


# Pan-Cancer Survival Classification With Clinicopathological and Targeted Gene Expression Features

Cancer Informatics  
Volume 20: 1–8  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/11769351211035137



Daniel Zhao<sup>1,\*</sup>, Daniel Y Kim<sup>2,\*</sup> , Peter Chen<sup>3,\*</sup>, Patrick Yu<sup>4,5</sup>, Sophia Ho<sup>6</sup>, Stephanie W Cheng<sup>7</sup>, Cindy Zhao<sup>6</sup>, Jimmy A Guo<sup>4,8,9</sup> and Yun R Li<sup>10</sup>

<sup>1</sup>School of Medicine, New York Medical College, Valhalla, NY, USA. <sup>2</sup>Molecular Pathology, Massachusetts General Hospital, Charlestown, MA, USA. <sup>3</sup>Raytheon Technologies, Brooklyn, NY, USA. <sup>4</sup>Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>5</sup>College of Engineering, University of Illinois at Urbana-Champaign, Champaign, IL, USA. <sup>6</sup>Northeastern University, Boston, MA, USA. <sup>7</sup>Harvard College, Harvard University, Cambridge, MA, USA. <sup>8</sup>Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. <sup>9</sup>School of Medicine, UCSF, San Francisco, CA, USA. <sup>10</sup>Department of Radiation Oncology, UCSF, San Francisco, CA, USA.

**ABSTRACT:** Prognostication for patients with cancer is important for clinical planning and management, but remains challenging given the large number of factors that can influence outcomes. As such, there is a need to identify features that can robustly predict patient outcomes. We evaluated 8608 patient tumor samples across 16 cancer types from The Cancer Genome Atlas and generated distinct survival classifiers for each using clinical and histopathological data accessible to standard oncology workflows. For cancers that had poor model performance, we deployed a random-forest-embedded sequential forward selection approach that began with an initial subset of the 15 most predictive clinicopathological features before sequentially appending the next most informative gene as an additional feature. With classifiers derived from clinical and histopathological features alone, we observed cancer-type-dependent model performance and an area under the receiver operating curve (AUROC) range of 0.65 to 0.91 across all 16 cancer types for 1- and 3-year survival prediction, with some classifiers consistently outperforming those for others. As such, for cancers that had poor model performance, we posited that the addition of more complex biomolecular features could enhance our ability to prognose patients where clinicopathological features were insufficient. With the inclusion of gene expression data, model performance for 3 select cancers (glioblastoma, stomach/gastric adenocarcinoma, ovarian serous carcinoma) markedly increased from initial AUROC scores of 0.66, 0.69, and 0.67 to 0.76, 0.77, and 0.77, respectively. As a whole, this study provides a thorough examination of the relative contributions of clinical, pathological, and gene expression data in predicting overall survival and reveals cancer types for which clinical features are already strong predictors and those where additional biomolecular information is needed.

**KEYWORDS:** Pan-cancer, survival classification, machine learning, TCGA, cancer prognosis, gene expression, transcriptomics

**RECEIVED:** March 21, 2021. **ACCEPTED:** July 5, 2021.

**TYPE:** Original Research

**FUNDING:** The author(s) received no financial support for the research, authorship, and/or publication of this article.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: D.Y.K. is a paid consultant at Verve Therapeutics and SeQure Dx, unrelated to this research. D.Y.K.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict-of-interest

policies. P.C. was an employee of Raytheon Technologies at the time of this study. No Raytheon intellectual property, equipment, or funding was used in this research. None of these affiliations represent a conflict of interest with respect to the design or execution of this study or interpretation of data presented in this manuscript. All other authors declare no competing interests.

**CORRESPONDING AUTHORS:** Jimmy A Guo, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02141, USA. Email: jguo@broadinstitute.org

Yun R Li, Department of Radiation Oncology, UCSF, 505 Parnassus Avenue, San Francisco, CA 94143, USA. Email: yun.li2@ucsf.edu

## Introduction

The dynamic range of survival within many cancer types can be broad, as many features influence tumor aggressiveness. Molecular subtyping efforts for many cancers have yielded distinctions between those which are more or less lethal, but stratification in this regard generally does not provide accurate survival classification at the individual patient level. As such, there would be high value in predicting survival for each patient using routinely collected clinical and/or pathological features.

The advent and application of machine learning over the past decade have enabled the development of survival classifiers trained on input data sets containing potentially predictive features. Through this, the relative importance of any given predictor can be quantified and thereby suggestive of its

saliency for prognostication. This triaging of information can be beneficial when it is infeasible to collect a large number of clinical data points that do not necessarily inform the actual management or expectations of the patient's disease. In addition, over the past decade, massive biomarker data sets of methylation, gene expression, or mutational status have demonstrated enormous promise as inputs for machine learning algorithms,<sup>1-3</sup> but these modes of tumor profiling are not yet integrated into the vast majority of clinical workflows and therefore do not offer a scalable or resource-efficient approach for prognostication in its current state.

Although prior studies have leveraged publicly available databases to model prognosis for individual cancer types, few have systematically employed the same techniques on multiple cancer types to predict overall survival (OS) outcomes at varying time points.<sup>3,4</sup> This precludes a direct comparison of the

\* D.Z., D.Y.K. and P.C. provided equal contribution.



most predictive features across various malignancies. We therefore orient this investigation first toward the prediction of OS for each of 16 different cancer types, using routinely used clinical and histopathological data. In doing so, we aim to provide a pragmatic classification blueprint compatible with most clinical oncology data collection workflows.

Unsurprisingly, clinical features alone resulted in a bottleneck of prediction performance for many cancer types, and we further inquired whether the addition of more complex biomolecular information could enhance survival classification. As such, we incorporated bulk RNA sequencing data from The Cancer Genome Atlas (TCGA) to evaluate whether gene expression patterns could confer any additional value. This study provides a thorough look at the contributions of clinical, pathological, and, later, gene expression data in predicting OS, providing guidance in areas where clinical features are already strong predictors and where more complex biomolecular information may be needed.

## Methods

### *Patient cohort*

Data were extracted from a total of 8608 patients across 16 different cancer types (bladder urothelial carcinoma, renal clear cell carcinoma, uterine corpus endometrial carcinoma, pancreatic adenocarcinoma (PAAD), breast invasive carcinoma (BRCA), lung squamous cell carcinoma, hepatocellular carcinoma, papillary thyroid carcinoma (THCA), colorectal adenocarcinoma, cutaneous melanoma, glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), stomach/gastric adenocarcinoma (STAD), lung adenocarcinoma, prostate adenocarcinoma, and serous adenocarcinoma) from publicly available studies by TCGA on cBioPortal.<sup>5</sup>

The inclusion criteria for TCGA patients/samples are as follows: primary untreated tumor, frozen and sufficiently sized resection samples, and sufficient percentage of tumor nuclei (60% as a general guideline, although exceptions were made for cancer types with low neoplastic cellularity).<sup>6</sup> As such, neoadjuvant therapies were not provided to patients in our cohort. Adjuvant therapies (delivered after treatment or surgery), on the contrary, were administered to most patients, but the type, dose, and length of such treatments were highly variable depending on the cancer type and not completely annotated for each patient.

### *Features used for classification*

Our data set included clinical and pathological features (which varied by cancer type), as well as unnormalized gene counts from bulk RNA-seq (HTSeq-Counts). We excluded clinical features that were missing in more than 40% of patients or functionally served as proxies for OS (“Disease Free Status,” “Overall Survival Status,” etc). Missing predictive data were imputed using 3 different imputation techniques: XGBoost’s built-in imputation, mean/median, and K-nearest neighbors

(KNN). The primary outcomes were 1- and 3-year OS from the date of diagnosis. The full list of clinicopathological features used in this study has been provided in Supplementary Table 1.

### *Classification of survival using clinical data alone*

Random forest and XGBoost models leveraging a variety of imputation methods were derived from clinical training data for each cancer type. Specifically, a sequential feature selection strategy was used: beginning with an empty set of clinical features, the next most informative feature was added until 15 clinical features were appended. A maximum of 15 features was chosen to optimize for parsimony.

The analysis was performed using an 80/20 split, in which 80% of patients were reserved for training and the remaining 20% for validation using 5-fold cross-validation. A grid search on 2 different machine learning models (XGBoost, Random Forest) and 3 different imputation methods (KNN, mean/median, XGBoost’s built-in imputation) was performed for each of the 16 cancer types; a total of 80 models were ultimately evaluated. Analyses were performed using open-source libraries (scikit-learn, xgboost, mlxtend) in Python 3.7. The performances of these classifiers were then evaluated by prediction accuracy and area under the receiver operating curve (AUROC).

### *Differential gene expression analysis*

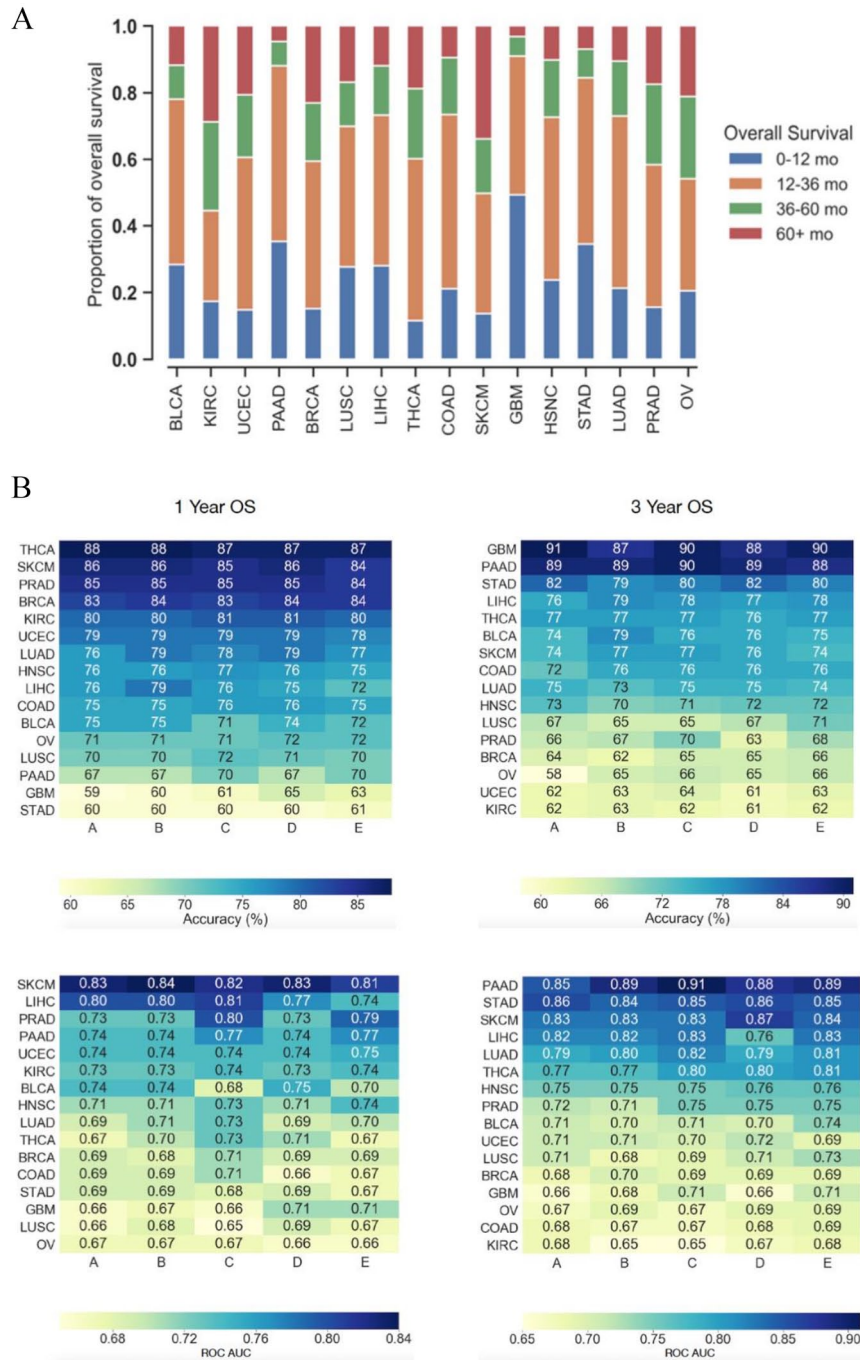
Differential gene expression analyses were performed on the following subsets of patients: > vs <1 year STAD, > vs <1-year GBM, and > vs <3-year ovarian serous carcinoma (OV). These specific cancer types were chosen because clinicopathological features were insufficient to generate strong model performance. As such, we reasoned they would stand to benefit the most from the addition of other biomolecular features. Genes with an adjusted *P* value (false discovery rate) of less than .005, with the exception of an adjusted *P* value threshold of less than .001 for OV to generate a more parsimonious list, were inputted as additional features for survival classification for STAD (at 1 year), GBM (at 1 year), and OV (at 3 years).

### *Classification of survival using clinical and bulk transcriptomic data*

XGBoost (with sequential feature selection) was performed on the differentially expressed genes to augment prediction of STAD survival at 1 year, GBM at 1 year, and OV at 3 years. This was carried out for up to 25 genes.

### *Cox regression analysis for OS and disease-free survival*

Cox proportional-hazards analyses were conducted for both OS and disease-free survival (DFS) for all 16 cancer types



**Figure 1.** Survival prediction scores derived from clinicopathological features. (A) Stacked proportion bar plots of gene expression profiles of patients who survived less than or greater than 1 year, 3 years, and 5 years. (B) Heat maps depicting survival prediction scores of 1-year OS and 3-year OS using 5 different machine learning models: (1) XGBoost’s built-in imputation and XGBoost, (2) imputation with median and XGBoost, (3) imputation with median and Random Forest, (4) imputation with K-nearest neighbors and XGBoost, and (5) imputation with K-nearest neighbors and Random Forest. OS indicates overall survival.

using the top clinical features identified by our machine learning approach.

## Results

### Classification of OS for each of 16 cancer types using clinicopathological features

Cancer prognostication is commonly performed at the 5-year mark, but we sought to define appropriate time points

for survival classification using primarily clinicopathological features in our specific cohort of TCGA patients. As such, we evaluated the proportion of patients in each cancer type either alive or deceased/censored at 1, 3, and 5 years (Figure 1A). This revealed a well-balanced proportion of patients with either outcome at 1 year, but became progressively skewed toward death or censorship at years 3 and 5 for most cancers (Figure 1A). Our results are consistent with markedly poor prognoses for certain

cancer types such as pancreatic cancer, GBM, and STAD, each of which had fewer than 10% of patients alive after 5 years in our cohort. We therefore decided to focus our classification analysis at the 1- and 3-year mark to preclude inflated model performance that could conceivably result from skewing predictions toward 1 outcome at year 5 (eg, deceased/censored)

In total, we assessed 8608 patient tumor samples across 16 cancer types from TCGA (Table 1). As many values for predictive features were incompletely annotated, we carried out imputation strategies such as KNN, median, and the built-in XGBoost imputation function that leverages sparsity-aware split finding. Using these imputed and nonimputed data sets, we then evaluated the performance of our classification models with a cross-validated 80-20 train-test split, separately for each of the 16 cancer types at 1 and 3 years. This analysis was carried out separately for each cancer type as opposed to a one-size-fits-all approach, given that there are many features that are not uniformly annotated across malignancies. We used both prediction accuracy and AUROC to evaluate model performance

As a whole, we noticed a large variance across cancer types in the capacity of clinicopathological features to predict patient survival at 1 year and 3 years (Figure 1B). At both time points, OS for skin cutaneous melanoma (SKCM 1-year OS: AUROC=0.83; 3-year OS: AUROC=0.84), liver hepatocellular carcinoma (LIHC 1-year OS: AUROC=0.78; 3-year OS: AUROC=0.81), and pancreatic ductal adenocarcinoma (PAAD 1-year OS: AUROC=0.75; 3-year OS: AUROC=0.88) could be effectively classified, whereas that for GBM (1-year OS: AUROC=0.68; 3-year OS: AUROC=0.68), OV (1-year OS: AUROC=0.67; 3-year OS: AUROC=0.68), and BRCA (1-year OS: AUROC=0.69; 3-year OS: AUROC=0.69) consistently could not. Other cancers such as STAD were predictable at the 3-year time point but not at 1 year. We initially posited that the strong model performance for a highly lethal cancer such as pancreatic could be explained by the skewing of patients toward 1 outcome, but this does not explain why the model does not perform similarly well for other aggressive cancers such as GBM. Upon further examination, however, we observed that cancers devoid of well-annotated clinical or pathological markers (eg, GBM) had less predictable outcomes relative to those with organ/disease-specific features such as liver fibrosis and albumin levels (eg, LIHC) (Supplementary Tables 2 and 3). This suggests that cancer types with currently hard-to-predict survival times may benefit from a prognostication standpoint from the collection of additional putative clinical or pathological correlates. Finally, for some cancer types, we noted significant discrepancies between accuracy and AUROC, the latter of which factors in both the sensitivity and the specificity of the model. Thyroid cancer, for instance, contains the most accurate 1-year survival prediction model but only the 10th best performing model by AUROC. Similarly, GBM is the most accurately predicted cancer for 3-year survival but exhibits poor performance by AUROC (rank: 13/16).

### *Orthogonal assessment of prognostic features using Cox proportional-hazards models*

In addition to our machine learning approach, we conducted univariate Cox proportional-hazards analyses on each of the 16 cancer types to orthogonally assess the impact of prognostic features on both OS (Supplementary Tables 4 and 5) and DFS (Supplementary Tables 6 and 7). We included variables that have well-established prognostic implications (eg, basal-like/triple-negative subtype for breast cancer) and many other features that have been more seldom discussed, including those which were prioritized by our aforementioned machine learning algorithms. Indeed, our analyses revealed several clinicopathological features which greatly influence OS in their respective patient cohorts. Patients with breast cancer harboring the basal-like molecular subtype (often ER-, PgR-, HER2-),<sup>7</sup> as expected, were associated with poorer OS: hazard ratio=1.55; 95% confidence interval = 1.08-2.23;  $P < .05$ . Further of note, we observed that stage IV/T4a for colorectal, lymph node stage N3 for melanoma, and age at diagnosis for GBM were the most prominent associated features with poor survival (Supplementary Tables 4 and 5). Overall, our study provides a thorough map of prognostic features using both machine learning methods and Cox proportional-hazards models.

### *Integration of molecular data to enhance survival classification for poor-performing cancers*

Given the poor performance of our models for cancer types such as GBM, STAD, and OV, we posited that addition of more complex biomolecular features with continuous input values could enhance our ability to prognose patients where clinicopathological features were insufficient. This may become increasingly relevant as commercially available transcriptomic profiling or targeted gene expression panels are integrated into clinical oncology workflows. Targeted gene expression panels facilitate multiplexed measurements of expression of select genes without having to perform whole transcriptome (WTX) sequencing and are anticipated to enter clinical oncology workflows.<sup>8</sup> Unlike targeted mutation profiling, which often has a binarized outcome, gene expression panels can provide dynamic information on tissue subtypes and states and can form the basis of more complex and quantitative predictive models.

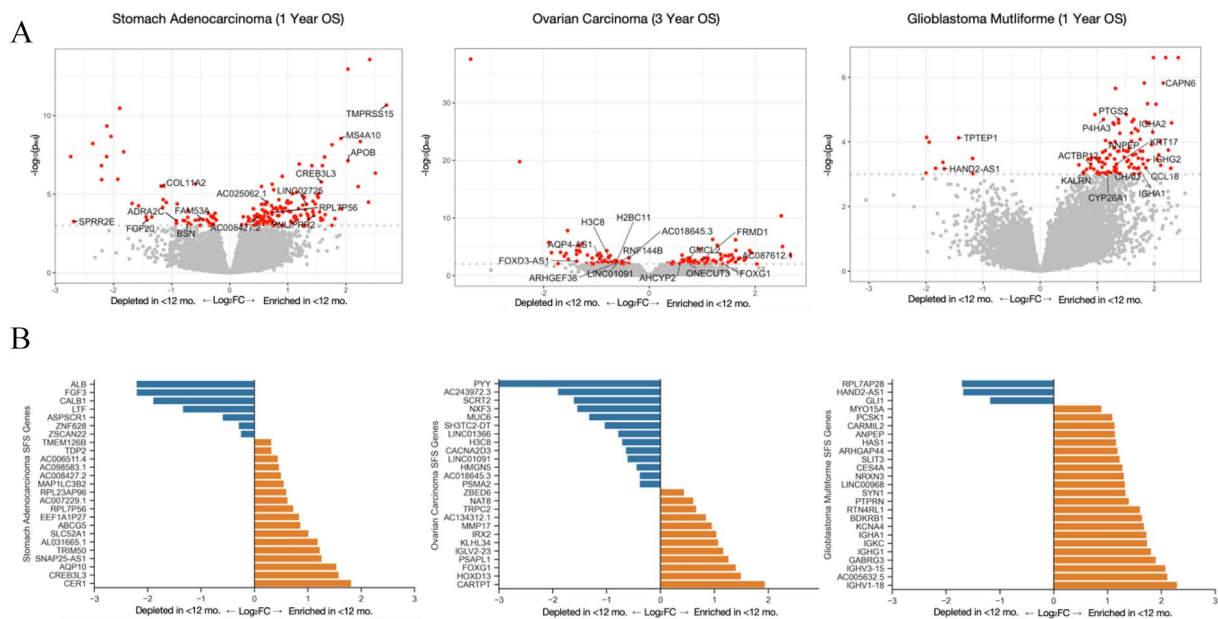
Beyond gene expression data, we also assessed the value of genomic sequence-level aberrations as classifier features, but this generally did not confer any significant advantage relative to clinicopathological information alone in the vast majority of cancer types examined (data not shown). As such, we focused our analysis on data derived from bulk RNA sequencing.

One challenge in using expression-level data for survival classification is the sheer number of genes, or features, across the entire transcriptome. As such, for each of GBM, STAD, and OV, we performed a differential expression (DE) analysis between

**Table 1.** Demographics of patient cohort assessed in this study.

	BLCA (N=413)	KIRC (N=537)	UCEC (N=548)	PAAD (N=176)	BRCA (N=1108)	LUSC (N=511)	LHC (N=440)	THCA (N=508)	COAD (N=636)	SKCM (N=471)	GBM (N=606)	HSNC (N=528)	STAD (N=478)	LUAD (N=548)	PRAD (N=500)	OV (N=600)
Age, y	68.0 (10.6)	60.6 (12.1)	63.9 (11.1)	64.8 (11.0)	58.4 (13.2)	67.3 (8.6)	59.5 (13.5)	47.2 (15.8)	66.2 (12.8)	58.1 (15.6)	57.7 (14.3)	61.0 (11.9)	65.6 (10.7)	65.3 (10.0)	61.0 (6.8)	59.6 (11.5)
Sex																
Male	304	346	0	98	12	373	255	136	335	290	366	386	285	242	500	0
Female	108	191	548	78	1085	131	122	371	294	180	230	142	158	280	0	587
NA	0	0	0	0	4	7	63	0	7	1	10	0	35	62	0	13
Race/Ethnicity																
White	327	466	374	153	757	351	187	334	296	447	507	452	278	393	147	498
B/AA	23	56	109	7	183	31	17	27	65	1	51	48	13	53	7	34
Asian	44	8	20	11	61	9	161	52	12	12	13	11	89	8	2	20
AI/AN	0	0	4	0	1	0	2	1	1	0	0	2	0	1	0	3
NH/PI	0	0	9	0	0	0	0	0	0	0	0	0	1	0	0	0
NA	18	7	32	5	99	120	73	94	262	11	35	15	97	129	344	44
Survival data																
OS	26.8	44.3	37.5	18.6	41.0	31.9	26.9	39.6	27.6	59.9	16.4	30.0	18.9	29.7	35.9	39.2

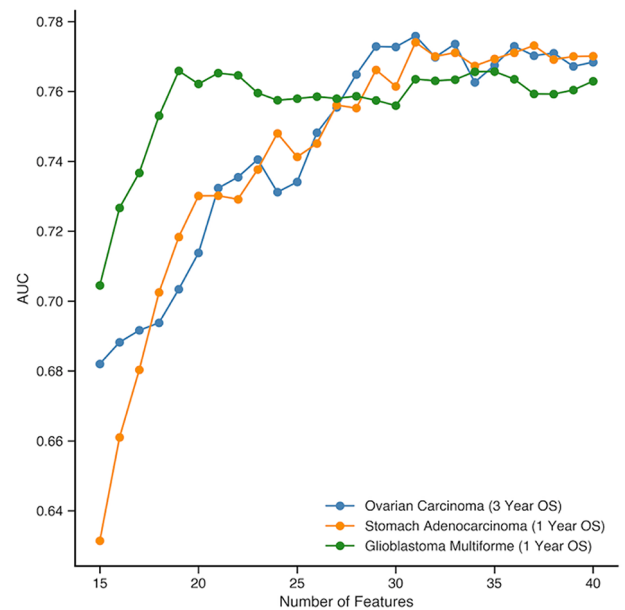
Abbreviations: AI/AN, American Indian or Alaska native; B/AA, Black or African American; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; COAD, colon adenocarcinoma; GBM, glioblastoma multiforme; HSNC, head and neck squamous cell carcinoma; KIRC, kidney clear cell carcinoma; LHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; NA, not applicable; NH/PI, Native Hawaiian or other Pacific Islander; OS, overall survival; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma; SKCM, skin cutaneous melanoma; STAD, stomach/gastric adenocarcinoma; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma.  
 Age, sex, race, and survival data were assessed across 16 different cancer types.  
 Categorical data are presented as n and continuous data as mean ± SD.



**Figure 2.** Gene expression classifiers associated with lower performing cancer subtypes. (A) Volcano plots of gene upregulated or downregulated in 3 different patient cohorts (1-year OS stomach/gastric adenocarcinoma, 3-year OS ovarian carcinoma, and 1-year OS glioblastoma multiforme). (B) Genes derived from the differential gene expression analysis and their corresponding  $\log_2$  fold change. OS indicates overall survival.

patients who survived above and below the respective time points we were interested in examining (Figure 2A). For instance, for STAD, a DE between patients surviving longer than 1 year and shorter than 1 year was performed, and genes with a  $P$  value of less than .001 were funneled into downstream analysis. This yielded genes enriched in the lower surviving comparator group that have been described in association with malignant and aggressive behavior; for instance, *PTGS2*, a prostaglandin synthase and cyclooxygenase, has been reported to facilitate resistance of GBM to radiotherapy, a commonly used treatment modality for this disease. Beyond genes that are upregulated in the poorer surviving group, there were also genes whose downregulation was associated with reduced OS (Figure 2B)

To observe whether integration of our DE analysis would improve survival prediction as evaluated by AUROC, we deployed a random-forest-embedded sequential forward selection (SFS) approach that begins with an initial subset of the 15 most predictive clinicopathological features (as already identified in the prior analysis) before sequentially appending the next most informative gene as an additional feature. We extended the SFS analysis for up to a total of 40 features to optimize for parsimony, split between 15 that were clinicopathological and 25 that were gene-expression-based. With clinicopathological features alone, the AUROCs for GBM, STAD, and OV survival prediction were 0.66, 0.69, and 0.67, respectively. However, following the inclusion of gene expression data, these increased to 0.76, 0.77, and 0.77, respectively (Figure 3). We therefore conclude that bulk gene expression data could serve to complement clinical features in predicting both 1-year and 3-year OS across numerous cancer types



**Figure 3.** Gene expression classifiers improve survival prediction scores. Line graph showing successive increases to AUROC as the 25 most relevant genes are included in the sequential feature selection algorithm. AUROC indicates area under the receiver operating curve; OS, overall survival.

## Discussion

Prediction of cancer survival for individual patients is challenging given the large number of features that contribute to intrinsic tumor behavior, treatment response, and the patient's ability to tolerate therapy or withstand tumor burden. However, using random forest and gradient boosting machine learning approaches, we model and predict OS across each of 16 cancer

types using both clinical features and bulk transcriptomic data. This is advantageous to performing a pan-cancer classification in aggregate, given that some of our highest performing features are highly cancer-specific (eg, liver fibrosis score in liver cancer). Indeed, performing an alternative aggregated analysis would have precluded the identification of robust predictors for individual cancer types. Overall, our results suggest that clinicopathological features alone can robustly predict 1-year and 3-year OS in a cancer-type-dependent manner, and that for more poorly performing cancers, the integration of gene expression can improve classification performance. However, it should be noted that there were discrepant results between prediction accuracy and AUROC for certain cancer types, such as GBM at 3 years and THCA at 1 year. To facilitate interpretation of these results, we suggest that greater emphasis be placed on the AUROC results for each cancer type, given that this metric takes into account both the sensitivity and specificity of the model. Indeed, it is possible for a classifier to achieve arbitrarily high prediction accuracies if the data are disproportionately skewed toward a particular outcome, which we posit occurred with the inflated prediction accuracy for GBM at 3 years. These challenges, such as overfitting models and having unbalanced outcomes groups, prompted us to further perform an orthogonal analysis of similar features with a Cox proportional-hazards model. Taken together, we believe this provides a broad atlas of prognostic features within the various TCGA patient cohorts.

For some cancer types, our survival classifiers exhibit similar performance as other previously described machine learning methods, albeit leveraging distinct data sets. As an example, for GBM, prior studies have used radiomic features and machine learning to predict OS for long survivors (>900 days), short survivors (<300 days), and mid-survivors (300-900 days), with AUROC scores of 0.784, 0.817, and 0.709, respectively.<sup>9</sup> Similarly, mean kurtosis (MK) and mean diffusivity (MD), both derived from diffusion kurtosis imaging (DKI), predict 2-year survival of patients with GBM with AUROC scores of 0.841 and 0.772, respectively.<sup>10</sup> For OV, a gradient boosting machine learning model was developed to predict survival of epithelial ovarian carcinoma and produced an AUROC score of 0.830 when validated on an external cohort.<sup>11</sup> Finally, for gastric adenocarcinoma, prognostic classifiers derived from immunohistochemistry biomarkers and support vector machine (SVM)-based methods have been shown to predict 5-year OS with an AUROC score of 0.834.<sup>12</sup> It remains to be seen how features used in other studies may be combined with those in ours to generate better models overall.

Although prior studies have reported the use of machine learning to predict survival for individual cancer types,<sup>13,14</sup> few to our knowledge have systematically deployed consistent approaches on as many disease sites as we have, thereby preventing a robust understanding of the diseases in which clinical and pathological features are capable of predicting patient outcomes and which still require more investigation. In light of this, we performed the analyses on clinicopathological and

biomolecular features separately, which allowed us to decipher the value of more accessible features first before determining the added benefit of more complex information.

One limitation of the study was the availability of data that could be inputted into our model. Although there were many other features that theoretically could have served as model inputs, incomplete annotation in a large proportion of patients precluded their use. Furthermore, while the integration of gene expression information to supplement clinicopathological features did provide substantial benefit to prediction for multiple cancer types, there was still significant room for improvement. This suggests that the complex nature of cancer behavior and patient prognosis cannot be entirely encapsulated through bulk transcriptomics alone, and it remains to be seen whether the integration of alternative modes of molecular profiling such as mutations, DNA methylation, and chromatin accessibility and even spatial/radiographic resolution of these biomarkers can enhance prognostication. Ultimately, however, the benefit of prognostication lies in better insights for clinical management and planning for patients. With more complete clinicopathological annotations and the advent of accessible technological advances, we may be poised to move beyond binarizing patients by their likelihood of survival at 5 years after diagnosis.

## Conclusions

Cancer prognostication is important for guiding clinical management and treatment, but it is challenging due to the sheer number of features that can affect outcomes. In this study, we use machine learning tools to characterize and examine the relative importance of clinical, pathological, and gene expression features in predicting OS across 16 different cancer types. For some cancer types like PAAD and SKCM, we observe that clinical features alone are strong predictors of OS with average 3-year AUROC scores of 0.88 and 0.84, respectively. In contrast, for other cancer types like GBM, STAD, and OV, clinical features are not robust predictors of OS (AUROC scores of 0.66, 0.69, and 0.67, respectively), and additional transcriptomic information is incorporated to improve survival prediction. With the inclusion of the 25 gene-expression-based features, AUROC scores for the 3 cancer types increased to 0.76, 0.77, and 0.77, respectively, emphasizing the prognostic impact of transcriptomic information. Finally, using an orthogonal approach, we leverage Cox proportional-hazards models to assess the impact of prognostic features on OS and DFS. Through this study, we provide a robust and pragmatic approach that uses routinely collected clinical oncology data and provides accurate survival classification at the individual patient level.

## Author Contributions

J.A.G., D.Z., and P.C. developed the concept and designed the study. D.Z. and D.Y.K. collected data from cBioPortal. D.Z., P.C., P.Y., and S.H. developed machine learning models and analyzed the RNA-seq data. J.A.G., D.Z., and P.C. guided and

performed statistical analyses. D.Z., P.C., and P.Y. generated figures and tables. J.A.G. and Y.R.L. supervised the research. J.A.G., D.Y.K., and C.Z. wrote the manuscript. All authors reviewed the manuscript.

## ORCID iD

Daniel Y Kim  <https://orcid.org/0000-0002-6273-7310>

## Availability of Data and Materials

All data is publicly available on cBioPortal (<https://www.cbioportal.org/>) and GDC Data Portal (<https://portal.gdc.cancer.gov/>).

## Supplemental Material

Supplemental material for this article is available online.

## REFERENCES

- Xu W, Xu M, Wang L, et al. Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. *Signal Transduct Target Ther*. 2019;4:55. doi:10.1038/s41392-019-0081-6.
- Zheng C, Xu R. Predicting cancer origins with a DNA methylation-based deep neural network model. *PLoS ONE*. 2020;15:e0226461. doi:10.1371/journal.pone.0226461.
- Jiao W, Atwal G, Polak P, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat Commun*. 2020;11:728. doi:10.1038/s41467-019-13825-8.
- Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17. doi:10.1016/j.csbj.2014.11.005.
- Cerami E, Gao J, Dogrusoz U, et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2:401-404. doi:10.1158/2159-8290.CD-12-0095.
- The Cancer Genome Atlas—Cancers Selected for Study. National Cancer Institute, n.d. <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers>.
- Badowska-Kozakiewicz AM, Budzik MP. Immunohistochemical characteristics of basal-like breast cancer. *Contemp Oncol (Poznan, Poland)*. 2016;20:436-443. doi:10.5114/wo.2016.56938.
- Hayette S, Grange B, Vallee M, et al. Performances of targeted RNA sequencing for the analysis of fusion transcripts, gene mutation, and expression in hematological malignancies. *Hemasphere*. 2021;5:e522. doi:10.1097/HS9.0000000000000522.
- Baid U, Rane SU, Talbar S, et al. Overall survival prediction in glioblastoma with radiomic features using machine learning. *Front Comput Neurosci*. 2020;14:61. doi:10.3389/fncom.2020.00061.
- Zhang J, Jiang J, Zhao L, et al. Survival prediction of high-grade glioma patients with diffusion kurtosis imaging. *Am J Transl Res*. 2019;11:3680-3688.
- Paik ES, Lee JW, Park JY, et al. Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. *J Gynecol Oncol*. 2019;30:e65. doi:10.3802/jgo.2019.30.e65.
- Jiang Y, Xie J, Han Z, et al. Immunomarker support vector machine classifier for prediction of gastric cancer survival and adjuvant chemotherapeutic benefit. *Clin Cancer Res*. 2018;24:5574-5584. doi:10.1158/1078-0432.CCR-18-0848.
- Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak*. 2019;19:48. doi:10.1186/s12911-019-0801-4.
- She Y, Jin Z, Wu J, et al. Development and validation of a deep learning model for non-small cell lung cancer survival. *JAMA Netw Open*. 2020;3:e205842. doi:10.1001/jamanetworkopen.2020.5842.