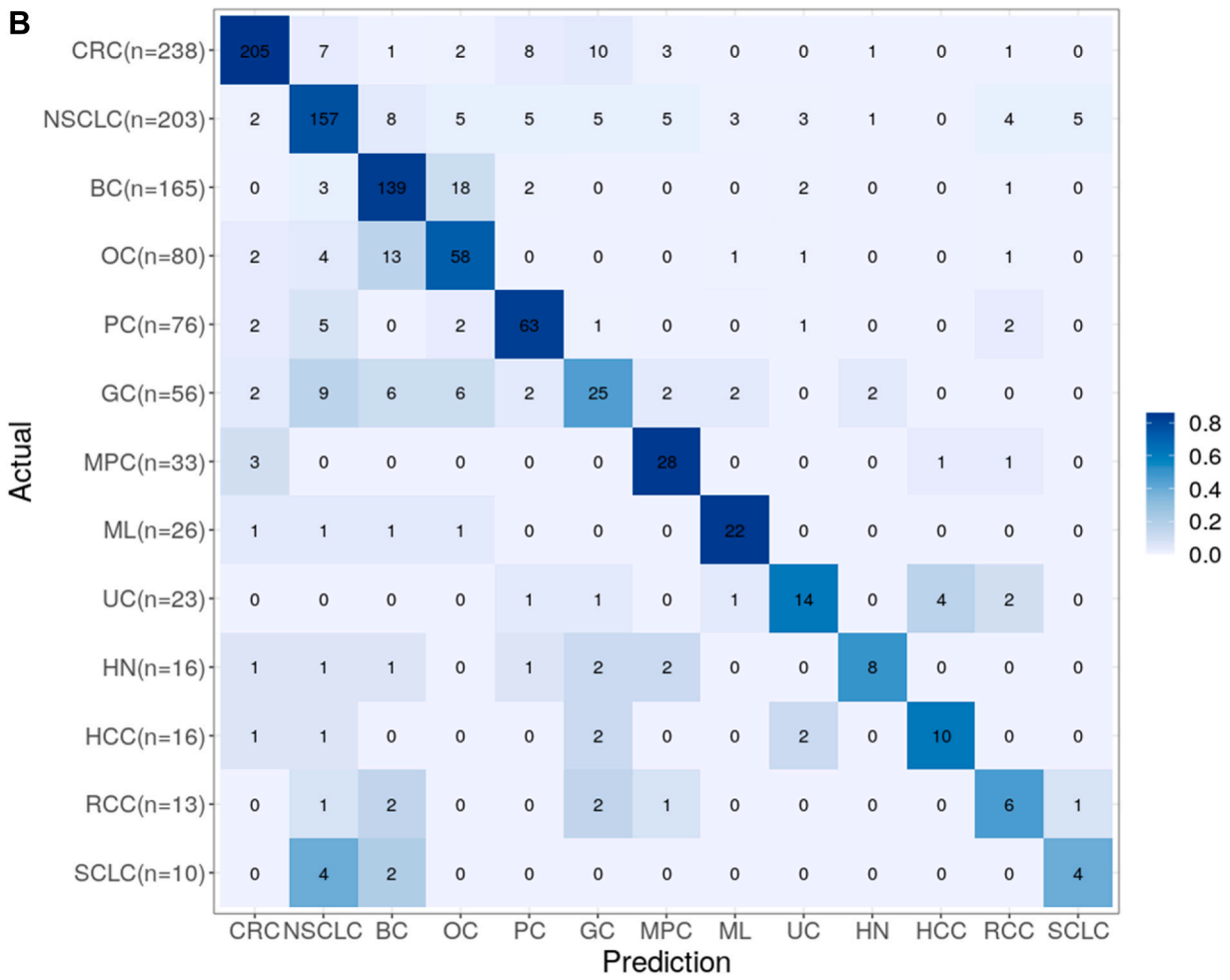


Fig. 2. (*continued*).

not as good as in patients with known primary. One potential reason might be errors in abstracted cancer diagnosis from EHR for patients with missing primary, as diagnostic errors existed across different tumor types [3]. Another potential explanation might be the intrinsic differences between the cohort with missing primary and that with known primary, which was used to develop the model. For example, among advanced NSCLC patients, EGFR short variant alterations, the 2nd top classifier for NSCLC and 23th overall important predictor, were less prevalent in patients with missing primary, compared to those with known primary (3.9% vs. 14.8%, $P < 0.001$). Since patients with missing primary at the time of CGP testing might be a unique population, it would be interesting to train models directly in this cohort and explore the potential improvement of model performance.

While XC-GeM showed good performance, and it could be included in routine care as an assistance tool for diagnosis if clinically validated, it is important to note that the model performance varies by cancer types, and there are potential downsides to wrong prediction, hence traditional methods should not be replaced nor stopped prematurely. In addition, to be used as part of standard care, XC-GeM might require continued improvements and investigations to achieve and expand clinical utility. First, the model was trained in solid tumor biopsies with only 13 cancers from FH US-based hospitals, where comprehensive cancer centers are under-represented. As the patient cohort and region coverage grow in the FH network, rare or other cancer types from a more generalizable population could be included. Second, the validated FoundationOne CGP test was a CLIA certified version, a prior iteration of the FoundationOne CDx assay that was approved by the FDA in November 2017 but had a limited sample size at the time of model development. However, these two CGP versions are comparable and have high concordance for all alterations (ranging from 94.3% to 100%) [41]. Future studies should be considered to validate XC-GeM using FoundationOne CDx data. Moreover, trials should be conducted to clinically validate XC-GeM. Third, despite observing worse survival in patients with missing primary than those with known primary, it is important to note that the study was not designed nor powered to establish causal relationship between missing primary diagnosis at time of CGP and survival. Other factors, including challenging histology [42], might also be related to patient survival outcomes. A randomized design study for clinically-validated XC-GeM and survival of patients with missing primary might be helpful to evaluate the causation. Fourth, it would be invaluable to expand the usage of XC-GeM with less invasive liquid biopsies to benefit patients with no available tissue samples and to reduce the time to obtain tissue-based testing results. Lastly, the clinical utility of XC-GeM in assisting primary tumor type diagnosis could be enhanced via earlier CGP testing (i.e. at the time of diagnosis), as it is currently not the case for most patients [43].

Several other limitations should be considered when interpreting our findings. First, sample sizes are imbalanced across different cancers, contributing to the relatively lower performance in the minority classes. However, while the prevalence of cancers largely varies, we incorporated class weights and evaluated models with AUC and MCC to overcome the imbalance at the statistical level. Second, ~14% CGP assays failed QC (Supplemental Fig. 1). Third, for those patients with CGP that passed QC, we were unable to fully account for all predictors related to tumor type classification. The database lacked information on other possible influential factors, including sample purity, disease

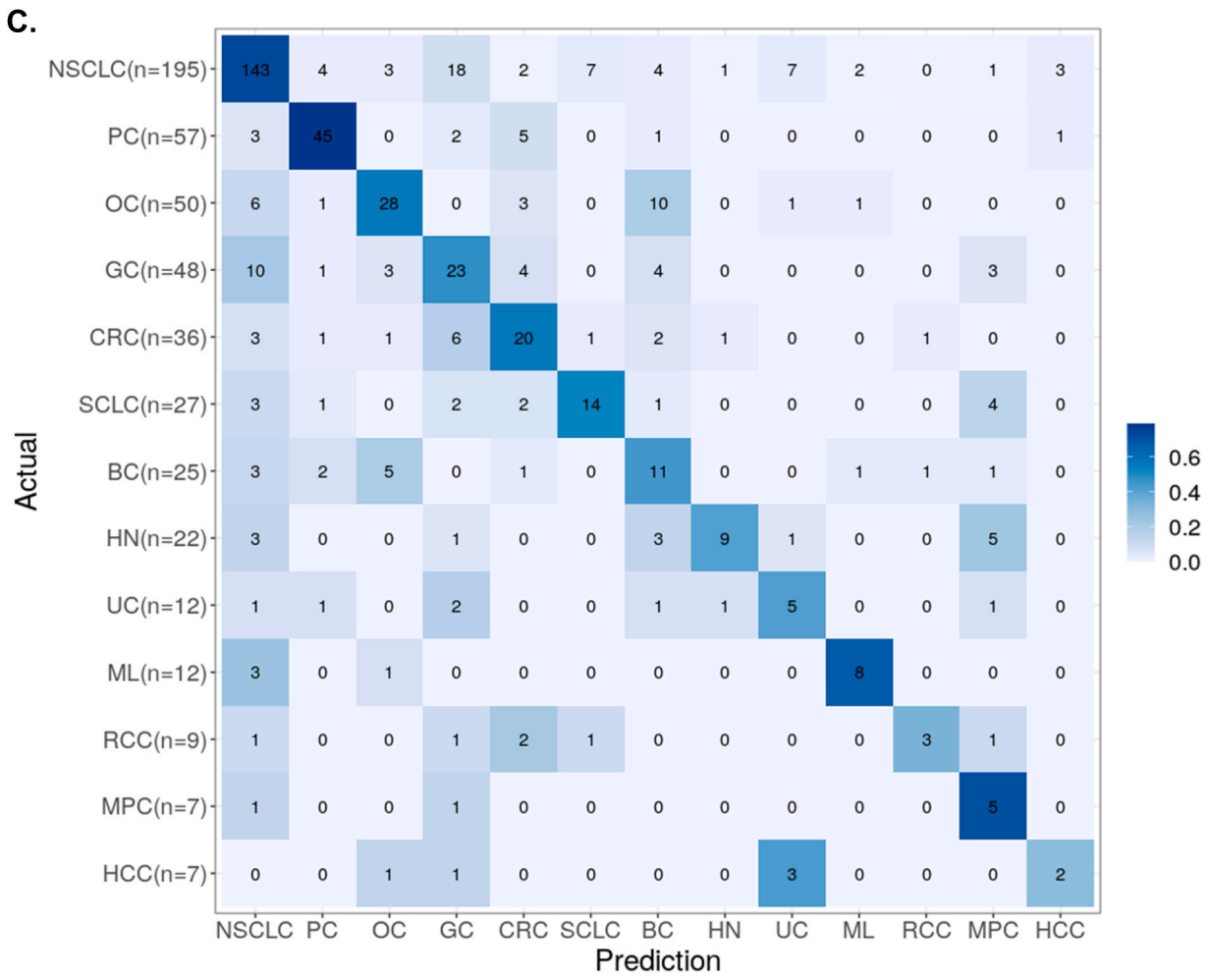


Fig. 2. (*continued*).

comorbidity, and smoking status across all tumor types. Moreover, causal non-human variants, including Hepatitis B in HCC, and HPV in HN cancer [44–46], are currently only exploratory in the CGDB; and the "known" or "likely" function designation for alterations is proprietary information in CGDB and is evolving with ongoing research. In addition, almost 8% of patients were missing data on "race", and 2.2% were missing BMI, which might lead to bias in the results. Nevertheless, we applied MICE imputation, which has been proven to be a very powerful technique in handling missing data [47]. Due to small sample size, detailed race information for "Asian" (2.6%), "Black or African American" (6.6%), "Hispanic or Latino" (0.2%), and "Other race" group (13.6%) has been recategorized to "non-white". It would be interesting to assess model performance in "white" and "non-white" separately, and future study with larger sample size is encouraged to be conducted using more detailed race categories. Fourth, misclassifications exist, weakening the classification. However, some might be interpretable because of biological similarities in tumor types at the molecular level. For example, since both reside in the lung, there is a possibility of misclassification between SCLC and NSCLC (15–40%). Additionally, XC-GeM might rely on the "female" predictor for sex-specific cancers, causing difficulty differentiating between OC and BC. On the other hand, both OC and BC patients may have *BRCA* alterations and could benefit from PARP inhibitors [48].

5. Conclusions

In summary, XC-GeM used data from a US nationwide CGDB, consisting of genomic information from a routine clinical CGP test, to predict 13 primary tumor types and showed good performance in both patients with known primary and missing primary. It may facilitate faster and more accurate cancer diagnosis, especially in patients with tumors whose primary origin is challenging to diagnose using traditional methods.

Author contributions

Yunru Huang conducted the analyses and drafted the manuscript. Shannon Pfeiffer worked on the exploratory version of the analyses and manuscript and assisted in the analysis interpretation. Qing supervised the study. All authors contributed substantially to the study design, manuscript revision, approved the final manuscript as submitted, and agreed to be accountable for all aspects of the work. [1]Shannon Pfeiffer's work on this project was during her time at Genentech as an intern.

Funding

This work was sponsored by F. Hoffmann-La Roche Ltd./Genentech, Inc.

Declaration of Competing Interest

Yunru Huang: Genentech/Roche employee; owns Roche, and 23andMe stocks. Shannon Pfeiffer: this work was done while Shannon Pfeiffer was an Intern at Genentech/Roche. Qing Zhang: Genentech/

Table 3
Selected Top Predictors for Each Cancer Type Identified by SHAP.

Cancers	top predictors (ordered by importance)*
All 13 cancer types	TMB score, sex, *KRAS*, *TERT*, BMI, *APC*, Age, *RB1*, *BRAF*, *HPV-16*
BC	Female, younger diagnosis age, absence of *KRAS*, absence of *APC*, higher BMI, earlier initial stage, *PIK3CA*, *CDH1,* absence of *BRAF*
CRC	*APC,* younger diagnosis age, *RNF43*, *BRAF*, absence of *CDKN2A*, *KRAS, SMAD4*, and absence of *TERT*
GC	Male, *GATA6, CDK6,* absence of *TERT*, *TP53, KRAS*
HCC	*TERT*, *CTNNB1,* lower TMB, *ATRX*, Male, absence of *TP53*
HN	*HPV-16*, lower BMI, lower TMB, *CDKN2A, BCL2L2, TERT, FGF3,* Male, absence of *KRAS*
ML	*TERT,* Higher TMB, *BRAF, NRAS, KIT, higher* BMI, *NF1*, absence of *KRAS* and absence of *TP53*
MPC	Male, *TMPRSS2, PTEN, ERG,* absence of *KRAS*, *AR* and absence of *CDKN2A*
NSCLC	Higher TMB, *EGFR*, absence of *APC*, *STK11,* absence of *TERT, ALK, KRAS*
OC	Female, lower TMB, absence of *APC,* initial stage III, absence of *KRAS*, absence of *CDKN2A*
PANC	*KRAS*, lower TMB, absence of *APC*, lower BMI, absence of *PIK3CA, SMAD4,* and *CDKN2A*
RCC	*VHL,* lower TMB, *PBRM1*, absence of *TP53, BAP1*, earlier initial stage, and *MET*
SCLC	*RB1*, higher TMB, later initial stage, absence of *TERT,* absence of *KRAS,* and *PREX2*
UC	*TERT,* higher BMI, *FGFR3, MLL2/KMT2D, KDM6A*, absence of *BRAF,* and absence of *NRAS*

BC, breast cancer; CRC, colorectal cancer; GC, gastroesophageal cancer; HCC, hepatocellular carcinoma; HN, head and neck cancer; ML, advanced melanoma; MPC, metastatic prostate cancer; NSCLC, non-small cell lung cancer; OC, ovarian cancer; PANC, pancreatic cancer; RCC, renal cell carcinoma; SCLC, small cell lung cancer; UC, advanced urothelial cancer;
APC, adenomatous polyposis coli; ALK, anaplastic lymphoma kinase; AR: androgen receptor; ATRX: alpha-thalassemia chromatin remodeler; BAP1, BCRA1 associated protein-1; BCL2L2, CBackspaceNBackspaceBCL2 apoptosis regulator like 2; BMI, body mass index; BRAF, proto-oncogene B-Raf; CDH1, cadherin 1; CDK, cyclin dependent kinase; CDKN2A, cyclin-dependent kinase inhibitor 2A; CTNNB1, catenin β 1; EGFR, epidermal growth factor receptor; ERG, ETS-related gene; FGF: fibroblast growth factor; FGFR, fibroblast growth factor receptor; GATA, GATA-binding protein; HPV, human papillomavirus; KDM6A, lysine demethylase 6A; KIT, KIT proto-oncogene, receptor tyrosine kinase; KMT2D, also known as MLL2, lysine methyltransferase 2D; KRAS, Kirsten rat sarcoma viral oncogene homolog; MET, MET proto-oncogene, receptor tyrosine kinase; NF, neurofibromin; NRAS, neuroblastoma RAS viral oncogene homolog; PBRM1, polybromo 1; PIK3CA, phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit α; PREX2, phsophatidylinositol-3,4,5-triphosphate dependent TBackspaceRac exchange factor 2; PTEN, phosphatase and tensin homolog; RB1, retinoblastoma protein; RNF43, ring finger protein 43; SMAD4, small mothers against decapentaplegic family member 4; STK11, serine/threonine kinase 11; TERT, telomerase reverse transcriptase; TMB, tumor mutational burden; TMPRSS2, transmembrane serine protease 2; TP53, tumor protein P53; VHL, von Hippel-Lindau tumor suppressor.

* All listed genes refer to gene alterations; All predictors were supported by previous literature.

Roche employee; owns Roche, Regeneron, BMS, AbbVie, Pfizer, BioNTech and AC Immune stocks.

Data availability

The data that support the findings of this study originated from Flatiron Health, Inc and Foundation Medicine, Inc. These de-identified data may be made available upon request and are subject to a license agreement with Flatiron Health and Foundation Medicine; interested researchers should contact dataaccess@flatiron.com to determine licensing terms.

Acknowledgements

The authors acknowledge their colleagues for their helpful discussions during preparation of this manuscript: Derrek Hibar, Dominik Heinzmann, Marlene Thomas, Melanie Huntley, Jamie Clendening, Svetlana Lyalina, Chris Harbron, Sarah McGough, M.K. Downer, Tricia Luhn, Ryan Gan, Robson Machado, Chris Bolen, and Kieran Mace.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.07.036.

References

[1] Ferlay J, et al. Global Cancer. Observatory: Cancer Today. *Lyon*: International Agency for Research on Cancer,; 2020. ⟨http://gco.iarc.fr/today/home⟩.
[2] Helsper CCW, van Erp NNF, Peeters PPH, de Wit NNJ. Time to diagnosis and treatment for cancer patients in the Netherlands: room for improvement. Eur J Cancer 2017;87:113–21.
[3] Newman-Toker DE, et al. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "Big Three". Diagnosis 2021;8:67–84.
[4] Jiao W, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun 2020;11:728.
[5] Varghese AM, et al. Clinical and molecular characterization of patients with cancer of unknown primary in the modern era. Ann Oncol 2017;28:3015–21.
[6] Martínez-Jiménez F, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer 2020;20:555–72.
[7] Reiter JG, et al. Minimal functional driver gene heterogeneity among untreated metastases. Science 2018. https://doi.org/10.1126/science.aat7171.
[8] Xu Q, et al. Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. Mod Pathol 2016;29:546–56.
[9] Penson A, et al. Development of genome-derived tumor type prediction to inform clinical cancer care. JAMA Oncol 2020;6:84–91.
[10] Agarwala V, et al. Real-world evidence in support of precision medicine: clinico-genomic cancer data as a case study. Health Aff 2018;37:765–72.
[11] Swaminathan A, et al. Abstract 864: changes over time in real-world next-generation sequencing (NGS) test use in patients (pts) with advanced non-small cell lung cancer (aNSCLC). 864–864 Cancer Res 2021;81. 864–864.
[12] Buuren S van, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67.
[13] Goel MK, Khanna P, Kishore J. Understanding survival analysis: kaplan-Meier estimate. Int J Ayurveda Res 2010;1:274–8.
[14] George B, Seals S, Aban I. Survival analysis and regression models. J Nucl Cardiol 2014;21:686–94.
[15] Candido-dos-Reis FJ, et al. Germline Mutation in BRCA1 or BRCA2 and ten-year survival for women diagnosed with epithelial ovarian cancer. Clin Cancer Res 2015;21:652–7.
[16] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery,; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
[17] Liu L, et al. An interpretable boosting model to predict side effects of analgesics for osteoarthritis. BMC Syst Biol 2018;12:105.
[18] Chang W, et al. A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics 2019;9:178.
[19] Hand DJ, Till RJ. A simple generalisation of the area under the roc curve for multiple class classification problems. Mach Learn 2001;45:171–86.
[20] Wang S, Yao X. Multiclass imbalance problems: analysis and potential solutions. IEEE Trans Syst Man Cyber Part B Cyber 2012;42:1119–30.
[21] Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. PLOS ONE 2012;7:e41882.
[22] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;vol. 30.
[23] Søkilde R, et al. Efficient identification of miRNAs for classification of tumor origin. J Mol Diagn 2014;16:106–15.
[24] Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol 2020;31:745–59.
[25] Lu MY, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature 2021;594:106–10.
[26] Ross JS, et al. Comprehensive genomic profiling of carcinoma of unknown primary origin: retrospective molecular classification considering the CUPISCO study design. Oncologist 2021;26:e394–402.
[27] Qaseem A, Usman N, Jayaraj JS, Janapala RN, 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 2019;11.
[28] Rassy E, et al. The role of site-specific therapy for cancers of unknown of primary: a meta-analysis. Eur J Cancer 2020;127:118–22.
[29] Hayashi H, et al. Site-specific and targeted therapy based on molecular profiling by next-generation sequencing for cancer of unknown primary site: a nonrandomized phase 2 clinical trial. JAMA Oncol 2020;6:1931–8.

[30] Yarchoan, M. et al. PD-L1 expression and tumor mutational burden are independent biomarkers in most cancers. *JCI Insight* **4**, e126908.
[31] Nassar AH, Adib E, Kwiatkowski DJ. Distribution of KRASG12C somatic mutations across race, sex, and cancer type. N Engl J Med 2021. https://doi.org/10.1056/NEJMc2030638.
[32] Ali R, Wendt MK. The paradoxical functions of EGFR during breast cancer progression. Signal Transduct Target Ther 2017;2:1–7.
[33] Giordano G, Remo A, Porras A, Pancione M. Immune resistance and EGFR antagonists in colorectal cancer. Cancers 2019;11:1089.
[34] Pécuchet N, et al. Different prognostic impact of STK11 mutations in non-squamous non-small-cell lung cancer. Oncotarget 2015;8:23831–40.
[35] Westcott PMK, To MD. The genetics and biology of KRAS in lung cancer. Chin J Cancer 2013;32:63–70.
[36] Jung S-J, et al. Mutation of the TERT promoter leads to poor prognosis of patients with non-small cell lung cancer. Oncol Lett 2017;14:1609–14.
[37] S. Kim S. Jeong Mutation Hotspots in the β-Catenin Gene: Lessons from the Human Cancer Genome Databases 42 2019 8 16.
[38] Amisaki M, Tsuchiya H, Sakabe T, Fujiwara Y, Shiota G. Identification of genes involved in the regulation of TERT in hepatocellular carcinoma. Cancer Sci 2019;110:550–60.
[39] Pinato DJ, et al. Immune-based therapies for hepatocellular carcinoma. Oncogene 2020;39:3620–37.
[40] Khemlina G, Ikeda S, Kurzrock R. The biology of Hepatocellular carcinoma: implications for genomic and immune therapies. Mol Cancer 2017;16:149.
[41] FoundationOne CDx technical information. (2017).
[42] Ota T, et al. Validity of using immunohistochemistry to predict treatment outcome in patients with non-small cell lung cancer not otherwise specified. J Cancer Res Clin Oncol 2019;145:2495–506.
[43] Nesline MK, et al. Oncologist uptake of comprehensive genomic profile guided targeted therapy. Oncotarget 2019;10:4616–29.
[44] Rheinbay E, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature 2020;578:102–11.
[45] Schulze K, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet 2015;47:505–11.
[46] Shuai S, Gallinger S, Stein L. Combined burden and functional impact tests for cancer driver discovery using driverpower. Nat Commun 2020;11:734.
[47] Lodder, P. *To Impute or not Impute: That's the Question*. (2014).
[48] Casaubon JT, Kashyap S, Regan J-P. BRCA 1 and 2. in *StatPearls*. (StatPearls Publishing,; 2022.