Learning Individual Survival Models from PanCancer Whole Transcriptome Data

Neeraj Kumar[1], Daniel Skubleny[2], Michael Parkes[3], Ruchika Verma[1], Sacha Davis[1], Luke Kumar[4], Amira Aissiou[5], and Russell Greiner[1,3]

ABSTRACT

Purpose: Personalized medicine attempts to predict survival time for each patient, based on their individual tumor molecular profile. We investigate whether our survival learner in combination with a dimension reduction method can produce useful survival estimates for a variety of patients with cancer.

Patients and Methods: This article provides a method that learns a model for predicting the survival time for individual patients with cancer from the PanCancer Atlas: given the (16,335 dimensional) gene expression profiles from 10,173 patients, each having one of 33 cancers, this method uses unsupervised nonnegative matrix factorization (NMF) to reexpress the gene expression data for each patient in terms of 100 learned NMF factors. It then feeds these 100 factors into the Multi-Task Logistic Regression (MTLR) learner to produce cancer-specific models for each of 20 cancers (with >50 uncensored instances); this produces "individual survival distributions" (ISD), which provide survival probabilities at each future time for each individual patient, which provides a patient's risk score and estimated survival time.

Results: Our NMF-MTLR concordance indices outperformed the VAECox benchmark by 14.9% overall. We achieved optimal survival prediction using pan-cancer NMF in combination with cancer-specific MTLR models. We provide biological interpretation of the NMF model and clinical implications of ISDs for prognosis and therapeutic response prediction.

Conclusions: NMF-MTLR provides many benefits over other models: superior model discrimination, superior calibration, meaningful survival time estimates, and accurate probabilistic estimates of survival over time for each individual patient. We advocate for the adoption of these cancer survival models in clinical and research settings.

Introduction

Cancer is a molecular disease with an estimated 19.3 million new cases and 10.0 million deaths in 2020 (1). Personalized medical oncology aims to improve cancer treatment by acknowledging that each patient's tumor has a distinct molecular profile, leading to its own cancer prognosis and treatment response (2, 3). Accurately estimating a patient's survival time could support end-of-life decision making, and inform personalized treatment plans. In support of personalized care, individual survival distributions (ISD) describe survival probabilities at all future time points for an individual patient (e.g., **Fig. 3A**). This article describes a way to learn survival models that accurately estimate a patient's ISD from transcriptomic data, learned from many different types of tumors.

Our model produces a patient's ISD based on the whole-transcriptome RNA-sequencing data from their tumor biopsy. Bulk tissue sequencing has influenced several aspects of cancer research, including biomarker discovery (4, 5), characterization of molecular mechanisms of drug resistance (6), tumor heterogeneity (7), and differences in survival outcomes among molecular subtypes of cancers (8, 9). Our study uses data from the PanCancer Atlas Initiative of The Cancer Genome Atlas (TCGA), which is one of the most comprehensive biomolecular cancer datasets including biopsies from 33 cancer types for 10,173 patients (10).

Each patient in the PanCancer dataset is represented by the expression levels of 16,335 genes. In such datasets, where the number of features (genes) is far larger than the number of training instances (biopsies), simple learning procedures often produce models that perform poorly, as it can be difficult to establish meaningful relationships between features and outcomes. This motivated us to seek low-dimensional representations of our high-dimensional data (11). Although cancer is typically characterized as many different diseases (each associated with its organ of origin), the modern hallmarks of cancer perspective view neoplasia as a disease involving numerous complex biologic disruptions, which may be common across organs (12). This viewpoint motivated us to derive a low-dimensional "pan-cancer" representation that captures information about all 33 cancers in the PanCancer dataset, and to use this representation for ISD modeling.

Our NMF-MTLR approach first uses the unsupervised nonnegative matrix factorization (NMF) algorithm to derive the low-dimensional representation of PanCancer gene expression data (**Fig. 1**)—a common approach in bioinformatics (11, 13, 14). NMF decomposed the entire PanCancer gene expression data matrix (16,355 genes by 10,173 biopsies from 33 cancers) into two nonnegative matrices—one gene-specific (16,355 by 100) and the other patient-specific (100 by 10,173; see Materials and Methods section). The gene-specific matrix expresses how each gene contributes to each of a small set of 100 "factors" (each often viewed as a metagene; ref. 15). A factor is a linear combination of genes; a gene with a high-magnitude coefficient relates strongly to that factor. The patient-specific matrix expresses each patient in terms of those 100 factors. Our approach then uses the

[1]Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada. [2]Department of Surgery, University of Alberta, Edmonton, Alberta, Canada. [3]Computing Science Department, University of Alberta, Edmonton, Alberta, Canada. [4]Microsoft, Vancouver, British Columbia, Canada. [5]Alberta Health Services, Edmonton, Alberta, Canada.

Corresponding Author: Russell Greiner, Athabasca Hall, University of Alberta, Edmonton, Alberta T6G2E8, Canada. Phone: 587-415-9622; Fax: 780-492-1071; E-mail: rgreiner@ualberta.ca

Clin Cancer Res 2023;XX:XX–XX

doi: 10.1158/1078-0432.CCR-22-3493

Translational Relevance

This article describes a new way to predict a patient's survival time, using a model learned from the PanCancer Atlas transcriptome data: we first encode a patient's transcriptome as 100 learned "factors," then use these factors to learn cancer-specific models that produce individual survival distributions (ISD)— each a personalized survival curve, which is more flexible than traditional Cox regression or nomograms. Even though the factors were derived without influence from the patients' survival status, we demonstrate that they also reflect biologically coherent phenomena, and thus they may be relevant for future biomarker discovery. At the bedside, ISDs have the potential to provide accurate, intuitive, and visual predictions that can facilitate collaborative and informed decision making between patients and physicians.

Multi-Task Logistic Regression (MTLR; ref. 16) to train 20 different cancer-specific ISD models using that cancer's biopsies' 100 factor scores as features.

This article shows (i) how to learn meaningful organ-specific ISDs from PanCancer whole transcriptome data alone; (ii) that survival models based on the low-dimensional pan-cancer NMF representations are better than the models learned instead from cancer-specific representations; (iii) that ISDs learned from pan-cancer NMF representations using MTLR (ref. 16; NMF-MTLR) outperform the benchmark variational autoencoder (VAE)–based VAECox model (17) for most cancers; and (iv) that the NMF representation has biologically coherent interpretations in the context of patient survival. We also demonstrate advantages ISDs hold over other common modeling strategies.

Materials and Methods

Dataset description

We obtained pan-cancer mRNA count data from the NIH Genomic Data Commons PanCanAtlas (18): TCGA IlluminaHiSeq mRNA expression profiles of 20,531 genes in 33 cancer types from tumor biopsies of 10,173 patients. We removed all genes with any missing values, leaving 16,335 genes that were used for unsupervised representation learning by NMF and for unsupervised pre-training of VAE.

Survival data were obtained from the TCGA-Clinical Data Resource (CDR) Outcome data file (10). Please note that we used overall survival to be consistent with the prior benchmark publication (17). Furthermore, other survival metrics in the dataset, such as disease-free survival or progression-free interval, were identical to overall survival data or contained missing data, respectively. For learning cancer-specific survival models using the low-dimensional pan-cancer representation, we used only the 20 cancers that included at least 50 uncensored individuals. This provided sufficient uncensored outcomes to compute quantitative measures of survival prediction performance. Supplementary Table S1 shows the number of patients, censoring status, and whether they were used for survival prediction. For complete clinical information of all patients in 33 cancer types, please see ref. 10.

We retained only those patients from the 20 cancer types for whom both the mRNA expression profiles and survival information were available. Thus, our final data matrix had mRNA expression profiles of 7,920 patients, each with 16,335 genes, from 20 cancer types. **Table 1** summarizes the clinical characteristics of patients included in this final data matrix. We created ten cross-validation folds separately within each cancer-type, balanced to ensure that event times and censoring rates are consistent across all folds, for reporting average cross-validation survival prediction results in this article.

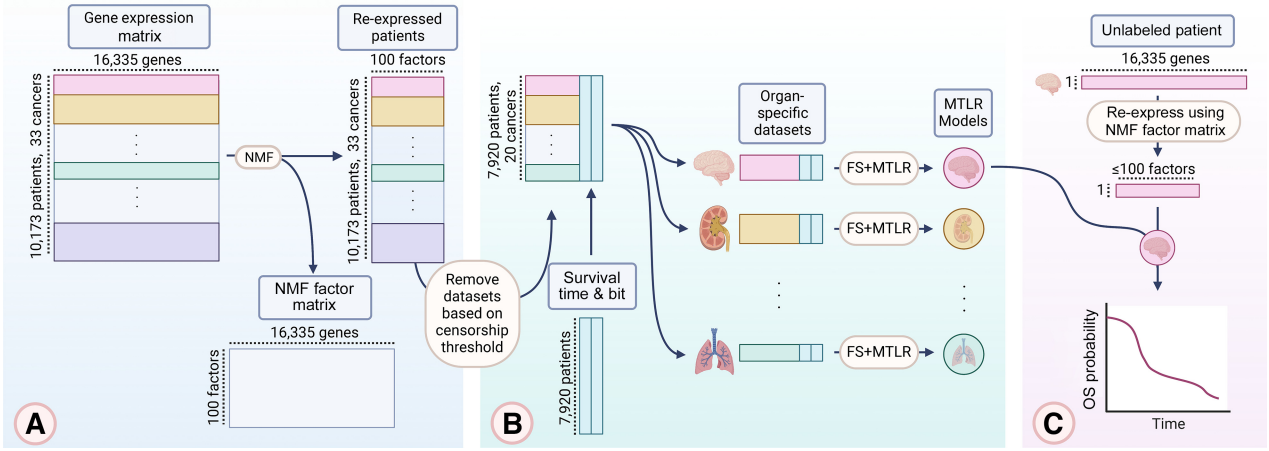


Figure 1.

Our proposed two-step process for individual survival prediction. **A,** The first step used NMF (57) to derive a low-dimensional pan-cancer representation of each patient in the PanCancer dataset. **B,** The second step trained cancer-specific MTLR (16) survival models (one model per cancer) that predicted a patient's survival distribution using the low-dimensional pan-cancer representations from the first step. Following VAECox (17), to ensure there were enough uncensored patients to evaluate model performance, we restricted survival modeling to the 20 cancers that each included at least 50 uncensored patients (Supplementary Table S1). This also involved a Feature Selection step, to report those 100 factors to a learned smaller cancer-specific subset (see Materials and Methods section for details). **C,** We can use the learned model to produce a survival curve for a novel patient, by first reexpressing that patient's 16,355 gene values in terms of the PanCancer representation of 100 factors; this uses the "NMF Factor Matrix" from **A**. We then extract just the subset of factors associated with this cancer, and apply this to the learned MTLR model associated with this cancer. The Materials and Methods section describes how we can use these learned models in general, and other issues related to evaluation. Images created with BioRender.com.

Table 1. Patient characteristics of TCGA PanCancer data.

Cancer Type[a]	No. of patients (Uncensored %)	Sex[b] M/F	Mean age ±std	Race White/Black/ Other/NA	Stage I/II/III/IV/NA	Grade 1/2/3/4/NA
BLCA	407 (43.7%)	300/107	68.1 ± 10.6	323/23/44/17	2/130/140 /133/2	0/0/0/0/407
BRCA	1,094 (13.8%)	12/1,082	58.4 ± 13.2	755/182/62/95	181/620/249/20/24	0/0/0/0/1,094
CESC	304 (23.4%)	0/304	48.2 ± 13.8	209/30/29/36	162/69/45/21/7	18/135/118/1/32
COAD	449 (22.7%)	236/213	66.8 ± 13.0	213/59/12/165	74/175/125/64/11	0/0/0/0/449
ESCA	184 (41.9%)	158/26	62.5 ± 11.9	113/5/46/20	18/82/62 /16/6	19/76/49/0/40
GBM	159 (80.5%)	103/56	59.5 ± 13.6	142/11/5/1	0/0/0/0/159	0/0/0/0/159
HNSC	520 (42.3%)	384/136	60.9 ± 11.9	445/48/13/14	27/83/94/316/0	62/304/125/7/22
KIRC	533 (32.8%)	345/188	60.6 ± 12.1	462/56/8/7	267/57/123/83/3	14/229/206/76/8
LAML	161 (64.0%)	87/74	55.8 ± 16.3	147/10/2/2	0/0/0/0/161	0/0/0/0/161
LGG	514 (24.3%)	285/229	42.9 ± 13.4	474/21/9/10	0/0/0/0/514	0/248/265/0/1
LIHC	370 (35.1%)	249/121	59.4 ± 13.5	184/17/159/10	171/85/85/5/24	55/177/121/12/5
LUAD	507 (36.1%)	236/271	65.4 ± 9.9	389/52/9/57	272/120/81/26/8	0/0/0/0/507
LUSC	495 (42.8%)	366/129	67.2 ± 8.6	348/29/9/109	242/159/83/7/4	0/0/0/0/495
MESO	86 (84.9%)	70/16	63.1 ± 9.8	84/1/1/0	10/16/44/16/0	0/0/0/0/86
OV	304 (60.5%)	0/304	59.2 ± 11.0	253/26/15/10	1/21/242/38/2	1/34/260/1/8
PAAD	178 (52.3%)	98/80	64.6 ± 10.9	157/6/11/4	21/147/3/4/3	31/95/48/2/2
SARC	259 (37.8%)	118/141	60.7 ± 14.6	226/18/6/9	0/0/0/0/259	0/0/0/0/259
SKCM	454 (46.9%)	281/173	58.3 ± 15.8	432/0/12/10	77/140/169/22/46	0/0/0/0/454
STAD	411 (38.4%)	262/149	65.7 ± 10.7	258/12/86/55	54/122/168/42/25	12/144/246/0/9
UCEC	531 (16.4%)	0/531	63.7 ± 11.0	360/106/33/32	330/50/122/29/0	99/119/302/0/11
Total	**7920 (34.9%)**	**3,590/4,330**	**60.4 ± 13.9**	**5974/712/571/663**	**1,909/2,076/1,835/842/1,258**	**311/1,561/1,740/99/4,209**

Abbreviation: NA, not available.
[a]We used TCGA study abbreviations for each cancer type; see Supplementary Table S1 for details.
[b]Sex is labeled as gender in TCGA source, but is better characterized as biological sex.

To facilitate interpretation of NMF factors, clinical features of interest were parsed from a variety of sources including TCGAbiolinks, maftools, cBioPortal, Genomic Data Commons and Supplementary Material of published PanCancer Atlas papers (see Supplementary Section E, Data File 1 for detailed list of datasets and variables used; refs. 18–21).

NMF

We used NMF to learn our pan-cancer low-dimensional representation from the input mRNA matrix $X \in R_+^{m \times n}$, where $m = 16{,}335$ (respectively, $n = 10{,}173$) is the number of genes (respectively, patients) and the subscript "+" indicates that all entries in this input matrix are positive real value (note this just uses the gene expression values, but not the label of each patient; hence it is considered "unsupervised"). We scaled mRNA expression values to be within (0,1) before applying NMF. NMF decomposes the mRNA expression matrix into two nonnegative matrices as $X \approx WH$, such that $W \in R_+^{m \times r}$ is the gene-specific factor, $H \in R_+^{r \times n}$ is the patient-specific factor, and r is the factorization rank (22). Each column of W represents a meta-gene (15), which is a nonnegative linear combination of m genes, while each column of H is a r-dimensional (note $r \ll m$) latent representation of the corresponding patient. For details regarding our NMF calculation, see Supplementary Methods.

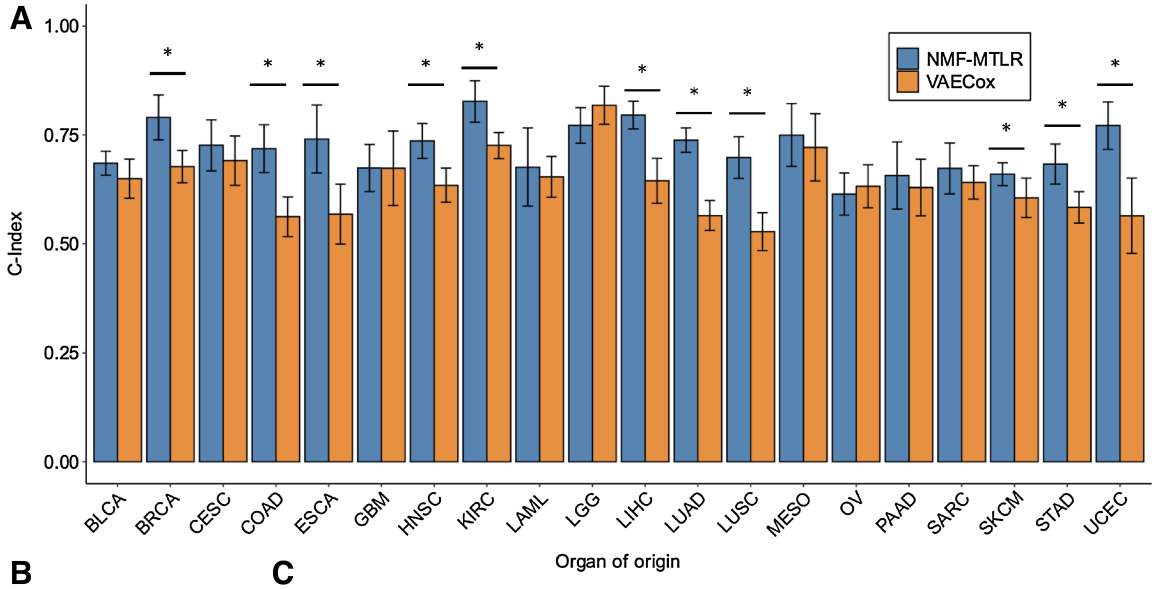
In our experiments, our analysis selected the best NMF rank as $r = 100$. Thus, our NMF of pan-cancer mRNA data encoded each patient with 16,335 genes expression values as a 100-dimensional column vector, so the set of n patients is $H \in R_+^{100 \times n}$.

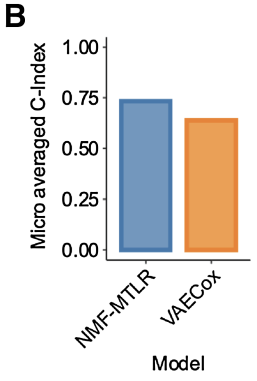
MTLR

After computing the pan-cancer NMF representation, we trained cancer-specific MTLR models to compute a patient's ISD using their 100-dimensional feature vector as input. Specifically, for each of the 20 cancer types selected for survival prediction (Supplementary Table S1), we first performed feature selection (in-fold) with a univariate Cox proportional hazards (Cox-PH) model by using 100-dimensional feature vectors of all patients of that cancer type, obtained from the patient-specific factor ($H \in R_+^{100 \times n}$) of previously computed pan-cancer NMF, as input followed by training a cancer-specific MTLR model for that cancer (we did this separately for each of the 20 cancer types). Note that the NMF representations were used as-is, without fine-tuning each of the 20 cancer-specific survival models (in contrast to variational autoencoder) and that every NMF feature with $P < 0.1$ in univariate Cox-PH models was included as input for ISD models. Because NMF feature selection was completed prior to NMF-based survival prediction, the different ISD models (NMF–MTLR, NMF–RSF, and NMF-CoxKP) all used the same NMF features. A brief overview of the MTLR algorithm is presented in the Supplementary Methods section; further technical details can be followed from ref. 16, and an R implementation of the algorithm can be found at https://github.com/haiderstats/MTLR (23).

VAECox

In general, VAE is a deep-learning–based generative model for learning the low-dimensional latent space distributions from the high-dimensional input data (24). Kim and colleagues combined a VAE with the Cox-PH model to learn cancer-specific survival prediction models (VAECox) from the TCGA PanCancer mRNA data (17). For a fair comparison with our NMF-MTLR approach, we reimplemented Kim and colleagues' algorithm using the VAECox architecture with the (hyper-) parameter settings as described in their original article (17). Details of this implementation are in the Supplementary Materials and Methods section. Please note that (i) this output is a single value for each patient, designed to optimize C-index (see below); and (ii) the VAE representations were fine-tuned separately for each cancer during supervised training of the cancer-specific VAECox





C

Micro-averaged concordance indices comparing pan-cancer versus cancer-specific perspectives for personalized survival prediction

		Non-negative matrix factorization	
		Pan-cancer	Cancer-specific
Multi-task logistic regression	PanCancer	0.681	0.587
	Cancer-specific	**0.733 (NMF–MTLR)**	0.618 (NMF-1-MTLR)

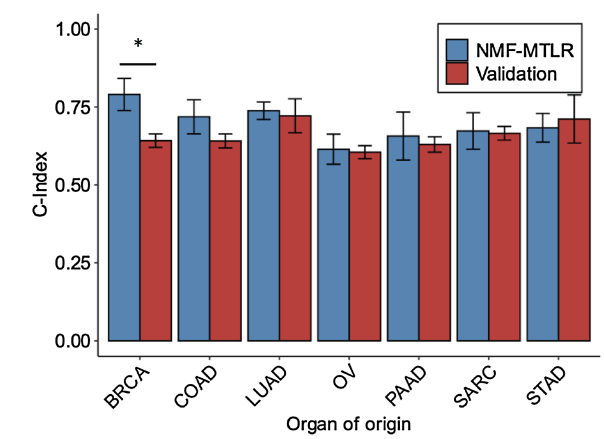


Figure 2.

NMF–MTLR survival prediction provides superior C-index. **A,** Comparing average 10-fold cross-validation C-index of NMF-MTLR and VAECox models for each of the 20 cancer types. Error bars represent 95% confidence intervals (CI). The significance of a paired two-sided *t* test is denoted by * $P < 0.05$. **B,** Comparison of micro-averaged C-index of NMF–MTLR versus the reference VAECox. **C,** Micro-averaged C-index comparison of the models developed using pan-cancer versus cancer-specific NMF dimensionality reduction and multi-task logistic regression techniques. **D,** Comparing average 10-fold cross-validation C-index of NMF–MTLR models to C-index of 5 imputed external validation datasets. Error bars represent 95% CIs. The significance of a two-sided Wilcoxon test is denoted by * $P < 0.05$.

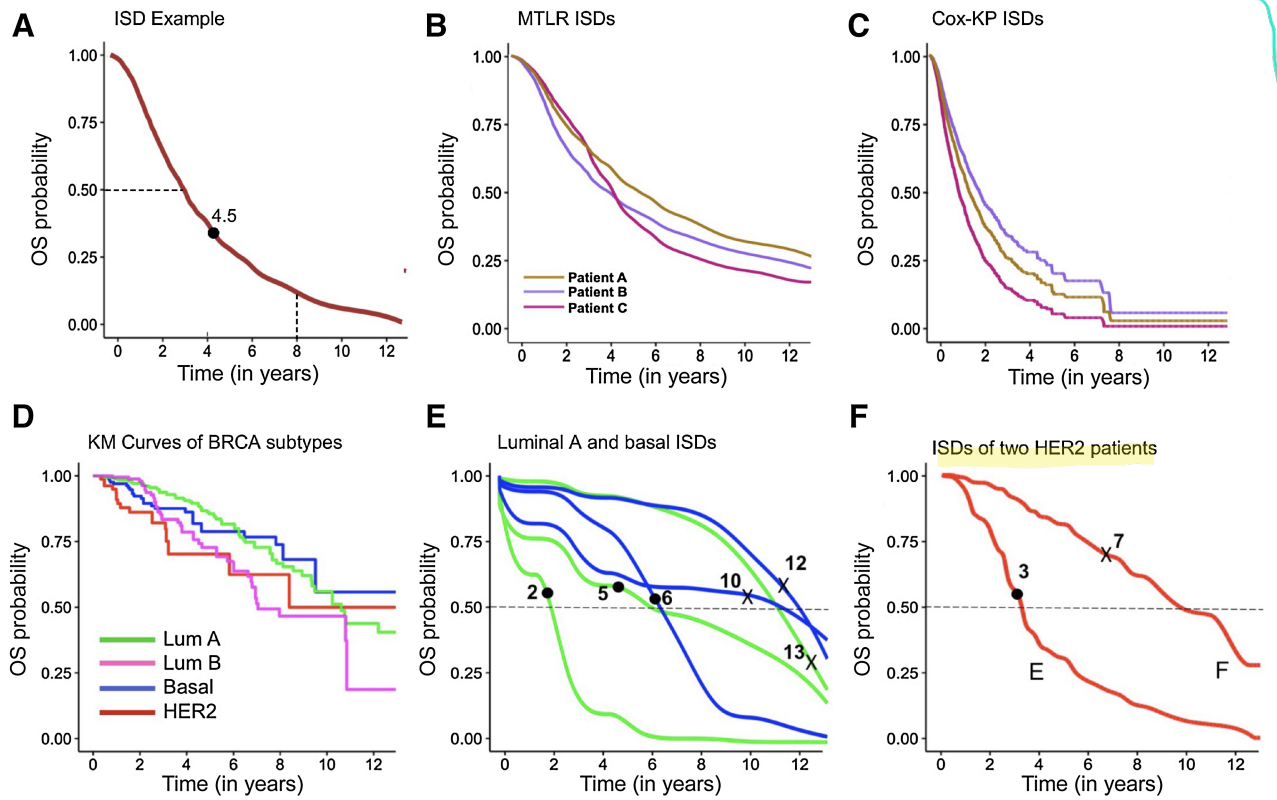


Figure 3.

NMF-MTLR survival prediction provides superior C-index and distinct functional advantages over competing survival models. **A,** ISD of an example patient in the PanCancer dataset computed using our NMF-MTLR model. Some useful statistics that can be computed from this ISD: the intersection of a patient's ISD with the median probability line (the horizontal line at 0.5) is the median survival time for that patient (here, around 3.9 years); the intersection of an ISD with the vertical line at a given time-point (e.g., at 8 years) provides the survival probability at that specific time-point (here, 0.18). To compute the risk score for C-index computations, we used the negative of the predicted median survival time—which is -3.9 here. **B,** ISDs of three example patients computed using the MTLR model; notice these curves have different shapes and cross over each other while ISDs of the Cox-KP model (**C**) do not cross over one another, because of Cox's proportional hazards assumption. **D,** KM curves for 4 different PAM50 subtypes of breast cancer. **E,** ISDs of three PAM50 Luminal A (green lines) and three PAM50 Basal (blue lines) patients show inter- and intra-subtype heterogeneity in survival outcomes. [Note the green line in **D** is the average over all Luminal A patients (including the 3 shown in **E**), and **D**'s blue line is the average over all Basal.] Note that we indicate when an uncensored patient actually died with a solid circle point, the censoring time for censored patients with "x," and the point where the horizontal dotted line (at probability = 0.5) intersects the ISDs in **E** and **F** represents the predicted median survival time. **F,** Individual survival distributions of two stage II, HER2$^+$ patients who were treated with trastuzumab: patient E died after 3 years but F lived for at least (i.e., was censored at) 7 years from their respective diagnosis dates. OS, overall survival.

models by using cancer-specific survival information. We also combined VAE with MTLR to implement a VAE-MTLR model for comparison. Noticeably, VAECox (and by extension VAE-MTLR) did not have a feature selection step.

Survival Prediction Performance Evaluation

Concordance-index

Concordance-index (C-index) is one of the most widely used metrics to report the performance of survival prediction algorithms (25). For reference, a description is placed in the Supplementary Methods. The C-index value ranges between 0 and 1, with a higher value indicating better model performance. In this article, we reported an average C-index of 10-fold cross-validation (10 CV) for each cancer-type; we also fixed the patients in each fold (see data description section) to avoid sampling bias for fair comparison of survival prediction models.

Micro-averaged C-index

To combine the average 10CV C-indices of 20 cancer-types, we computed the micro-averaged C-index,

$$\frac{\sum_{i=1}^{20} n_i c_i}{\sum_{i=1}^{20} n_i},$$

where n_i is the number of patients and c_i is the average 10CV C-index for ith cancer type. We will use this micro-averaged C-index wherever we need one number to compare the average performance of a survival approach, across all 20 cancer-types.

Calibration

In our experiments, we assessed the calibration of ISD computing models (NMF-CoxKP, NMF-RSF, and NMF-MTLR) using Distribution calibration (D-Calibration; ref. 26). Because VAECox and NMF-Cox models estimated time-invariant risk scores instead of ISDs, we assessed their calibration using calibration plots (1-calibration measures;

ref. 27). These calibration measures are described in detail in the Supplementary Methods.

External validation

Transcriptome data from seven cancer types with overall survival data were identified from four external datasets (Supplementary Table S4). To account for technological differences between mRNA platforms, external transcriptome data were normalized to the reference pan-cancer data distribution using Feature Specific Quantile Normalization (28). Missing genes from external datasets were then imputed using predictive mean matching with the multivariate imputation by chained equations package in R (29). All datasets were imputed five times with five iterations each. The transcriptome data of each patient was then represented as a 100-dimensional vector in the patient-specific NMF matrix by keeping the learned unsupervised NMF pan-cancer gene-specific matrix fixed. The NMF representation of each cancer type from external datasets were next used as input to the cancer specific MTLR models to generate a C-index for comparison.

Interpretation of NMF Factors

Random forest models

For biological interpretation of the NMF factors (**Fig. 4**), we used the Random Forest model implemented as "rf" with "caret" in R (30). Models were trained using 5-fold cross-validation with default tuning parameters. We used mean accuracy and Cohen Kappa to assess model performance (31). NMF factors were deemed to contribute relevant information to a given biological characteristic if model accuracy was statistically greater than the no information rate (NIR) using a one-sided binomial test ($\alpha > 0.05$). The NIR is defined as the largest proportion of observed classes, which is the accuracy of the "constant" classifier that predicts the majority class for any instance.

Generalized additive models

A Generalized Additive Model (GAM) was constructed to assess the nonlinear association of NMF factor expression with median survival times predicted by the benchmark MTLR models. See the Supplementary Materials and Methods section for details of model construction. The GAM was developed using the "mgcv" package in R (32). We fit each continuous variable, including age and the 100 NMF factors, using penalized cubic regression splines. Smoothing parameters were estimated using restricted maximum likelihood (33). Concurvity and model diagnostics were assessed for the final GAM model (Supplementary Data S3).

Relative NMF factor gene expression with limma

The R package "limma" was used to identify top genes associated with each of the 100 NMF factors from the pan-cancer model (34). The expression level of each gene was modeled as a function of NMF patient-factor coefficient for the factor of interest, using $\log_2(1 +$ count) of the normalized RNA-seq count data. Thus, the models give change in expression of a gene per unit change in the factor's coefficient. The top 500 genes were ranked by B-statistic, which is the posterior odds of differential expression.

Gene-set enrichment analysis

Gene-set enrichment analysis (GSEA) from the "clusterProfiler" R package was used to assess the biological function of NMF factors (35–37). The Hallmarks gene set from the Molecular Signature Database (MsigDb; ref. 38) used a list of all 16,335 genes ranked by the $\log_2$ fold change (FC) from our differential expression analysis using limma. GSEA uses the numeric value of differential expression, which allows detection of small but coordinated signals related to a particular functional gene set. The Benjamini–Hochberg method for multiple comparisons correction adjusted P values to a cutoff <0.05.

Gene ontology analysis

The online Database for Annotation, Visualization, and Integrated Discovery (DAVID) was used to identify Gene Ontology Biological Process (GO-BP) annotations of genes that were associated with each of the NMF factors. We compiled separate lists of overexpressed and underexpressed genes that were among the top 500 genes from the limma analysis for each factor of interest and supplied their ENTREZ IDs to DAVID. In the results, we only described GO-BP terms that were statistically significantly enriched in the input gene list after Benjamini–Hochberg correction for multiple comparisons.

Heterogeneity of Treatment Effects Analysis

Patient and factor selection

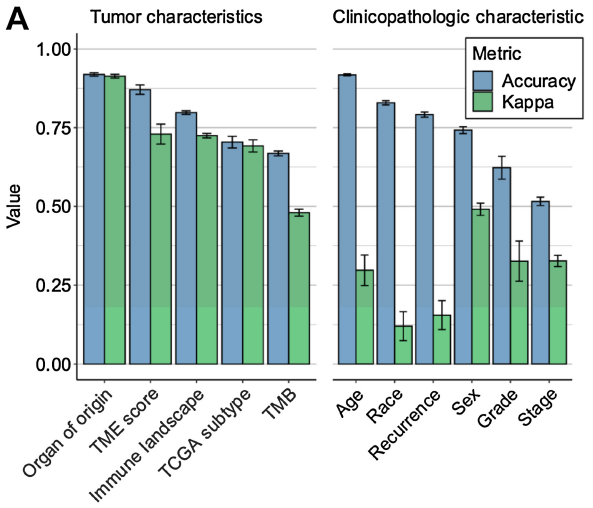
TCGA clinical therapy data were retrieved using TCGA biolinks (19). We parsed pharmaceutical annotations and identified a number of targeted therapies oriented against angiogenesis and vascular endothelial growth factor (VEGF) signaling. To identify potential factors that could serve as a predictor of therapy efficacy, we parsed the top 500 genes for each of the 100 NMF factors identified by our limma differential expression regression. We targeted factors with significant differential expression of *VEGFA* and other angiogenesis-associated genes. Factor 82 was identified as a candidate using these criteria.

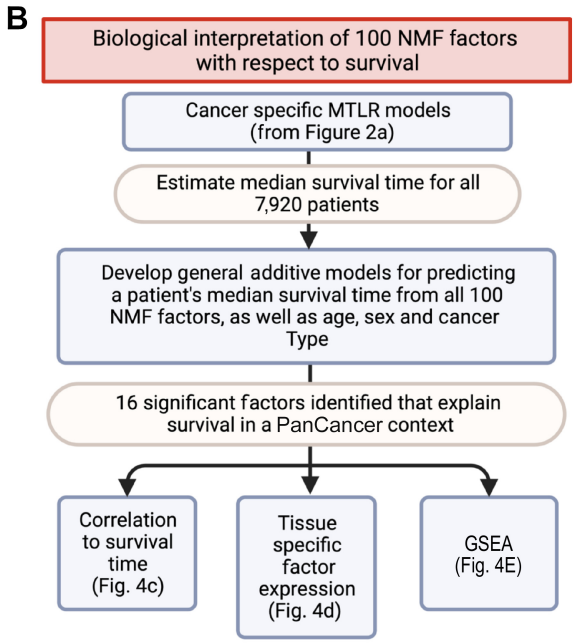
Individual treatment effect

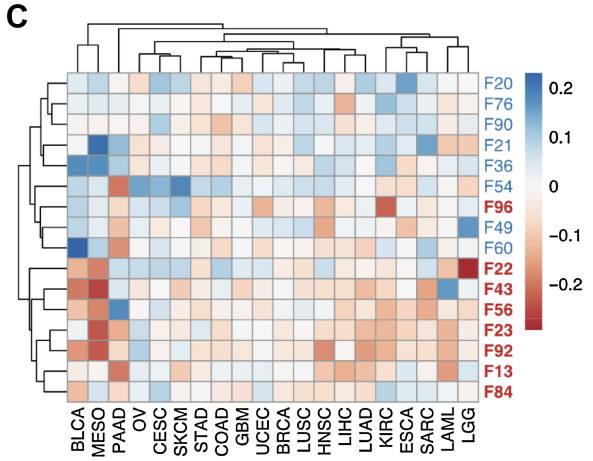
Heterogeneity of treatment effects were tested using interaction terms in an MTLR model. We modeled the survival effect of anti-VEGF therapy with respect to factor 82 levels in 3,257 patients with known clinical therapy data (97 anti-VEGF, 3,160 Other). The reference model covariates included treatment status and the values of our 100 NMF factors, for each patient. The interaction model included an interaction term between factor 82 and treatment status. Significance of the interaction term in the MTLR model was established using a likelihood ratio test assessing the difference in negative log likelihood between the reference model and the model containing the interaction term.

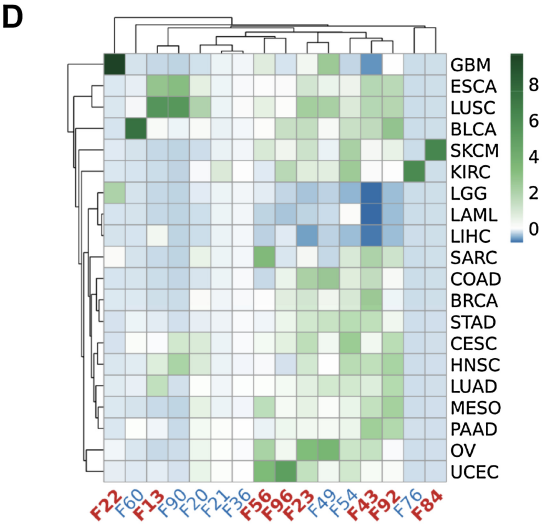
Model interaction effects were graphically evaluated using GAM created with the "mgcv" package in R (32). We fit cubic regression splines using generalized cross-validation to assess the predicted mean survival times from the MTLR model stratified by treatment status as a function of factor 82. The 95% simultaneous confidence interval (39–41) was calculated and plotted by sampling 10,000 simulations from the multivariate normal distribution using a mean of zero with the Bayesian covariance matrix generated from the GAM.

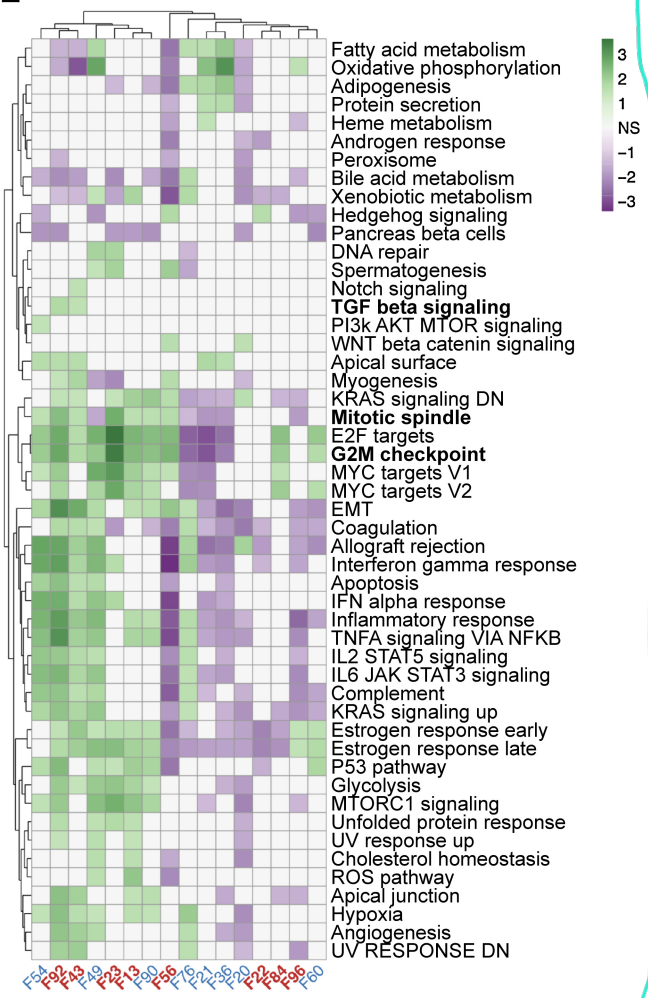
To demonstrate the potential use of ISDs for personalized medicine, we modeled the individual treatment effect (ITE) derived from the MTLR model. For each patient, the ITE was calculated as the difference between the predicted mean survival time under the observed treatment ("ground truth") and the counterfactual treatment (i.e., the alternate treatment that was not administered). The survival difference was calculated as the predicted survival of all patients if they were to receive anti-VEGF therapy minus the predicted survival of all patients if they were to receive other therapies.


A
Tumor characteristics
Clinicopathologic characteristic
Metric
Accuracy
Kappa
Value
1.00
0.75
0.50
0.25
0.00
Organ of origin
TME score
Immune landscape
TCGA subtype
TMB
Age
Race
Recurrence
Sex
Grade
Stage


B
Biological interpretation of 100 NMF factors with respect to survival
Cancer specific MTLR models (from Figure 2a)
Estimate median survival time for all 7,920 patients
Develop general additive models for predicting a patient's median survival time from all 100 NMF factors, as well as age, sex and cancer Type
16 significant factors identified that explain survival in a PanCancer context
Correlation to survival time (Fig. 4c)
Tissue specific factor expression (Fig. 4d)
GSEA (Fig. 4E)


C
F20
F76
F90
F21
F36
F54
F96
F49
F60
F22
F43
F56
F23
F92
F13
F84
0.2
0.1
0
−0.1
−0.2
BLCA
MESO
PAAD
OV
CESC
SKCM
STAD
COAD
GBM
UCEC
BRCA
LUSC
HNSC
LIHC
LUAD
KIRC
ESCA
SARC
LAML
LGG


D
GBM
ESCA
LUSC
BLCA
SKCM
KIRC
LGG
LAML
LIHC
SARC
COAD
BRCA
STAD
CESC
HNSC
LUAD
MESO
PAAD
OV
UCEC
8
6
4
2
0
F22
F60
F13
F90
F20
F21
F36
F56
F96
F23
F49
F54
F43
F92
F76
F84


E
Fatty acid metabolism
Oxidative phosphorylation
Adipogenesis
Protein secretion
Heme metabolism
Androgen response
Peroxisome
Bile acid metabolism
Xenobiotic metabolism
Hedgehog signaling
Pancreas beta cells
DNA repair
Spermatogenesis
Notch signaling
TGF beta signaling
PI3k AKT MTOR signaling
WNT beta catenin signaling
Apical surface
Myogenesis
KRAS signaling DN
Mitotic spindle
E2F targets
G2M checkpoint
MYC targets V1
MYC targets V2
EMT
Coagulation
Allograft rejection
Interferon gamma response
Apoptosis
IFN alpha response
Inflammatory response
TNFA signaling VIA NFKB
IL2 STAT5 signaling
IL6 JAK STAT3 signaling
Complement
KRAS signaling up
Estrogen response early
Estrogen response late
P53 pathway
Glycolysis
MTORC1 signaling
Unfolded protein response
UV response up
Cholesterol homeostasis
ROS pathway
Apical junction
Hypoxia
Angiogenesis
UV RESPONSE DN
3
2
1
NS
−1
−2
−3
F54
F92
F43
F49
F23
F13
F90
F56
F76
F21
F36
F20
F22
F84
F96
F60

Data availability statement

Gene expression data are publicly available from the NIH Genomic Data Commons at https://gdc.cancer.gov/about-data/publications/pancanatlas [File name: RNA (Final) - EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv] and survival data from the TCGA-Clinical Data Resource (CDR) Outcome data file (10). Additional data used in this study are available within the article and Supplementary Data S1E. The data and code used to generate this study is also available in a public repository at https://github.com/neerajkumarvaid/PanCancer_ISDs.

Results

Survival modeling

NMF-MTLR outperforms VAECox benchmark

Like NMF-MTLR, the VAECox algorithm also applied a survival prediction learner to a dimension-reduced dataset (17). VAECox, however, used deep-learning (a VAE) to reduce the dimensions (not NMF), then Cox proportional hazards (not MTLR) to predict survival. The original VAECox article reported that low-dimensional representations learned across many cancer types led to better survival predictions than cancer-specific representations. We corroborated this finding with NMF. **Figure 2A** compares discrimination (C-index) of NMF-MTLR with the benchmark VAECox model, showing that NMF-MTLR performed statistically better on 11 of 20 cancer types (paired two-sided *t* test, $P < 0.05$), marginally outperformed VAECox for six cancer types [bladder urothelial carcinoma (BLCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), acute myeloid leukemia (LAML), mesothelioma (MESO), pancreatic adenocarcinoma (PAAD), and sarcoma (SARC)], essentially tied for one [glioblastoma multiforme (GBM)], and marginally underperformed for two [brain lower grade glioma (LGG) and ovarian serous cystadenocarcinoma (OV)]. Overall, NMF-MTLR outperformed the VAECox model by 14.9% in terms of the micro-average C-index (**Fig. 2B**; Supplementary Table S2). Furthermore, while NMF-MTLR models were D-calibrated (26) for all 20 cancer types, VAECox was not D-calibrated for esophageal carcinoma (ESCA), GBM, lung squamous cell carcinoma (LUSC), and stomach adenocarcinoma (STAD). NMF-Cox also outperformed VAECox (micro-average C-index 0.679 vs. 0.638; Supplementary Table S3). Thus, NMF was a superior dimensionality reduction strategy for this application.

Superior performance is achieved with pan-cancer NMF representation in combination with cancer-specific MTLR models

Recall our NMF-MTLR approach first learns NMF representations on the basis of all 10,274 cancer transcriptomes (from 33 different cancers), then produces cancer-specific survival models by running MTLR on patients (with this cancer) encoding in this representation. We compared this to the NMF-1-MTLR approach, where the "1" signifies learning a cancer-specific NMF representation for each cancer, individually, based on just the biopsies from that cancer. **Figure 2C** shows that pan-cancer NMF representations outperform cancer-specific NMF representations, but cancer-specific MTLR survival models outperform pan-cancer MTLR survival models.

NMF-based MTLR outperforms other ISD modeling techniques

Having established that NMF performs better than VAE in Cox modeling, we compared the performance of three different ISD-modeling strategies in combination with NMF: the Kalbfleisch–Prentice extension of Cox (NMF-CoxKP), random survival forests (NMF–RSF), and MTLR (NMF–MTLR), where each ISD model was trained on a cancer-specific basis using the 100-dimensional patient representations obtained from pan-cancer NMF. NMF–MTLR outperformed both NMF-CoxKP and NMF–RSF (Supplementary Table S2), and was calibrated for all cancer types, whereas NMF-CoxKP (respectively, NMF-RSF) was not calibrated for 4 (respectively, 10) cancers.

We also considered replacing the (unsupervised) NMF first step with (supervised deep-learning) VAE—that is, VAE–MTLR. We found, however, that NMF-MTLR was superior to VAE–MTLR (micro-averaged C-index of 0.733 vs. 653), which suggests that NMF is well suited to reduce these complex data (see Supplementary Table S2).

External validation of NMF-MTLR models

External validation was performed to assess model generalizability of the internally cross-validated NMF-MTLR models. Transcriptomes for breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), and STAD were derived from microarray data from METABRIC, GSE 39582, and ACRG studies, respectively (42–44). RNA-seq transcriptomes from the pan-cancer analysis of advanced cancers, which includes tumor samples from primary and metastatic sites, were included for cancers with greater than 25 cases [lung adenocarcinoma (LUAD), OV, PAAD, and SARC; ref. 45]. **Figure 2D** demonstrates that C-index values for the external validation of COAD, STAD, LUAD, OV, PAAD, and SARC datasets were similar to internal validation C-indices (two-sided Wilcox test, $P > 0.05$). External validation of BRCA from the METABRIC dataset performed worse than the reference NMF-MTLR model (C-Index 79.0% vs. 64.2%, $P < 0.05$).

Arguments for using the ISD and MTLR approach

An MTLR-derived ISD has a direct, meaningful interpretation: it predicts the probability that a patient will survive *at least* until time *t* for all possible times. It can be used to answer questions like "what is the probability that this patient will survive at least one year?" and "how long should this patient expect to live?" Contrast this with simple risk scores (e.g., Cox-based OncotypeDx, SurvExpress or Deepsurv; refs. 46–48), which are only meaningful in relation to another patient—here, given two patients, the one with a higher risk score is predicted to die first. Risk

Figure 4.

Biological interpretation of NMF factors. **A,** Plot showing how well various learned random forest models can predict tumor and clinicopathologic characteristics using all 100 NMF factors. Grouped bars represent mean accuracy (blue) and Cohen's Kappa (green) values over 5-fold cross-validation for each target characteristic. Error bars represent 95% CIs calculated with 1,000 bootstraps. TMB, tumor mutational burden. **B,** Flow chart illustrating the modeling process and subsequent downstream analysis used to assess the biologic importance of NMF factors with respect to OS. **C,** Heat map of Spearman ρ correlation between estimated median survival time derived from MTLR models and z-score normalized factor values (row labels) for 20 cancer types (column labels). In **C**–**E**, column factor labels associated with decreased or increased survival are shown in red and blue, respectively, on the basis of the coefficient values from a negative binomial linear model. **D,** Heat map of z-score normalized factor values for the 90th percentile stratified by cancer type (row labels). The colored scale bar represents z-score normalized factor scores. Green cells indicate cancer types that possess a given 90th percentile factor score greater than the overall mean factor score. Blue cells indicate cancer types with a given 90th percentile factor score less than the overall mean factor score. **E,** Heat map of normalized enrichment scores derived from GSEA of Hallmark gene sets (row labels) for significant factors (column labels) from the nonlinear multivariable GAM. Colored cells passed the Benjamini–Hochberg FDR < 0.05 threshold for statistical significance of enrichment. Green cells indicate that the gene set was overenriched among the genes most associated with the indicated factor. Violet cells indicate that the gene set was underenriched. Clustering for all heat maps was performed using complete linkage on Euclidean distance.

predictions are accurate whether patient A dies in 30 years and patient B in 31, or A in 3 days and B in 4, but they offer no sense of when death might occur. But risk can be inferred from probabilistic ISD survival models, too: a patient with a 90% chance of surviving at least one year has lower one-year risk than a patient with an 80% chance. Compared with purely risk-based models, ISDs have the advantage of telling us that both of these patients will likely survive longer than one year.

There are many ISD models. The standard Cox model can also be used to produce an ISD (viz., its Kalbfleisch–Prentice extension; ref. 49). But unlike Cox, MTLR does not rely on the potentially invalid assumption that the effect of prognostic features is time-invariant (the proportional hazards assumption). Thus, MTLR is robust in cases where Cox would fail (note the curves cross in **Fig. 3B** and not **Fig. 3C**).

ISD models base their predictions on patient-specific features. Therefore, they better facilitate patient-tailored care than modeling strategies such as Kaplan–Meier (KM) that conflate the individual with a group. KM curves capture average survival of patient groups (**Fig. 3D**), but do not reveal heterogeneity within groups. If desired, MTLR-derived ISDs can provide similar population-level information as KM by averaging its survival curves' predictions over groups of patients, while additionally illustrating variability in outcomes between individuals within groups (**Fig. 3E** and **F**).

The area under an ISD curve visually provides a sense of a patient's projected trajectory. Single-time models, such as nomograms, cannot do this (50). Consider two patients whom we would expect to have similar outcomes on the basis of their clinical characteristics: Patient E, a 62 year old with Stage II, HER3$^+$ breast cancer (TCGA-AR-A1AT), and Patient F, a 58 year old with Stage II, HER2 2$^+$/FISH positive breast cancer (TCGA-AR-A250), who both appropriately received trastuzumab. Patient E passed away much earlier than Patient F, and NMF-MTLR produced ISDs that reflect the difference in outcomes (**Fig. 3F**). In contrast to these results generated from a single ISD model, multiple independent single-time nomogram models would need to be constructed in order to achieve a similar conclusion. In **Figure 3E** several curves intersect. A single-time model might not predict differences in the patients' outcomes at one of these time points, yet the AUCs suggest different trajectories that vary with time. Conceivably, the slope of an ISD could support clinical decision making regarding, for example, the aggressiveness of intervention, or end-of-life planning. For a more comprehensive review and analysis, refer to Haider and colleagues (26).

Biological Interpretation of NMF Representation

NMF captures clinicopathologic and molecular characteristics of tumors

Next, we explored whether our 100 NMF factors capture patients' clinical data and standard tumor- and immune-based classifications of their biopsies. We trained 11 different random forest classifiers, each described using these 100 NMF factors, along with the relevant label—e.g., the organ site of the cancer, or the sex of the patient. (**Fig. 4A**). Supplementary Data S1 provides details about the source and definitions of the dependent variables used, while Supplementary Table S5 provides the summary statistics for each dependent variable used in the random forest classifiers.

NMF factors reliably captured tumor molecular characteristics with known implications for prognosis and/or therapy. Overall, they contributed statistically meaningful information for nine of the outcomes of interest (accuracy < NIR, one-sided binomial test, $P < 0.05$) but not for race and recurrence. Note NMF factors predicted the organ-of-origin among 20 different cancer types with a mean predictive accuracy of 91.9% [95% confidence interval (CI), 91.4–92.5]. Factor scores also effectively represented cancer-specific immunity explained by the tumor microenvironment score (TME; 87.0% accuracy; 95% CI, 86.1–88.0) and Immune Landscape of Cancer classification (ILC; 79.8% accuracy; 95% CI, 79.2–80.3).

Biological importance of NMF factors with respect to overall survival

Given that NMF factors approximated relevant biological and clinical phenomena, we next sought to characterize the relationship of factors to overall survival. We were motivated to assess if any factors were consistently associated with improved or reduced survival across all cancer types—a universal hallmark of cancer. Because feature selection was used to train the MTLR models (see Materials and Methods section) and no single factor was selected as a feature for all cancer types, it was not possible to assess the relationship of each factor with survival across all cancer types directly from the MTLR models. To work around this, we used a nonlinear negative binomial GAM to evaluate the relationship of all factors to the median survival times predicted by the MTLR ISD models (see Materials and Methods section for details about the model selection process). The effect of cancer types and nonlinear relationships of factors, and age, on survival is illustrated in Supplementary Figs. 2 and 3. We found that, for all 20 cancer types, patient age and 16 factors—F13, F20, F21, F22, F23, F36, F43, F49, F54, F56, F60, F76, F84, F90, F92, and F96—were significantly associated with median overall survival (Supplementary Data S3). We also performed a sensitivity analysis that included cancer types with cancer stage and grade data, which demonstrated that our approach to use NMF factor values across all cancer types was not limited by confounding (see Supplementary Data B.1). The next sections summarize the biological phenomena associated with these factors.

Biological characterization of the most prognostic NMF factors

As illustrated in **Fig. 4C**, no factor was consistently associated with better or worse outcomes across all cancer types in the GAM. Factors F20, F21, F36, F49, F54, F60, F76, and F90 were generally correlated with better survival on average. Factors F13, F22, F23, F43, F56, F84, F92, and F96 were correlated with worse survival on average. To understand the cellular processes associated with these factors at a high level, we studied their 90th percentile values in each cancer type (to assess their relative abundance; **Fig. 4D** and 50th percentile values in Supplementary Fig. S7) as well as the enrichment of all 50 Hallmarks MSigDB gene sets (**Fig. 4E**) among the genes that were most associated with each of these factors (38). Here, we identified that some factors are highly expressed in only one or a handful of cancers such as factor 60 in BLCA or factor 90 in LUSC, CESC, ESCA, and head and neck squamous cell carcinoma (HNSC), while other factors are highly expressed across more than 10 cancers. The factors that exhibited high expression values across the majority of cancers were also those that tended to possess significantly enriched immune-related signatures such as factors F43, F54, and F92. Additional fine-grained examination of individual genes and Gene Ontology Biological Process (GO-BP) annotations (data not shown; see Materials and Methods section) associated with the factors provided deeper insight. Below we describe four of these factors from that figure; the remainder are described in the Supplementary Data B.2.

F23, one of the factors that was associated with worse outcomes, was enriched with genes in the Hallmarks sets for unfolded protein response and cell-cycle progression (E2F targets, G2M checkpoint), among others. This factor may correspond to an abundance of actively growing neoplastic cells, as suggested by their overexpression of genes

associated with glycolysis and mitotic spindle GO-BP terms. It was widely expressed by many cancer types.

F43 was distinctly associated with genes involved in tissue remodeling, signified by GO-BP terms including extracellular matrix organization, angiogenesis, basement membrane organization, cell migration, and so on. Naturally, these processes include genes that are also part of immune-related Hallmarks gene sets, as well as the epithelial–mesenchymal transition (EMT) gene set. Other associated GO-BP terms suggested tissue differentiation, for example: ossification, osteoblast differentiation, chondrocyte development, and lung, heart, and palate development. It had very low scores in GBM, LGG, LAML, and liver hepatocellular carcinoma (LIHC), but higher F43 scores seemed protective in LAML. It was widely expressed but generally deleterious in everything else, especially BLCA and MESO.

F54 was strongly associated with infiltration and activation of lymphoid cells (T cells, NK cells). Its top associated gene was *IFNG* itself. Interestingly, it was strongly associated with worse survival in PAAD, but protective in almost everything else.

F92 was associated with poor outcomes in most cancer types. This factor reflected the immunologically mediated aspect of tissue response to wounding, rooted in proinflammatory and immunomodulatory roles of the innate immune system. Innate immune response and tissue response to wounding appear to differentiate F92 from F54, which was skewed towards adaptive immune responses and correlated with better outcomes.

These results show that several factors have biologically coherent characteristics. Their relationships with survival appear largely context-dependent, and in many cases, appear to reflect subtly different aspects of similar sets of core biological activities such as tissue response to wounding, immune responses, and metabolic regulation. Although it is beyond the scope of this article, further understanding of these subtleties in a clinicopathologic context will be an essential step toward understanding cancer progression.

Pan-cancer NMF factors predict prognostic benefits of anticancer therapy

Beyond purely prognostic applications, we postulated that some NMF factors could predict whether certain types of anticancer therapy would be effective. TCGA clinical therapy data identified 3,257 patients with annotated pharmacologic-based therapy, of which 97 patients, from 12 cancer types, were treated with targeted anti-VEGF therapy—including bevacizumab, sorafenib, pazopanib, regorafenib, and cabozantinib (51).

Differential gene expression results for each factor were evaluated for potential biology related to cancer angiogenesis. Factor 82 overexpressed key activators of the angiogenic switch including *VEGFA* and the endogenous VEGF receptors *KDR, FLT1*, and *FLT4* (**Fig. 5A**; refs. 52, 53). Several other critical angiogenic mediators were also overexpressed (e.g., *PGF, PDGFB, PECAM1*, and *ANGPT2*; refs. 54–57). Overexpression of *DLL4*, *NOTCH4* and vascular-endothelial cadherin (*CDH5)* suggests factor 82 also facilitates the selection, regulation, and maturation of tip and stalk cells responsible for the formation of new blood vessels (53, 58, 59). Hallmark gene sets annotated in tissue hypoxia and epithelial to mesenchymal transition (i.e., metastasis and fibrosis) were increased in association with factor 82 (**Fig. 5B**).

We evaluated the possible survival effect of factor 82 on anti-VEGF–oriented therapies using MTLR ISD models with and without an interaction term between factor 82 and treatment and found that the interaction was significantly associated with survival (likelihood ratio $P < 0.001$). To visualize the interaction, we extracted the estimated mean survival times from the MTLR model and plotted them in **Fig. 5C**, which illustrated a clear nonlinear interaction effect between treatment and factor 82. **Figure 5D** shows the Individual treatment effect, defined as the difference in (predicted) survival time for each patient given the counterfactual scenario where they actually received the alternative therapy. For example, this counterfactual argument suggests that patients with a scaled factor 82 score of approximately 0.5 who received other therapies would have survived approximately 3 additional years if they had instead received targeted anti-VEGF therapy. Together, these analyses suggest that factor 82 may serve as a potential biomarker to predict treatment efficacy of anti-VEGF therapy.

Discussion

This article describes our NMF-MTLR–based approach for modeling individual survival curves (ISD) from whole transcriptome data of a tumor biopsy. Each ISD is similar to a KM survival curve, as it provides the survival probability at each future time, but while KM applies only to a group of patients, an ISD is personalized for each individual patient (26). A patient's ISD can be used to estimate their "time until death," risk score, and survival probability at a specific future time point (see **Fig. 3**). ISD visualizations could also facilitate meaningful conversation between doctors and their patients about disease prognosis.

We demonstrated that our NMF-MTLR ISD model outperformed the benchmark VAECox model by 14.9% in terms of C-index micro-averaged over all 20 cancer types modeled in this article. We also demonstrated that our MTLR performed better than other ISD modeling strategies (NMF-Cox-KP and NMF-RSF).

In addition to superior discrimination (concordance), NMF-MTLR was the only ISD model to achieve calibration in all 20 cancer survival models. A calibrated model has obvious clinical utility as its predicted probability of events reflects the real-world probability of the event actually occurring. For example, over all the times when a well-calibrated model informs a meteorologist that there is a 30% chance of rain, rain actually occurs 30% of the time. Many celebrated machine learning algorithms, such as random forests or deep-learning neural networks, are notoriously poorly calibrated (27, 60), which can lead to poor decisions, and so limit their application to real-world problems, such as predicting probability of death for specific times (26, 27, 61). Indeed, we confirmed that our Random Survival Forest model variant was uncalibrated in half of the cancers it modeled.

Our pan-cancer representation of gene expression led to better survival predictions than the alternative of cancer-specific representations. This finding is consistent with the Hallmarks of Cancer perspective, as it suggests that the molecular footprints of cancer are best revealed when it is viewed as a common disease (which is not to say that all cancer shares the same characteristics—see below; ref. 12). Notably, the pan-cancer representation performed best when used for cancer-specific MTLR survival models (as opposed to a single survival model applied to patients with any kind of cancer).

Although our results obtained by viewing cancer as a common disease through NMF were successful, they nevertheless underscore several caveats to the Hallmarks of Cancer theory: (i) no single factor was universally favorable or unfavorable to survival; (ii) some factors exhibited heterogeneous expression levels in different cancer types; and (iii) biological processes specific to cancer, such as G2M checkpoint or EMT, were functionally enriched in multiple factors but also associated with heterogeneous survival effects. These findings argue that the actionable utility and clinical effect of a given factor must be

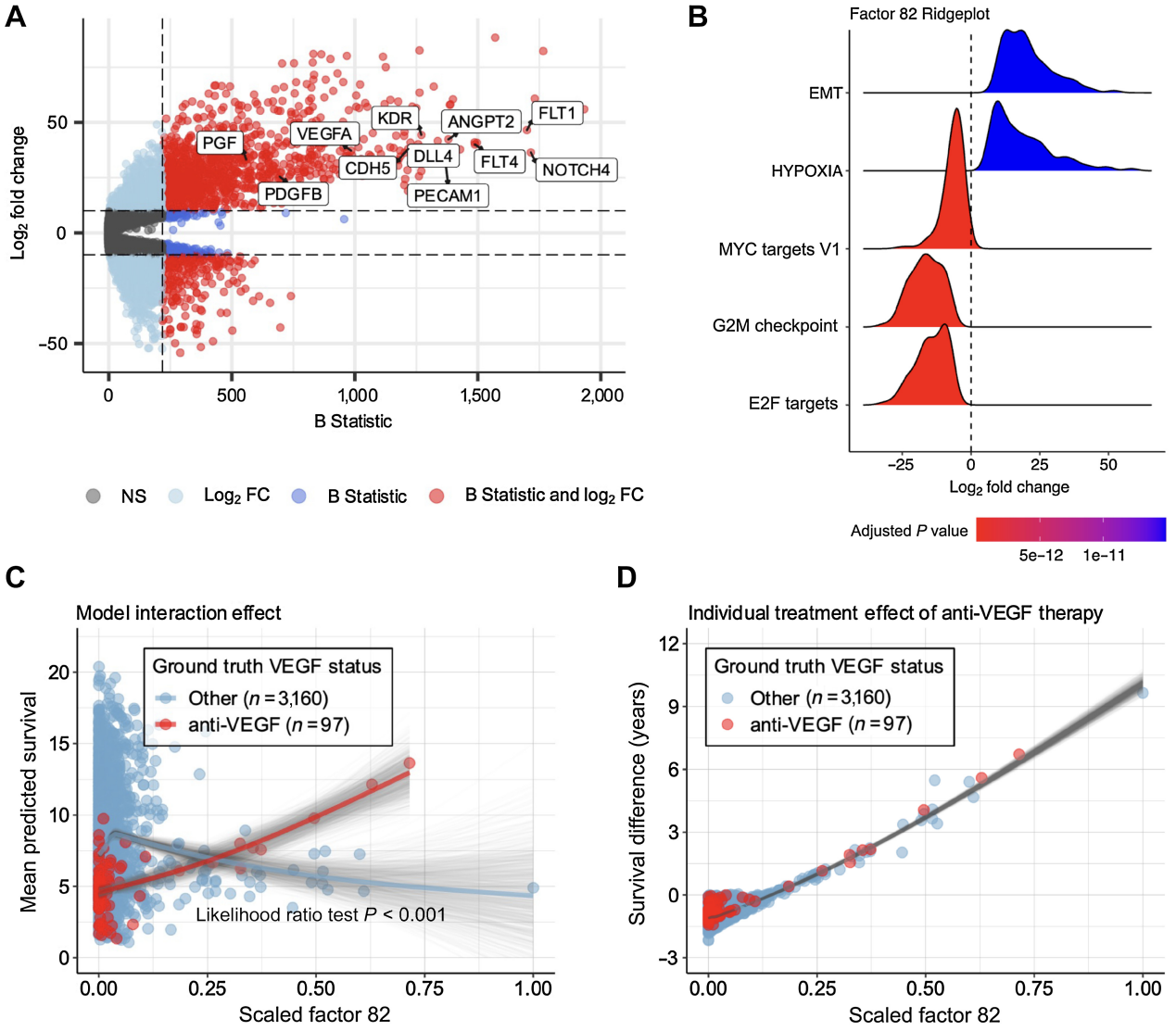


Figure 5.

NMF factor 82 represents a neovascularization program and is associated with improved survival with anti-VEGF therapy. **A,** Volcano plot of differential regression of NMF factor genes. The *y*-axis shows $\log_2$ FC in gene expression per unit change in factor 82 score. The *x*-axis shows the B Statistic, which is the odds of a gene being differentially expressed. Each circle represents an individual gene. The vertical dashed line provides a B statistic cutoff correlating to an adjusted *P* of 1×10^{-100}. The horizontal dashed lines provide a $\log_2$ FC cutoff of 10. Significance cutoff of genes is designated by color according to the legend below the graph (ns, not significant). White labels show genes associated with neovascularization. **B,** Factor 82 ridgeplot derived from GSEA of MSigDb Hallmarks gene set. Probability density functions represent the distribution for expression levels of core enriched genes corresponding to $\log_2$ FC on the *x*-axis. "–" denotes the top 5 enriched Hallmark features. Benjamini–Hochberg adjusted *P* values were calculated from a permutation test with 10,000 permutations. **C,** Interaction plot of mean predicted survival versus scaled factor 82 score derived from the MTLR model. GAMs were fit for anti-VEGF (red spline) and other therapies (blue spline) to model the effect. 95% simultaneous CIs for both models are represented by shaded gray lines. **D,** Individual Treatment Effect (ITE) expressed as the difference in survival time (in years) when receiving anti-VEGF therapy versus other therapy as a function of factor 82 score. See Methods regarding the calculation of ITE. Factor 82 (*x*-axis) was scaled between 0 and 1. The gray smoothed spline and 95% CI (shaded gray adjacent to line) were calculated using a generalized additive model. Each point represents a single patient and is colored according to the actual treatment received.

considered at both the pan-cancer and at the cancer-specific level. These phenomena may be related to the degree that a given factor can be expressed in a given tissue due to local epigenetic regulation or represent distinct organ-specific mechanisms of cancer pathogenesis (62, 63). The heterogeneity observed in the relationship between factor expression and survival for different cancer types is also likely augmented by differences in cancer diagnosis, screening, and treatment for each respective cancer (i.e., surgical, pharmacologic and/or diagnostic challenges due to anatomic location or tissue type). Such differences can impact both molecular and survival phenotypes.

At the bedside, ISDs have the potential to provide accurate, intuitive and visual predictions that can facilitate collaborative and informed decision making between patients and physicians. This is critical, as many standard survival models provide only risk scores; worse,

physicians are known to provide inaccurate risk estimates in end-of-life scenarios with meta-analyses reporting physician C-index values ranging from 23% to 78% (64–67). Given the advantages of ISDs, a model such as MTLR can aid in arriving at a "correct" therapy for each patient that can range from curative intent therapeutic intervention to the initiation of appropriate palliative measures (68, 69).

There are some limitations with our study. The findings listed above must be explored in external datasets to further validate the NMF-MTLR approach. Although the PanCancer Atlas database is the most comprehensive multi-omic database with annotated clinical and therapeutic data, many of its variables have missing data. Our use of overall survival introduces noise, as it includes deaths for noncancer causes; however, we do not have the information needed to identify more desirable outcomes such as disease-free survival or progression-free interval outcomes. Furthermore, the dataset does not contain detailed comorbidity status to gauge the degree to which overall survival is related to cancer-related death. Our analysis is also based on retrospective data and thus cannot establish causal relationships. Our analyses of heterogeneity of treatment effects, such as those demonstrated in our factor 82 analysis, makes strong assumptions related to counterfactual individual treatment effects. Once again, while these findings are intriguing, they do require additional study and external validation.

Personalized medicine means providing the right therapy for the right patient at the right time. In this study, we established NMF-MTLR as a new benchmark for personalized cancer survival prediction on publicly available transcriptome data. NMF-MTLR provides many benefits over other models: superior model discrimination, superior calibration, and accurate probabilistic estimates of survival over time for each individual patient. We anticipate that the application of ISDs to future research and clinical practice will drive advancement in personalized medicine. We advocate for further exploration of these cancer survival models in clinical and research settings, as well as recommend our general NMF-MTLR approach for other diseases.

Authors' Disclosures

N. Kumar reports grants from Alberta Machine Intelligence Institute during the conduct of the study. D. Skubleny reports other support from BOLD Therapeutics outside the submitted work. R. Greiner reports personal fees from Amii outside the submitted work. No disclosures were reported by the other authors.

Authors' Contributions

N. Kumar: Data curation, formal analysis, investigation, writing–original draft, project administration. **D. Skubleny:** Formal analysis, investigation, writing–original draft, writing–review and editing. **M. Parkes:** Investigation, writing–review and editing. **R. Verma:** Data curation, validation, investigation, methodology, writing–review and editing. **S. Davis:** Data curation, formal analysis, writing–review and editing. **L. Kumar:** Formal analysis, methodology, writing–original draft, writing–review and editing. **A. Aissiou:** Writing–review and editing. **R. Greiner:** Conceptualization, resources, supervision, funding acquisition, writing–review and editing.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

Note

Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Received November 19, 2022; revised February 11, 2023; accepted July 11, 2023; published first July 18, 2023.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–49.
2. Kalia M. Personalized oncology: recent advances and future challenges. Metabolism 2013;62:S11–4.
3. Schilsky RL. Personalized medicine in oncology: the future is now. Nat Rev Drug Discov 2010;9:363–6.
4. Zhang X, Marjani SL, Hu Z, Weissman SM, Pan X, Wu S. Single-cell sequencing for precise cancer research: progress and prospects. Cancer Res 2016;76:1305–12.
5. Yang X, Kui L, Tang M, Li D, Wei K, Chen W, et al. High-throughput transcriptome profiling in drug and biomarker discovery. Front Genet 2020;11:19.
6. Cohen KA, Manson AL, Desjardins CA, Abeel T, Earl AM. Deciphering drug resistance in mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges. Genome Med 2019;11:45.
7. Levitin HM, Yuan J, Sims PA. Single-cell transcriptomic analysis of tumor heterogeneity. Trends Cancer 2018;4:264–8.
8. Collisson EA, Bailey P, Chang DK, Biankin AV. Molecular subtypes of pancreatic cancer. Nat Rev Gastroenterol Hepatol 2019;16:207–20.
9. Alwers E, Jia M, Kloor M, Bläker H, Brenner H, Hoffmeister M. Associations between molecular classifications of colorectal cancer and patient survival: a systematic review. Clin Gastroenterol Hepatol 2019;17:402–10.
10. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 2018;173:400–16.
11. Ray P, Reddy SS, Banerjee T. Various dimension reduction techniques for high dimensional data analysis: a review. Artif Intell Rev 2021;54:3473–515.
12. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.
13. Zhu X, Ching T, Pan X, Weissman SM, Garmire L. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. PeerJ. 2017;5:e2888.
14. Leo P, Lee G, Shih NNC, Elliott R, Feldman MD, Madabhushi A. Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. J Med Imaging (Bellingham) 2016;3:047502.
15. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 2004;101:4164–9.
16. Yu C-N, Greiner R, Lin H-C, Baracos V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2011. Available from: https://proceedings.neurips.cc/paper/2011/file/1019c8091693ef5c5f55970346633f92-Paper.pdf.
17. Kim S, Kim K, Choe J, Lee I, Kang J. Improved survival analysis by learning shared genomic information from pan-cancer data. Bioinformatics 2020;36:i389–98.
18. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med 2016;375:1109–12.
19. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016;44:e71.
20. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. Genome Res 2018;28:1747–56.
21. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:pl1.
22. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–91.
23. Haider H. MTLR: Survival Prediction with Multi-Task Logistic Regression [Internet]. 2019 [cited 2023 Jul 27]. Available from: https://CRAN.R-project.org/package=MTLR.
24. Kingma DP, Welling M. Auto-Encoding Variational Bayes; 2013. Available from: http://arxiv.org/abs/1312.6114.

25. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30:1105–17.
26. Haider H, Hoehn B, Davis S, Greiner R. Effective ways to build and evaluate individual survival distributions. Journal of Machine Learning Research 2020;21: 1–63.
27. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inform 2019;125: 55–61.
28. Franks JM, Cai G, Whitfield ML. Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data. Bioinformatics 2018;34:1868–74.
29. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67.
30. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26.
31. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22: 276–82.
32. Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J Am Stat Assoc 2004;99:673–86.
33. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Series B Stat Methodol 2011;73:3–36.
34. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
35. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
36. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003;34:267–73.
37. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation 2021;2:100141.
38. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst 2015;1:417–25.
39. Marra G, Wood SN. Coverage properties of confidence intervals for generalized additive model components. Scand J Stat 2012;39:53–74.
40. Nychka D. Bayesian confidence intervals for smoothing splines. J Am Stat Assoc 1988;83:1134–43.
41. Gavin Simpson ML. gratia: Graceful ggplot-Based Graphics and Other Functions for GAMs Fitted using mgcv [Internet]; 2023 [cited 2023 Feb 9]. Available from: https://gavinsimpson.github.io/gratia/.
42. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486:346–52.
43. Cristescu R, Lee J, Nebozhyn M, Kim K-M, Ting JC, Wong SS, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 2015;21:449–56.
44. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med 2013;10:e1001453.
45. Pleasance E, Titmuss E, Williamson L, Kwan H, Culibrk L, Zhao EY, et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. Nat Cancer 2020;1:452–68.
46. Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Chacolla-Huaringa R, Rodriguez-Barrientos A, et al. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. PLoS One 2013;8:e74250.
47. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
48. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 2018;18:24.
49. Kalbfleisch JD, Prentice RL. Marginal likelihoods based on cox's regression and life model. Biometrika 1973;60:267–78.
50. Iasonos A, Schrag D, Raj Gv Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol 2008;26:1364–70.
51. Clarke JM, Hurwitz HI. Understanding and targeting resistance to anti-angiogenic therapies. J Gastrointest Oncol 2013;4:253–63.
52. Bergers G, Benjamin LE. Tumorigenesis and the angiogenic switch. Nat Rev Cancer 2003;3:401–10.
53. Adams RH, Alitalo K. Molecular regulation of angiogenesis and lymphangiogenesis. Nat Rev Mol Cell Biol 2007;8:464–78.
54. Delisser HM, Christofidou-Solomidou M, Strieter RM, Burdick MD, Robinson CS, Wexler RS, et al. Involvement of endothelial PECAM-1/CD31 in angiogenesis. Am J Pathol 1997;151:671–7.
55. O'Brien CD, Cao G, Makrigiannakis A, DeLisser HM. Role of immunoreceptor tyrosine-based inhibitory motifs of PECAM-1 in PECAM-1-dependent cell migration. Am J Physiol Cell Physiol 2004;287:C1103–13.
56. Saharinen P, Eklund L, Alitalo K. Therapeutic targeting of the angiopoietin–TIE pathway. Nat Rev Drug Discov 2017;16:635–61.
57. Raica M, Cimpean AM. Platelet-derived growth factor (PDGF)/PDGF receptors (PDGFR) axis as target for antitumor and antiangiogenic therapy. Pharmaceuticals 2010;3:572–99.
58. Potente M, Gerhardt H, Carmeliet P. Basic and therapeutic aspects of angiogenesis. Cell 2011;146:873–87.
59. Sauteur L, Krudewig A, Herwig L, Ehrenfeuchter N, Lenard A, Affolter M, et al. Cdh5/VE-cadherin promotes endothelial cell interface elongation via cortical actin polymerization during angiogenic sprouting. Cell Rep 2014;9: 504–13.
60. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks [abstract]. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, New South Wales, Australia. New York (NY): 2017.
61. D'Agostino RB, Nam B. Evaluation of the performance of survival analysis models: discrimination and calibration measures. Advances in Survival Analysis 2003.
62. Nolan DJ, Ginsberg M, Israely E, Palikuqi B, Poulos MG, James D, et al. Molecular signatures of tissue-specific microvascular endothelial cell heterogeneity in organ maintenance and regeneration. Dev Cell 2013;26:204–19.
63. Kashyap MP, Sinha R, Mukhtar MS, Athar M. Epigenetic regulation in the pathogenesis of non-melanoma skin cancer. Semin Cancer Biol 2022;83: 36–56.
64. Glare P, Virik K, Jones M, Hudson M, Eychmuller S, Simes J, et al. A systematic review of physicians survival predictions in terminally ill cancer patients. BMJ 2003;327:195–8.
65. Gwilliam B, Keeley V, Todd C, Roberts C, Gittins M, Kelly L, et al. Prognosticating in patients with advanced cancer—observational study comparing the accuracy of clinicians' and patients' estimates of survival. Ann Oncol 2013;24: 482–8.
66. White N, Reid F, Harris A, Harries P, Stone P. A systematic review of predictions of survival in palliative care: how accurate are clinicians and who are the experts? PLoS One 2016;11:e0161407.
67. Farinholt P, Park M, Guo Y, Bruera E, Hui D. A comparison of the accuracy of clinician prediction of survival versus the palliative prognostic index. J Pain Symptom Manage 2018;55:792–7.
68. Weeks JC, Catalano PJ, Cronin A, Finkelman MD, Mack JW, Keating NL, et al. Patients' expectations about effects of chemotherapy for advanced cancer. N Engl J Med 2012;367:1616–25.
69. Weeks JC, Cook EF, O'Day SJ, Peterson LM, Wenger N, Reding D, et al. Relationship between cancer patients' predictions of prognosis and their treatment preferences. JAMA 1998;279:1709–14.