

Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Safoora Yousefi

Date

Neural Networks for Cancer Survival Analysis Using High-Dimensional Data

By

Safoora Yousefi
Doctor of Philosophy

Computer Science and Informatics

Lee Cooper
Advisor

Daniel J. Brat
Committee Member

Joyce Ho
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Neural Networks for Cancer Survival Analysis Using High-Dimensional Data

By

Safoora Yousefi
B.A., University of Tehran, Iran, 2013
M.Sc., Emory University, GA, 2018

Advisor: Lee Cooper

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2019

Abstract

Neural Networks for Cancer Survival Analysis Using High-Dimensional Data By Safoora Yousefi

Since the emergence of high throughput experiments such as Next Generation Sequencing, the volume of genomic data produced has been increasing exponentially. This data holds the key to accurate predictions of clinical outcomes and mapping patients to the optimal treatment. However, analyzing genomic data is challenged by its high-dimensionality. Many prediction methods face limitations in learning from high-dimensional data generated by these platforms, and rely on experts to hand-select a small number of features for training prediction models. In this thesis, we demonstrate how the latest advances in neural networks methods that have been remarkably successful in general high-dimensional prediction tasks can be leveraged to the problem of predicting cancer outcomes. We perform an extensive comparison of deep survival models and other state of the art machine learning methods for survival analysis. We appreciate that interpretability is of great importance in adapting neural networks in bioinformatics, and propose a framework for interpreting deep survival models using a risk back-propagation technique that can lead to new understanding of diseases. Finally, we illustrate that deep survival models can successfully transfer information across heterogeneous data sources to improve prognostic accuracy, and describe an adversarial multi-task learning approach that outperforms traditional multi-task learning methods. We provide an open-source software implementation of these frameworks that enables automatic training, evaluation and interpretation of deep survival models.

Neural Networks for Cancer Survival Analysis Using High-Dimensional Data

By

Safoora Yousefi

B.A., University of Tehran, Iran, 2013

M.Sc., Emory University, GA, 2018

Advisor: Lee Cooper

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2019

Contents

1	Introduction	1
1.1	Survival Analysis	2
1.1.1	Censoring	2
1.1.2	Survival and Hazard Functions	3
1.1.3	Cox’s Proportional Hazards Model	6
1.2	Machine Learning and Genomic Data: Challenges	7
1.2.1	Genomic Data, Dimensionality, and Heterogeneity	8
1.2.2	Interpretability	10
1.3	Model Selection and Evaluation	13
1.3.1	Performance Metrics	13
1.3.2	Hyper-parameter Optimization	14
2	Previous Work	18
2.1	Machine Learning for Survival Analysis	18
2.2	Deep Learning	20
2.2.1	Background	20
2.2.2	Representation Learning	23
2.2.3	Convolutional Networks	24
2.2.4	Adversarial Learning	26
2.2.5	Interpretable Deep Learning	28

2.3	Multi-task Learning	29
3	Learning Genomic Representations to Predict Clinical Outcomes in Cancer	34
3.1	Learning Genomic Representations to Predict Clinical Outcomes in Cancer, ICLR-workshop, 2016	35
3.2	Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models, Nature Scientific Reports, 2017	40
3.3	Predicting cancer outcomes from histology and genomics using convolutional networks, PNAS, 2018	52
3.4	Multi-faceted computational assessment of risk and progression in oligodendroglioma implicates NOTCH and PI3K pathways, NPJ Precision Oncology, 2018	63
4	Learning Clinical Outcomes from Heterogeneous Data Sources	73
4.1	Learning Clinical Outcomes from Heterogeneous Genomic Data Sources: An Adversarial Multi-task Learning Approach, ICML, 2019 Adaptive and Multitask Learning Workshop	74
4.2	Learning clinical outcomes from heterogeneous genomic datasets using adversarial and multi-task learning, Manuscript in Progress	83
5	Transfer Learning From Nucleus Detection To Classification In Histopathology Images, ISBI 2019	106
	Bibliography	108

List of Figures

1.1	Right censoring in time-to-event data. x indicates event occurrence while o marks either loss to follow-up or long term survival, resulting in censoring. Subject 4 is censored due to follow-up termination while other censored subjects are lost to follow-up.	3
1.2	Example of Kaplan-Meier curves on random synthetically generated data comparing survival in two groups.	4
1.3	Weibull survival and hazard functions using random synthetically generated data.	4
1.4	Interpretation of survival neural networks using back-propagation and gene set enrichment analysis.	12

Chapter 1

Introduction

This dissertation focuses on improving survival analysis as applied to the prediction of cancer outcomes from high dimensional genomic data, by leveraging the ability of neural networks to map high-dimensional inputs such as gene expression data to outputs, in this case, risk of death or disease progression. In the following sections, we introduce survival analysis, and genomic data and the unique challenges it poses to prediction models. In section 1.1, we introduce survival analysis and some traditional approaches to modeling survival. In section 1.2, an overview is given of the promises and challenges of learning from high dimensional genomic data, including dimensionality, data insufficiency and the difficulty of interpretation of complex models that work well in high dimensional settings. In order to set the stage to compare existing survival analysis models with our proposed models, in section 1.3, we introduce performance metrics and model selection procedures. In chapter 2, we provide background and previous work in areas related to this dissertation, including machine learning in survival analysis, deep learning, and multi-task learning. Finally, our contributions are presented in chapters 3, 4, and 5.

1.1 Survival Analysis

Survival analysis involves predicting the time to some event of interest, such as the time until cardiovascular death, time until failure of a light bulb, or time until death or progression of disease in cancer. Survival analysis problems differ from binary classification of events in that the time until the occurrence of the event, also known as survival time, event time, or failure time, is important to us. It differs from ordinary regression due to a missing data issue known as incomplete followup or censoring, described in the following section.

1.1.1 Censoring

Survival analysis allows for outcomes to be incompletely determined. For example, consider the case where after a five-year follow-up study of survival after surgery, some patients are still alive. All that is known about the survival time of these cases is that they exceed five years. Another cause of censoring is loss to follow-up, for instance, due to subjects moving out of town (See Figure 1.1). Incomplete or censored observations could provide critical information about long-term survivors and are therefore important to incorporate into the model. In survival analysis, it is usually assumed that censoring is random and non-informative, meaning that it is statistically independent of risk of event. Censoring is non-informative when caused by planned follow-up termination, or by patients moving out of town and lost to follow-up. If patients are removed from follow-up due to factors related to risk of event, such as worsening conditions, then censoring is informative and assuming otherwise would lead to inaccurate statistical inference about survival.

The type of censoring described here is known as *random right censoring* and is the only type of censoring we deal with in this dissertation. Other types of censoring can be observed in survival data, such as interval censoring which happens in presence

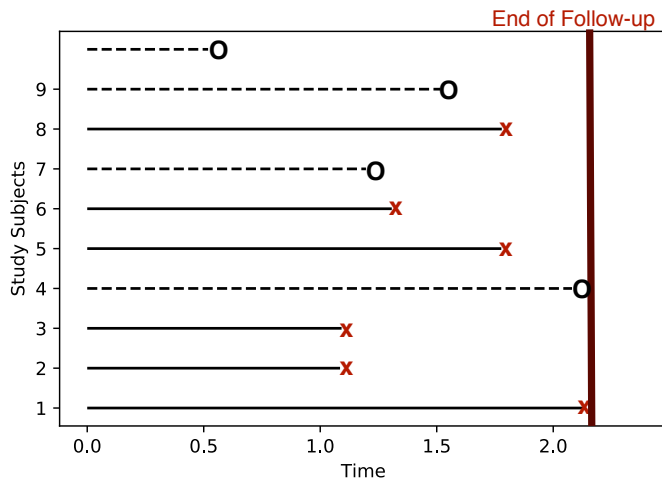


Figure 1.1: Right censoring in time-to-event data. x indicates event occurrence while o marks either loss to follow-up or long term survival, resulting in censoring. Subject 4 is censored due to follow-up termination while other censored subjects are lost to follow-up.

of periodic exams, or left censoring which means the event is only known to have happened before a certain point in time. Censoring is the reason we need models specifically designed to handle time-to-event data and is what differentiates survival regression from regular regression.

1.1.2 Survival and Hazard Functions

Let T be the random variable of waiting time until outcome of interest, and t a specific value of this random variable. The *survival function* is given by:

$$S(t) = 1 - F(t) = P\{T > t\} \quad (1.1)$$

where $F(t)$ is the cumulative distribution function of T . The survival function is the probability that the event of interest occurs after a certain point in time. The *hazard function* $\lambda(t)$ on the other hand, is the instantaneous probability of event at any time t given that the event has not happened up to time t :

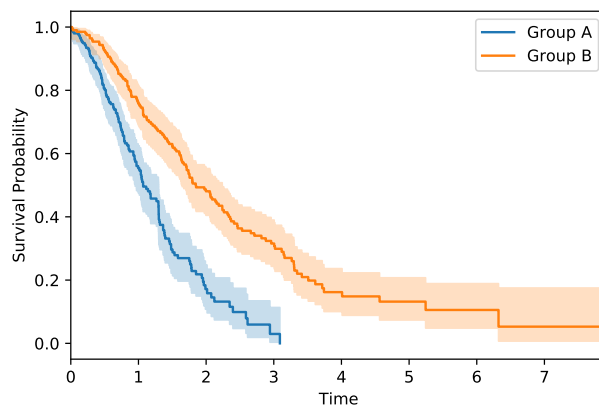


Figure 1.2: Example of Kaplan-Meier curves on random synthetically generated data comparing survival in two groups.

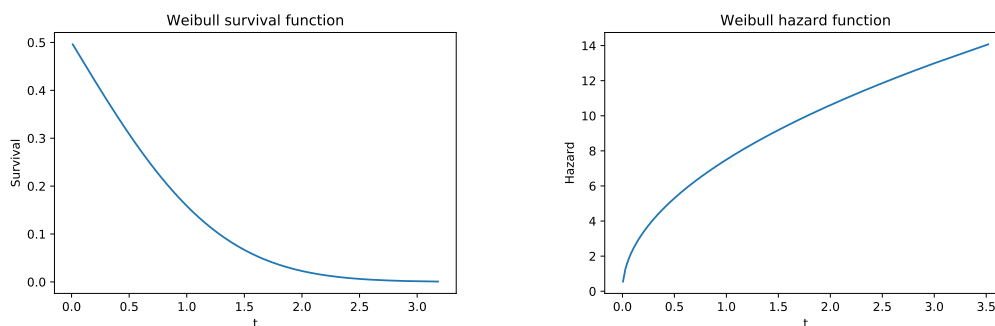


Figure 1.3: Weibull survival and hazard functions using random synthetically generated data.

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P\{t < T < t + \Delta t | T > t\}}{\Delta t} \quad (1.2)$$

The true form of $S(t)$ is almost always unknown and various assumptions have been used in the literature to describe the distribution of event times. The exponential and Weibull distributions are the two most popular parametric survival distributions, assuming specific shapes for the distribution of survival times T . The hazard function for exponentially distributed T is simply a constant γ , and the Weibull hazard function is a generalization of the exponential case:

$$\lambda(t) = \alpha \gamma t^{\gamma-1} \quad (1.3)$$

Unlike the exponential and Weibull distributions that make strong assumptions about the form of the survival function, non-parametric models like Kaplan-Meier estimate the distribution without making any assumptions. Kaplan-Meier product-limit estimator estimates $S(t)$ by calculating the probability of surviving at time t as the product of the probability of surviving up to time t and the probability of surviving at time t after surviving $t - 1$ days. Let k denote number of unique event times, t_1, t_2, \dots, t_k denote unique event times, d_i denote the number of events at t_i and n_i the number of *at-risk* subjects at time t_i (that is subjects with event or censoring times greater than t_i). Then the Kaplan Meier estimator is:

$$S_{KM} = \prod_{i:t_i < t} (1 - d_i/n_i) \quad (1.4)$$

The above traditional parametric and non-parametric models do not differentiate between subjects based on predictors $x = x_1, x_2, \dots, x_p$ and assume the same hazard function for all subjects. x can be a combination of p continuous, binary and categorical predictors for a subject that affects survival. In order to take these predictors into account, we need to generalize survival models to survival regression models. One

way to do so is the *proportional hazards* model that multiplies the hazard function by $\exp(x\beta)$, generalizing from a hazard function for a time t to a hazard function for a time t given predictors x :

$$\lambda(t|x) = \lambda(t)\exp(x\beta) \quad (1.5)$$

The proportional hazards assumption implies no interaction between t and x , in other words, the predictors have the same effect on the hazard at all points in time. Specific survival models such as the exponential model can be used in the above proportional hazards formulation to create specific survival regression models.

1.1.3 Cox's Proportional Hazards Model

The most widely used approach to survival analysis is the semi-parametric Cox proportional hazards model (Cox, 1972). Unlike the exponential or Weibull models, it makes no assumptions about $\lambda(t)$ but assumes a parametric relationship between the predictors and the hazard function. It is considered a linear model since the predictors are linearly related to the log hazard. In cases where $\lambda(t)$ is not primarily interesting to one's research purposes, the Cox model allows one to completely ignore it by showing how to estimate parameters β without knowledge of $\lambda(t)$. The parameters of the model are estimated by optimizing Cox's partial log-likelihood. Note that $\lambda(t)$ is dropped out of the likelihood function:

$$l(f_\beta(X), Y) = - \sum_{x_i \in U} \left(f_\beta(x_i) - \log \sum_{j \in R_i} e^{f_\beta(x_j)} \right) \quad (1.6)$$

where $X = \{x_1, \dots, x_N\}$ are the samples, and $Y = \{e, t\}$ represents label vectors of overall survival $t = \{t_1, \dots, t_N\}$ and event status $e = \{e_1, \dots, e_N\}$. Function f is a linear function with parameters β , U is the set of uncensored samples, and R_i is the set of at-risk samples with survival or follow-up times $t_j \geq t_i$. The above notation for

Cox's partial log-likelihood inherently uses Breslow's approximation (Breslow, 1974) in cases with tied event times.

1.2 Machine Learning and Genomic Data: Challenges

Since the emergence of high throughput experiments such as Next Generation Sequencing, the volume of genomic data produced has been increasing exponentially, with the volume of sequence data doubling every seven months over the last decade (Stephens et al., 2015). A single biopsy can generate tens of thousands of transcriptomic, genetic, proteomic, or epigenetic features. This fascinating growth rate poses challenges to the storage, distribution, and analysis of genomic data (Kahn, 2011). One of the technological needs for this exponentially growing genomic data is the development of large-scale machine learning methods to translate these data into clinically actionable information. So far, the rate of genomic data generation has far exceeded the ability to design predictive machine learning models that can map this high-dimensional data to optimal treatment options, as usually only a handful of known genetic features are used in clinical decision making.

Machine-learning has emerged as a powerful tool for analyzing high-dimensional data, with open software tools such as Tensorflow (Abadi et al., 2015) that enable scalable and distributed data analysis. A sub-field of machine learning, known as deep learning, has recently achieved remarkable success in learning from high dimensional images and sequences (LeCun et al., 2015). Some machine learning approaches have been employed for genomic applications and survival analysis (Wang et al., 2017b; Leung et al., 2015). Additional research is needed to recognize and address the specific challenges of applying the quickly growing techniques of deep learning to survival analysis, and to develop interpretable models that can improve understanding of

patient outcomes and disease biology.

1.2.1 Genomic Data, Dimensionality, and Heterogeneity

A major challenge in applying machine learning to genomic data is *curse of dimensionality* (Bellman, 1961). Curse of dimensionality in machine learning refers to the fact that the feature space expands exponentially with the dimensionality of the space, leading to the need for an exponentially larger training set in order to train a model that can generalize. One way to demonstrate this is in terms of sample density. If N is a dense sample of a single-dimensional space ($d = 1$), then we need N^d samples to densely represent a d -dimensional space. This renders all feasible datasets sparse in high dimensions.

Another way to look at this is that when the number of features exceeds the number of observations (p greater than n scenario), we are dealing with an under-determined system that has many solutions, not all of which will generalize to unseen data. This becomes an issue particularly in complex models like neural networks that have great representation power (large hypothesis sets) and are prone to over-fitting in such scenarios. Generally, the more parameters a machine learning model has, the more independent samples it requires for being able to differentiate meaningful patterns from noise in data (Abu-Mostafa, 1989) in order to avoid over-fitting. Deep learning models have many parameters due to their layered nature. Cancer genomic datasets, on the other hand, usually consist of only hundreds of samples compared to, for instance, 23K features in the case of gene expression data. This data insufficiency issue is further pronounced in survival analysis, where large fractions (e.g. 90% in TCGA breast cancer data) of available samples have incomplete labels (see section 1.1.1).

Several approaches have been employed to alleviate this data insufficiency including data augmentation and transfer learning (Ching et al., 2018). Data augmentation,

although very helpful in image recognition tasks and successfully applied to mammograms (Dhungel et al., 2015) and histopathology images (Litjens et al., 2016), cannot be trivially applied to gene expression data. Augmented samples are generated by introducing small changes to training samples that do not change the underlying meaning, such as mirroring in images. This requires domain expertise of gene regulatory pathways in the case of gene expression data.

In Chapter 4, we use multi-task learning (a type of transfer learning) and adversarial representation learning to tackle this challenge by enabling learning from pools of several datasets, increasing training set size for each individual problem. Pooling data from multiple studies and hospitals is indeed a promising solution to the problem of sample size limitation. But the heterogeneity of available genomic datasets due to technical and sample biases poses challenges to integrating multiple data sources, and ignoring this heterogeneity can lead to incorrect conclusions. Cohorts from multiple sources typically have different demographic or disease stage distributions, may be subject to different signal capture calibration, post-processing artifacts, and naming conventions. This problem is also referred to as *batch effects* in the literature and means that naively combining heterogeneous cohorts is both difficult and may degrade model accuracy (Tom et al., 2017). Batch effects are a common challenge in high throughput experiments and are caused by laboratory conditions, personnel differences, and other factors that are not of clinical interest but could mislead us to incorrect conclusions if they are confounded with the outcome of interest.

A significant amount of work has been done in the area of normalizing datasets for integration and removing batch effects. Many of these methods are based on linear regression and singular value decomposition, and make numerous assumptions such as orthogonality of the batch effect and biological variation, the ability of humans to distinguish between batch effects and biological effects, and assumptions on the batch structure (see for example Leek et al. (2010); Haghverdi et al. (2018)). Another

limitation of such methods is that they do not take a learning objective into account to distinguish between relevant variations and batch effects. In chapter 4, we propose two multi-task learning solutions to handle learning from heterogeneous genomic data sources without letting generalization suffer from batch effects.

1.2.2 Interpretability

The most important challenge in employing complex machine learning methods such as deep learning in critical decision-making problems is interpretability. Despite the recent success of such models, we still have little understanding of how decisions are made by them. The concerns about lack of transparency behind these models have hindered their wide application to critical decision-making applications such as cancer care. Interpretability and explainability can be used to validate and gain confidence in machine learning systems and encourage wider adaptation of them. Moreover, they help researchers and developers understand the problem better, and discover causes of failure, eventually leading to better models.

Specifically, by interpreting survival neural networks, we could validate them by comparing what we already know about important factors in survival and what the model finds important. Moreover, highly ranked biomarkers in our model could be investigated by biologists to understand their role or function, or could be targeted in therapies. In our work described in section 3.3, novel patterns that are suggested by the network could be incorporated into diagnostic criteria, and we could train pathologists to recognize these novel patterns.

In this dissertation, we design and implement techniques to explain the predictions made by neural networks about disease outcomes in both single task (chapter 3) and multi-task (chapter 4) settings. This approach is based on the calculation of partial derivatives of the risk prediction with respect to the input features as first proposed by Dimopoulos et al. (1995), and provides a ranking of input features with respect

to their importance in the predictions made by the model. Since the outputs of our models are always risk predictions, we refer to the partial derivative of the model predictions with respect to features as the *risk scores* of the features. In order to motivate the use of partial derivatives as a measure of feature importance, let us start by looking at a linear model of risk:

$$f_{\beta}(x) = x \top \beta \tag{1.7}$$

In the case of the above linear model, it is easy to take elements of the parameter vector β corresponding to each feature in x as a measure of the importance of that feature. In a neural network, however, we are dealing with a highly non-linear function, which we can approximate using the first-order Taylor expansion in the neighborhood of a given sample x_0 :

$$g_W(x) = g_W(x_0) + \left. \frac{\partial g_W(x)}{\partial x} \right|_{x=x_0} (x - x_0)$$

$$g_W(x) = \left. \frac{\partial g_W(x)}{\partial x} \right|_{x=x_0} x + c$$

Where c is a constant with respect to x . The partial derivative of the neural network predictions with respect to the input act as the feature coefficients in the linear function f_{β} .

The risk scores of transcriptional features are later used in Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) to identify pre-specified gene sets enriched with genes that highly contribute to model predictions. Gene Set Enrichment Analyses analyses the risk scores of genes on the gene-set level, by taking into account domain knowledge of genetic pathways and previously discovered co-expressions, and determines whether members of each gene set occur at the top or bottom of the gene ranking as opposed to being randomly distributed. If a gene set S is enriched, then

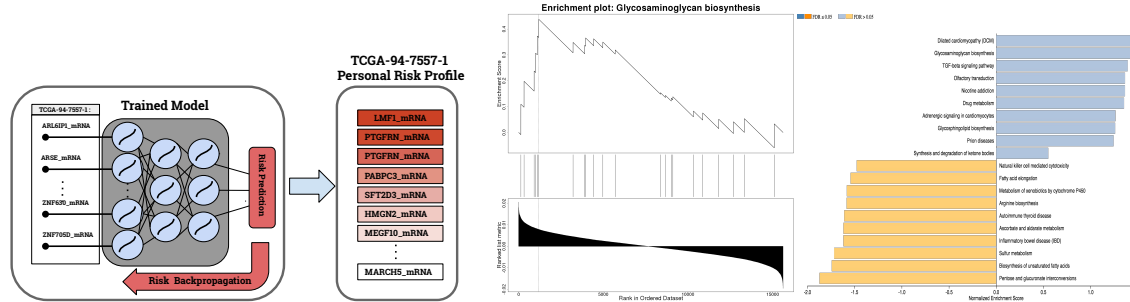


Figure 1.4: Interpretation of survival neural networks using back-propagation and gene set enrichment analysis.

we can conclude that the model is basing its predictions on the biological phenotype represented by S .

An enrichment score for S is calculated by walking down the list L , increasing the score when we encounter genes in S and decreasing it when we encounter genes that are not in S . The enrichment score is the maximum deviation from zero encountered in the walk. After calculating an enrichment score for each gene set, a permutation test is performed to determine the statistical significance of the score. Importantly, in the original GSEA method, the permutation is performed on the phenotype level as opposed to the individual gene level to preserve the correlation structure of gene expression data. But in our application, since we are not dealing with different phenotypes, we use pre-ranked GSEA that uses gene-set level permutations for significance analysis. We use this interpretation procedure in sections 3.2, 3.4 and 4.2. In order to make comparisons between the decision-making mechanisms of two models, Gene Set Enrichment Analyses were performed on the risk scores for each model to identify differences in pathway enrichment between the two models.

1.3 Model Selection and Evaluation

1.3.1 Performance Metrics

For survival models, discrimination means separation between survival curves for individuals groups. Kaplan Meier curves of different prognostic groups have been used to provide an informal and visual measure of discrimination, but the use of this analysis is challenged by possible residual confounding when the categorization of samples into prognostic groups does not exactly capture the relationship between input and outcomes. Several measures of discrimination have been proposed (Royston and Altman, 2013), including but not limited to Harrell’s concordance index (c-index) (Harrell Jr et al., 1982).

Concordance Index

Throughout this dissertation, we measure model performance using Harrell’s *concordance index* (CI) that captures the rank correlation of predicted and actual survival. Denoting the i th patient with X_i and the set of all patients with X , where t_i represents either the time of death or the time of last follow-up of the i th patient, CI was calculated in the following way:

$$CI(\beta, X) = \sum_P \frac{I(i,j)}{|P|} \quad (1.8)$$

$$I(i,j) = \begin{cases} 1, & \text{if } Risk_j > Risk_i \text{ and } t_j > t_i \\ 0, & \text{otherwise} \end{cases} \quad (1.9)$$

Where P is the set of orderable pairs. A pair of samples (X_i, X_j) is orderable if either the event is observed for both X_i and X_j , or X_j is censored and $t_j > t_i$. In other words, CI measures the proportion of all orderable pairs of samples i and j where the prognostic scores $Risk_i, Risk_j$ predicted by the model and the actual

times of death t_i and t_j are concordant. Due to the censoring of event times, not all possible pairs are orderable. A pair of samples is orderable only if the sample with smaller t is not censored.

Attempts have been made to propose differentiable versions of CI based on sigmoid and exponential functions and optimize it directly. But recent studies show that optimizing the cox partial likelihood is equivalent to optimizing CI (Steck et al., 2008). In order to evaluate the performance during a follow-up period, Heagerty and Zheng (2005) defined a time-dependent version of CI for a fixed follow-up time period as the weighted average of AUC values at all possible observation time points.

1.3.2 Hyper-parameter Optimization

The use of deep learning techniques requires tuning of several hyper-parameters; parameters that are not learned during training but are configuration decisions made by the developer. The number of layers, regularization rate, learning rate, size of layers, and type of non-linearity used in hidden units are examples of such hyper-parameters.

Deep learning researchers typically use a mixture of random search, grid search, and expert knowledge to pick values for hyper-parameters, which emphasizes the lack of a systematic scalable method of hyper-parameter tuning. The difficulty of finding the optimal configuration, makes parameter-free machine learning models appealing. A more flexible approach to this issue is to automate the hyper-parameter optimization procedure.

Random Search and Grid Search

Random search and grid search are common alternatives to manual search. In grid-search, all possible values within a predefined range of each hyperparameter are evaluated on a validation set, and the set of hyper-parameters with the smallest loss on

the validation set is selected for final model evaluation. In random search, a pre-defined number of random hyperparameter values are evaluated. It has been shown that random search is often able to reach similar or close local minima in a fraction of loss function evaluations required with grid search (Bergstra and Bengio, 2012). For smaller models that can be evaluated in seconds, grid search could still be a reasonable hyper-parameter search method.

We use grid search for model selection in Chapter 4. Only a couple of hyperparameters (learning rate, ℓ_2 regularization rate) proved to be of significant effect so a smaller grid would lead to the same loss function value as a larger one. We selected a smaller grid in exchange for more randomized experiments in order to be able to measure the statistical significance of the results.

Bayesian Optimization

Bayesian Optimization is an Estimation of Distribution Algorithm (ADE) that was proposed as a method of globally optimizing expensive black-box functions (Kushner, 1964; Moćkus, 1975) and was later applied to hyperparameter tuning of neural networks. It works by assuming the cost function is a sample of a Gaussian process and maintaining the posterior distribution of the assumed distribution as it makes function evaluations. Nearly all previous Bayesian optimization approaches make one evaluation at a time. This is one of the major reasons Bayesian Optimization was mainly ignored by the machine learning community.

Let's use f to denote a prior measure of possible functions that describe how the neural network cost function changes with hyper-parameters. Mathematically we are considering the problem of finding a global maximizer (or minimizer) of an unknown objective function f .

$$x^* = \mathit{arg\,max}(f(x))$$

Where x belongs to R^d , and d denotes the number of hyperparameters to tune.

The Bayesian optimization framework has two components. The first component is a probabilistic surrogate model, which consists of a prior distribution that captures our beliefs about the behavior of the unknown objective function, and a likelihood model that describes the data generation mechanism. The second ingredient is a loss function that describes how optimal a sequence of queries is. The expected loss, or acquisition function as we will call it from now on to be consistent with the literature, is then minimized to select an optimal sequence of queries. After observing the output of each query of the objective, the prior is updated to produce a more informative posterior distribution over the space of objective functions. Acquisition functions trade-off between exploration (evaluate in places where variance is high.) and exploitation (evaluate in places where mean is low) in order to decide where to evaluate next.

Expected improvement (Jones et al., 1998) is an acquisition function that we will define shortly. Let

$$f_n^* = \min[y^{(1)}, \dots, y^{(n)}]$$

denote the current best value of the cost function value. Before we evaluate point x^{n+1} , we do not know what $y(x^{n+1})$ is. But we can model the uncertainty around $y(x^{n+1})$ by treating it as the realization of a normally distributed random variable Y with mean and standard deviation given by a stochastic process model. The improvement at the point x^{n+1} is

$$I = \max(f_n^* - Y, 0).$$

This expression is a random variable because Y is a random variable. To obtain the expected improvement we simply take the expected value

$$EI(x) = E[\max(f_n^* - Y, 0)].$$

The above acquisition function can be written in closed form in terms of standard

normal density and cumulative distribution function and can be optimized using any gradient-based optimization to find the most promising point to evaluate next. We can use any optimization method for the acquisition function since it is not as expensive to evaluate as our original cost function.

A parallel formulation of this problem, called q-point EI or q-EI is introduced in Ginsbourger et al. (2007), where expected improvement (EI) is measured based on a set of observations, and the next samples are selected to maximize the EI:

$$q - EI(x_1, \dots, x_q) = E[\max(f_n^* - \max_{i:1..q} f(x_i), 0)], \quad (1.10)$$

where f_n^* is the best evaluation so far. One of the challenges involved with parallel hyperparameter tuning of deep learning models is the variable time of cost function evaluations; it takes longer to evaluate a model with 1000 hidden units than a model with 10 hidden units. Snoek et al. (2012) propose a parallelization scheme that takes this variable evaluation time into account.

In sections 3.1 and 3.2 we use Bayesian optimization as implemented in (Martinez-Cantin, 2014) for systematic and unbiased model selection.

Chapter 2

Previous Work

2.1 Machine Learning for Survival Analysis

High-dimensional learning problems are commonly dealt with in machine-learning, but survival analysis has been largely overlooked by the community. Some machine-learning approaches have been applied to predicting survival or time to progression (Kourou et al., 2015; Wang et al., 2017b). Feature engineering solutions such as dimensionality reduction based on prior knowledge have been used by learning gene signatures of cancer hallmarks to generate intermediate features that successfully predict outcomes (Li et al., 2010; Gao et al., 2016). Regularization methods for Cox models like elastic net have been developed to perform objective and data-driven feature selection with time-to-event data (Park and Hastie, 2007; Simon et al., 2011).

Support Vector Machine (SVM) (Cortes and Vapnik, 1995) is a very successful supervised learning algorithm used mostly for classification. It can be efficiently applied to problems of non-linearly separable classes using the kernel method. Although originally proposed to solve classification problems, SVM can be modified for regression problems (SVR) (Drucker et al., 1997) and has also been successfully adapted in survival analysis (Van Belle et al., 2011). Incorporating censored observations has

been done via updating the SVR loss function and constraints as in Khan and Zubek (2008); Shivaswamy et al. (2007). The difference between these two models lies in the penalty for incorrect predictions. In `shivaswamy2007support`, incorrect predictions are penalized irrespective of whether the prediction was sooner or later than the observed failure time, and whether the sample is censored or observed. Incorrect predictions for right-censored data are penalized only if the prediction is sooner than the observed censoring time. On the contrary, `khan2008support` applies different penalties for the four possible cases. A major drawback of the latter method is the large number of hyper-parameters. The time complexity of SVM-based survival models usually follows the complexity of the original SVM and is another drawback of such models.

Decision tree algorithms have been adapted to censored data. Tree models work by recursively partitioning the data based on a splitting criterion so that similar samples will be placed under the same leaf node. By treating time as a covariate, we can convert the survival analysis problem to a classification problem and solve it with decision trees. Regression trees have been adapted to censored data using different splitting and pruning techniques. In Gordon and Olshen (1985), the Wasserstein distance between Kaplan Meier curves was used as a splitting criterion and CART pruning was used. Segal (1988) propose a new goodness-of-split measure based on two-sample statistics for censored data. In Ture et al. (2009) bagging was applied to survival trees. Random forests, a variation on bagged decision trees, have been adapted to survival modeling (Ishwaran et al., 2008). There are mainly four steps in RSF: (i) Draw several random bootstrap samples from the data. (ii) For each sample, build a survival tree by splitting each node using the best candidate feature from a random subset of features (iii) continue until every terminal node has greater than or equal to a certain number of events. (iv) Using the non-parametric Nelson-Aalen estimator, calculate the ensemble cumulative hazard function (CHF) of out-of-bag

data by taking the average of the CHF of each tree.

Neural network-based approaches have been used in low-dimensional survival prediction problems (Faraggi and Simon, 1995) using a neural network with a linear output layer and a single logistic hidden layer, but subsequent evaluation of these methods found no performance improvement over linear Cox regression (Xiang et al., 2000). With the goal of exploiting complex feature interactions, authors in (De Laurentiis and Ravdin, 1994) propose three neural network models: a network that uses time as an input variable and is trained on the status of the patient at that time, with several training points for each patient at different follow-up times; a single time point model that produces predictions at only one specific time; and a multiple time point model that essentially consists of multiple single time point models running in parallel. The authors in (Biganzoli et al., 1998) propose the partial logistic artificial neural network (PLANN), a single hidden layer neural network with logistic activation for grouped discrete hazard prediction that takes time as one of its inputs. In (Lisboa et al., 2003), the PLANN was extended to a Bayesian neural framework with covariate-specific regularization to carry model selection using automatic relevance determination.

2.2 Deep Learning

2.2.1 Background

Since it was demonstrated in 1969 that the 2-layer perceptron is incapable of representing functions outside a very special class, researchers have been exploring the theoretical capability of multi-layer neural networks to represent general functions. In (Hornik et al., 1989), the authors establish that multi-layer feed-forward neural networks are capable of approximating any measurable function to any desired degree of accuracy, and lack of success in applications must be due to insufficient model

complexity and lack of deterministic input-output relationship. Authors in (Blum and Rivest, 1988) also address the question of computational complexity and show that for a 3 node neural network, the training problem is NP-complete. They define the training problem as a decision problem of whether weights exist for a given neural net and training sample to always output the correct label. They extend their proof to several other classes of networks including 2 layer networks with more than two hidden nodes. They show that one way to get around this intractability is to expand input representation.

Recently, deep neural networks have been successfully applied to various domains and have shattered performance benchmarks. One of the remarkable successes of deep learning was to achieve one of AI’s grand challenges by beating the human expert in the full-sized game of Go (Silver et al., 2016). The difficulty of Go lies in the evaluation of board positions and the enormous search space. Traditional search space reduction (approximate position evaluation and action sampling) approaches do not help much with Go. Most successful previous works regarding Go rely on Monte Carlo Tree Search using shallow policies or value functions. Based on recent successes of deep reinforcement learning, the authors train a supervised 13-layer convolutional network with SGD to predict expert moves. The input to this network is the 19×19 image of board positions. They then train a fast policy that can rapidly sample actions during MC roll-outs, a reinforcement learning policy network to adjust the fast policy network toward the goal of winning rather than predicting, and a reinforcement learning value network that predicts the winner of self-play.

Another work that showed the potential of deep learning was image caption generation. Vinyals et al. (2015) combine convolutional neural networks and LSTMs to learn image representations, and generate each word of the image caption given the image representation and all previously generated words. The authors employ a variety of methods such as Inception (Szegedy et al., 2015), batch normalization (Ioffe

and Szegedy, 2015), and encoder-decoder networks (Cho et al., 2014).

The influence of initialization and saturating and non-saturating activation functions on the behavior of feedforward neural networks (FFNNs) is studied in (Glorot and Bengio, 2010), mainly by monitoring activations and gradients. Experiments show that random initialization with gradient descent performs poorly in deep nets, so the authors propose a novel initialization scheme, *normalized initialization*, to tackle this issue by maintaining the flow of activations (forward) and gradients (backward). They experiment on several datasets with the goal of detecting saturation and overly linear units and discover that with *sigmoid* units, saturation in the final layer happens quickly but can be escaped if the network is not too deep as apposed to *tanh* units where saturation start from the first layer and propagates up to the top layer. The proposed initialization technique maintains the forward and backward flow and successfully prevents saturation. Normalized initialization and both saturating and non-saturating activation functions are used in this dissertation.

Overfitting is a common concern with neural networks and is defined as the problem of learning complicated relationships between input and output due to sampling noise. Deep neural networks are particularly prone to overfitting when training data is limited, because of their great expressiveness. Many methods have been proposed to reduce overfitting in learning in general, including early stopping and ℓ_1 and ℓ_2 norms. A couple of more recent regularization methods proposed specifically for neural networks are described here.

Dropout is a technique for both regularization and bagging (Srivastava et al., 2014). Dropping a unit out means temporarily removing it from the network along with its input and output connections. At training time, each unit is dropped with a probability p which should be determined using a validation set. At test time, an approximate averaging method is employed to use a single model without dropout; the outputs of all the units are multiplied by $1 - p$ to ensure that the expected output

is the same as it was at training time. By doing this 2^n networks are integrated into a single one.

Dropout can be used to fine-tune networks that have been pre-trained unsupervised. After pre-training, the weights of the network should be multiplied by $1/(1-p)$ to maintain the expected outputs. The learning rate for fine-tuning should be picked carefully in order not to wipe out the information obtained from unlabeled data during pre-training. One reason offered by the authors on why dropout leads to performance improvements is that it prevents co-adaptation of features: each hidden unit is encouraged to learn a meaningful feature without relying on other units (Hinton et al., 2012).

DropConnect is a generalization of Dropout where weights are dropped instead of hidden units. The authors in (Wan et al., 2013) derive an upper bound on the complexity of the model and show that it is a linear function of the dropout rate p . They demonstrate on four datasets that sometimes DropConnect outperforms Dropout. We use a combination of Dropout and ℓ norms in this dissertation.

2.2.2 Representation Learning

The idea of representation learning comes up frequently in this dissertation. Representation learning is the idea of moving from the engineering of hand-crafted features to letting the AI decide what features are explanatory of the data. In his review of deep representation learning (Bengio, 2013), Bengio explains the importance of representation learning to the success of machine learning algorithms, and how recent successes in training deep networks promise advances in learning good representations. Computational scalability, optimization, and feature disentanglement are named as challenges facing representation learning, and current approaches to these challenges are reviewed. In chapter 3, we train neural networks to transform raw high-dimensional genomic and histology data into predictive features of survival, guided

only by the objective function as opposed to pre-processing, feature-engineering or explicit feature selection. In Chapter 4, we focus on learning a shared representation that is predictive for multiple tasks.

2.2.3 Convolutional Networks

One of the major breakthroughs in applying neural networks to images is the idea of convolutional networks, first applied to handwritten digit classification (Le Cun, 1989). Since their introduction, convolutional networks have been applied to many computer vision applications such as image classification (Krizhevsky et al., 2012), general object detection (Ren et al., 2015), object segmentation (Long et al., 2015), facial recognition (Lawrence et al., 1997), nucleus detection (Xie et al., 2018), cancer diagnosis (Esteva et al., 2017; Cruz-Roa et al., 2014), genetic variant calling (Poplin et al., 2018), and many more applications.

In section 3.3, we demonstrate the utility of convolutional neural networks in cancer prognosis from a combination of histopathology and genomic data. In chapter 5, we use general-purpose object detection convolutional neural networks for detecting and clustering nuclei in histopathology images. In the rest of this subsection, we provide some background on convolutional networks and the theory and assumptions behind them.

Le Cun (1989) introduces the motivation and technique for using local convolutional feature maps. Minimal preprocessing i.e. size normalization, removal of extraneous marks, and scaling of grayscale images are done and the remainder of the recognition is done by the adaptive layers of the network. To avoid dealing with a lot of parameters, they use the following prior knowledge from shape recognition:

- Parameter sharing based on statistical stability: if a feature detector (filter) is useful in one location of the image, it is probably useful in other locations as well. So we apply the same filter with the same parameters over all of the

image.

- Learning local features: instead of looking at the whole picture they look at local blocks and combine them later.

The first application of a large and deep convolutional to the Imagenet dataset was by Krizhevsky et al. (2012). They used ReLU activation units along with dropout to stack and train 5 convolutional layers and 1 feedforward layer and achieved winning top-1 and top-5 performance in the ILSVRC 2010 task.

Long et al. (2015) use fully convolutional classification networks (with no fully connected layers) with skip connections for semantic image segmentation. The output of their model is a heatmap for each possible object which is normally very low resolution because of pooling in the intermediate layers, but they develop their network further by using skip connections from intermediate layers to capture high-resolution edges while preserving the big picture obtained at the final layer.

Typically, researchers have used multiple layers of convolutional layers, increasing the number of filters in each layer, topped with one or more fully connected layers. In order to build deeper models from larger datasets such as the Imagenet, we need to move toward sparse designs to avoid computational intractability. Inception (Szegedy et al., 2015) is a successful architecture proposed to participate in the ILSVRC 2014. The main idea of Inception is to approximate the local sparse structure of human vision with regularly used dense components. They use a concatenation of outputs from different resolution filters and apply 1×1 convolutions to before applying their 3×3 or 5×5 filters in order to reduce the dimensionality. This allows us to increase the number of filters as we go deeper, without facing a computational complexity blow-up.

2.2.4 Adversarial Learning

It was discovered in 2014 (Szegedy et al., 2013) that neural networks are vulnerable to examples that are only slightly different from examples from the data distribution that the model correctly labels. The fact that many different machine learning models fail on these *adversarial examples* is an intriguing discovery. Goodfellow et al. (Goodfellow et al., 2014b) show that linear behavior in high dimensional spaces is a sufficient explanation of vulnerability to adversarial examples, and propose the *fast gradient sign method* to generate adversarial training samples for additional regularization. The adversarial examples generated by this method cause the error rate to increase dramatically, supporting the linearity explanation. They argue that although all machine learning models are vulnerable to adversarial examples, the universal approximator theorem (Hornik et al., 1989) guarantees that deep neural networks are at least capable of representing functions that are robust to adversarial perturbations. Other machine learning methods including linear methods such as SVM and logistic regression were later found to be vulnerable to adversarial examples, but deep neural networks benefit the most from training on adversarial examples. Adversarial training acts as a regularizer and mitigates overfitting in deep neural networks.

Generative adversarial neural networks (GANs) (Goodfellow et al., 2014a) were proposed to generate data that resembles the training data. GANs consist of a generator (G with parameters θ_G) and a discriminator (D with parameters θ_D) component. The discriminator's job is to tell between synthesized samples from the generator and real samples from the training data, while the generator's objective is to deceive the discriminator. As a result of this competition, the generator learns to match its representation with the real data representation leading to realistic generated data.

The generator maps random noise z drawn from a distribution $p_z(z)$ to the training data space. Let's call the generated data space p_g . The discriminator takes a vector x that is either sampled from the real data distribution $p_{data}(x)$ or the generated data

distribution $p_z(z)$, and outputs a probability prediction of the input coming from the real data distribution. The objective function of a GAN is, therefore, a min-max game between the generator and discriminator with the value function $V(\theta_D, \theta_G)$:

$$V(\theta_D, \theta_G) = \mathbb{E}_{x \sim p_{data}(x, \theta_D)} \log(D(x)) + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z, \theta_G), \theta_D))] \quad (2.1)$$

Ideally, we want D to output 0.5 for every input at the equilibrium point, meaning that it can not distinguish between real and generated data. (Goodfellow et al., 2014a) proves that if G and D have enough capacity and D is allowed to reach its optimal point at every step of training G , then convergence to $p_g = p_{data}$ is theoretically guaranteed.

In practice, however, the main practical challenge in training adversarial networks is convergence. Unlike optimizing a single loss function, optimizing two agents that are playing a game against each other may never converge. An example of this non-convergence challenge is mode collapse, where the generator visits isolated modes of the distribution that maximize discriminator’s loss, instead of visiting multiple modes (Metz et al., 2017). Solving the mode-collapse issue is an active area of research (Salimans et al., 2016).

Adversarial training has found many applications and extensions since its introduction. (Abadi and Andersen, 2016), for instance, train a multi-agent adversarial system where the objective of two agents (Alice and Bob) is to communicate clearly and confidentially, and the objective of a third agent (Eve) is to eavesdrop. They show that with sufficient training, Alice and Bob learn to counter the objective of not a fixed Eve, but the best possible version of Eve.

Adversarial learning of deep domain-invariant features has been used for domain adaptation in recent years (Ganin et al., 2016; Hoffman et al., 2017; Tzeng et al., 2017, 2015). In this direction of research, a domain confusion loss is maximized

by the model, while at the same time a domain discriminator tries to differentiate between samples from different domains. Minimization is done with respect to parameters of the domain classifier, while maximization updates parameters of the learned representation. All this is done simultaneously with training for the original purpose of the model, so a useful task-related domain-invariant representation is learned. For example in (Ganin et al., 2016), an adversarial discrimination loss is used in a neural network for domain adaptation in document sentiment analysis and image classification. In (Bousmalis et al., 2017) a similar model with a generative base is used for pixel-level domain adaptation, which changes images of the source domain to look like they come from the target domain. In (Kamnitsas et al., 2017) the same idea is used in domain adaptation for brain lesion segmentation. The same idea has been applied to biomedical relation extraction from text (Rios et al., 2018). Authors in (Tzeng et al., 2017) present a unified framework to describe a wide range of domain-adversarial approaches, and proposed a novel adversarial domain-adaptation method with untied source and domain weights. In chapter 4, we describe a similar model to learn from heterogeneous genomic data sources without suffering from batch effects that are common in high-throughput experiments.

2.2.5 Interpretable Deep Learning

As discussed in section 1.2.2, the most important challenge in employing complex machine learning methods such as deep learning is interpretability. A recent survey (Du et al., 2018) summarized the progress in achieving interpretable machine learning and categorized proposed interpretable models into two main categories: intrinsic interpretability and post-hoc interpretability.

Intrinsic interpretability refers to self-explanatory models that incorporate interpretability directly into the model structure, such as linear regression and decision trees. Cox’s proportional hazards model provides a representative example for this

category, as the coefficient of each feature in the trained model indicates the prognostic value of that feature. An example of intrinsic interpretability in deep learning is the use of attention mechanisms in sequential models (Xu et al., 2015).

Post-hoc interpretability aims to shed light on the mechanisms of prediction after the model is trained. *Permutation feature importance* is a traditional model-agnostic approach to post-hoc interpretation of machine learning models which was originally proposed to explain the prediction mechanism of random forests (Breiman, 2001). In this approach, the performance of the model is re-measured after shuffling the values of each feature. The difference between the original performance and the new performance achieved after shuffling the values of a feature provides a measure of the importance of that feature.

Modern neural networks are commonly explained using gradient-based methods after training, where the gradient of the model predictions (or some variant of the gradient) is back-propagated to the feature space (Dimopoulos et al., 1995; Simonyan et al., 2013; Zeiler and Fergus, 2014). The magnitude of the gradient then provides a measure of importance for each feature. In our application, in addition to the magnitude of the gradient of risk with respect to features, the sign of the gradient is also important, with negative gradients indicating genes that have a positive effect on survival, and vice versa.

2.3 Multi-task Learning

Every time we are optimizing more than one loss function, we are doing multi-task learning. Learning multiple related tasks simultaneously has been both empirically and theoretically shown to significantly improve performance relative to learning each task independently (Argyriou et al., 2007). In his presentation at the Constructive Induction Workshop at the 1994 International Conference on Machine Learning, Rich

Sutton emphasizes the importance of multi-task learning:

”The standard machine learning methodology is to consider a single concept to be learned. That itself is the crux of the problem... Instead we should look to natural learning systems, such as people, to get a better sense of the real task facing them. When we do this, I think we find the key difference that, for all practical purposes, people face not one task, but a series of tasks. The different tasks have different solutions, but they often share the same useful representations.”

Multi-task learning is particularly helpful when only a few data per task are available and with multi-task learning, each task has more data to learn from. In high-dimensional problems where it is particularly difficult to find relevant features, additional tasks provide additional evidence for the relevance of certain features helping focus the model’s attention on meaningful patterns in data. Another way to explain why multi-task learning works is in terms of inductive bias. Just as ℓ_1 inductive bias constraints learning to the hypotheses that are sparse, multi-task learning constraints it to the hypotheses that can explain more than one task (Ruder, 2017).

One of the earliest examples of multi-task learning was the use of *hints* by Abu-Mustafa (Abu-Mostafa, 1990) where invariance hints were used to incorporate in learning any information we already have about the predictive function. Caruana argues that when it is impractical to include some features as input to the model (such as in-hospital measurements for predicting hospital admittance), it may still be useful to include those features as outputs in a multi-task learning setting (Caruana and De Sa, 1997). Caruana takes multi-task learning to an extreme by arguing that even multi-task learning of small variations of the exact same task can be helpful empirically (Montavon et al., 2012). Since then, multi-task learning has developed and been used across all applications of machine learning including computer vision

(Zhang et al., 2014), natural language processing (Collobert and Weston, 2008), and survival analysis (Wang et al., 2017a; Li et al., 2016).

Following Pan and Yang (2010), we provide a classification of multi-task learning problem settings in cancer survival analysis. Let us first define the terms *domain* and *task*. A domain is a pair $\{\mathcal{X}, P(X)\}$ which includes a feature space and a marginal probability distribution where $X = \{x_1, \dots, x_n\} \in \mathcal{X}$. A task $\{\mathcal{Y}, P(Y|X)\}$ consists of a label space and a conditional probability distribution function. $P(Y|X)$ is the ultimate predictive function that is not observed but can be learned from training data. Multi-task learning, by definition, involves different tasks, i.e. different $P(Y|X)$, or even different label spaces. With that in mind, we classify multi-task survival analysis problems as follows:

1. Different $P(X)$: Data for the tasks come from different distributions. Examples include:
 - Standard gene expression data and progression-free survival labels are available for all cohorts, but the cohorts are diagnosed with different cancer types.
 - Standard gene expression data and progression-free survival labels are available for all cohorts, and the cohorts are diagnosed with the same cancer types but belong to different studies/hospitals.
- 2 Different \mathcal{X} : Data for the tasks come from different feature spaces. Note that this automatically leads to different $P(X)$. An example of different feature spaces is gene expression data and mutation data.
3. Different $P(Y|X)$: All tasks are the same in nature, but the conditional distribution of labels are different. For example, learning overall survival and progression-free survival simultaneously for the same cohort of patients falls under this category.

4 Different \mathcal{Y} : This class of multi-task problems involves different prediction tasks (such as survival analysis and classification).

The general form of the loss function when learning T tasks simultaneously is:

$$L(Y, X, W) = \sum_{t=1}^T L_t(y^t, g^t(W^t, X^t)) + \gamma\lambda(Y, X, W) \quad (2.2)$$

l_t and W^t , respectively, are the loss function and the parameters of task t . $Y = \{Y^1, \dots, Y^T\}$ and $X = \{X^1, \dots, X^T\}$ are the combined input data of all t tasks. g^t indicates the prediction function corresponding to task t , and λ is a regularization or auxiliary function that captures task relatedness assumptions, examples of which include cluster norm (Jacob et al., 2009), and $\ell_{2,1}$ norm (Argyriou et al., 2007). γ is a weight parameter controlling the importance of the auxiliary function.

One of the main ideas in the literature is enforcing sparsity across tasks through regularization. Argyriou et al. (2007) assume only a small subset of features are shared across tasks and enforces this assumption via $\ell_{q,1}$ norm. The ℓ_q norm of each feature across tasks is computed first, and the ℓ_1 norm of the result is then minimized, leading to all but a few features to have ℓ_q norms close to zero.

The above approach assumes all tasks are related. In order to exploit task relationships in cases where this assumption does not hold, Evgeniou et al. (2005) propose to enforce a clustering constraint among tasks, by penalizing the variance of task parameters from the mean task parameter in each cluster.

Previous work involving multi-task learning in deep neural networks can be categorized into two main categories: soft and hard parameter sharing (Ruder, 2017). It has been shown that hard parameter sharing reduces the risk of overfitting by an order of T , T being the number of tasks, compared to overfitting in task-specific parameters (Baxter, 1997). In soft parameter sharing, each model has its own parameters while the distance between model parameters is regularized by different methods, mostly

inspired by the same regularization methods used in traditional multi-task learning.

In chapter 4, we train a multitask neural network with hard parameter sharing to predict survival in cancer, specifically addressing scenarios 1 and 3 listed above.

Chapter 3

Learning Genomic Representations to Predict Clinical Outcomes in Cancer

This chapter is a collection of articles on adapting artificial neural networks to survival analysis from high dimensional genomic data. All articles included in this chapter are open access and available online.

Sections 3.1 and 3.2 present SurvivalNet, a software framework that enables predicting risk of event using a fully-connected artificial neural network that is trained by maximizing Cox's proportional hazards model likelihood. A method for interpreting SurvivalNet based on partial derivatives of risk with respect to input features is introduced that enables the extraction of biological insights from the trained neural networks. A rigorous comparison of SurvivalNet with state-of-the-art survival analysis models is performed. Finally, preliminary experiments with combinations of cohorts with different cancer types lead to mixed but promising results and inspire a line of research that we pursue in Chapter 4.

Section 3.3 presents an article that describes the application of Cox's proportional hazards model likelihood to a different type of artificial neural network known as convolutional networks. The proposed framework enables the integration of histology and genomic data for prediction of outcomes in cancer, and is shown to outperform the WHO standard used in the classification of gliomas.

Finally, in section 3.4, SurvivalNet and its interpretation mechanism are applied to investigating markers of progression in oligodendrogliomas, specifically the role of the Notch signaling pathway.

3.1 Learning Genomic Representations to Predict Clinical Outcomes in Cancer, ICLR-workshop, 2016

This section is an exact copy of the following open-access conference paper:

Safoora Yousefi, Congzheng Song, Nelson Nauata, and Lee Cooper.
Learning genomic representations to predict clinical outcomes in cancer.
In International Conference on Learning Representations (ICLR), 2016.

Abstract. Genomics is rapidly transforming medical practice and basic biomedical research, providing insights into disease mechanisms and improving therapeutic strategies, particularly in cancer. The ability to predict the future course of a patient's disease from high-dimensional genomic profiling will be essential in realizing the promise of genomic medicine, but presents significant challenges for state-of-the-art survival analysis methods. In this abstract, we present an investigation in learning genomic representations with neural networks to predict patient survival in cancer. We demonstrate the advantages of this approach over existing survival analysis methods using brain tumor data.

LEARNING GENOMIC REPRESENTATIONS TO PREDICT CLINICAL OUTCOMES IN CANCER

Safoora Yousefi[†] Congzheng Song[†] Nelson Nauata[‡] Lee Cooper^{‡,*}

[†]Department of Mathematics & Computer Science Emory University

[‡]Department of Biomedical Informatics Emory University School of Medicine

*Department of Biomedical Engineering Georgia Institute of Technology

{safoora.yousefi, congzheng.song, nnauata, lee.cooper}@emory.edu

ABSTRACT

Genomics are rapidly transforming medical practice and basic biomedical research, providing insights into disease mechanisms and improving therapeutic strategies, particularly in cancer. The ability to predict the future course of a patient’s disease from high-dimensional genomic profiling will be essential in realizing the promise of genomic medicine, but presents significant challenges for state-of-the-art survival analysis methods. In this abstract we present an investigation in learning genomic representations with neural networks to predict patient survival in cancer. We demonstrate the advantages of this approach over existing survival analysis methods using brain tumor data.

1 INTRODUCTION

Genomics provide a window into the complex molecular workings of disease. In the treatment of cancer, genomic analysis of a tissue biopsy can reveal specific molecular vulnerabilities that can be matched to targeted therapies, or to prognosticate the future behavior of a patient’s disease and expected survival in order to better inform clinical interventions including surgery and radiation therapy. Although genomic analysis generates rich high-dimensional signals that contain hundreds to hundreds-of-thousands of variables, typically only several variables are used for prognostication for any given cancer type. Typically, these variables are used to assign patients into discrete disease classes or “subtypes” that associate with response to specific therapies, or with varying degrees of disease aggressiveness. Learning the underlying latent prognostic variables from high-dimensional genomic profiles can extract additional prognostic value from unused variables, and is critical in realizing the promise of genomic medicine. This problem presents significant challenges, ranging from the familiar “large p small N ”, to how to adapt developments in the machine learning domain to the analysis of time-to-event survival data.

In this abstract we present an investigation in building survival prediction neural networks to learn representations from genomic data for survival prediction. We use backpropagation to train neural networks to maximize the Cox proportional hazards likelihood of time-to-event data, and apply these predictive models to molecular profiles of brain tumor patients from The Cancer Genome Atlas where survival ranges from 6 months to 10+ years. We compare our methods to state-of-the-art survival analysis algorithms based on elastic-net (linear combination of L1 and L2) regularization of Cox hazard models and random forest based methods, and demonstrate improvements in survival prediction accuracy for neural network approaches.

2 BACKGROUND AND RELATED WORK

2.1 HAZARD MODELS AND LIKELIHOOD FUNCTIONS

Survival analysis involves predicting the time to some event of interest, which in cancer is often death or progression of disease. It differs from ordinary regression due to *incomplete followup*, where a death or relapse event is not observed at or before the final encounter with the patient. These censored observations provide critical information, and often represent an important population of long-term survivors or treatment responders that are very important to incorporate into the model. The most commonly used regression approach to survival analysis is the Cox proportional hazards

model proposed by Cox (1972). At time t , the hazard for a sample with covariates x is given by the following hazard function:

$$\lambda(t|x) = \lambda_0(t)e^{\beta x}, \quad (1)$$

where $\lambda_0(t)$ is baseline hazard. The hazard function is an exponential linear function of the covariates x and model coefficients β , with the effect of any covariate $x^{(i)}$ assumed to be the same over time. Since the evaluation criteria of the models in this paper is based on ranking predicted survival times, the baseline hazard is left unspecified and we maximize the partial likelihood function during training:

$$l(\beta, X) = - \sum_{i \in U} \left(X_i \beta - \log \sum_{j \in R_i} e^{X_j \beta} \right) \quad (2)$$

where U is the set of all uncensored patients, and R_i is the set of patients whose time of death or last follow-up is later than time of death of i .

We measured model performance using *concordance index* (CI) that captures the rank correlation of predicted and actual survival. Denoting the i th patient with X_i and the set of all patients with X , where t_i represents either the time of death or the time of last follow-up of the i th patient, CI was calculated in the following way:

$$CI(\beta, X) = \sum_P \frac{I(i, j)}{|P|} \quad (3)$$

$$I(i, j) = \begin{cases} 1, & \text{if } Risk_j > Risk_i \text{ and } t_j > t_i \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Where P is the set of orderable pairs. A pair of samples (X_i, X_j) is orderable if either the event is observed for both X_i and X_j , or X_j is censored and $t_j > t_i$. Intuitively, CI measures the pairwise agreement of the prognostic scores $Risk_i, Risk_j$ predicted by the model and the actual time of death for all orderable pairs. Attempts have been made to propose differentiable versions of CI based on sigmoid and exponential functions and optimize it directly. But recent studies show that optimizing the cox partial likelihood is equivalent to optimizing CI (Steck et al., 2008).

2.2 RELATED WORKS

Regularization techniques have been proposed for feature selection in survival analysis of high dimensional data (Zou & Hastie, 2005). Efforts have been made to introduce successful machine learning algorithms such as random forests to survival analysis (Ishwaran et al., 2008). Deep learning techniques have been employed for cancer diagnosis using genomic data and medical images, such as Fakoor et al. (2013) and Esteva et al. To the best of our knowledge, representation learning techniques have not been applied to survival prediction from genomic data, and the previous work investigating neural networks for survival analysis dealt with low dimensional data and different cost functions (Lisboa et al., 2003).

3 SURVIVAL PREDICTION NEURAL NETWORK

3.1 PRETRAINING AND FINE-TUNING

In this work we trained an autoencoder to represent genomic data and fine tune this representation using partial log Cox likelihood. In training, we employ stacked denoising autoencoders proposed in Vincent et al. (2008). We train the auto-encoders using 183-dimensional genomic features, then we add a risk prediction output layer as shown in Figure 1-a.

We use survival times and censoring status to calculate the Cox partial log likelihood given by equation 2 and differentiate it with respect to X :

$$\frac{\partial l(\beta, X)}{\partial X_i} = c_i \beta - \sum_{j \in U: i \in R_j} \frac{\beta e^{X_i \beta}}{\sum_{k \in R_j} e^{X_k \beta}} \quad (5)$$

where c_i is 1 if sample i is not censored, and is 0 otherwise, and β denotes the parameters of the risk prediction layer. This derivative is then back-propagated through the network to fine tune the learned representation specifically for the task of survival analysis.

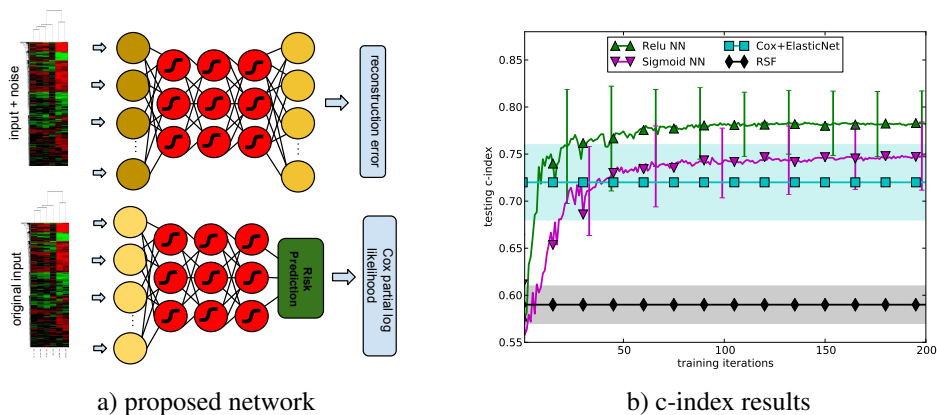


Figure 1: a) Survival network model. b) Comparison of the proposed survival prediction neural network to two competing methods: Elastic-net (L1 and L2 combined) regularized Cox regression and Random Survival Forest (RSF). Average testing CI trends are shown for neural networks with two different activation functions: sigmoid and rectified linear units. The error bars and shaded areas indicate standard deviation of CI over 10 cross validation sets (See section 3.3 for details.)

3.2 MODEL SELECTION

The training of a neural network involves many hyperparameters: type of nonlinearities used, number of layers, number of hidden units in each layer, learning rates for pretraining and fine tuning and regularization parameters. Since this is the first work where deep neural networks are used to address survival analysis, we could not look at existing literature for a conventional choice of hyperparameters. Unlike areas such as image classification, no rule of thumb has been developed for setting hyper-parameters in survival analysis. Therefore we employed bayesian optimization (Martinez-Cantin, 2014) with Gaussian prior to decrease the number of objective function evaluations needed to reach a decent choice of hyperparameters. More shallow networks demonstrate superior performance over deeper architectures in our experiments. This could be justified considering the small number of training samples (628) and the scarcity of labels within the available samples. Our average choice of configuration is 2 fully connected layers of 250 hidden units each. On average, we use a learning rate .001 for pre-training and .0009 for fine-tuning.

3.3 EVALUATION

Due to the small size of available training data, performance of the model might considerably depend on the partitioning of the data into testing and training. To mitigate this, we randomly sampled from the data set 10 times without replacement to have 10 permutations of the same data set. Then in each of the 10 sets, we used the first %70 of the data for training, half of the remaining %30 of data for model selection and the other half for model assessment. We performed training, model selection and testing on these 10 permutation datasets separately. The reported CI in Figure 1-b is averaged over these 10 experiments to represent the generalization error of the model. The exact same setting was used for hyper-parameter tuning and assessment for competing methods. We picked the learning rate and elastic-net mixture coefficient for regularized Cox regression (Hastie & Qian (2014)) based on performance on the same validation sets we used for the neural networks. We tuned number of trees, leaf size, and number of split points for random survival forest in the same fashion. Our experiments reveal that Random Survival Forests do not adapt well to high dimensionality and are markedly outperformed by survival neural networks (See Figure 1-b). Neural networks also achieve %5 absolute improvement over regularized Cox regression with ReLU activation and %3 with sigmoid activation.

ACKNOWLEDGMENTS

This work was supported by US Public Health Service National Institutes of Health (NIH) grants K22LM011576-03 and U24CA194362-01.

REFERENCES

- David R Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220, 1972.
- Andre Esteva, Brett Kopley, and Sebastian Thrun. Deep networks for early stage skin disease and skin cancer classification.
- Rasool Fakoor, Faisal Ladhak, Azade Nazi, and Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. In *The 30th International Conference on Machine Learning (ICML 2013), WHEALTH workshop*, 2013.
- Trevor Hastie and Junyang Qian. Glmnet vignette. Technical report, Technical report, Stanford, 2014.
- Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pp. 841–860, 2008.
- Paulo JG Lisboa, H Wong, P Harris, and Ric Swindell. A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial intelligence in medicine*, 28(1):1–25, 2003.
- Ruben Martinez-Cantin. Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits. *The Journal of Machine Learning Research*, 15(1):3735–3739, 2014.
- Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pp. 1209–1216, 2008.
- Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103. ACM, 2008.
- Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

3.2 Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models, Nature Scientific Reports, 2017

This section is an exact copy of the following open-access journal paper:

Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E Lewis, Congzheng Song, David A Gutman, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, and Lee AD Cooper. *Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models*. Scientific Reports, 7(1):11707, 2017.


Abstract. Translating the vast data generated by genomic platforms into accurate predictions of clinical outcomes is a fundamental challenge in genomic medicine. Many prediction methods face limitations in learning from the high-dimensional profiles generated by these platforms, and rely on experts to hand-select a small number of features for training prediction models. In this paper, we demonstrate how deep learning and Bayesian optimization methods that have been remarkably successful in general high-dimensional prediction tasks can be adapted to the problem of predicting cancer outcomes. We perform an extensive comparison of Bayesian optimized deep survival models and other state-of-the-art machine learning methods for survival analysis, and describe a framework for interpreting deep survival models using a risk backpropagation technique. Finally, we illustrate that deep survival models can successfully transfer information across diseases to improve prognostic accuracy. We provide an open-source software implementation of this framework called Survival-Net that enables automatic training, evaluation, and interpretation of deep survival models.

SCIENTIFIC REPORTS



OPEN

Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models

Safoora Yousefi¹, Fatemeh Amrollahi¹, Mohamed Amgad¹, Chengliang Dong², Joshua E. Lewis³, Congzheng Song⁴, David A. Gutman⁵, Sameer H. Halani⁶, Jose Enrique Velazquez Vega⁷, Daniel J. Brat^{7,8} & Lee A. D. Cooper^{1,3,8} 

Received: 23 May 2017

Accepted: 30 August 2017

Published online: 15 September 2017

Translating the vast data generated by genomic platforms into accurate predictions of clinical outcomes is a fundamental challenge in genomic medicine. Many prediction methods face limitations in learning from the high-dimensional profiles generated by these platforms, and rely on experts to hand-select a small number of features for training prediction models. In this paper, we demonstrate how deep learning and Bayesian optimization methods that have been remarkably successful in general high-dimensional prediction tasks can be adapted to the problem of predicting cancer outcomes. We perform an extensive comparison of Bayesian optimized deep survival models and other state of the art machine learning methods for survival analysis, and describe a framework for interpreting deep survival models using a risk backpropagation technique. Finally, we illustrate that deep survival models can successfully transfer information across diseases to improve prognostic accuracy. We provide an open-source software implementation of this framework called *SurvivalNet* that enables automatic training, evaluation and interpretation of deep survival models.

Advanced molecular platforms can generate rich descriptions of the genetic, transcriptional, epigenetic and proteomic profiles of cancer specimens, and data from these platforms are increasingly utilized to guide clinical decision-making. Although contemporary platforms like sequencing can provide thousands to millions of features describing the molecular states of neoplastic cells, only a small number of these features have established clinical significance and are used in prognostication^{1–4}. Making reliable and accurate predictions of clinical outcomes from high-dimensional molecular data remains a major challenge in realizing the potential of precision genomic medicine.

Traditional Cox proportional hazards models require enormous cohorts for training models on high-dimensional datasets containing large numbers of features. Consequently, a small set of features is selected in a subjective process that is prone to bias and limited by imperfect understanding of disease biology. High-dimensional learning problems are common in the machine-learning community, and many machine-learning approaches have been adapted to predicting survival or time to progression⁵. Prior knowledge has been used to reduce dimensionality by learning gene signatures of cancer hallmarks to generate intermediate features that successfully predict outcomes^{6,7}. Regularization methods for Cox models like elastic net have been developed to perform objective and data-driven feature selection with time-to-event data⁸. Random forests are reputed to resist overfitting in high-dimensional prediction problems, and have been adapted to survival modeling⁹. Neural network based approaches have been used in low-dimensional survival prediction problems¹⁰, but subsequent evaluation of these methods found no performance improvement over ordinary Cox regression¹¹. The difficulty of deconstructing these black-box models to gain insights into disease progression or biology remains

¹Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, 30322, USA. ²Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, 10032, USA. ³Department of Biomedical Engineering, Georgia Institute of Technology/Emory University School of Medicine, Atlanta, GA, 30322, USA. ⁴Department of Computer Science, Cornell University, Ithaca, NY, 14850, USA. ⁵Department of Neurology, Emory University School of Medicine, Atlanta, GA, 30322, USA. ⁶Emory University School of Medicine, Atlanta, GA, 30322, USA. ⁷Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, 30322, USA. ⁸Winship Cancer Institute, Emory University, Atlanta, GA, 30322, USA. Correspondence and requests for materials should be addressed to L.A.C. (email: lee.cooper@emory.edu)

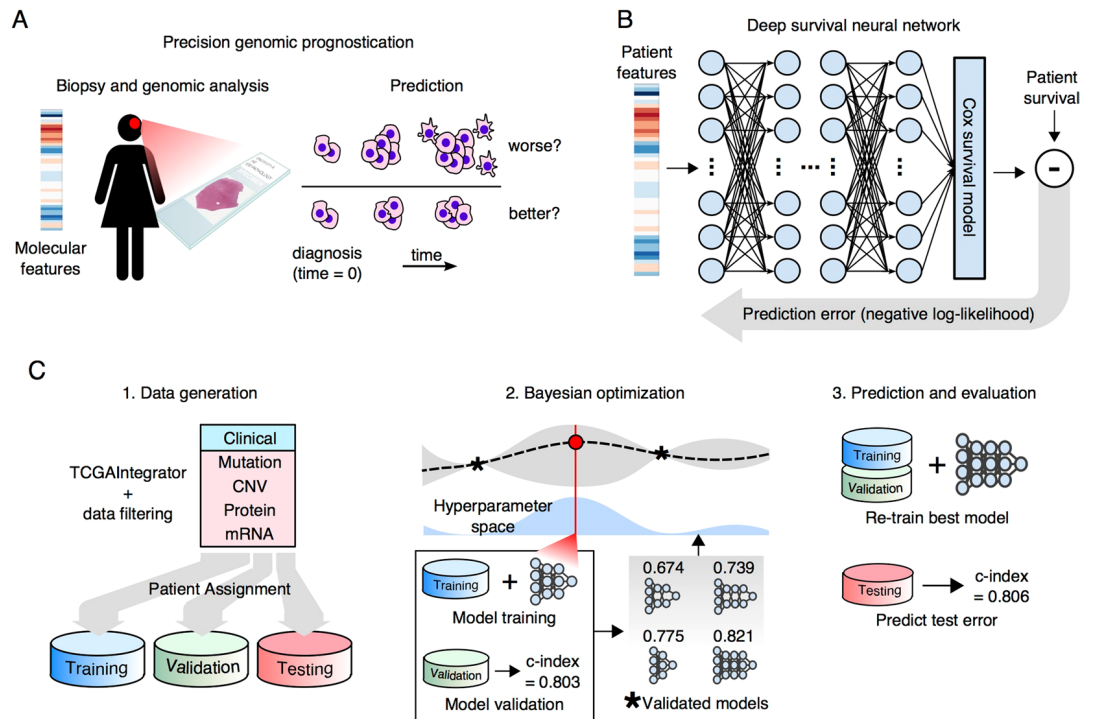


Figure 1. Overview of the SurvivalNet framework. **(A)** Accurate prognostication is crucial to clinical decision making in cancer treatment. Molecular platforms produce data that can be used for precision prognostication with learning algorithms. **(B)** Deep survival models are neural networks composed of layers of non-linear transformations, driven by a Cox survival model at the output layer. Model likelihood is used to adaptively train the network to improve the statistical likelihood of the overall survival prediction. **(C)** The SurvivalNet framework enables automatic design optimization and validation of deep survival models. Molecular profiles obtained from TCGA datasets are randomized, assigning patients to training, testing and validation sets. Bayesian optimization searches the space of hyperparameters like the number of network layers to optimize the model design. Each selected design is trained and evaluated using validation samples to update the Bayesian optimizer. The best model design is then evaluated on the independent testing set to measure the final optimized model accuracy.

a key challenge in their adoption. Deep neural networks were combined with input-level feature selection to identify promoters and enhancers of gene regulation, with the goal of creating interpretable nonlinear models¹².

Advances in neural networks broadly described as *deep learning* have shattered performance benchmarks in general machine-learning tasks, enabled by improvements in methodology, computing hardware, and datasets¹³. These networks are composed of densely interconnected layers that sequentially transform the inputs into more predictive features through adaptive learning of the interconnection parameters (see Fig. 1). Deep networks composed of many layers perform *feature-learning* on high dimensional datasets to extract latent explanatory features¹⁴, and have been successfully applied to biomedical problems including image classification¹⁵, transcription factor binding site prediction¹⁶, and medication dosing control¹⁷. A fundamental challenge in deep learning is determining the network design that provides the best prediction accuracy, a process that involves choosing network *hyperparameters* including the number of layers, transformation types, and training parameters. Searching the vast space of network designs quickly becomes intractable, given the considerable time required to train a single deep network. *Bayesian optimization* techniques have been developed to automate the search of the hyperparameter space, and provide measurable gains in accuracy over expert tuning¹⁸ or random search¹⁹, and identify optimal models with fewer experiments^{19,20}. Advanced deep learning techniques including dropout regularization, unsupervised pre-training, and Bayesian optimization were first applied to build unbiased deep models from high-dimensional genomic data in ref.²¹ where deep networks were trained to optimize proportional hazards likelihood. A subsequent study applied deep networks to model survival in breast cancer using a low-dimensional dataset (14 features) that were selected with a priori disease knowledge²². This study did not evaluate prediction using high-dimensional data or compare to state-of-the-art methods like regularized Cox regression that perform unbiased feature selection.

This paper extends the preliminary studies exploring deep learning for survival modeling, and presents a software package called *SurvivalNet* (SN) that enables users to train and interpret deep survival models. SurvivalNet uses Bayesian optimization to identify optimal hyperparameter settings, saving users considerable time and effort in choosing model parameters. We also illustrate how backpropagation methods can be modified to interpret deep survival models, scoring individual features for their contribution to risk, and show how feature risk scores can be used with pathway analysis tools to uncover higher-order biological themes associated with patient survival. Using clinical and molecular data from The Cancer Genome Atlas (TCGA), we show that Bayesian-optimized

deep survival models provide comparable performance to Cox elastic net regression, and superior performance to random survival forests when analyzing high-dimensional genomic data. Finally, we show how deep survival models can learn prognostic information from multi-cancer datasets to improve prognostication through transfer learning.

Results

Automatic training and validation of deep survival models. An overview of the SurvivalNet framework is presented in Fig. 1. SurvivalNet is implemented as an open-source Python module (<https://github.com/CancerDataScience/SurvivalNet>) using Theano and is available as a pre-built Docker software container. A deep survival model uses the Cox partial log likelihood to train the weights of neural network to transform molecular features into explanatory factors that explain survival. The partial log likelihood serves as a feedback signal to train the model weights using backpropagation. Deep neural networks have many hyperparameters that impact prediction accuracy including the number of layers, number and type of activation functions in each layer, and choices for optimization/regularization procedures. The time needed to train a deep survival model prohibits exhaustive hyperparameter search, and so SurvivalNet employs a Bayesian optimization strategy to identify hyperparameters that optimize prediction accuracy including the number of network layers, the number of elements in each layer, the activation function, and the dropout fraction. Bayesian optimization enables users who lack experience tuning neural networks to optimize model designs automatically, and results in considerable savings in time and effort as previously reported¹⁹. Data is first split into training (60%), validation (20%), and testing (20%) sets. Training samples are used to train the model weights with backpropagation using the network design suggested by Bayesian optimization. The prediction accuracy of the trained deep survival model is then estimated using the validation samples, and is used to maintain a probabilistic model of performance as a function of hyperparameters. Based on the probabilistic model, the design with the best expected accuracy is inferred as the next design to test. After the Bayesian optimization process is finished (typically after a prescribed number of experiments), the best network design is used to re-train a deep survival model using the training and validation samples, and the accuracy of this best model is reported using the held-out testing samples.

Comparing deep survival networks with Cox elastic net and random survival forests. We compared the performance of SurvivalNet models with Cox elastic net (CEN) and random survival forest (RSF) models using data from multiple TCGA projects: pan-glioma (LGG/GBM), breast (BRCA), and pan-kidney (KIPAN) which consists of chromophobe, clear cell, and papillary carcinomas. Datasets were selected based on the availability of molecular and clinical data and for extent of complete clinical follow up. Performance was evaluated with two feature-sets: 1) a “transcriptional” feature set containing 17,000 + gene expression features obtained by RNA-sequencing, and 2) an “integrated” feature set containing 3–400 features describing clinical features, mutations, gene and chromosome arm-level copy number variations, and protein expression features. Details of these datasets are presented in Methods and Tables S1 and S2. Optimization procedures for CEN and RSF hyperparameters are described in Methods.

In each experiment, samples were randomized to training (60%), validation (20%), and testing (20%) sets, and the performance of optimized SN, CEN, and RSF models was assessed. Performance was calculated using Harrell’s *c*-index, a non-parametric statistic that measures concordance between predicted risks and actual survival²³. A *c*-index of 1 indicates perfect concordance, and a *c*-index of 0.5 corresponds to random chance. Experiments were repeated for 20 randomizations to account for variations due to sample assignment. Differences in performance between methods were evaluated through rank-sum statistical testing of *c*-index values. Results are presented in Fig. 2 (extended results are presented in Table S3).

Both SN and CEN significantly outperform RSF models in most experiments. All methods perform markedly better than random, with median *c*-index scores ranging from: 0.75–0.84 in LGG/GBM; 0.52–0.68 in BRCA; and 0.73–0.79 in KIPAN. In the transcriptional feature set (Fig. 2B), SN models have a slight advantage over CEN models in LGG/GBM (Wilcoxon rank-sum $p = 2.39e-2$) and KIPAN ($p = 0.0565$). In the integrated feature set (Fig. 2A), SN and CEN performance were indistinguishable in the BRCA dataset ($p = 0.770$), but CEN models have a slight advantage over SN models in the LGG/GBM ($p = 1.78e-3$) and KIPAN ($p = 0.0699$) datasets. Performance is generally better on the integrated feature set than the transcriptional feature set for all methods. One exception to this is the performance of SN on the LGG/GBM feature sets, where performance on the transcriptional feature set exceeds the integrated feature set (*c*-index 0.841 versus 0.818). RSF models have the worst performance generally, and are severely challenged in learning from the BRCA transcriptional feature set, with a median *c*-index of 0.520 (slightly better than random guess). Comparing performance across diseases, we noticed that prediction accuracy generally decreases as the proportion of right-censored samples in a dataset increases. This pattern holds for all prediction methods. Glioma had the highest overall prediction accuracy, being a uniformly fatal disease that has relatively fewer long-term survivors and incomplete follow-up (62–64%). Breast carcinoma had the lowest overall prediction accuracy with more than 86–91% of BRCA samples being right-censored.

Finally, we observed that CEN model execution routinely fails with some randomizations, producing a segmentation fault software error. In these instances, we generated a new randomization for CEN and repeated the experiments. The performance accuracy of SN and RSF models on these failed randomizations does not suggest that they present particularly difficult learning problems, but we cannot exclude the possibility of introducing a performance bias for CEN by generating new randomizations when CEN execution fails.

Interpreting deep survival models with risk backpropagation. Linear survival models weight individual features based on their contribution to overall risk, providing a clear interpretation of the prognostic significance of individual features, and insights into the biology of disease progression. The complex transformations

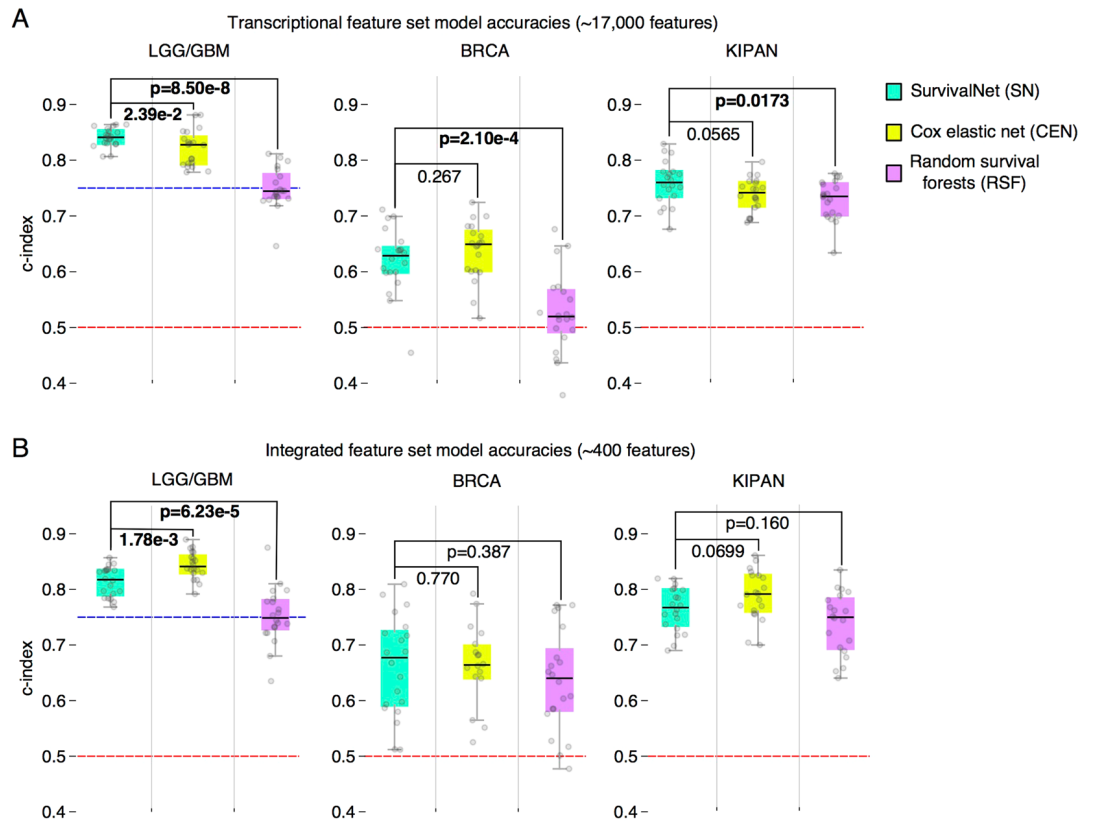


Figure 2. Performance comparison of SurvivalNet, Cox elastic net, and random survival forest models. The prognostic accuracy of these methods was evaluated in different diseases/datasets (GBMLGG, BRCA, KIPAN) using a high-dimensional transcriptional feature set and a lower-dimensional integrated feature set that combines clinical, genetic, and protein expression features. Patients were randomized to 20 training/validation/testing sets that were used to train, optimize, and evaluate models in each case. **(A)** SurvivalNet models have an advantage over Cox elastic net in predicting survival using high-dimensional transcriptional features. **(B)** Cox elastic net has an advantage in predicting survival using lower-dimensional integrated features. Dashed red lines correspond to a random prediction ($c\text{-index} = 0.5$). Dashed blue lines correspond to $c\text{-index}$ of molecular classification of gliomas.

that machine-learning methods apply to input features makes interpreting these models more difficult. This is especially true for deep learning where the input features are subjected to multiple sequential nonlinear transformations. To enable interpretation of deep survival models, we implemented a technique that we describe as *risk backpropagation*. In the same way that backpropagation can propagate prediction errors back through the layers of a deep model for training, backpropagation can also propagate predicted risks back to the input layer to assess how individual features contribute to risk (see Fig. 3). Partial derivatives were first used to analyze variable importance in ref.²⁴

A linear survival model is defined by a static set of weights that represent the importance of features in predicting patient risk. In the linear model the predicted risk can be conceptualized as a plane that has a uniform gradient for any input feature values. The slope of this plane is defined by the model weights and represents the rate of change of risk with respect to each feature. Partial derivatives in SurvivalNet are directly analogous to model weights in a linear model, yet the weights differ depending on the values of the features. In the nonlinear SurvivalNet, the prediction can be conceptualized instead as a nonlinear surface where the risk gradients change depending on a patient's feature values, and so these feature weights are calculated separately for each patient.

We applied risk backpropagation to our LGG/GBM integrated feature set model to investigate the prognostic significance of features (see Fig. 4). Risk backpropagation was applied to each patient to generate feature risk scores, and then each feature was ranked using its median score across patients as a measure of overall prognostic significance (see Fig. 4A). Among the top-ranked features indicative of poor prognosis are: increased age at diagnosis (rank 3); histologic classification as de novo grade IV glioblastoma (rank 5); loss of chromosome arms 10p and 10q (ranks 2, 4); and deletions of tumor suppressor genes *CDKN2A* and *PTEN* (ranks 1, 8). The top-ranked features associated with better prognosis included mutations in *SMARCA4* (rank 6), *IDH1/IDH2* (ranks 9, 10) and in *CIC* (rank 17). We note that many of these features are either incorporated or highly correlated with the recently published World Health Organization genomic classification of gliomas²⁵.

To investigate molecular pathways related to glioma prognosis, we also performed a risk backpropagation gene-set enrichment analysis of our LGG/GBM transcriptional model. Median risk scores from the

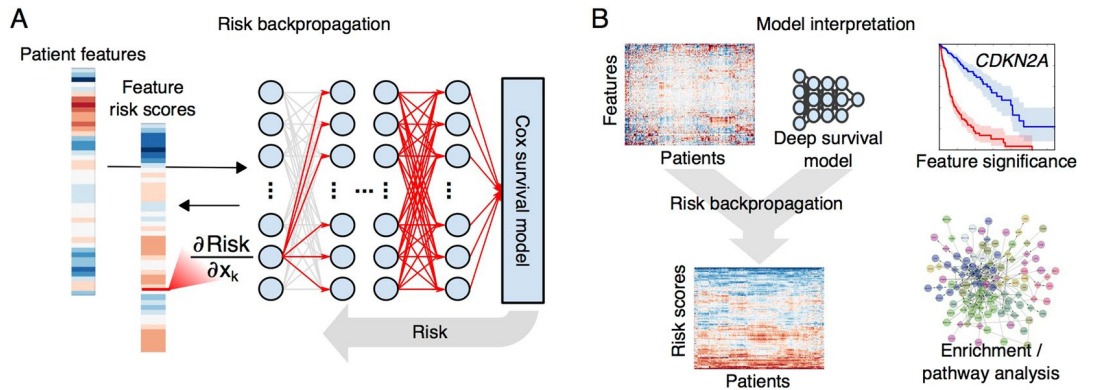


Figure 3. Interpreting deep survival models with risk backpropagation. **(A)** Backpropagation was used to calculate the sensitivity of predicted risk to each input feature, generating feature risk scores for each feature and patient. **(B)** Feature risk scores can be analyzed to gain insights into the deep survival model. Risk scores can be used to evaluate the prognostic significance of individual features, or to identify gene sets or molecular pathways that are enriched with high-risk or low-risk features.

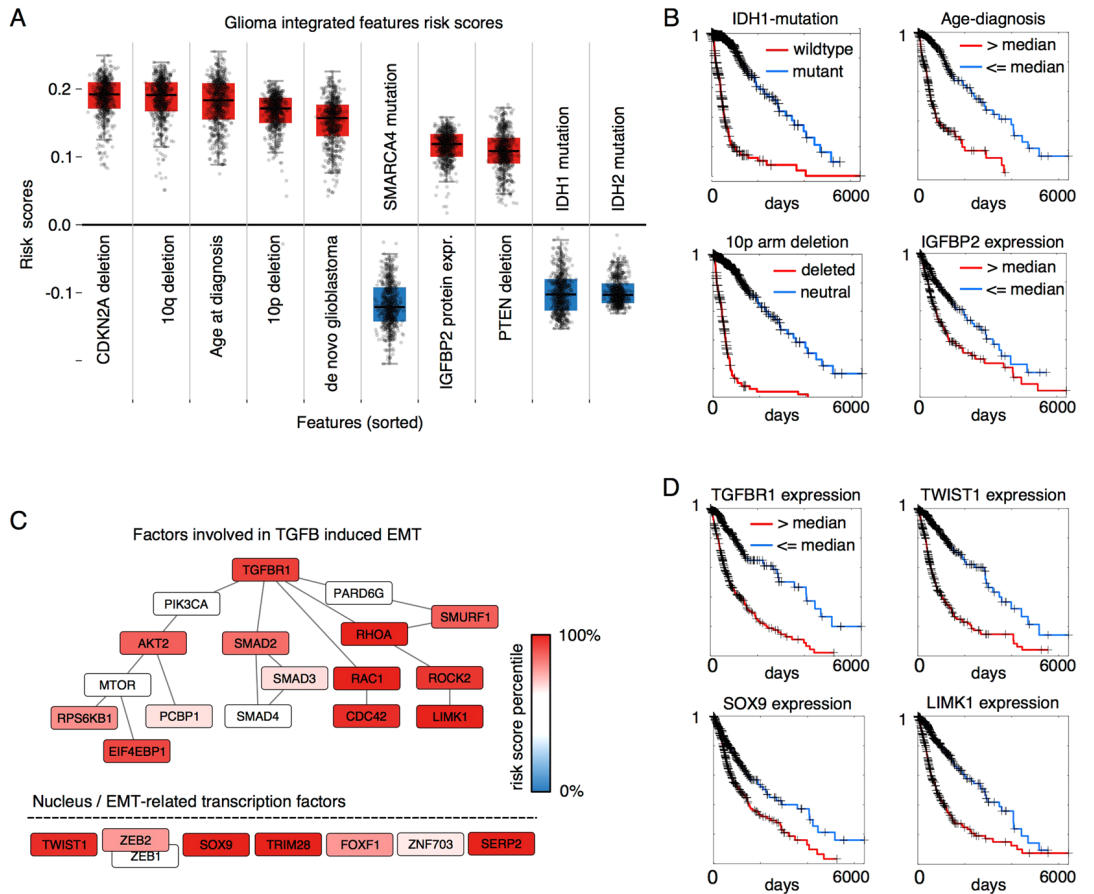


Figure 4. Interpretation of glioma deep survival models. **(A)** SurvivalNet learns features that are definitional (IDH mutation) or strongly associated (CDKN2A deletion, SMARCA4 mutation) with WHO genomic classification of diffuse gliomas. Feature risk scores for the top 10 of 399 features in the integrated model are shown here, in order. Each boxplot represents the risk scores for one feature across all patients. Features were ranked by median absolute risk score. **(B)** Kaplan-Meier plots for select features from **(A)**. **(C)** A gene set enrichment analysis of transcriptional feature risk scores identified the TGF-Beta 1 signaling and epithelial-mesenchymal transition (EMT) gene sets as enriched with features associated with poor prognosis. **(D)** Kaplan-Meier plots for select features from **(C)**.

transcriptional model were calculated for each transcript, and gene set enrichment analysis was performed on these scores to identify pathways enriched with prognosis-associated transcripts²⁶ (See Table S4). Pathways and gene sets associated with poor-prognosis include cell cycle (G2M checkpoint, E2F targets), apoptosis, angiogenesis, inflammation (Interferon alpha, gamma responses) and epithelial to mesenchymal transition (EMT and TGF-Beta signaling). EMT has received significant attention in cancer²⁷, and also specifically in glioma^{28–30} as being associated with aggressive phenotypes and poor clinical outcomes. The TGF-Beta signaling hallmark gene set was significantly enriched ($p = 6e-3$, FDR $q = 2.7e-2$) with genes having high risk scores including *RHOA*, *TGFB1*, *TGFBRI*, *SERPINE1*, *JUNB1* and *ARID4B*. The Epithelial to Mesenchymal Transition gene set was also significantly enriched ($p = 1.6e-1$, FDR $q = 1.55e-1$) with genes having high risk scores including *MMP12/3*, *IL6*, *ECM1*, and *VCAM1*. TGF-Beta signaling is understood to be one of the main pathways involved in EMT, and our results support the importance of EMT in determining glioma patient outcomes. The feature risk scores of the EMT-related transcription factors TGF-Beta induced EMT signaling as described in ref.²⁷ are visualized in Fig. 4C. Major TGF-Beta-EMT inducing factors (*RHOA*, *RAC1*, *ROCK2*, *CDC43* and *LIMK1*) and EMT transcription factors (*TWIST1*, *SOX9*, *TRIM28* and *SERP2*) have among the highest risk scores in our glioma transcriptional model.

Extended feature risk scores for the LGG/GBM integrated and transcriptional models are presented in Table S4. The procedure for obtaining models used for interpretation is described in Methods.

Transfer learning with multi-cancer datasets. We performed a series of *transfer learning* experiments to evaluate the ability of deep survival models to benefit from training with data from multiple cancer types. The transfer learning paradigm is illustrated in Fig. 5A. Survival models were trained using three different datasets: BRCA-only, BRCA + OV (ovarian serous carcinoma), and BRCA + OV + UCEC (corpus endometrial carcinoma), and were evaluated for their accuracy in predicting BRCA outcomes. The large proportion of right-censored cases in the BRCA dataset (90%) makes training accurate models difficult, and so we hypothesized that augmenting BRCA training data with samples from other hormone-driven cancers could improve BRCA prognostication. BRCA samples were randomized to training, validation, and testing and full Bayesian optimization was performed to measure c-index on BRCA testing samples for 20 randomizations. For the integrated feature set, we combined datasets by discarding disease-specific clinical features.

Adding samples from the OV and UCEC datasets provides measurable improvements in BRCA prognostic accuracy for both integrated and transcriptional feature set deep survival models (see Fig. 5B). For integrated models, training with BRCA + OV samples increases median c-index from 0.588 to 0.643 (rank-sum $p = 2.92e-3$), and training with BRCA + OV + UCEC improves this further to 0.710 ($p = 3.10e-5$). For the transcriptional feature set, training with BRCA + OV does not produce a measurable improvement over BRCA-alone ($p = 0.978$), but training with BRCA + OV + UCEC provides a marginal 3.5% improvement ($p = 0.168$).

We also evaluated the ability of Cox elastic net to benefit from transfer learning, and found significant performance degradation with transfer learning in transcriptional feature set (see Figure 5C). Training with BRCA + OV reduces the median c-index to from 0.664 to 0.599 ($p = 0.0699$), and training with BRCA + OV + UCEC reduces this further to 0.59335 ($p = 0.0165$). Performance improvements with the integrated feature set for CEN were similar to those observed with deep survival models.

Risk backpropagation analysis of transfer learning. To understand the information that OV and BRCA samples provide in predicting BRCA prognosis, we performed analysis of the BRCA and BRCA + OV + UCEC deep survival models using risk backpropagation. Risk backpropagation analysis was applied independently to the BRCA and BRCA + OV + UCEC transcriptional models to generate features risk scores, and gene set enrichment analyses were performed on these risk scores for each model to identify differences in pathway enrichment between the two models. Gene set enrichment scores for the BRCA + OV + UCEC model show increased emphasis on inflammatory pathways (particularly IL2-STAT5 signaling, IL6-JAK-STAT3 signaling and Interferon gamma response) as well as the apical junction gene set (known for its relevance to cell adhesion and metastasis). KRAS signaling and MYC targets v1 gene sets were de-emphasized in the BRCA + OV + UCEC model, pointing to a less prominent role of these pathways in determining breast cancer disease progression (See Fig. 5D and Tables S5 and S6).

Discussion

We created a software framework for Bayesian optimization and interpretation of deep survival models, and evaluated the ability of optimized models to learn from high-dimensional and multi-cancer datasets. Our software enables investigators to efficiently construct deep survival models for their own applications without the need for expensive manual tuning of design hyperparameters, a process that is time consuming and that requires considerable technical expertise. We also provide methods for model interpretation, using the backpropagation of risk to assess the prognostic significance of features and to gain insights into disease biology. Our analysis shows the ability of deep learning to extract important prognostic features from high-dimensional genomic data, and to effectively leverage multi-cancer datasets to improve prognostication. It also reveals limitations in deep learning for survival analysis and the value of complex and deeply layered survival models that need to be further investigated.

SN models have slightly better prognostic accuracy on two of three learning tasks using 17,000 + transcriptional features (GBMLGG and KIPAN), where CEN performed better using the lower-dimensional 300–400 integrated features. The high dimensionality of the transcriptional feature set presents a more challenging prediction problem where algorithms are more likely to overfit training data noise. CEN models are regularized linear models that use data-driven feature selection to identify a core subset of informative features for linear prediction. Their linearity does not appear to limit performance in our experiments, as their accuracy is similar to deep learning models and surpasses RSF models. While the deep models can effectively learn survival from

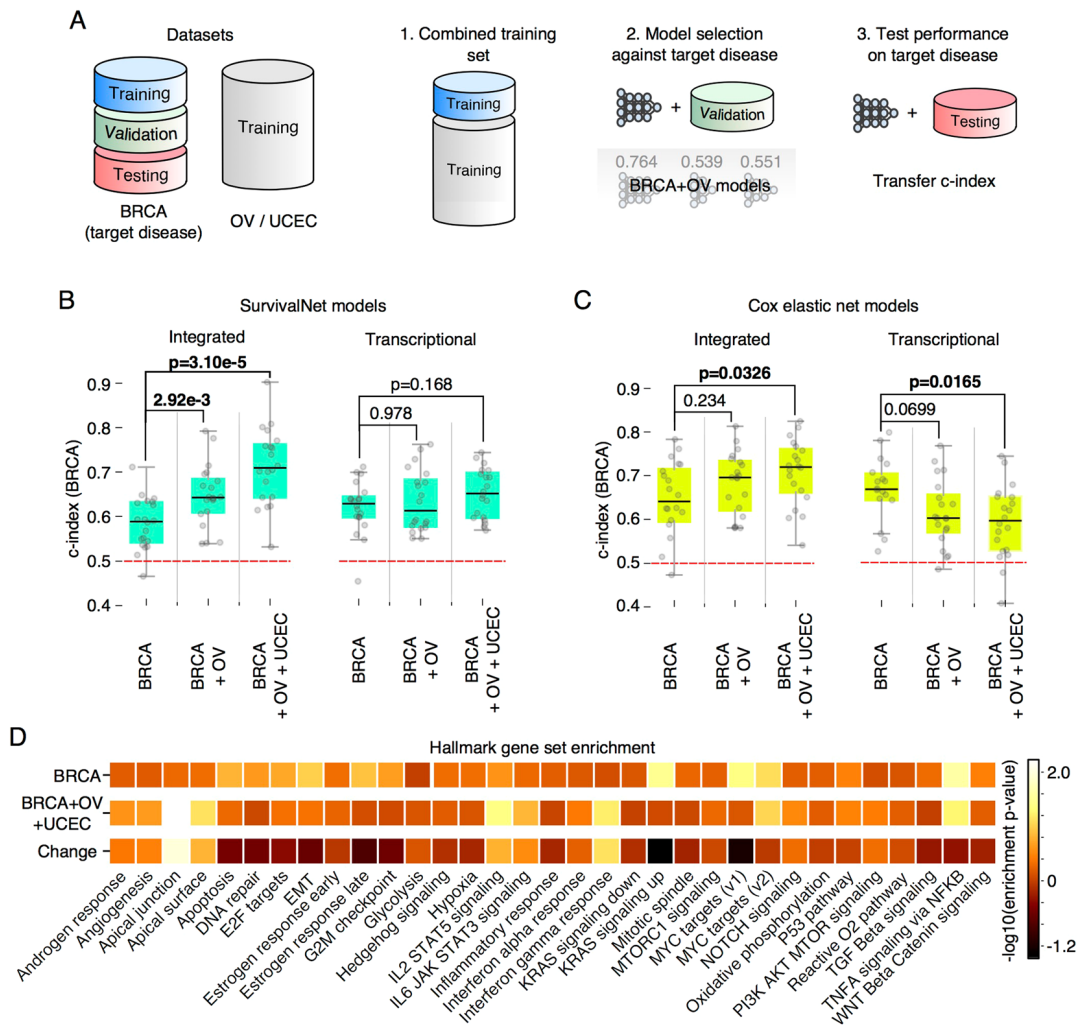


Figure 5. Learning with data from multiple cancer types improves deep survival models. **(A)** Data from the BRCA dataset was partitioned into training, validation, and testing sets. The BRCA training set was augmented with samples from the OV and UCEC and used to construct models for BRCA survival prediction. **(B)** Augmented training sets significantly improve the performance of SurvivalNet models for the integrated feature set. For the transcriptional feature set, marginal improvement was observed when training with BRCA + OV + UCEC data, but training with BRCA + OV data provides no improvement. **(C)** For Cox elastic net, augmentation significantly degrades performance for the high-dimensional transcriptional feature set. **(D)** Gene set enrichment analysis of feature risk scores from the BRCA and BRCA + OV + UCEC transcriptional models. The model trained with BRCA + OV + UCEC samples emphasizes different biological concepts than the BRCA-only model.

high-dimensional data, the feature-learning capabilities of layered nonlinear transformations did not translate into significant gains as has been demonstrated in general image classification or language processing tasks¹³. Larger datasets may be needed to overcome overfitting issues and to reveal anticipated performance benefits of deep learning. Deep learning methods typically require large amounts of training data to effectively learn their many parameters³¹, although empirical results in some applications have demonstrated otherwise³². In our experiments data requirements were exacerbated by the need to allocate validation samples for hyperparameter optimization. Smaller testing sets also introduced considerable variance in performance measurements.

Risk backpropagation analysis of gliomas demonstrated that SurvivalNet models could identify key features in high-dimensional datasets, recovering important genetic alterations that currently used to classify gliomas in clinical practice. Survival of patients diagnosed with infiltrating glioma depends largely on age, histologic grade and classification into three molecular subtypes defined by mutations in the Krebs cycle enzyme *isocitrate dehydrogenase* (*IDH1/IDH2*) and co-deletion of chromosome arms 1p and 19q¹: 1. Gliomas with wild-type *IDH* (astrocytoma) have an expected survival of 18 months, and are overwhelmingly diagnosed as advanced grade IV glioblastoma 2. Gliomas with co-deletion of 1p and 19q and mutations in *IDH* (oligodendroglioma) have the best outcomes, with some patients surviving 10 years or more and 3. Gliomas with *IDH* mutations that lack co-deletions (*IDH*-mutant astrocytoma) have intermediate outcomes. Risk backpropagation analysis of our model identified *IDH1* and *IDH2* mutations (ranks 9, 10) as strongly associated with better prognosis, consistent

with the role of these mutations as the primary feature in classifying gliomas. While our analysis did not explicitly identify 1p and 19q deletions as strongly associated with better prognosis (ranks 45, 233), it did identify *CIC* mutations, a signature of oligodendrogliomas (*CIC* mutations occur in more than 50% of oligodendrogliomas), and *SMARCA4* mutations, that occur frequently in both the less aggressive oligodendroglioma and IDH-mutant astrocytoma subtypes. The top-ranked feature associated with poor prognosis in our analysis was deletion of *CDKN2A* which is strongly associated with the aggressive astrocytomas, as well as with a subset of poor prognosis IDH-mutant astrocytomas that lack broad DNA hypermethylation (GCIMP-low)³³. Loss of *PTEN* (rank 8) is also characteristic of astrocytomas, has been shown to be an early event in gliomagenesis, and related to the loss of its parent chromosome 10 (10q and 10p were ranked 2 and 4, respectively)³⁴. Similarly, enrichment analysis of our transcriptional glioma model risk scores identified molecular pathways and processes related to epithelia to mesenchymal transition, a process that is associated with poor prognosis in cancers generally and specifically in gliomas.

Transfer learning experiments showed that deep survival models could benefit from training with multi-cancer datasets in the high-dimensional transcriptional feature set. Training with combined BRCA, OV and UCEC transcriptional data significantly degraded the accuracy of Cox elastic net models in predicting BRCA outcomes, but provided a small benefit to deep survival models (3.5% improvement). Both methods benefit significantly from training on multi-cancer integrated feature sets. Given that the integrated feature sets contain a much smaller number of samples than the transcriptional datasets (see Figure S1), it is reasonable that they would benefit more from additional training data. A similar rationale could explain the performance difference between SurvivalNet and Cox elastic net on the transcriptional feature set: SurvivalNet likely requires more training data and so it would be more likely to benefit from additional cancer types. Additional experiments are needed to investigate if SurvivalNet has a real advantage in transfer learning common prognostic signals across cancer types. Although genetic alterations and expression patterns are often strongly associated with primary disease site, common mechanisms of progression are likely shared by many cancers, and deep survival models can benefit from training with augmented datasets that provide additional evidence of these mechanisms. Enrichment analysis of risk scores from the BRCA-only and BRCA + OV + UCEC transcriptional models showed changes in the biological themes associated with highly prognostic transcripts, with increased emphasis on inflammatory response and cell adhesion in the BRCA + OV + UCEC model.

Although our study provides important insights into the use of deep learning for survival modeling, it has some limitations. Larger genomic datasets with clinical follow-up are needed to determine if the feature learning and nonlinearity of deep learning methods can provide substantial benefits in predicting survival. Secondly, our risk backpropagation analysis was simplified by averaging feature risk scores across patients. With nonlinear models, feature risk scores can vary significantly from patient to patient, and an in-depth analysis of these variations could yield insights into alternative paths for disease progression.

Methods

Data. All datasets were created using TCGAIntegrator (<https://github.com/cooperlab/TCGAIntegrator>), a Python module for assembling integrated TCGA genomic and clinical datasets with the Broad Institute Firehose (<https://gdac.broadinstitute.org/>). Datasets were filtered to remove patients lacking essential data platforms required in each experiment. Clinical variables including age and stage were required for each experiment, with missing radiation treatment status (binary) being mean-imputed to reflect prior likelihood in receiving radiation therapy. Features with categorical or ordinal values (i.e. stage) were expanded to a series of binary variables for model training. Copy number features were derived from the Affymetrix Genome-Wide Human SNP Array 6.0 platform. Gene expression features were taken as RSEM values from the Illumina HiSeq. 2000 RNA Sequencing V2 platform. Protein expression measurements were taken from the MD Anderson Reverse Phase Protein Array (RPPA) Core platform that measures expression of cancer-relevant proteins and phosphoproteins. Sparse missing values in protein or gene expression features were 1 nn-imputed (<20% missing values), where features exceeding this missing value threshold were discarded. Significant mutations were identified for inclusion in each dataset (LGG/GBM, KIPAN, BRCA) using a MutSig2CV ≤ 0.1 q-value threshold. Gene-level copy number features were filtered using a GISTIC ≤ 0.25 q-value threshold to identify focal events, and were further filtered using the Sanger Cancer Gene Census³⁵. All clinical and molecular features were standardized to zero-mean unit-variance to comply with best practices for training deep-learning algorithms. All datasets used to create this paper, along with the TCGAIntegrator commands used to generate these datasets are available on request.

Software and hardware. All software used in training deep survival models, bayesian optimization, and model interpretation are provided as an installable python package at <https://github.com/CancerDataScience/SurvivalNet>. We have also provided a Docker container containing an installation of the package and all dependencies that provides access to SurvivalNet functionality without the need for software installations. SurvivalNet is implemented on top of the Numpy (v1.11)/SciPy (v0.18) stack using Theano (v0.8.2). Bayesian optimization was performed using the BayesOpt package (<https://github.com/rmcantin/bayesopt>). Survival analysis statistics like Kaplan Meier analysis and logrank testing were performed using the Python lifelines package (v0.8.0.1). Cox elastic net models were trained using Glmnet for Matlab (http://web.stanford.edu/~hastie/glmnet_matlab/). Random survival forest models were trained using the RandomForestSRC (2.2.0) R package. Experiments were performed on a workstation equipped with two Intel Xeon E5-2620 v3 six-core processors, 64GB RAM, and two Titan-X GTX graphics processing units.

Training, model selection and validation procedures. Deep survival models are multi-layer feed forward artificial neural networks with a Cox proportional hazards output layer that calculates negative log partial likelihood

$$l(\beta, X) = - \sum_{i \in U} \left(X_i \beta - \log \sum_{j \in R_i} e^{X_j \beta} \right) \quad (1)$$

where X_i are the inputs to the output layer, β are the Cox model parameters, U is the set of uncensored samples and R_i is the set of “at-risk” samples with survival or follow-up times $Y_j \geq Y_i$.

This likelihood was optimized using backpropagation and line-search gradient descent. In each backpropagation iteration, the log partial likelihood is backpropagated throughout the network layers to update the interconnecting weights. The derivative used in backpropagation is

$$\frac{\partial l(\beta, X)}{\partial X_i} = c_i \beta - \sum_{j \in U, i \in R_j} \frac{\beta e^{X_j \beta}}{\sum_{k \in R_j} e^{X_k \beta}} \quad (2)$$

where X_i is the input to the output/Cox layer. This derivative is multiplied by derivatives of the hidden layers using the chain rule to update all the network parameters back to the first network layer. Training was performed by combining all samples into a single batch, and updating the model once per epoch, due to the dependence between samples in calculating the Cox partial likelihood (Equations 1, 2). We note that mini-batch training can be performed with SurvivalNet by fitting the likelihood to smaller batches of samples, but this approach was not used in our experiments. Regularization of the network during training was performed using random dropout of network weights.

Bayesian optimization was performed by splitting samples into training (60%), validation (20%) and testing (20%) sets. The training and validation sets were used by Bayesian optimization to determine the optimal model hyperparameters, namely number of layers (1–5), layer width (10–1000), dropout fraction (0–0.9) and activation function (Rectified-linear or hyperbolic tangent). The optimal model architecture was then applied to the testing set to evaluate c-index of the selected model. We repeated this procedure on 20 randomized assignments of the samples to training/validation/testing.

Cox elastic net models contain two hyperparameters, λ which controls the overall degree of regularization and the mixture coefficient α that controls the balance between L2 and L1 norm penalties. Grid search over λ , α was performed to optimize the choice of these parameters. For each choice of α , a separate λ sequence was generated by Glmnet since the range of λ depends strongly on the α . A model was trained for each α/λ pair using the training set, and the model with the best performance on the validation set was then evaluated on the testing set. The same validation procedure was used to tune RSF hyperparameters including the number of trees (50, 100, 500, 1000), node size (1, 3, 5, 7, 9), and random splitting based on the recommendations in the randomForestSRC R package.

Risk backpropagation and model interpretation. The models used for risk backpropagation and interpretation were created by identifying the best performing model configuration from the 20 randomized experiments. These configurations were then used to re-train a model using all available samples. Risk backpropagation was implemented using Theano to calculate the partial derivatives of risk with respect to each input variable using the multivariable chain rule. Given a deep survival model with H hidden layers that operates on an N -dimensional feature vector f to predict risk R , the feature risk scores are calculated as the partial derivative of the model with respect to inputs

$$\frac{\partial R}{\partial f} = \beta \times \prod_{h=1}^H J_h \quad (3)$$

where J_h is the Jacobian matrix of the h -th hidden layer with respect to its inputs, and β is the vector of parameters of the final layer that is a linear transformation (note the exponential is not applied since we are dealing with risk). This partial derivative is evaluated using the features of each patient f_i to generate an N -dimensional feature risk score vector for each patient. Features were ranked by calculating the median risk score for each feature across all patients.

For transcriptional models, feature risk scores were analyzed using the Preranked Gene Set Enrichment Analysis (GSEAPrerankedv1) module in GenePattern. The Hallmark gene set³⁶ from the MSigDB database (<http://software.broadinstitute.org/gsea/msigdb/>) was used for enrichment analysis. The HUGO Gene Nomenclature Committee database was used to harmonize gene symbols between gene sets and model features prior to GSEA analysis (<http://www.genenames.org/>).

Transfer learning experiments. Datasets were combined using their shared features. For transcriptional and molecular features this merging is trivial, although many of the mutations and copy-number variations are dataset specific since they are filtered by GISTIC and MutSig to identify frequent alterations for each disease (integrated feature sets used in transfer learning are considerably smaller as a result). Pathologic stage and clinical stage were merged as a single “stage” variable where necessary, since their definitions of stage are similar (although the method of determining this stage differs). No additional normalization measures were employed to remove disease-specific biases.

Data availability. This paper was produced using large volumes of publicly available genomic data. The authors have made every effort to make available links to these resources as well as making publicly available the software methods used to produce the datasets, analyses, and summary information. All data not published in the tables and supplements of this article are available from the corresponding author on request.

References

1. Cancer Genome Atlas Research, N. *et al.* Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med* **372**, 2481–2498, doi:<https://doi.org/10.1056/NEJMoa1402121> (2015).
2. Solin, L. J. *et al.* A multigene expression assay to predict local recurrence risk for ductal carcinoma *in situ* of the breast. *J Natl Cancer Inst* **105**, 701–710, doi:<https://doi.org/10.1093/jnci/djt067> (2013).
3. Cardoso, F. *et al.* 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N Engl J Med* **375**, 717–729, doi:<https://doi.org/10.1056/NEJMoa1602253> (2016).
4. Bartlett, J. M. *et al.* Mammostrat as a tool to stratify breast cancer patients at risk of recurrence during endocrine therapy. *Breast Cancer Res* **12**, R47, doi:<https://doi.org/10.1186/bcr2604> (2010).
5. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* **13**, 8–17, doi:<https://doi.org/10.1016/j.csbj.2014.11.005> (2015).
6. Gao, S. *et al.* Identification and Construction of Combinatory Cancer Hallmark-Based Gene Signature Sets to Predict Recurrence and Chemotherapy Benefit in Stage II Colorectal Cancer. *JAMA Oncol* **2**, 37–45, doi:<https://doi.org/10.1001/jamaoncol.2015.3413> (2016).
7. Li, J. *et al.* Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun* **1**, 34, doi:<https://doi.org/10.1038/ncomms1033> (2010).
8. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw* **39**, 1–13 (2011).
9. Ishwaran, H. *et al.* Random survival forests for competing risks. *Biostatistics* **15**, 757–773, doi:<https://doi.org/10.1093/biostatistics/kxu010> (2014).
10. Faraggi, D. & Simon, R. A neural network model for survival data. *Stat Med* **14**, 73–82 (1995).
11. Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J. & Azen, S. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational Statistics & Data Analysis* **34**, 243–257, doi:[https://doi.org/10.1016/S0167-9473\(99\)00098-5](https://doi.org/10.1016/S0167-9473(99)00098-5) (2000).
12. Li, Y., Chen, C. Y. & Wasserman, W. W. Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters. *J Comput Biol* **23**, 322–336, doi:<https://doi.org/10.1089/cmb.2015.0189> (2016).
13. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444, doi:<https://doi.org/10.1038/nature14539> (2015).
14. Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *Ieee T Pattern Anal* **35**, 1798–1828, doi:<https://doi.org/10.1109/TPAMI.2013.50> (2013).
15. Turkki, R., Linder, N., Kovanen, P. E., Pellinen, T. & Lundin, J. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *J Pathol Inform* **7**, 38, doi:<https://doi.org/10.4103/2153-3539.189703> (2016).
16. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**, 831–838, doi:<https://doi.org/10.1038/nbt.3300> (2015).
17. Nemati, S. *et al.* Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. *Conf Proc IEEE Eng Med Biol Soc* **2016**, 2978–2981, doi:<https://doi.org/10.1109/EMBC.2016.7591355> (2016).
18. Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2960–2968 (2012).
19. Martinez-Cantin, R. BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits. *Journal of Machine Learning Research* **15**, 3735–3739 (2014).
20. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. in *25th Annual Conference on Neural Information Processing Systems* (2011).
21. Yousefi, S., Song, C., Nauata, N. & Cooper, L. Learning Genomic Representations to Predict Clinical Outcomes in Cancer. *ArXiv e-prints* 1609, arXiv:1609.08663 (2016).
22. Katzman, J. *et al.* Deep Survival: A Deep Cox Proportional Hazards Network. *ArXiv e-prints* **1606**, arXiv:1606.00931 (2016).
23. Harrell, F. E. Jr., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).
24. Dimopoulos, Y., Bourret, P. & Lek, S. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters* **2**, 1–4, doi:<https://doi.org/10.1007/bf02309007> (1995).
25. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803–820, doi:<https://doi.org/10.1007/s00401-016-1545-1> (2016).
26. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550, doi:<https://doi.org/10.1073/pnas.0506580102> (2005).
27. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* **15**, 178–196, doi:<https://doi.org/10.1038/nrm3758> (2014).
28. Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325, doi:<https://doi.org/10.1038/nature08712> (2010).
29. Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110, doi:<https://doi.org/10.1016/j.ccr.2009.12.020> (2010).
30. Bhat, K. P. *et al.* The transcriptional coactivator TAZ regulates mesenchymal differentiation in malignant glioma. *Genes Dev* **25**, 2594–2609, doi:<https://doi.org/10.1101/gad.176800.111> (2011).
31. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *ArXiv e-prints* **1611**, arXiv:1611.03530 (2016).
32. Fakoor, R., Ladhak, F., Nazi, A. & Huber, M. Using deep learning to enhance cancer diagnosis and classification in *Proceedings of the WHEALTH ICML Workshop*, 129–133 (2011).
33. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550–563, doi:<https://doi.org/10.1016/j.cell.2015.12.028> (2016).
34. Ozawa, T. *et al.* Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like precursor glioma. *Cancer Cell* **26**, 288–300, doi:<https://doi.org/10.1016/j.ccr.2014.06.005> (2014).
35. Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177–183, doi:<https://doi.org/10.1038/nrc1299> (2004).
36. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425, doi:<https://doi.org/10.1016/j.cels.2015.12.004> (2015).

Acknowledgements

This work was supported by the National Brain Tumor Society Oligo Research Fund, U.S. National Institutes of Health, National Library of Medicine Career Development Award K22LM011576, and National Cancer Institute grant U24CA194362, National Institutes of Health CTSI grants UL1TR000454 and TL1TR000456, and with funds from the Emory Winship Cancer Institute.

Author Contributions

S.Y. and F.A. developed the primary method. S.Y., F.A. curated datasets. S.Y., F.A., C.D., J.L., D.A.G. performed experiments. C.S. assisted with software development. M.A., S.H., J.E.V.V., and D.J.B. assisted with interpretation of experimental results. L.A.D.C. conceived of the ideas and supervised the work.

Additional Information

Supplementary information accompanies this paper at doi:[10.1038/s41598-017-11817-6](https://doi.org/10.1038/s41598-017-11817-6)

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

3.3 Predicting cancer outcomes from histology and genomics using convolutional networks, PNAS, 2018

This section is an exact copy of the following open-access journal paper:

Pooya Mobadersany, **Safoora Yousefi**, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, Jos E Velzquez Vega, Daniel J Brat, Lee AD Cooper. *Predicting cancer outcomes from histology and genomics using convolutional networks*. Proceedings of the National Academy of Sciences, 115(13), E2970-E2979.

Abstract. Cancer histology reflects underlying molecular processes and disease progression and contains rich phenotypic information that is predictive of patient outcomes. In this study, we show a computational approach for learning patient outcomes from digital pathology images using deep learning to combine the power of adaptive machine learning algorithms with traditional survival models. We illustrate how these survival convolutional neural networks (SCNNs) can integrate information from both histology images and genomic biomarkers into a single unified framework to predict time-to-event outcomes and show prediction accuracy that surpasses the current clinical paradigm for predicting the overall survival of patients diagnosed with glioma. We use statistical sampling techniques to address challenges in learning survival from histology images, including tumor heterogeneity and the need for large training cohorts. We also provide insights into the prediction mechanisms of SCNNs, using heat map visualization to show that SCNNs recognize important structures, like microvascular proliferation, that are related to prognosis and that are used by pathologists in grading. These results highlight the emerging role of deep learning in precision medicine and suggest an expanding utility for computational analysis of histology in the future practice of pathology.

Candidate's contribution. Development, testing, and optimization of Cox's proportional hazards survival analysis layer in Tensorflow.



Predicting cancer outcomes from histology and genomics using convolutional networks

Pooya Mobadersany^a, Safoora Yousefi^a, Mohamed Amgad^a, David A. Gutman^b, Jill S. Barnholtz-Sloan^c, José E. Velázquez Vega^d, Daniel J. Brat^e, and Lee A. D. Cooper^{a,f,g,1}

^aDepartment of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322; ^bDepartment of Neurology, Emory University School of Medicine, Atlanta, GA 30322; ^cCase Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH 44106; ^dDepartment of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA 30322; ^eDepartment of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611; ^fWinship Cancer Institute, Emory University, Atlanta, GA 30322; and ^gDepartment of Biomedical Engineering, Emory University and Georgia Institute of Technology, Atlanta, GA 30322

Edited by Bert Vogelstein, Johns Hopkins University, Baltimore, MD, and approved February 13, 2018 (received for review October 4, 2017)

Cancer histology reflects underlying molecular processes and disease progression and contains rich phenotypic information that is predictive of patient outcomes. In this study, we show a computational approach for learning patient outcomes from digital pathology images using deep learning to combine the power of adaptive machine learning algorithms with traditional survival models. We illustrate how these survival convolutional neural networks (SCNNs) can integrate information from both histology images and genomic biomarkers into a single unified framework to predict time-to-event outcomes and show prediction accuracy that surpasses the current clinical paradigm for predicting the overall survival of patients diagnosed with glioma. We use statistical sampling techniques to address challenges in learning survival from histology images, including tumor heterogeneity and the need for large training cohorts. We also provide insights into the prediction mechanisms of SCNNs, using heat map visualization to show that SCNNs recognize important structures, like microvascular proliferation, that are related to prognosis and that are used by pathologists in grading. These results highlight the emerging role of deep learning in precision medicine and suggest an expanding utility for computational analysis of histology in the future practice of pathology.

artificial intelligence | machine learning | digital pathology | deep learning | cancer

Histology has been an important tool in cancer diagnosis and prognostication for more than a century. Anatomic pathologists evaluate histology for characteristics, like nuclear atypia, mitotic activity, cellular density, and tissue architecture, incorporating cytologic details and higher-order patterns to classify and grade lesions. Although prognostication increasingly relies on genomic biomarkers that measure genetic alterations, gene expression, and epigenetic modifications, histology remains an important tool in predicting the future course of a patient's disease. The phenotypic information present in histology reflects the aggregate effect of molecular alterations on cancer cell behavior and provides a convenient visual readout of disease aggressiveness. However, human assessments of histology are highly subjective and are not repeatable; hence, computational analysis of histology imaging has received significant attention. Aided by advances in slide scanning microscopes and computing, a number of image analysis algorithms have been developed for grading (1–4), classification (5–10), and identification of lymph node metastases (11) in multiple cancer types.

Deep convolutional neural networks (CNNs) have emerged as an important image analysis tool and have shattered performance benchmarks in many challenging applications (12). The ability of CNNs to learn predictive features from raw image data is a paradigm shift that presents exciting opportunities in medical imaging (13–15). Medical image analysis applications have heavily relied on feature engineering approaches, where algorithm pipelines are used to explicitly delineate structures of interest using segmentation algorithms to measure predefined features of

these structures that are believed to be predictive and to use these features to train models that predict patient outcomes. In contrast, the feature learning paradigm of CNNs adaptively learns to transform images into highly predictive features for a specific learning objective. The images and patient labels are presented to a network composed of interconnected layers of convolutional filters that highlight important patterns in the images, and the filters and other parameters of this network are mathematically adapted to minimize prediction error. Feature learning avoids biased a priori definition of features and does not require the use of segmentation algorithms that are often confounded by artifacts and natural variations in image color and intensity. While feature learning has become the dominant paradigm in general image analysis tasks, medical applications pose unique challenges. Large amounts of labeled data are needed to train CNNs, and medical applications often suffer from data deficits that limit performance. As “black box” models, CNNs are also difficult to deconstruct, and therefore, their prediction mechanisms are difficult to interpret. Despite these

Significance

Predicting the expected outcome of patients diagnosed with cancer is a critical step in treatment. Advances in genomic and imaging technologies provide physicians with vast amounts of data, yet prognostication remains largely subjective, leading to suboptimal clinical management. We developed a computational approach based on deep learning to predict the overall survival of patients diagnosed with brain tumors from microscopic images of tissue biopsies and genomic biomarkers. This method uses adaptive feedback to simultaneously learn the visual patterns and molecular biomarkers associated with patient outcomes. Our approach surpasses the prognostic accuracy of human experts using the current clinical standard for classifying brain tumors and presents an innovative approach for objective, accurate, and integrated prediction of patient outcomes.

Author contributions: P.M., S.Y., M.A., D.A.G., D.J.B., and L.A.D.C. designed research; P.M., S.Y., J.E.V.V., and L.A.D.C. performed research; P.M., J.S.B.-S., and L.A.D.C. analyzed data; and P.M., M.A., D.A.G., J.S.B.-S., J.E.V.V., D.J.B., and L.A.D.C. wrote the paper.

Conflict of interest statement: L.A.D.C. leads a research project that is financially supported by Ventana Medical Systems, Inc. While this project is not directly related to the manuscript, it is in the general area of digital pathology.

This article is a PNAS Direct Submission.

This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Data deposition: Software and other resources related to this paper have been deposited at GitHub, <https://github.com/CancerDataScience/SCNN>.

¹To whom correspondence should be addressed. Email: Lee.Cooper@Emory.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717139115/-DCSupplemental.

Published online March 12, 2018.

challenges, CNNs have been successfully used extensively for medical image analysis (9, 11, 16–26).

Many important problems in the clinical management of cancer involve time-to-event prediction, including accurate prediction of overall survival and time to progression. Despite overwhelming success in other applications, deep learning has not been widely applied to these problems. Survival analysis has often been approached as a binary classification problem by predicting dichotomized outcomes at a specific time point (e.g., 5-y survival) (27). The classification approach has important limitations, as subjects with incomplete follow-up cannot be used in training, and binary classifiers do not model the probability of survival at other times. Time-to-event models, like Cox regression, can utilize all subjects in training and model their survival probabilities for a range of times with a single model. Neural network-based Cox regression approaches were explored in early machine learning work using datasets containing tens of features, but subsequent analysis found no improvement over basic linear Cox regression (28). More advanced “deep” neural networks that are composed of many layers were recently adapted to optimize Cox proportional hazard likelihood and were shown to have equal or superior performance in predicting survival using genomic profiles containing hundreds to tens of thousands of features (29, 30) and using basic clinical profiles containing 14 features (31).

Learning survival from histology is considerably more difficult, and a similar approach that combined Cox regression with CNNs to predict survival from lung cancer histology achieved only marginally better than random accuracy (0.629 c index) (32). Time-to-event prediction faces many of the same challenges as other applications where CNNs are used to analyze histology. Compared with genomic or clinical datasets, where features have intrinsic meaning, a “feature” in an image is a pixel with meaning that depends entirely on context. Convolution operations can learn these contexts, but the resulting networks are complex, often containing more than 100 million free parameters, and thus, large cohorts are needed for training. This problem is intensified in time-to-event prediction, as clinical follow-up is often difficult to obtain for large cohorts. Data augmentation techniques have been adopted to address this problem, where randomized rotations and transformations of contrast and brightness are used to synthesize additional training data (9, 11, 14, 15, 17, 19, 25, 26, 33). Intratumoral heterogeneity also presents a significant challenge in time-to-event prediction, as a tissue biopsy often contains a range of histologic patterns that correspond to varying degrees of disease progression or aggressiveness. The method for integrating information from heterogeneous regions within a sample is an important consideration in predicting outcomes. Furthermore, risk is often reflected in subtle changes in multiple histologic criteria that can require years of specialized training for human pathologists to recognize and interpret. Developing an algorithm that can learn the continuum of risks associated with histology can be more challenging than for other learning tasks, like cell or region classification.

In this paper, we present an approach called survival convolutional neural networks (SCNNs), which provide highly accurate prediction of time-to-event outcomes from histology images. Using diffuse gliomas as a driving application, we show how the predictive accuracy of SCNNs is comparable with manual histologic grading by neuropathologists. We further extended this approach to integrate both histology images and genomic biomarkers into a unified prediction framework that surpasses the prognostic accuracy of the current WHO paradigm based on genomic classification and histologic grading. Our SCNN framework uses an image sampling and risk filtering technique that significantly improves prediction accuracy by mitigating the effects of intratumoral heterogeneity and deficits in the availability of labeled data for training. Finally, we use

heat map visualization techniques applied to whole-slide images to show how SCNNs learn to recognize important histologic structures that neuropathologists use in grading diffuse gliomas and suggest relevance for patterns with prognostic significance that is not currently appreciated. We systematically validate our approaches by predicting overall survival in gliomas using data from The Cancer Genome Atlas (TCGA) Lower-Grade Glioma (LGG) and Glioblastoma (GBM) projects.

Results

Learning Patient Outcomes with Deep Survival Convolutional Neural Networks. The SCNN model architecture is depicted in Fig. 1 (Fig. S1 shows a detailed diagram). H&E-stained tissue sections are first digitized to whole-slide images. These images are reviewed using a web-based platform to identify regions of interest (ROIs) that contain viable tumor with representative histologic characteristics and that are free of artifacts (*Methods*) (34, 35). High-power fields (HPFs) from these ROIs are then used to train a deep convolutional network that is seamlessly integrated with a Cox proportional hazards model to predict patient outcomes. The network is composed of interconnected layers of image processing operations and nonlinear functions that sequentially transform the HPF image into highly predictive prognostic features. Convolutional layers first extract visual features from the HPF at multiple scales using convolutional kernels and pooling operations. These image-derived features feed into fully connected layers that perform additional transformations, and then, a final Cox model layer outputs a prediction of patient risk. The interconnection weights and convolutional kernels are trained by comparing risk predicted by the network with survival or other time-to-event outcomes using a backpropagation technique to optimize the statistical likelihood of the network (*Methods*).

To improve the performance of SCNN models, we developed a sampling and risk filtering technique to address intratumoral heterogeneity and the limited availability of training samples (Fig. 2). In training, new HPFs are randomly sampled from each ROI at the start of each training iteration, providing the SCNN model with a fresh look at each patient’s histology and capturing heterogeneity within the ROI. Each HPF is processed using standard data augmentation techniques that randomly transform the field to reinforce network robustness to tissue orientation and variations in staining (33). The SCNN is trained using multiple transformed HPFs for each patient (one for each ROI) to further account for intratumoral heterogeneity across ROIs. For prospective prediction, we first sample multiple HPFs within each ROI to generate a representative collection of fields for the patient. The median risk is calculated within each ROI, and then, these median risks are sorted and filtered to predict a robust patient-level risk that reflects the aggressiveness of their disease while rejecting any outlying risk predictions. These sampling and filtering procedures are described in detail in *Methods*.

Assessing the Prognostic Accuracy of SCNN. To assess the prognostic accuracy of SCNN, we assembled whole-slide image tissue sections from formalin-fixed, paraffin-embedded specimens and clinical follow-up for 769 gliomas from the TCGA (*Dataset S1*). This dataset comprises lower-grade gliomas (WHO grades II and III) and glioblastomas (WHO grade IV), contains both astrocytomas and oligodendrogliomas, and has overall survivals ranging from less than 1 to 14 y or more. A summary of demographics, grades, survival, and molecular subtypes for this cohort is presented in *Table S1*. The Digital Slide Archive was used to identify ROIs in 1,061 H&E-stained whole-slide images from these tumors.

The prognostic accuracy of SCNN models was assessed using Monte Carlo cross-validation. We randomly split our cohort into paired training (80%) and testing (20%) sets to generate 15 training/testing set pairs. We trained an SCNN model using

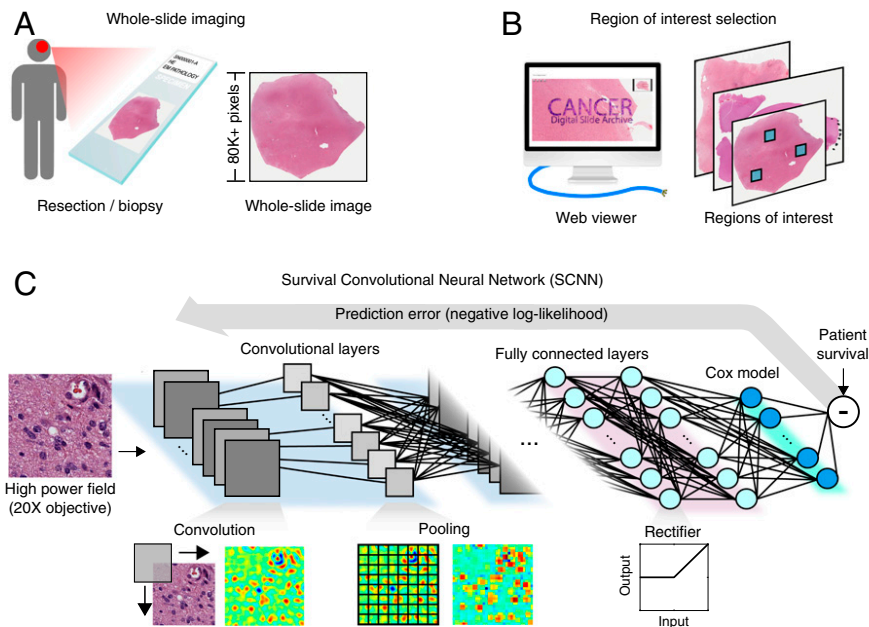


Fig. 1. The SCNN model. The SCNN combines deep learning CNNs with traditional survival models to learn survival-related patterns from histology images. (A) Large whole-slide images are generated by digitizing H&E-stained glass slides. (B) A web-based viewer is used to manually identify representative ROIs in the image. (C) HPFs are sampled from these regions and used to train a neural network to predict patient survival. The SCNN consists of (i) convolutional layers that learn visual patterns related to survival using convolution and pooling operations, (ii) fully connected layers that provide additional nonlinear transformations of extracted image features, and (iii) a Cox proportional hazards layer that models time-to-event data, like overall survival or time to progression. Predictions are compared with patient outcomes to adaptively train the network weights that interconnect the layers.

each training set and then, evaluated the prognostic accuracy of these models on the paired testing sets, generating a total of 15 accuracy measurements (*Methods* and *Dataset S1*). Accuracy was measured using Harrell's *c* index, a nonparametric statistic that measures concordance between predicted risks and actual survival (36). A *c* index of 1 indicates perfect concordance between

predicted risk and overall survival, and a *c* index of 0.5 corresponds to random concordance.

For comparison, we also assessed the prognostic accuracy of baseline linear Cox models generated using the genomic biomarkers and manual histologic grades from the WHO classification of gliomas (Fig. 3A). The WHO assigns the diffuse gliomas

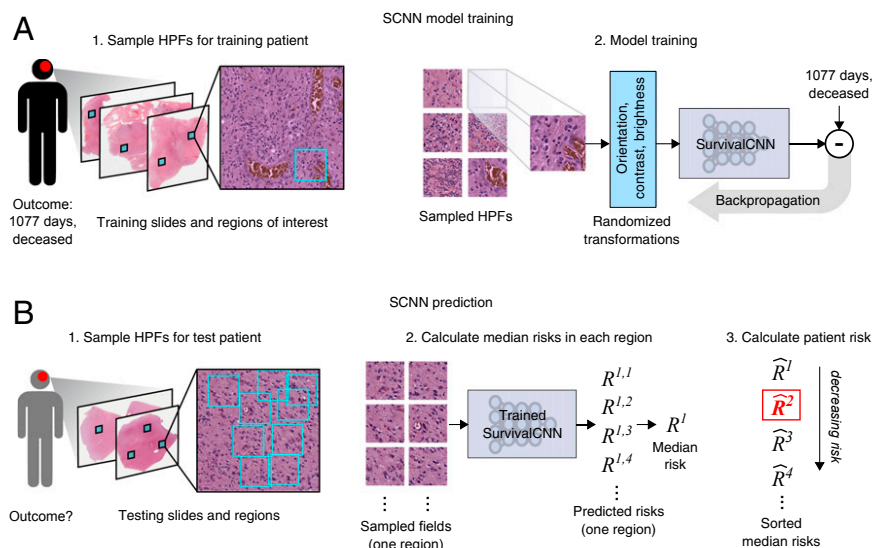


Fig. 2. SCNN uses image sampling and filtering to improve the robustness of training and prediction. (A) During training, a single 256×256 -pixel HPF is sampled from each region, producing multiple HPFs per patient. Each HPF is subjected to a series of random transformations and is then used as an independent sample to update the network weights. New HPFs are sampled at each training epoch (one training pass through all patients). (B) When predicting the outcome of a newly diagnosed patient, nine HPFs are sampled from each ROI, and a risk is predicted for each field. The median HPF risk is calculated in each region, these median risks are then sorted, and the second highest value is selected as the patient risk. This sampling and filtering framework was designed to deal with tissue heterogeneity by emulating manual histologic evaluation, where prognostication is typically based on the most malignant region observed within a heterogeneous sample. Predictions based on the highest risk and the second highest risk had equal performance on average in our experiments, but the maximum risk produced some outliers with poor prediction accuracy.

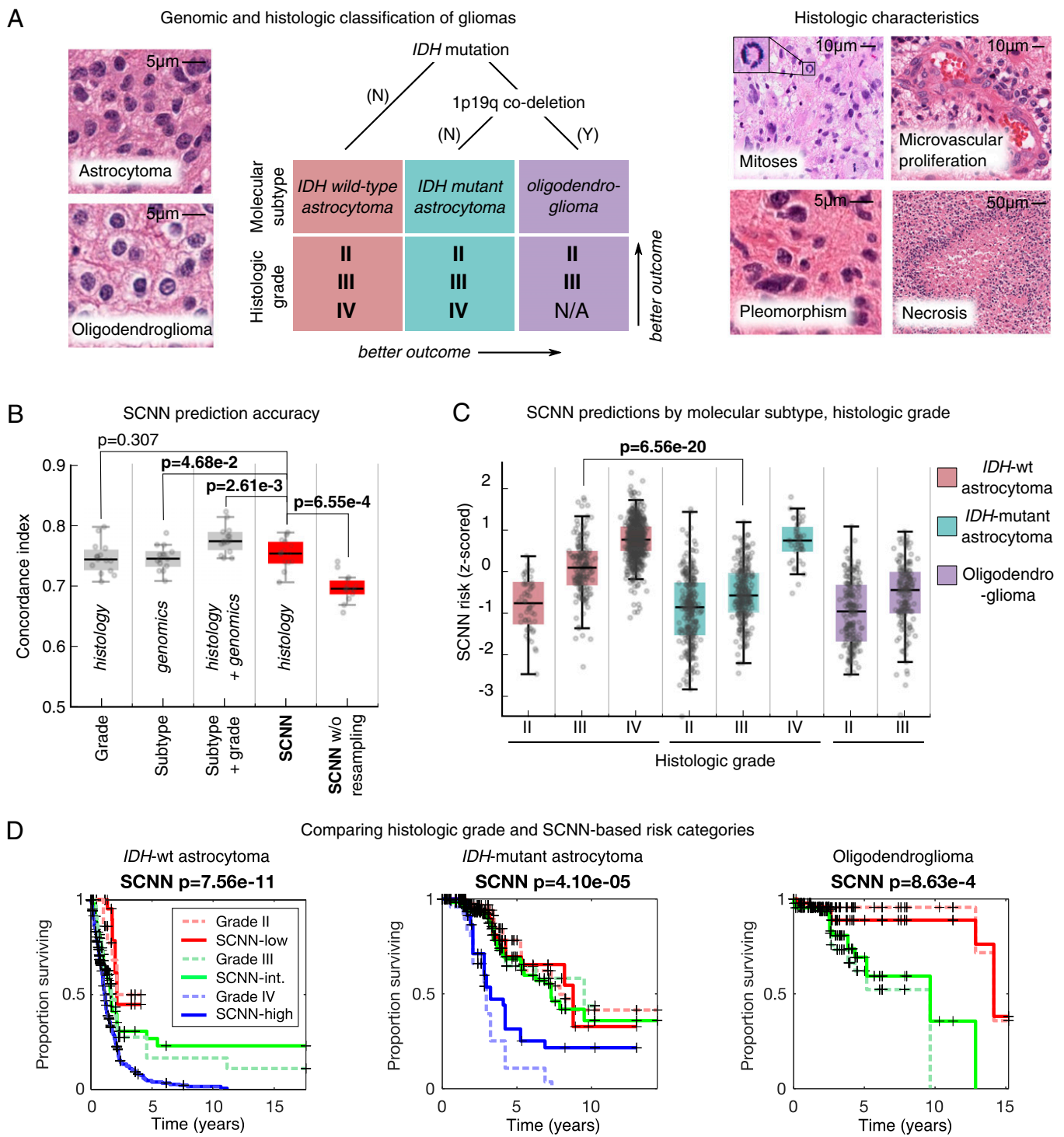


Fig. 3. Prognostication criteria for diffuse gliomas. (A) Prognosis in the diffuse gliomas is determined by genomic classification and manual histologic grading. Diffuse gliomas are first classified into one of three molecular subtypes based on *IDH1/IDH2* mutations and the codeletion of chromosomes 1p and 19q. Grade is then determined within each subtype using histologic characteristics. Subtypes with an astrocytic lineage are split by *IDH* mutation status, and the combination of 1p/19q codeletion and *IDH* mutation defines an oligodendroglioma. These lineages have histologic differences; however, histologic evaluation is not a reliable predictor of molecular subtype (37). Histologic criteria used for grading range from nuclear morphology to higher-level patterns, like necrosis or the presence of abnormal microvascular structures. (B) Comparison of the prognostic accuracy of SCNN models with that of baseline models based on molecular subtype or molecular subtype and histologic grade. Models were evaluated over 15 independent training/testing sets with randomized patient assignments and with/without training and testing sampling. (C) The risks predicted by the SCNN models correlate with both histologic grade and molecular subtype, decreasing with grade and generally trending with the clinical aggressiveness of genomic subtypes. (D) Kaplan–Meier plots comparing manual histologic grading and SCNN predictions. Risk categories (low, intermediate, high) were generated by thresholding SCNN risks. N/A, not applicable.

to three genomic subtypes defined by mutations in the isocitrate dehydrogenase (*IDH*) genes (*IDH1/IDH2*) and codeletion of

chromosomes 1p and 19q. Within these molecular subtypes, gliomas are further assigned a histologic grade based on criteria that vary

depending on cell of origin (either astrocytic or oligodendroglial). These criteria include mitotic activity, nuclear atypia, the presence of necrosis, and the characteristics of microvascular structures (microvascular proliferation). Histologic grade remains a significant determinant in planning treatment for gliomas, with grades III and IV typically being treated aggressively with radiation and concomitant chemotherapy.

SCNN models showed substantial prognostic power, achieving a median c index of 0.754 (Fig. 3B). SCNN models also performed comparably with manual histologic-grade baseline models (median c index 0.745, $P = 0.307$) and with molecular subtype baseline models (median c index 0.746, $P = 4.68\text{e-}2$). Baseline models representing WHO classification that integrate both molecular subtype and manual histologic grade performed slightly better than SCNN, with a median c index of 0.774 (Wilcoxon signed rank $P = 2.61\text{e-}3$).

We also evaluated the impact of the sampling and ranking procedures shown in Fig. 2 in improving the performance of SCNN models. Repeating the SCNN experiments without these sampling techniques reduced the median c index of SCNN models to 0.696, significantly worse than for models where sampling was used ($P = 6.55\text{e-}4$).

SCNN Predictions Correlate with Molecular Subtypes and Manual Histologic Grade. To further investigate the relationship between SCNN predictions and the WHO paradigm, we visualized how risks predicted by SCNN are distributed across molecular subtype and histologic grade (Fig. 3C). SCNN predictions were highly correlated with both molecular subtype and grade and were consistent with expected patient outcomes. First, within each molecular subtype, the risks predicted by SCNN increase with histologic grade. Second, predicted risks are consistent with the published expected overall survivals associated with molecular subtypes (37). *IDH* WT astrocytomas are, for the most part, highly aggressive, having a median survival of 18 mo, and the collective predicted risks for these patients are higher than for patients from other subtypes. *IDH* mutant astrocytomas are another subtype with considerably better overall survival ranging from 3 to 8 y, and the predicted risks for patients in this subtype are more moderate. Notably, SCNN risks for *IDH* mutant astrocytomas are not well-separated for grades II and III, consistent with reports of histologic grade being an inadequate predictor of outcome in this subtype (38). Infiltrating gliomas with the combination of *IDH* mutations and codeletion of chromosomes 1p/19q are classified as oligodendrogliomas in the current WHO schema, and these have the lowest overall predicted risks consistent with overall survivals of 10+ y (37, 39). Finally, we noted a significant difference in predicted risks when comparing the *IDH* mutant and *IDH* WT grade III astrocytomas (rank sum $P = 6.56\text{e-}20$). These subtypes share an astrocytic lineage and are graded using identical histologic criteria. Although some histologic features are more prevalent in *IDH*-mutant astrocytomas, these features are not highly specific or sensitive to *IDH* mutant tumors and cannot be used to reliably predict *IDH* mutation status (40). Risks predicted by SCNN are consistent with worse outcomes for *IDH* WT astrocytomas in this case (median survival 1.7 vs. 6.3 y in the *IDH* mutant counterparts), suggesting that SCNN models can detect histologic differences associated with *IDH* mutations in astrocytomas.

We also performed a Kaplan–Meier analysis to compare manual histologic grading with “digital grades” based on SCNN risk predictions (Fig. 3D). Low-, intermediate-, and high-risk categories were established by setting thresholds on SCNN predictions to reflect the proportions of manual histologic grades in each molecular subtype (Methods). We observed that, within each subtype, the differences in survival captured by SCNN risk categories are highly similar to manual histologic grading. SCNN risk categories and manual histologic grades have similar prognostic

power in *IDH* WT astrocytomas (log rank $P = 1.23\text{e-}12$ vs. $P = 7.56\text{e-}11$, respectively). In *IDH* mutant astrocytomas, both SCNN risk categories and manual histologic grades have difficulty separating Kaplan–Meier curves for grades II and III, but both clearly distinguish grade IV as being associated with worse outcomes. Discrimination for oligodendroglioma survival is also similar between SCNN risk categories and manual histologic grades (log rank $P = 9.73\text{e-}7$ vs. $P = 8.63\text{e-}4$, respectively).

Improving Prognostic Accuracy by Integrating Genomic Biomarkers.

To integrate both histologic and genomic data into a single unified prediction framework, we developed a genomic survival convolutional neural network (GSCNN model). The GSCNN learns from genomics and histology simultaneously by incorporating genomic data into the fully connected layers of the SCNN (Fig. 4). Both data are presented to the network during training, enabling genomic variables to influence the patterns learned by the SCNN by providing molecular subtype information.

We repeated our experiments using GSCNN models with histology images, *IDH* mutation status, and 1p/19q codeletion as inputs and found that the median c index improved from 0.754 to 0.801. The addition of genomic variables improved performance by 5% on average, and GSCNN models significantly outperform the baseline WHO subtype-grade model trained on equivalent data (signed rank $P = 1.06\text{e-}2$). To assess the value of integrating genomic variables directly into the network during training, we compared GSCNN with a more superficial integration approach, where an SCNN model was first trained using histology images, and then, the risks from this model were combined with *IDH* and 1p/19q variables in a simple three-variable Cox model (Fig. S2). Processing genomic variables in the fully connected layers and including them in training provided a statistically significant benefit; models trained using the superficial approach performed worse than GSCNN models with median c index decreasing to 0.785 (signed rank $P = 4.68\text{e-}2$).

To evaluate the independent prognostic power of risks predicted by SCNN and GSCNN, we performed a multivariable Cox regression analysis (Table 1). In a multivariable regression that included SCNN risks, subtype, grade, age, and sex, SCNN risks had a hazard ratio of 3.05 and were prognostic when correcting for all other features, including manual grade and molecular subtype ($P = 2.71\text{e-}12$). Molecular subtype was also significant in the SCNN multivariable regression model, but histologic grade was not. We also performed a multivariable regression with GSCNN risks and found GSCNN to be significant ($P = 9.69\text{e-}12$) with a hazard ratio of 8.83. In the GSCNN multivariable regression model, molecular subtype was not significant, but histologic grade was marginally significant. We also used Kaplan–Meier analysis to compare risk categories generated from SCNN and GSCNN (Fig. S3). Survival curves for SCNN and GSCNN were very similar when evaluated on the entire cohort. In contrast, their abilities to discriminate survival within molecular subtypes were notably different.

Visualizing Histologic Patterns Associated with Prognosis. Deep learning networks are often criticized for being black box approaches that do not reveal insights into their prediction mechanisms. To investigate the visual patterns that SCNN models associate with poor outcomes, we used heat map visualizations to display the risks predicted by our network in different regions of whole-slide images. Transparent heat map overlays are frequently used for visualization in digital pathology, and in our study, these overlays enable pathologists to correlate the predictions of highly accurate survival models with the underlying histology over the expanse of a whole-slide image. Heat maps were generated using a trained SCNN model to predict the risk for each nonoverlapping HPF in a whole-slide image. The predicted risks were used to generate a color-coded transparent

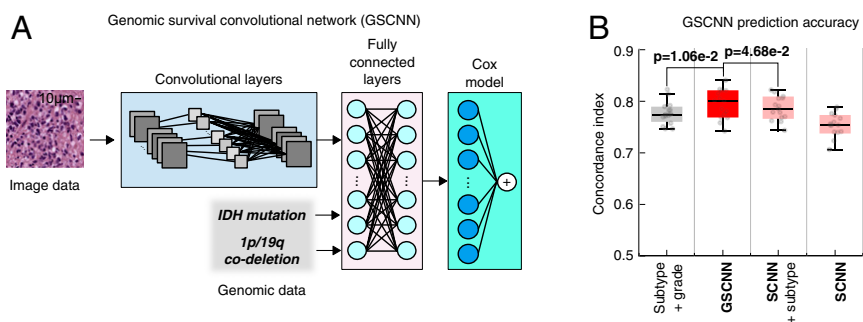


Fig. 4. GSCNN models integrate genomic and imaging data for improved performance. (A) A hybrid architecture was developed to combine histology image and genomic data to make integrated predictions of patient survival. These models incorporate genomic variables as inputs to their fully connected layers. Here, we show the incorporation of genomic variables for gliomas; however, any number of genomic or proteomic measurements can be similarly used. (B) The GSCNN models significantly outperform SCNN models as well as the WHO paradigm based on genomic subtype and histologic grading.

overlay, where red and blue indicate higher and lower SCNN risk, respectively.

A selection of risk heat maps from three patients is presented in Fig. 5, with inlays showing how SCNNs associate risk with important pathologic phenomena. For TCGA-DB-5273 (WHO grade III, *IDH* mutant astrocytoma), the SCNN heat map clearly and specifically highlights regions of early microvascular proliferation, an advanced form of angiogenesis that is a hallmark of malignant progression, as being associated with high risk. Risk in this heat map also increases with cellularity, heterogeneity in nuclear shape and size (pleomorphism), and the presence of abnormal microvascular structures. Regions in TCGA-S9-A7J0 have varying extents of tumor infiltration ranging from normal brain to sparsely infiltrated adjacent normal regions exhibiting satellitosis (where neoplastic cells cluster around neurons) to moderately and highly infiltrated regions. This heat map correctly associates the lowest risks with normal brain regions and can distinguish normal brain from adjacent regions that are sparsely infiltrated. Interestingly, higher risks are assigned to sparsely infiltrated regions (region 1, *Top*) than to regions containing relatively more tumor infiltration (region 2, *Top*). We observed a similar pattern in TCGA-TM-A84G, where edematous regions (region 1, *Bottom*) adjacent to moderately cellular tumor regions (region 1, *Top*) are also assigned higher risks. These latter examples provide risk features embedded within histologic sections that have been previously unrecognized and could inform and improve pathology practice.

Discussion

We developed a deep learning approach for learning survival directly from histological images and created a unified framework for integrating histology and genomic biomarkers for predicting time-to-event outcomes. We systematically evaluated the prognostic accuracy of our approaches in the context of the

current clinical standard based on genomic classification and histologic grading of gliomas. In contrast to a previous study that achieved only marginally better than random prediction accuracy, our approach rivals or exceeds the accuracy of highly trained human experts in predicting survival. Our study provides insights into applications of deep learning in medicine and the integration of histology and genomic data and provides methods for dealing with intratumoral heterogeneity and training data deficits when using deep learning algorithms to predict survival from histology images. Using visualization techniques to gain insights into SCNN prediction mechanisms, we found that SCNNs clearly recognize known and time-honored histologic predictors of poor prognosis and that SCNN predictions suggest prognostic relevance for histologic patterns with significance that is not currently appreciated by neuropathologists.

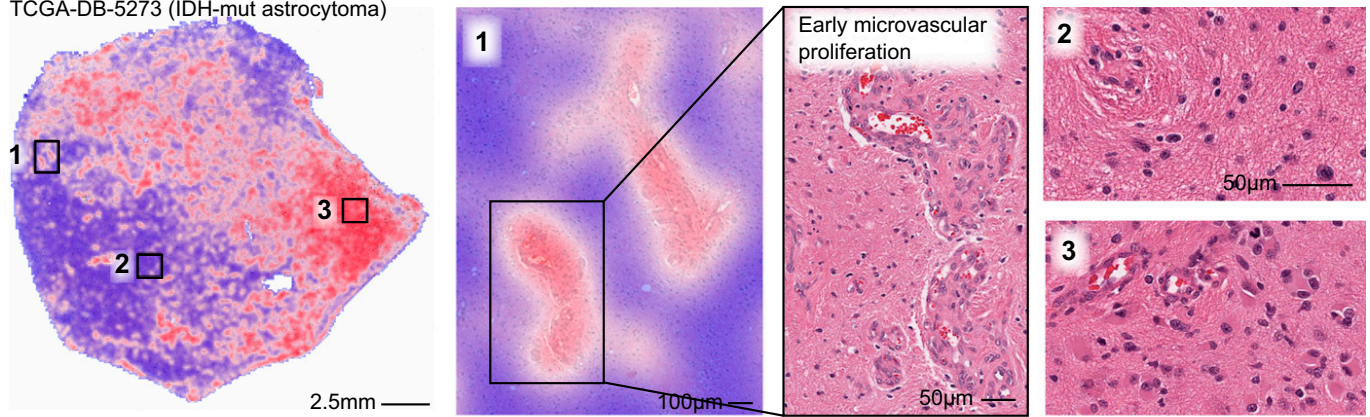
Our study investigated the ability to predict overall survival in diffuse gliomas, a disease with wide variations in outcomes and an ideal test case where histologic grading and genomic classifications have independent prognostic power. Treatment planning for gliomas is dependent on many factors, including patient age and grade, but gliomas assigned to WHO grades III and IV are typically treated very aggressively with radiation and concomitant chemotherapy, whereas WHO grade II gliomas may be treated with chemotherapy or even followed in some cases (41). Histologic diagnosis and grading of gliomas have been limited by considerable intra- and interobserver variability (42). While the emergence of molecular subtyping has resolved uncertainty related to lineage, criteria for grading need to be redefined in the context of molecular subtyping. For example, some morphologic features used to assess grade (e.g., mitotic activity) are no longer prognostic in *IDH* mutant astrocytomas (38). The field of neuro-oncology is currently awaiting features that can better discriminate more aggressive gliomas from those that are more indolent. Improving the accuracy and objectivity of grading will directly impact

Table 1. Hazard ratios for single- and multiple-variable Cox regression models

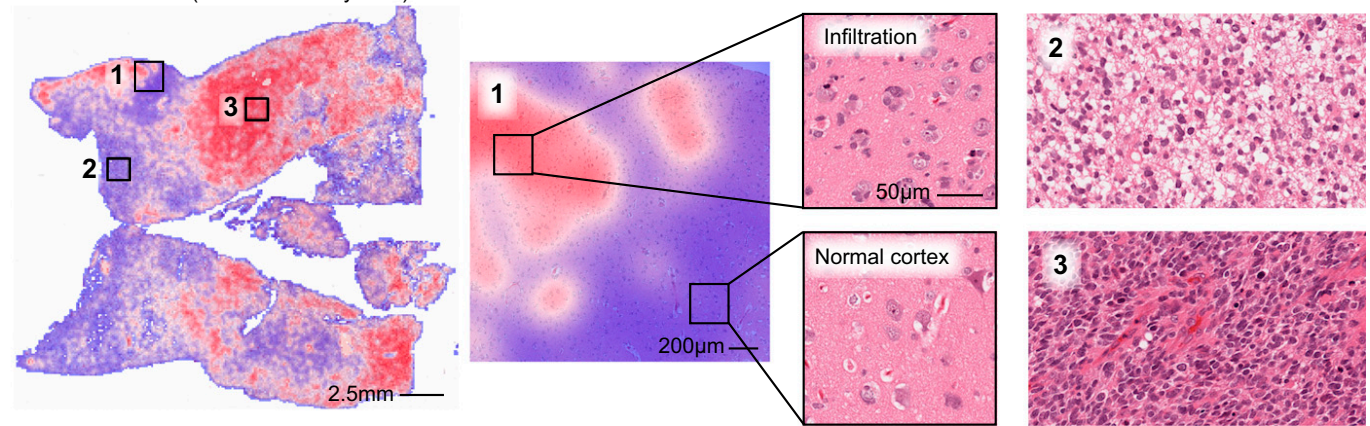
Variable	Single variable				Multivariable (SCNN)			Multivariable (GSCNN)		
	c Index	Hazard ratio	95% CI	P value	Hazard ratio	95% CI	P value	Hazard ratio	95% CI	P value
SCNN	0.741	7.15	5.64, 9.07	2.08e-61	3.05	2.22, 4.19	2.71e-12	—	—	—
GSCNN	0.781	12.60	9.34, 17.0	3.08e-64	—	—	—	8.83	4.66, 16.74	9.69e-12
<i>IDH</i> WT astrocytoma	0.726	9.21	6.88, 12.34	3.48e-52	4.73	2.57, 8.70	3.49e-7	0.97	0.43, 2.17	0.93
<i>IDH</i> mutant astrocytoma	—	0.23	0.170, 0.324	2.70e-19	2.35	1.27, 4.34	5.36e-3	1.67	0.90, 3.12	0.10
Histologic grade IV	0.721	7.25	5.58, 9.43	2.68e-51	1.52	0.839, 2.743	0.159	1.98	1.11, 3.51	0.017
Histologic grade III	—	0.44	0.332, 0.591	1.66e-08	1.57	0.934, 2.638	0.0820	1.78	1.07, 2.97	0.024
Age	0.744	1.77	1.63, 1.93	2.52e-42	1.33	1.20, 1.47	9.57e-9	1.34	1.22, 1.48	9.30e-10
Sex, female	0.552	0.89	0.706, 1.112	0.29	0.85	0.67, 1.08	0.168	0.86	0.68, 1.08	0.18

Bold indicates statistical significance ($P < 5e-2$).

TCGA-DB-5273 (IDH-mut astrocytoma)



TCGA-S9-A7J0 (IDH-mut astrocytoma)



TCGA-TM-A84G (Oligodendroglioma)

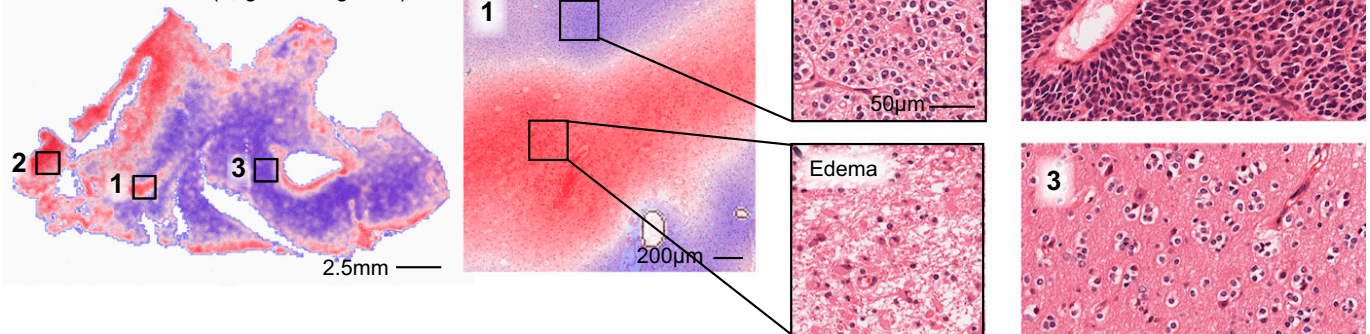


Fig. 5. Visualizing risk with whole-slide SCNN heat maps. We performed SCNN predictions exhaustively within whole-slide images to generate heat map overlays of the risks that SCNN associates with different histologic patterns. Red indicates relatively higher risk, and blue indicates lower risk (the scale for each slide is different). (Top) In TCGA-DB-5273, SCNN clearly and specifically predicts high risks for regions of early microvascular proliferation (region 1) and also, higher risks with increasing tumor infiltration and cell density (region 2 vs. 3). (Middle) In TCGA-S9-A7J0, SCNN can appropriately discriminate between normal cortex (region 1 in Bottom) and adjacent regions infiltrated by tumor (region 1 in Top). Highly cellular regions containing prominent microvascular structures (region 3) are again assigned higher risks than lower-density regions of tumor (region 2). Interestingly, low-density infiltrate in the cortex was associated with high risk (region 1 in Top). (Bottom) In TCGA-TM-A84G, SCNN assigns high risks to edematous regions (region 1 in Bottom) that are adjacent to tumor (region 1 in Top).

patient care by identifying patients who can benefit from more aggressive therapeutic regimens and by sparing those with less aggressive disease from unnecessary treatment.

Remarkably, SCNN performed as well as manual histologic grading or molecular subtyping in predicting overall survival in our dataset, despite using only a very small portion of each histology image for training and prediction. Additional investigation

of the associations between SCNN risk predictions, molecular subtypes, and histologic grades revealed that SCNN can effectively discriminate outcomes within each molecular subtype, effectively performing digital histologic grading. Furthermore, SCNN can effectively recognize histologic differences associated with *IDH* mutations in astrocytomas and predict outcomes for these patients accordingly. SCNNs correctly predicted lower

risks for WHO grade III *IDH* mutant astrocytomas compared with WHO grade III *IDH* WT astrocytomas, consistent with the considerably longer median survival for patients with *IDH* mutant astrocytoma (6.3 vs. 1.7 y). While there are histologic features of astrocytomas that are understood to be more prevalent in *IDH* mutant astrocytomas, including the presence of microcysts and the rounded nuclear morphology of neoplastic nuclei, these are not reliable predictors of *IDH* mutations (40).

To integrate genomic information in prognostication, we developed a hybrid network that can learn simultaneously from both histology images and genomic biomarkers. The GSCNN presented in our study significantly outperforms the WHO standard based on identical inputs. We compared the performance of GSCNN and SCNN in several ways to evaluate their ability to predict survival and to assess the relative importance of histology and genomic data in GSCNN. GSCNN had significantly higher *c* index scores due to the inclusion of genomic variables in the training process. Performance significantly declined when using a superficial integration method that combines genomic biomarkers with a pretrained SCNN model.

In multivariable regression analyses, GSCNN has a much higher hazard ratio than SCNN (8.83 vs. 3.05). Examining the other variables in the regression models, we noticed an interesting relationship between the significance of histologic-grade and molecular subtype variables. In the SCNN regression analysis, histologic-grade variables were not significant, but molecular subtype variables were highly significant, indicating that SCNN could capture histologic information from image data but could not learn molecular subtype information entirely from histology. In contrast, molecular subtype information was not significant in the GSCNN regression analysis. Interestingly, histologic-grade variables were marginally significant, suggesting that some prognostic value in the histology images remained untapped by GSCNN.

Kaplan–Meier analysis showed remarkable similarity in the discriminative power of SCNN and GSCNN. Additional Kaplan–Meier analysis of risk categories within molecular subtypes revealed interesting trends that are consistent with the regression analyses presented in Table 1. SCNN clearly separates outcomes within each molecular subtype based on histology. Survival curves for GSCNN risk categories, however, overlap significantly in each subtype. Since SCNN models do not have access to genomic data when making predictions, their ability to discriminate outcomes was worse in general when assessed by *c* index or multivariable regression.

Integration of genomic and histology data into a single prediction framework remains a challenge in the clinical implementation of computational pathology. Our previous work in developing deep learning survival models from genomic data has shown that accurate survival predictions can be learned from high-dimensional genomic and protein expression signatures (29). Incorporating additional genomic variables into GSCNN models is an area for future research and requires larger datasets that combine histology images with rich genomic and clinical annotations.

While deep learning methods frequently deliver outstanding performance, the interpretability of black box deep learning models is limited and remains a significant barrier in their validation and adoption. Heat map analysis provides insights into the histologic patterns associated with increased risk and can also serve as a practical tool to guide pathologists to tissue regions associated with worse prognosis. The heat maps suggest that SCNN can learn visual patterns known to be associated with histologic features related to prognosis and used in grading, including microvascular proliferation, cell density, and nuclear morphology. Microvascular prominence and proliferation are associated with disease progression in all forms of diffuse glioma, and these features are clearly delineated as high risk in the heat map presented for slide TCGA-DB-5273. Likewise, increases in

cell density and nuclear pleomorphism were also associated with increased risk in all examples. SCNN also assigned high risks to regions that do not contain well-recognized features associated with a higher grade or poor prognosis. In region 1 of slide TCGA-S9-A7J0, SCNN assigns higher risk to sparsely infiltrated cerebral cortex than to region 2, which is infiltrated by a higher density of tumor cells (normal cortex in region 1 is properly assigned a very low risk). Widespread infiltration into distant sites of the brain is a hallmark of gliomas and results in treatment failure, since surgical resection of visible tumor often leaves residual neoplastic infiltrates. Similarly, region 1 of slide TCGA-TM-A84G illustrates a high risk associated with low-cellularity edematous regions compared with adjacent oligodendroglioma with much higher cellularity. Edema is frequently observed within gliomas and in adjacent brain, and its degree may be related to the rate of growth (43), but its histologic presence has not been previously recognized as a feature of aggressive behavior or incorporated into grading paradigms. While it is not entirely clear why SCNN assigns higher risks to the regions in the sparsely infiltrated or edematous regions, these examples confirm that SCNN risks are not purely a function of cellular density or nuclear atypia. Our human interpretations of these findings provide possible explanations for why SCNN unexpectedly predicts high risks in these regions, but these findings need additional investigation to better understand what specific features the SCNN network perceives in these regions. Nevertheless, this shows that SCNN can be used to identify potentially practice-changing features associated with increased risk that are embedded within pathology images.

Although our study provides insights into the application of deep learning in precision medicine, it has some important limitations. A relatively small portion of each slide was used for training and prediction, and the selection of ROIs within each slide required expert guidance. Future studies will explore more advanced methods for automatic selection of regions and for incorporating a higher proportion of each slide in training and prediction to better account for intratumoral heterogeneity. We also plan to pursue the development of enhanced GSCNN models that incorporate additional molecular features and to evaluate the value added of histology in these more complex models. In our Kaplan–Meier analysis, the thresholds used to define risk categories were determined in a subjective manner using the proportion of manual histologic grades in the TCGA cohort, and a larger dataset would permit a more rigorous definition of these thresholds to optimize survival stratification. The interpretation of risk heat maps was based on subjective evaluation by neuropathologists, and we plan to pursue studies that evaluate heat maps in a more objective manner to discover and validate histologic features associated with poor outcomes. Finally, while we have applied our techniques to gliomas, validation of these approaches in other diseases is needed and could provide additional insights. In fact, our methods are not specific to histology imaging or cancer applications and could be adapted to other medical imaging modalities and biomedical applications.

Methods

Data and Image Curation. Whole-slide images and clinical and genomic data were obtained from TCGA via the Genomic Data Commons (<https://gdc.cancer.gov/>). Images of diagnostic H&E-stained, formalin-fixed, paraffin-embedded sections from the Brain LGG and the GBM cohorts were reviewed to remove images containing tissue-processing artifacts, including bubbles, section folds, pen markings, and poor staining. Representative ROIs containing primarily tumor nuclei were manually identified for each slide that passed a quality control review. This review identified whole-slide images with poor image quality arising from imaging artifacts or tissue processing (bubbles, significant tissue section folds, overstaining, understaining) where suitable ROIs could not be selected. In the case of grade IV disease, some regions include microvascular proliferation, as this feature was exhibited throughout tumor regions. Regions containing geographic necrosis

were excluded. A total of 1,061 whole-slide images from 769 unique patients were analyzed.

ROI images ($1,024 \times 1,024$ pixels) were cropped at $20\times$ objective magnification using OpenSlide and color-normalized to a gold standard H&E calibration image to improve consistency of color characteristics across slides. HPFs at 256×256 pixels were sampled from these regions and used for training and testing as described below.

Network Architecture and Training Procedures. The SCNN combines elements of the 19-layer Visual Geometry Group (VGG) convolutional network architecture with a Cox proportional hazards model to predict time-to-event data from images (Fig. S1) (44). Image feature extraction is achieved by four groups of convolutional layers. (i) The first group contains two convolutional layers with $64 \ 3 \times 3$ kernels interleaved with local normalization layers and then followed with a single maximum pooling layer. (ii) The second group contains two convolutional layers ($128 \ 3 \times 3$ kernels) interleaved with two local normalization layers followed by a single maximum pooling layer. (iii) The third group interleaves four convolutional layers ($256 \ 3 \times 3$ kernels) with four local normalization layers followed by a single maximum pooling layer. (iv) The fourth group contains interleaves of eight convolutional ($512 \ 3 \times 3$ kernels) and eight local normalization layers, with an intermediate pooling layer and a terminal maximum pooling layer. These four groups are followed by a sequence of three fully connected layers containing 1,000, 1,000, and 256 nodes, respectively.

The terminal fully connected layer outputs a prediction of risk $R = \beta^T X$ associated with the input image, where $\beta \in \mathbb{R}^{256 \times 1}$ are the terminal layer weights and $X \in \mathbb{R}^{256 \times 1}$ are the inputs to this layer. To provide an error signal for backpropagation, these risks are input to a Cox proportional hazards layer to calculate the negative partial log likelihood:

$$L(\beta, X) = - \sum_{i \in U} \left(\beta^T X_i - \log \sum_{j \in \Omega_i} e^{\beta^T X_j} \right), \quad [1]$$

where $\beta^T X_i$ is the risk associated with HPF i , U is the set of right-censored samples, and Ω_i is the set of "at-risk" samples with event or follow-up times $\Omega_i = \{j | Y_j \geq Y_i\}$ (where Y_i is the event or last follow-up time of patient i).

The adagrad algorithm was used to minimize the negative partial log likelihood via backpropagation to optimize model weights, biases, and convolutional kernels (45). Parameters to adagrad include the initial accumulator value = 0.1, initial learning rate = 0.001, and an exponential learning rate decay factor = 0.1. Model weights were initialized using the variance scaling method (46), and a weight decay was applied to the fully connected layers during training (decay rate = $4e-4$). Models were trained for 100 epochs (1 epoch is one complete cycle through all training samples) using minibatches consisting of 14 HPFs each. Each minibatch produces a model update, resulting in multiple updates per epoch. Calculation of the Cox partial likelihood requires access to the predicted risks of all samples, which are not available within any single minibatch, and therefore, Cox likelihood was calculated locally within each minibatch to perform updates (U and Ω_i were restricted to samples within each minibatch). Local likelihood calculation can be very sensitive to how samples are assigned to minibatches, and therefore, we randomize the minibatch sample assignments at the beginning of each epoch to improve robustness. Mild regularization was applied during training by randomly dropping out 5% of weights in the last fully connected layer in each minibatch during training to mitigate overfitting.

Training Sampling. Each patient has possibly multiple slides and multiple regions within each slide that can be used to sample HPFs. During training, a single HPF was sampled from each region, and these HPFs were treated as semi-independent training samples. Each HPF was paired with patient outcome for training, duplicating outcomes for patients containing multiple regions/HPFs. The HPFs are sampled at the beginning of each training epoch to generate an entirely new set of HPFs. Randomized transforms were also applied to these HPFs to improve robustness to tissue orientation and color variations. Since the visual patterns in tissues can often be anisotropic, we randomly apply a mirror transform to each HPF. We also generate random transformations of contrast and brightness using the "random_contrast" and "random_brightness" TensorFlow operations. The contrast factor was randomly selected in the interval [0.2, 1.8], and the brightness was randomly selected in the interval [-63, 63]. These sampling and transformation procedures along with the use of multiple HPFs for each patient have the effect of augmenting the effective size of the labeled training data. In tissues with pronounced anisotropy, including adenocarcinomas that exhibit prominent glandular structures, these mirror transformations are intended to improve

the robustness of the network to tissue orientation. Similar approaches for training data augmentation have shown considerable improvements in general imaging applications (33).

Testing Sampling, Risk Filtering, and Model Averaging. Sampling was also performed to increase the robustness and stability of predictions. (i) Nine HPFs are first sampled from each region j corresponding to patient m . (ii) The risk of the k th HPF in region j for patient m , denoted $R_m^{j,k}$, is then calculated using the trained SCNN model. (iii) The median risk $R_m^j = \text{median}_k \{R_m^{j,k}\}$ is calculated for region j using the aforementioned HPFs to reject outlying risks. (iv) These median risks are then sorted from highest to lowest $\widehat{R}_m^1 > \widehat{R}_m^2 > \widehat{R}_m^3 \dots$, where the superscript index now corresponds to the risk rank. (v) The risk prediction for patient m is then selected as the second highest risk $R_m^* = \widehat{R}_m^2$. This filtering procedure was designed to emulate how a pathologist integrates information from multiple areas within a slide, determining prognosis based on the region associated with the worst prognosis. Selection of the second highest risk (as opposed to the highest risk) introduces robustness to outliers or high risks that may occur due to some imaging or tissue-processing artifact.

Since the accuracy of our models can vary significantly from one epoch to another, largely due to the training sampling and randomized minibatch assignments, a model-averaging technique was used to reduce prediction variance. To obtain final risk predictions for the testing patients that are stable, we perform model averaging using the models from epochs 96 to 100 to smooth variations across epochs and increase stability. Formally, the model-averaged risk for patient m is calculated as

$$\overline{R}_m^* = \frac{1}{5} \sum_{\gamma=96}^{100} R_{m(\gamma)}^*, \quad [2]$$

where $R_{m(\gamma)}^*$ denotes the predicted risk for patient m in training epoch γ .

Validation Procedures. Patients were randomly assigned to nonoverlapping training (80%) and test (20%) sets that were used to train models and evaluate their performance. If a patient was assigned to training, then all slides corresponding to that patient were assigned to the training set and likewise, for the testing set. This ensures that no data from any one patient are represented in both training and testing sets to avoid overfitting and optimistic estimates of generalization accuracy. We repeated the randomized assignment of patients training/testing sets 15 times and used each of these training/testing sets to train and evaluate a model. The same training/testing assignments were used in each model (SCNN, GSCNN, baseline) for comparability. Prediction accuracy was measured using Harrell's c index to measure the concordance between predicted risk and actual survival for testing samples (36).

Statistical Analyses. The c indices generated by Monte Carlo cross-validation were performed using the Wilcoxon signed rank test. This paired test was chosen, because each method was evaluated using identical training/testing sets. Comparisons of SCNN risk values across grade were performed using the Wilcoxon rank sum test. Cox univariable and multivariable regression analyses were performed using predicted SCNN risk values for all training and testing samples in randomized training/testing set 1. Analyses of the correlation of grade, molecular subtype, and SCNN risk predictions were performed by pooling predicted risks for testing samples across all experiments. SCNN risks were normalized within each experiment by z score before pooling. Grade analysis was performed by determining "digital"-grade thresholds for SCNN risks in each subtype. Thresholds were objectively selected to match the proportions of samples in each histologic grade in each subtype. Statistical analysis of Kaplan-Meier plots was performed using the log rank test.

Hardware and Software. Prediction models were trained using TensorFlow (v0.12.0) on servers equipped with dual Intel(R) Xeon(R) CPU E5-2630L v2 @ 2.40 GHz CPUs, 128 GB RAM, and dual NVIDIA K80 graphics cards. Image data were extracted from Aperio .svs whole-slide image formats using OpenSlide (openslide.org). Basic image analysis operations were performed using HistomicsTK (<https://github.com/DigitalSlideArchive/HistomicsTK>), a Python package for histology image analysis.

Data Availability. This paper was produced using large volumes of publicly available genomic and imaging data. The authors have made every effort to make available links to these resources as well as make publicly available the software methods and information used to produce the datasets, analyses, and summary information.

ACKNOWLEDGMENTS. This work was supported by US NIH National Library of Medicine Career Development Award K22LM011576 and Na-

tional Cancer Institute Grant U24CA194362 and by the National Brain Tumor Society.

1. Kong J, et al. (2008) Computer-assisted grading of neuroblastic differentiation. *Arch Pathol Lab Med* 132:903–904, author reply 904.
2. Niazi MKK, et al. (2017) Visually meaningful histopathological features for automatic grading of prostate cancer. *IEEE J Biomed Health Inform* 21:1027–1038.
3. Naik S, et al. (2008) Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. *Proceedings of the 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (IEEE, Piscataway, NJ), pp 284–287.
4. Ren J, et al. (2015) Computer aided analysis of prostate histopathology images Gleason grading especially for Gleason score 7. *Conf Proc IEEE Eng Med Biol Soc 2015*: 3013–3016.
5. Kothari S, Phan JH, Young AN, Wang MD (2013) Histological image classification using biologically interpretable shape-based features. *BMC Med Imaging* 13:9.
6. Sertel O, et al. (2009) Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern Recognit* 42:1093–1103.
7. Fauzi MF, et al. (2015) Classification of follicular lymphoma: the effect of computer aid on pathologists grading. *BMC Med Inform Decis Mak* 15:115.
8. Dundar MM, et al. (2011) Computerized classification of intraductal breast lesions using histopathological images. *IEEE Trans Biomed Eng* 58:1977–1984.
9. Hou L, et al. (2016) Patch-based convolutional neural network for whole slide tissue image classification. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ), pp 2424–2433.
10. Kong J, et al. (2013) Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. *PLoS One* 8:e81049.
11. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH (2016) Deep learning for identifying metastatic breast cancer. arXiv:1606.05718.
12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
13. Greenspan H, van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Trans Med Imaging* 35:1153–1159.
14. Janowczyk A, Madabhushi A (2016) Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J Pathol Inform* 7:29.
15. Litjens G, et al. (2016) Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep* 6:26286.
16. Chen T, Chef'd'hotel C (2014) Deep learning based automatic immune cell detection for immunohistochemistry images. *Machine Learning in Medical Imaging* (Springer, Berlin), pp 17–24.
17. Cruz-Roa A, et al. (2017) Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci Rep* 7:46450.
18. Pereira S, Pinto A, Alves V, Silva CA (2016) Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging* 35:1240–1251.
19. Sirinukunwattana K, et al. (2016) Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans Med Imaging* 35:1196–1206.
20. Esteva A, et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115–118.
21. Gulshan V, et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316: 2402–2410.
22. Havaei M, et al. (2017) Brain tumor segmentation with deep neural networks. *Med Image Anal* 35:18–31.
23. Huynh BQ, Li H, Giger ML (2016) Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging (Bellingham)* 3:034501.
24. Kamnitsas K, et al. (2017) Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* 36:61–78.
25. Turkki R, Linder N, Kovanen PE, Pellinen T, Lundin J (2016) Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *J Pathol Inform* 7:38.
26. Bychkov D, Turkki R, Haglund C, Linder N, Lundin J (2016) Deep learning for tissue microarray image-based outcome prediction in patients with colorectal cancer. *SPIE Medical Imaging*, eds Gurcan MN, Madabhushi A (International Society for Optics and Photonics, Bellingham, WA), p 6.
27. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2014) Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 13:8–17.
28. Xiang A, Lapuerta P, Ryutov A, Buckley J, Azen S (2000) Comparison of the performance of neural network methods and Cox regression for censored survival data. *Comput Stat Data Anal* 34:243–257.
29. Yousefi S, et al. (2017) Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep* 7:11707.
30. Yousefi S, Congzheng S, Nelson N, Cooper LAD (2016) Learning genomic representations to predict clinical outcomes in cancer. arXiv:1609.08663.
31. Katzman J, et al. (2016) DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. arXiv:1606.00931.
32. Zhu X, Yao J, Huang J (2016) Deep convolutional neural network for survival analysis with pathological images. *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine* (IEEE, Piscataway, NJ), pp 544–547.
33. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, eds Pereira F, Burges CJC, Bottou L, Weinberger KQ (Neural Information Processing Systems Foundation, Inc., La Jolla, CA), pp 1097–1105.
34. Gutman DA, et al. (2013) Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc* 20:1091–1098.
35. Gutman DA, et al. (2017) The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research. *Cancer Res* 77: e75–e78.
36. Harrell FE, Jr, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. *JAMA* 247:2543–2546.
37. Brat DJ, et al.; Cancer Genome Atlas Research Network (2015) Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med* 372: 2481–2498.
38. Reuss DE, et al. (2015) IDH mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for WHO. *Acta Neuropathol* 129:867–873.
39. Leeper HE, et al. (2015) IDH mutation, 1p19q codeletion and ATRX loss in WHO grade II gliomas. *Oncotarget* 6:30295–30305.
40. Nguyen DN, et al. (2013) Molecular and morphologic correlates of the alternative lengthening of telomeres phenotype in high-grade astrocytomas. *Brain Pathol* 23: 237–243.
41. Wijnenga MMJ, et al. (2018) The impact of surgery in molecularly defined low-grade glioma: an integrated clinical, radiological, and molecular analysis. *Neuro-oncol* 20: 103–112.
42. van den Bent MJ (2010) Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol* 120:297–304.
43. Pope WB, et al. (2005) MR imaging correlates of survival in patients with high-grade gliomas. *AJNR Am J Neuroradiol* 26:2466–2474.
44. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
45. Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res* 12:2121–2159.
46. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision* (IEEE, Piscataway, NJ), pp 1026–1034.

3.4 Multi-faceted computational assessment of risk and progression in oligodendroglioma implicates NOTCH and PI3K pathways, NPJ Precision Oncology, 2018

This section is an exact copy of the following open-access journal paper:

Sameer H Halani, **Safoora Yousefi**, Jose Velazquez Vega, Michael R Rossi, Zheng Zhao, Fatemeh Amrollahi, Chad A Holder, Amelia Baxter-Stoltzfus, Jennifer Eschbacher, Brent Griffith, Jeffrey J Olson, Tao Jiang, Joseph R Yates, Charles G Eberhart, Laila M Poisson, Lee AD Cooper, Daniel J Brat. *Multi-faceted computational assessment of risk and progression in oligodendroglioma implicates NOTCH and PI3K pathways*. NPJ precision oncology. 2018 Nov 6;2(1):24.

Abstract. Oligodendrogliomas are diffusely infiltrative gliomas defined by IDH-mutation and co-deletion of 1p/19q. They have highly variable clinical courses, with survivals ranging from 6 months to over 20 years, but little is known regarding the pathways involved with their progression or optimal markers for stratifying risk. We utilized machine-learning approaches with genomic data from The Cancer Genome Atlas to objectively identify molecular factors associated with clinical outcomes of oligodendroglioma and extended these findings to study signaling pathways implicated in oncogenesis and clinical endpoints associated with glioma progression. Our multi-faceted computational approach uncovered key genetic alterations associated with disease progression and shorter survival in oligodendroglioma and specifically identified Notch pathway inactivation and PI3K pathway activation as the most strongly associated with MRI and pathology findings of advanced disease and poor clinical outcome. Our findings that Notch pathway inactivation and PI3K pathway activation are associated with advanced disease and survival risk will pave the way for clinically relevant markers of disease progression and therapeutic targets to improve clinical outcomes. Furthermore, our approach demonstrates the strength of machine learning and computational methods for identifying genetic events critical to disease progression in the era of big data and precision medicine.

Candidate's contribution. Training and interpretation of survival analysis neural networks on the TCGA glioma subtypes including Oligodendrogliomas.

ARTICLE OPEN

Multi-faceted computational assessment of risk and progression in oligodendroglioma implicates NOTCH and PI3K pathways

Sameer H. Halani¹, Safoora Yousefi², Jose Velazquez Vega³, Michael R. Rossi³, Zheng Zhao⁴, Fatemeh Amrollahi², Chad A. Holder⁵, Amelia Baxter-Stoltzfus¹, Jennifer Eschbacher⁶, Brent Griffith^{7,8}, Jeffrey J. Olson^{1,9,10}, Tao Jiang⁴, Joseph R. Yates¹¹, Charles G. Eberhart¹¹, Laila M. Poisson^{8,12}, Lee A. D. Cooper^{1,2,10,13} and Daniel J. Brat¹⁴

Oligodendrogliomas are diffusely infiltrative gliomas defined by *IDH*-mutation and co-deletion of 1p/19q. They have highly variable clinical courses, with survivals ranging from 6 months to over 20 years, but little is known regarding the pathways involved with their progression or optimal markers for stratifying risk. We utilized machine-learning approaches with genomic data from The Cancer Genome Atlas to objectively identify molecular factors associated with clinical outcomes of oligodendroglioma and extended these findings to study signaling pathways implicated in oncogenesis and clinical endpoints associated with glioma progression. Our multi-faceted computational approach uncovered key genetic alterations associated with disease progression and shorter survival in oligodendroglioma and specifically identified Notch pathway inactivation and PI3K pathway activation as the most strongly associated with MRI and pathology findings of advanced disease and poor clinical outcome. Our findings that Notch pathway inactivation and PI3K pathway activation are associated with advanced disease and survival risk will pave the way for clinically relevant markers of disease progression and therapeutic targets to improve clinical outcomes. Furthermore, our approach demonstrates the strength of machine learning and computational methods for identifying genetic events critical to disease progression in the era of big data and precision medicine.

npj Precision Oncology (2018)2:24; doi:10.1038/s41698-018-0067-9

INTRODUCTION

Oligodendrogliomas are diffuse gliomas characterized by *IDH*-mutation, co-deletion of 1p/19q and *TERT* promoter mutations. They have the least aggressive clinical course among this group, yet display widely variable outcomes—some patients survive 6 months while others live over 20 years.^{1–5} Aside from their defining genetic alterations, oligodendrogliomas also harbor other mutations, including: capicua transcriptional repressor (*CIC*) (62%), far upstream element binding protein 1 (*FUBP1*) (27–29%), *NOTCH1* (18–31%), catalytic and regulatory subunits of phosphoinositide-3-kinase (PI3K; *PIK3CA* (15–20%) and *PIK3R1* (7–9%), respectively), and others.^{1,6,7} Now that lower-grade gliomas are understood in objective, molecular terms, markers of progression and targets of therapy are being evaluated in a pure cohort, without the confounding contamination of dissimilar tumor types. Recent investigations by Aoki et al.⁸ for example, indicated that *NOTCH1* mutations were associated with poor clinical outcomes in patients with oligodendroglioma.

With the tremendous expansion of genomic data available for both investigation and potential clinical care, a need has

developed for novel computational approaches to investigate risk factors in a highly multidimensional and interdependent space.⁹ Machine-learning approaches are capable of using large genomic datasets in a manner that adds value to traditional risk modeling by identifying key prognostic factors among tens of thousands of possible variables. We employed machine-learning to identify molecular factors associated with clinical outcomes of oligodendroglioma using The Cancer Genome Atlas (TCGA) LGG dataset. We advanced and translated these findings using neuroimaging and pathology imaging features of progression to identify molecular biomarkers most closely related to advanced disease status, as defined by: (1) contrast-enhancement on magnetic resonance imaging (MRI); (2) high cellular density in digitized histopathologic images; and (3) increased cellular proliferation.^{10–12} In addition, our approach enabled us to identify key signaling pathways associated with more aggressive disease in addition to individual biomarkers. Our approach confirmed the association of *NOTCH1* mutations with disease progression and shorter survival in oligodendroglioma, and further uncovered

¹Emory University School of Medicine, Atlanta, GA, USA; ²Department of Biomedical Informatics, Emory University, Atlanta, GA, USA; ³Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA, USA; ⁴Department of Neurosurgery, Tiantan Hospital, Capital Medical University, Beijing, China; ⁵Department of Radiology, Emory University, Atlanta, GA, USA; ⁶Department of Neuropathology, Barrow Neurological Institute, Phoenix, AZ, USA; ⁷Department of Radiology, Henry Ford Health System, Detroit, MI, USA; ⁸Josephine Ford Cancer Institute, Henry Ford Health System, Detroit, MI, USA; ⁹Department of Neurosurgery, Emory University, Atlanta, GA, USA; ¹⁰Winship Cancer Institute, Emory University, Atlanta, GA, USA; ¹¹Divisions of Pathology, Ophthalmology, and Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; ¹²Department of Public Health Sciences, Henry Ford Hospital Systems, Detroit, MI, USA; ¹³Department of Biomedical Engineering, Emory University/Georgia Institute of Technology, Atlanta, GA, USA and ¹⁴Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Correspondence: Lee A. D. Cooper (lee.cooper@emory.edu) or Daniel J. Brat (daniel.bratt@northwestern.edu)

These authors contributed equally: Lee A.D. Cooper, Daniel J. Brat

Received: 25 February 2018 Revised: 18 September 2018 Accepted: 24 September 2018

Published online: 06 November 2018

aberrant regulation of Notch and PI3K pathways as most strongly associated with advanced disease.

RESULTS

Patient and tumor characteristics

The clinical factors from the 169 oligodendroglioma patients included in our study are presented in Table 1. *TERT* promoter mutations were present in 98% (86 of 88).¹³

Neural network analyses identifies molecular factors associated with outcomes

Analysis of the genetic-protein neural network model revealed multiple mutations, CNAs, and proteins associated with overall survival in oligodendrogliomas (see Fig. 1a). *NOTCH1* (rank #5), *BCOR* (rank #4), and *ZBTB20* (rank #1) mutations were among the most highly ranked factors associated with poor prognosis, along with loss of 15q (rank #3). Both *NOTCH1* mutations and 15q loss occur in a substantial subset of oligodendrogliomas and have previously been suggested as markers of poor prognosis in traditional risk models,¹⁴ providing support for our model. The complete list of ranked factors is in the Supplementary Materials (Data file S1). Among these factors, we focused on the Notch pathway since *NOTCH1* mutations are relatively specific to

oligodendroglioma among diffuse gliomas; occur in a substantial subset (18–31%) compared to *BCOR* and *ZBTB20*; and represent one component of the Notch signaling network that could be more generally relevant to disease progression. PI3K pathway subunit mutations were also of interest since they were heavily enriched among highly ranked negative prognostic factors (*PIK3R1*, #30; *PIK3CA*, #193).

Similar analysis of the gene expression neural network model was performed to determine the prognostic importance of mRNA transcripts, and a gene-set-enrichment analysis (GSEA) was then used to identify molecular pathways enriched with prognostic transcripts. GSEA identified the NOTCH1 Intracellular Domain Regulates Transcription pathway ($P = 0.004$) as highly enriched in transcripts associated with better prognosis, suggesting that Notch pathway inactivation is associated with poor outcomes (Fig. 1b). Regulation of KIT Signaling was also significantly enriched with positive prognosis transcripts ($P = 0.002$). The P38 / MKK3 ($P < 0.05$) and SMAD2 / SMAD3 pathways ($P = 0.002$) were also significantly enriched in transcripts associated with a poor prognosis, and notably, both interface directly with the PI3K pathway.^{15,16}

The results of Monte-Carlo cross validation of the genetic-protein and gene expression survival neural networks are presented in Supplementary Figure S1. The median c-index of the tested genetic-protein models was 0.8 (± 0.124), while the median c-index of the tested gene expression models was 0.752 (± 0.196).

Table 1. Patient demographics	
Characteristic	Total (N = 169)
Original histologic diagnosis (WHO 2007)—no. (%)	
Oligodendroglioma	
Grade II	62 (36.7)
Grade III	55 (32.5)
Oligoastrocytoma	
Grade II	17 (10.1)
Grade III	13 (7.7)
Astrocytoma	
Grade II	2 (1.2)
Grade III	2 (1.2)
Age at diagnosis (yrs)	
Mean \pm SD	45.8 \pm 12.8
Range	17–75
Male sex—no. (%)	
	84 (49.7)
White race—no./total no. (%)	
	155/164 (94.5)
Extent of resection—no./total no. (%)	
Open biopsy	1/164 (0.6)
Subtotal resection	59/164 (36.0)
Gross total resection	104/164 (63.4)
Tumor location—no./total no. (%)	
Frontal lobe	122/166 (73.5)
Occipital lobe	3/166 (1.8)
Parietal lobe	14/166 (8.4)
Temporal lobe	27/166 (16.3)
Laterality—no./total no. (%)	
Left	79/168 (47.0)
Midline	3/168 (1.8)
Right	86/168 (51.2)
Clinical characteristics of patients from The Cancer Genome Atlas database with confirmed diagnosis of oligodendroglioma (i.e., <i>IDH</i> -mutant, 1p19q co-deleted glioma).	

Radiographic and pathologic features are associated with aggressive clinical behavior

We next focused on mutations and CNAs with a > 5% incidence to assess their association with radiographic and pathologic measures of disease progression, including: mutations of *CIC* (ranked #107; 61.5% incidence) *NOTCH1* (ranked #5; 18.9%), *FUBP1* (ranked #20; 27.2%), both *PIK3* subunits (*PIK3R1* ranked #30 and *PIK3CA* ranked #193; 23.1%), and CNA's including gain of chromosomal arms 7p (ranked #300; 8.9%) and 11p (ranked #153; 11.2%), as well as loss of 14q (ranked #310; 11.8%) and 15q (ranked #3; 16.6%) (Fig. S2 illustrates a waterfall plot of the most frequent genetic alterations; Table S1).

Contrast-enhancement observed on MRI is a well-known marker of higher-grade disease (Fig. 2a). Among 55 patients with MRI images available, contrast-enhancing (CE+) tumors ($n = 35$) had worse overall survival (OS) (median, 154.3 vs. 62.0 months; $P = 0.10$) and progression-free survival (PFS) (median, 97.3 vs. 63.8 months; $P = 0.029$) compared to those lacking enhancement (CE-) ($n = 20$) (Figs. 2b, c). CE+ was highly enriched for histologic grade III tumors; 24 of 25 grade III tumors were CE+ ($P < 0.0001$).

Since cell density increases with disease progression, we used a computational nearest-neighbor analysis to quantify cellular density in tissue sections from 142 cases (Fig. 2d). Higher cell density trended towards worse OS (mean 152.8 vs. 126.1 months; $P = 0.076$) and worse PFS (median 142.8 vs. 95.9 months; $P = 0.14$) (Figs. 2e, f). High cell density cases were also enriched for histologic grade III tumors; 44 of 58 high density tumors were WHO grade III ($P < 0.0001$).

As a measure of proliferation, *MKI67* mRNA expression was analyzed for 169 tumors. *MKI67* expression was strongly correlated with Ki-67/MIB-1 proliferation indices based on immunohistochemistry (IHC) and listed in TCGA pathology reports ($P < 0.0001$) (Figs. 2g, h). Patients with high cellular proliferation ($n = 31$) had worse OS (median 154.3 vs. 62.0 months; $P = 0.001$); no significant difference was noted in PFS ($P = 0.38$) (Figs. 2i, j). Highly proliferative tumors were also enriched for histologic grade III tumors; 21 of 28 high proliferation tumors with grade information available were WHO grade III ($P = 0.001$).

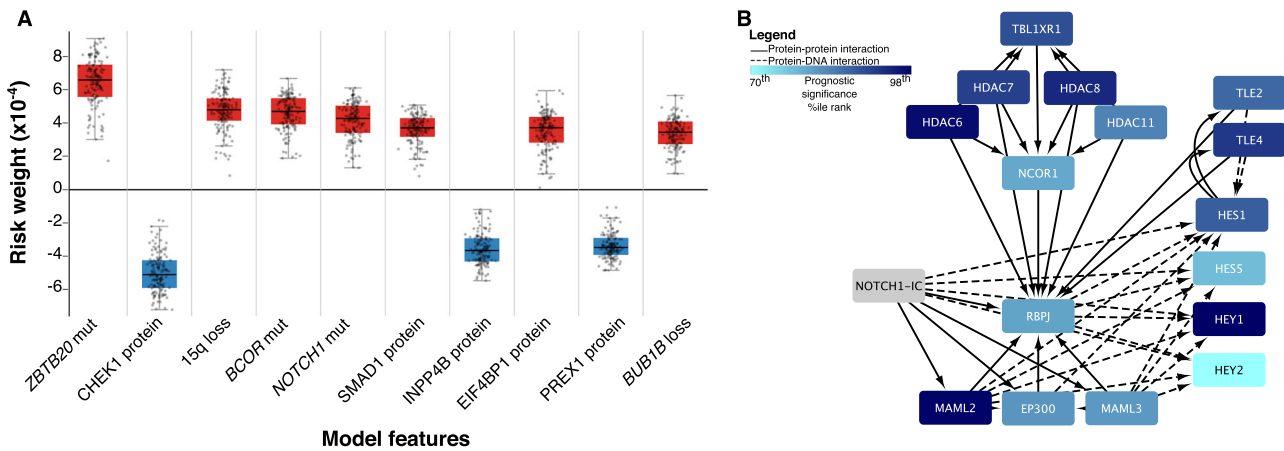


Fig. 1 **a** Neural network risk factors. A nonlinear Cox proportional hazards model was trained using a neural network to model survival in oligodendrogliomas using clinical, genetic and proteomic factors. Prognostic significance of each feature was assessed by determining how its changes impact prognosis. Positive scores indicate a negative impact on survival (red) while negative scores (blue) suggest a positive impact. The boxplot contains the top 10 factors ranked by median prognostic importance; complete results in Datafile S1. **b** Gene set enrichment analysis of Notch pathway members. A separate model based on mRNA expression weighed the prognostic significance of individual transcripts and used this data in a gene-set-enrichment analysis to identify pathways associated with prognosis. The canonical Notch pathway was highly enriched with significantly negatively scored transcripts (i.e., darker blue signifies negative scores). Increased expression of downstream targets, including *HES1*, *HES5*, and *HEY1*, were associated with improved prognosis. This model demonstrates Notch signaling inactivation is associated with poor prognosis

Genetic alterations associated with radiographic contrast enhancement, cellular density, and *MKI67* expression

Among 55 patients with MR imaging (Table S2), *NOTCH1* mutations were most strongly associated with CE+ tumors, with 13 of 14 *NOTCH1* mutants being CE+ ($P=0.008$) (Fig. 3a). The combined *PI3K* group mutants were mostly CE+ (14 of 18; $P=0.054$), and a similar trend was found among *FUBP1* mutants (14 of 17; $P=0.13$). All 9 tumors with 11p gain were CE+ ($P=0.019$). Although 5 of 5 tumors demonstrating loss of 14q were CE+, this did not reach statistical significance ($P=0.15$). Similar trends were found with 15q loss (9 of 10 CE+; $P=0.075$) and 7p gain (6 of 6 CE+; $P=0.076$).

NOTCH1 mutant oligodendrogliomas ($n=26$) had higher cellular density than *NOTCH1* wild-type tumors ($n=126$) and this difference was the most significant among all mutations and CNAs ($P=0.0015$) (Fig. 3b). *FUBP1* mutants ($n=40$) trended toward a higher cellular density compared to wild-type ($n=102$; $P=0.10$), and *CIC* ($n=88$) and *PIK3* ($n=33$) mutants did not show increased cell density (Fig. S3). Gains of 7p ($n=12$) or 11p ($n=17$) were significantly associated with higher cell densities ($P=0.006$ and 0.03, respectively), and loss of 15q ($n=21$) trended towards higher cellular density as well ($P=0.19$) (Fig. 3b).

NOTCH1 mutants ($n=32$) had higher *MKI67* expression and this association was the strongest among all mutations and CNAs tested ($P=0.095$) (Fig. 3c). *FUBP1*, *CIC*, and *PIK3* mutations were not strongly related to *MKI67* expression (Fig. S4). Although gain of 7p and 11p, and loss of 14q and 15q trended towards higher cellular proliferation, none reached statistical significance.

Inactivation of the canonical Notch pathway is associated with disease progression measures

Since *NOTCH1* mutations were consistently and strongly associated with radiologic, pathologic, and molecular measures of progression, we investigated downstream targets of the canonical Notch pathway, including family members of hairy/enhancer of split 1 (*HES*) and hairy/enhancer of split with YRPW motif (*HEY*). Since nearly all (93%) *NOTCH1* mutations were located within the epidermal growth factor (EGF) like region, where they inhibit Notch activation, we hypothesized these targets would be down regulated in *NOTCH1* mutants.^{17,18} Expression of *HES1*, *HEY1*, and *HEY2* was reduced in CE+ tumors, with *HES1* and *HEY2* reaching

statistical significance ($P=0.016$ and 0.050, respectively) (Fig. 4a and Fig. S5). *HEY2* (Pearson correlation = 0.230, $P=0.006$) was positively correlated with nearest-neighbor distance (Fig. 4b) and negatively correlated with cellular proliferation as approximated by *MKI67* expression (Pearson correlation = -0.353 , $P<0.0001$) (Fig. 4c). Negative correlations between *MKI67* expression and *HES1* (Pearson correlation = -0.152 , $P=0.048$) and *HEY1* (Pearson correlation = -0.082 , $P=0.288$) were also observed. Thus, among *HES* and *HEY* family members, *HES1*, *HEY1* and *HEY2* showed reduced expression with advanced disease, with *HEY2* showing the most consistent and statistically significant reductions.

Alternate mechanisms of Notch pathway inactivation in oligodendroglioma

Recombinant signal binding protein for immunoglobulin kappa-J region (*RBPJ*), the nuclear binding partner of activated *NOTCH1*'s intracellular binding domain (NICD), was mutated ($n=5$) or homodeleted ($n=1$) in 3% (6 of 169) of oligodendrogliomas. *RBPJ* aberrations were mutually exclusive with *NOTCH1* mutations and were not present in *IDH* mutant or *IDH* wild-type astrocytomas. *RBPJ* altered tumors had greater *MKI67* expression compared to wild-type ($P=0.001$) and showed a trend toward higher cell density ($P=0.20$), but were not enriched in CE+ tumors (Fig. S6A). When *RBPJ* and *NOTCH1* mutant tumors were grouped ($n=38$), *MKI67* expression and 1/nearest-neighbor distance showed stronger statistical significance in the combined group than in the group with *NOTCH1* mutants alone ($P=0.0030$ and 0.00039 for combined groups, respectively vs. $P=0.095$ and 0.002 for *NOTCH1* mutants alone) (Fig. S6B). Thus, *RBPJ* mutation likely represents an alternative mechanism for Notch pathway inactivation in oligodendroglioma.

Survival analysis reveals *PIK3* mutations and reduced Notch target expression are associated with worse prognosis

A comprehensive analysis of clinical and genetic factors associated with survival was performed using a Cox proportional hazards models (Table 2 and Table S3). Univariable analysis revealed age and grade as strong predictors of poor OS (Hazards ratio (HR) 3.64 per 10 years, $P<0.0001$; HR 6.61, $P=0.013$, respectively). After adjusting for age and grade, the combination of *PIK3* mutations

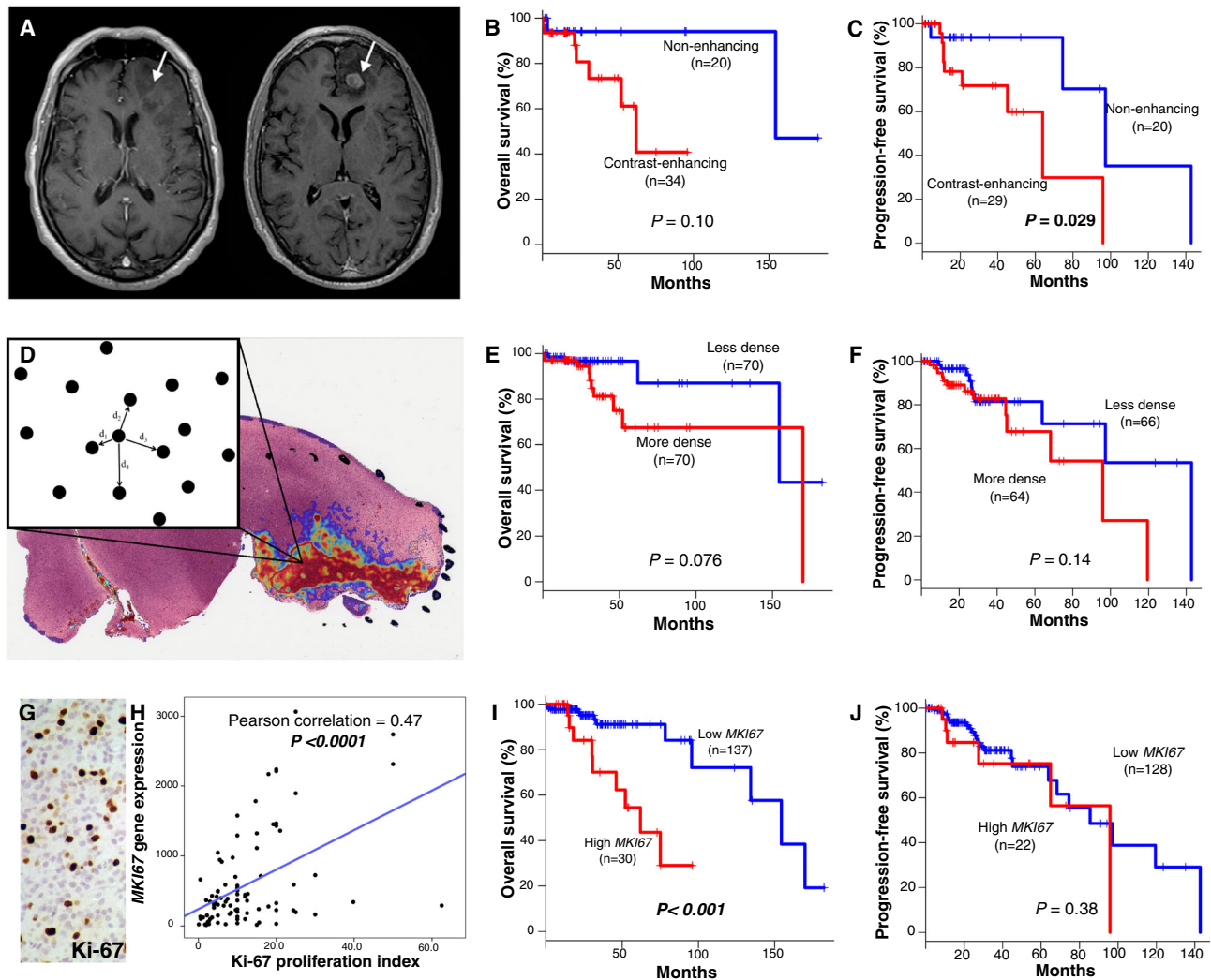


Fig. 2 Markers of disease progression in oligodendroglioma **a** T1-weighted axial MR images with gadolinium contrast demonstrating CE– (left) and CE+ (right) features of oligodendroglioma from The Cancer Imaging Archive. **b** Kaplan–Meier plots of overall survival (OS) for CE– vs. CE+. **c** Progression-free survival (PFS) for CE– vs. CE+. **d** Visual representation of a tumor heatmap showing regions of interest of cell density, with a schematic diagram of the nearest-neighbor algorithm. **e** OS for cellular density (less vs. more dense). **f** PFS for less vs. more dense. **g**. High Ki-67 proliferation index visualized with IHC. **h** Linear regression of *MKI67* expression and Ki-67 proliferation index approximated by IHC. **i** OS for high vs. low *MKI67*. **j** PFS for high vs. low *MKI67*. *P* values for survival plots determined using log-rank tests

were found to confer poor prognosis (HR 3.11, $P = 0.045$). Among the downstream Notch target genes, increased *HES5* expression had a significant protective effect (HR 0.74, $P = 0.024$) after accounting for age and grade.

Univariable analysis of PFS uncovered increased risk with grade III relative to grade II (HR 2.24, $P = 0.046$). *PIK3* (HR 1.98, $P = 0.092$) mutations trended toward increased risk of progression after accounting for tumor grade. Loss of 14q (HR 3.90, $P = 0.0035$) predicted more rapid time to progression after adjusting for grade. While *NOTCH1* mutants were not individually predictive of PFS, when combined with *RBPJ* altered tumors, the combined mutants predicted shorter time to first progression (HR 2.47, $P = 0.021$). After adjusting for grade, reduced *HEY1* (HR 0.48, $P = 0.018$) expression had a negative impact on PFS, while *HES5* trended in this direction (HR 0.86, $P = 0.120$). Complete survival analysis results in Table S3 and Fig. S7–S8.

Translation and validation in clinical cases

We investigated 51 newly diagnosed cases of oligodendroglioma, grades II and III, from hospital archives. Pre-operative imaging was available for 47. We focused our IHC analysis on *HEY2*, since its

gene expression showed greatest reduction in *NOTCH1* mutants, and pAkt, a downstream marker of PI3K activation (Figs. 4d, e).

Thirty-two tumors were WHO grade II and 19 were grade III; 21 tumors were CE– and 26 were CE+. By IHC analysis of *HEY2*, 20 tumors showed low expression and 31 showed high expression. Fourteen of 19 (73.7%) tumors with low *HEY2* were CE+. Tumors with low *HEY2* also had greater cell density ($P = 0.014$) and were more proliferative ($P = 0.0096$) than those with increased *HEY2* staining (Fig. 4f). IHC investigation of pAkt found 27 tumors had low expression; 22 showed high expression; 15 of 20 (75%) tumors with pre-operative imaging and high pAkt expression were CE+. Tumors with high pAkt expression had greater cell density and were more proliferative ($P < 0.0001$, for both) (Fig. 4f).

DISCUSSION

We used a multi-faceted, technologically advanced, computational approach to identify molecular events associated with aggressive disease within molecularly defined oligodendroglioma (*IDH* mutant, 1p/19q co-deleted) and uncovered Notch pathway inactivation and PI3K activation as critical events. Our deep

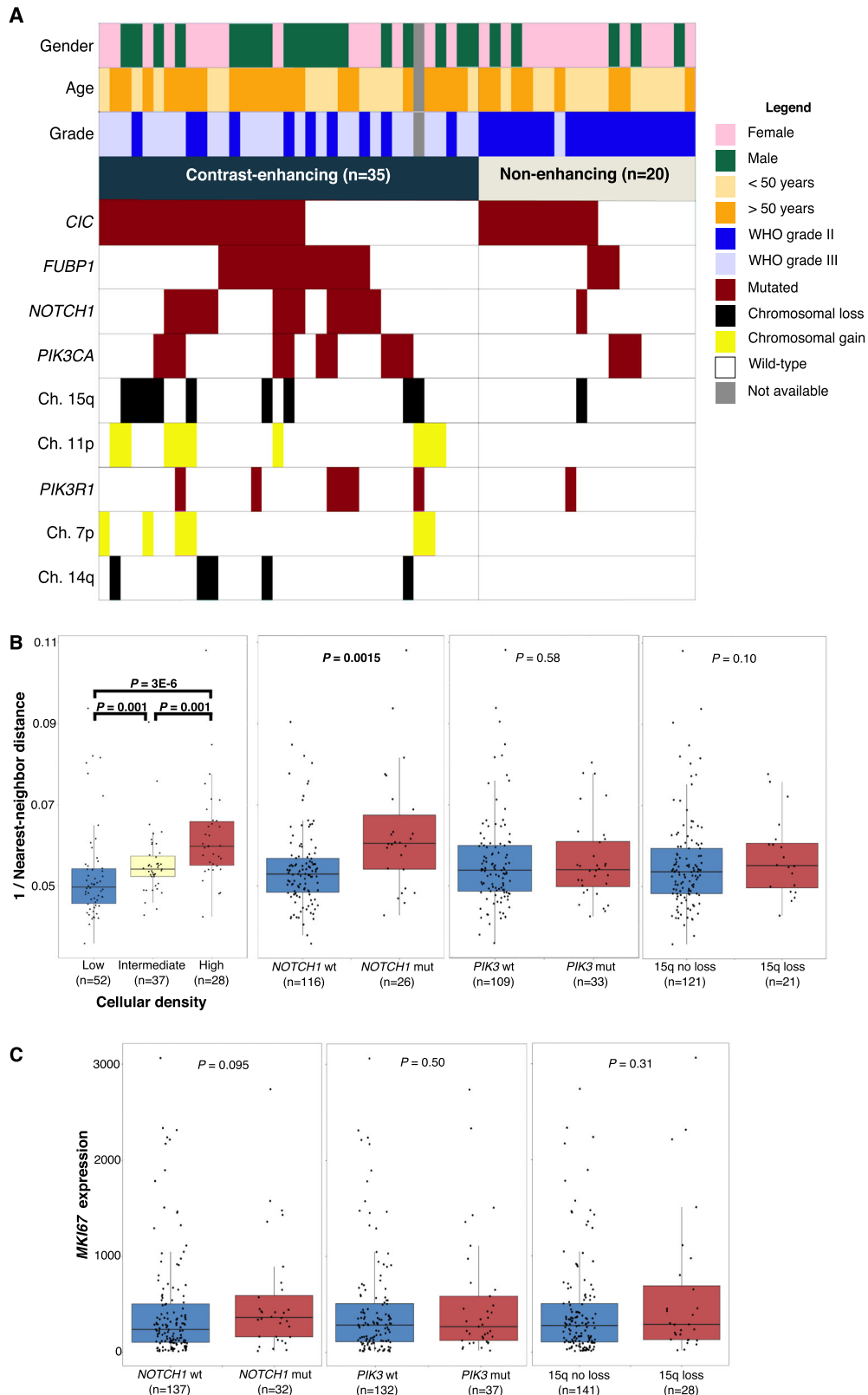


Fig. 3 Genetic alterations associated with advanced disease progression **a** Waterfall plot illustrating the mutational landscape of oligodendrogliomas based on radiographic features of progression. **b** Boxplots demonstrating nearest-neighbor validation, and differential 1/nearest-neighbor distances in key genetic alterations of oligodendroglioma. **c** Boxplots for differential *MKI67* expression in key genetic alterations of oligodendroglioma. *P* values determined using Wilcoxon rank sum tests

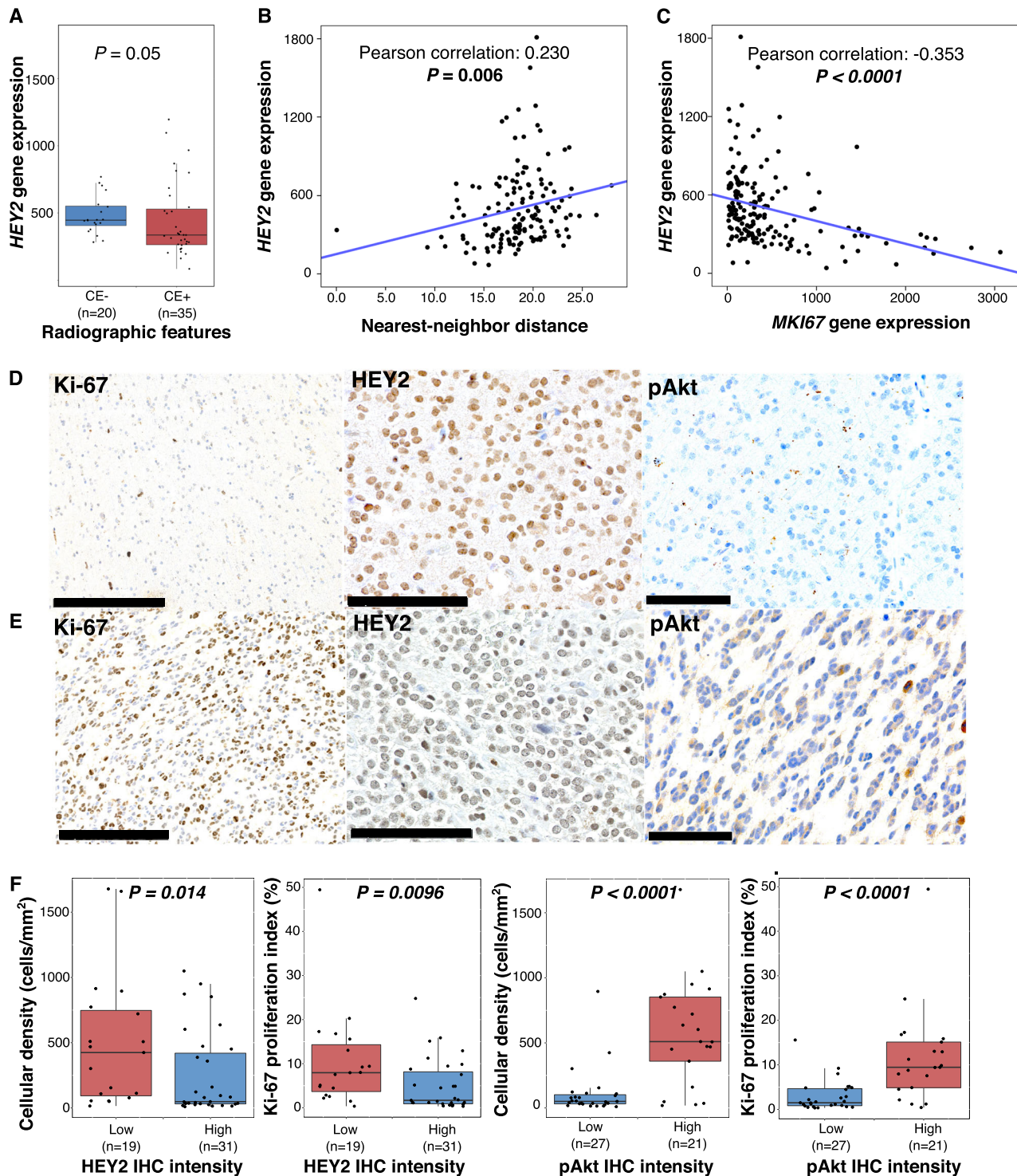


Fig. 4 HEY2 associations with advanced disease and validation cohort. **a** Boxplots demonstrating differential HEY2 gene expression in CE– and CE+; P value determined using Wilcoxon rank-sum test. **b** Linear regression of HEY2 gene expression and nearest-neighbor distance, demonstrating positive correlation. **c** Linear regression of HEY2 and MKI67 expression, demonstrating negative correlation. P values from Pearson correlation. **d** IHC showing high Ki-67 proliferation index (25%) (bar, 250 μ m), with corresponding absent HEY2 expression (bar, 100 μ m) and high pAkt expression (bar, 100 μ m). **e** IHC showing low Ki-67 proliferation index (1%) (bar, 250 μ m), with corresponding high HEY2 expression (bar, 100 μ m) and absent pAkt expression (bar, 100 μ m). **f** HEY2 and pAkt IHC intensity as related to cellular density and Ki-67 proliferation indices

learning neural network methods analyzed multiplatform TCGA molecular data to generate protein-genetic and gene expression models of overall survival, and provided an objective ranking of clinical and molecular risk factors. In concordance with recent

investigations,⁸ NOTCH1 mutations were identified as one of the most highly weighted risk factors in our deep learning prognostic model, and was the genetic event most associated with disease progression in each endpoint assessed (MRI contrast-

Table 2. Survival tables

Predictor	OS hazard ratio	P-value	Adjusted OS hazard ratio	P-value
^a Age (per 10 yrs)	3.64	<0.0001	–	–
^a Grade III (vs. II)	6.61	0.013	–	–
^{a, c} <i>MKI67</i> exp.	1.58	0.0029	1.12	0.42
<i>NOTCH1</i> mut.	1.71	0.28	1.10	0.87
^b <i>PIK3</i> mut.	1.97	0.15	3.11	0.045
<i>RBPJ</i> + <i>NOTCH1</i> mut.	1.81	0.210	0.85	0.76
^a 15q loss	3.52	0.007	1.48	0.47
^{b, c} <i>HES5</i> exp.	0.82	0.086	0.74	0.024
^{a, c} <i>HEY1</i> exp.	0.34	0.0009	0.86	0.72
^{a, c} <i>HEY2</i> exp.	0.35	0.0001	0.79	0.54

	PFS hazard ratio	P-value	Adjusted PFS hazard ratio	P-value
Age (per 10 yrs)	1.12	0.28	–	–
^a Grade III (vs. II)	2.24	0.046	–	–
^c <i>MKI67</i> exp.	1.04	0.71	0.97	0.81
^a <i>FUBP1</i> mut.	2.48	0.022	2.14	0.058
<i>NOTCH1</i> mut.	2.07	0.091	1.52	0.33
<i>PIK3</i> mut.	1.91	0.11	1.98	0.092
^a <i>RBPJ</i> + <i>NOTCH1</i> mut.	2.47	0.021	1.86	0.13
^{a, b} 14q loss	3.70	0.010	3.90	0.0035
^{a, b, c} <i>HEY1</i> exp.	0.41	0.0022	0.48	0.018

Cox proportional hazard models for overall survival (OS) and progression-free survival (PFS)
 Multivariable OS adjusted for grade and age
 Multivariable PFS adjusted for grade
Mut mutation
Exp. expression
^aSignificant on univariate analysis
^bSignificant after covariate adjustment
^cGene expression on a log₂ scale, such that the hazard ratio is for each doubling of gene expression

enhancement, cell density, and cellular proliferation). Therefore, inactivating point mutations of *NOTCH1* are one of the most clinically meaningful alterations in oligodendroglioma progression and might suggest that inactivation of the Notch pathway is more generally responsible for poor clinical outcomes.

The NOTCH family is an evolutionarily conserved set of transmembrane receptors that regulate numerous critical biological functions. Notch pathway is activated by extracellular ligand binding, followed by γ -secretase cleavage to release an active intracellular domain (NICD), which localizes to the nucleus and binds to its partner RBPJ to initiate transcription of downstream targets, including *HES* and *HEY* family members.^{19,20} Both activating and inactivating *NOTCH1* mutations have been described in cancer, including in oligodendroglioma.^{8,20–23} Inactivating mutations, such as those noted in oligodendroglioma and head and neck squamous cell carcinoma, are enriched within EGF-like regions and interfere with ligand-mediated pathway activation.^{1,17,20,22,24–26}

Our results suggest inactivation of Notch signaling may be more relevant to oligodendroglioma progression than *NOTCH1* mutations alone. For example, reduced expression of Notch targets, namely *HES1*, *HEY1*, and especially *HEY2*, was seen in clinically progressed oligodendroglioma, while *HES5* expression was most

associated with shorter survival on multivariable analysis. *HEY2* showed a strong positive correlation with cellular density and proliferation, beyond those of *NOTCH1* mutations alone, suggesting other Notch pathway members might be inactivated and lead to reduced downstream target activation.

Furthermore, we found mutations and deletions of *RBPJ*, the nuclear binding partner of NOTCH1 and a member of the canonical Notch pathway, are linked to advanced disease, providing additional evidence that Notch pathway inactivation may be a general progression mechanism. RBPJ normally recruits corepressor proteins and suppresses transcription of downstream targets, whereas active NOTCH1 binds RBPJ and initiates transcription.²⁷ Genetic aberrations of *RBPJ* likely prevent active NOTCH1 from binding to the transcriptional complex. However, Notch-independent functions of RBPJ have also been described.²⁷ *RBPJ* was mutated in 3% of our cohort and homozygously deleted in another case, which is relatively low, but consistent with other forms of cancer.^{18,28} Importantly, *RBPJ* alterations were mutually exclusive from *NOTCH1* mutations, showed strong trends of association with features of disease progression, and had reduced downstream target expression when considered independently. When cases with either *NOTCH1* mutations or *RBPJ* alterations were considered together, the combined group was more strongly associated with disease progression and pathway inactivation than either one alone, and was strongly associated with worse PFS, again raising the possibility that Notch pathway inactivation by multiple mechanisms may be associated with oligodendroglioma progression.

Other prognostically-significant chromosomal aberrations associated with disease progression uncovered by our analysis, including losses of 14q and 15q and gains of 7p, also harbor Notch pathway members, and may be mechanistically relevant to pathway inactivation and disease progression, but will require further investigation. Chromosome 14q contains genes that encode presenilin-1 (PSEN1), a component of the γ -secretase that activates Notch; NUMB, a Notch inhibitor; and jagged-2 (JAG2), a NOTCH receptor ligand. 15q, whose loss was nearly mutually exclusive with *NOTCH1* and *RBPJ* aberrations, contains genes coding for Delta-like 4 (DLL4), a NOTCH ligand; a disintegrin and metalloproteinase domain-containing protein 10 (ADAM10), a controller of NOTCH cleavage; and APH1B, a γ -secretase of NOTCH.²⁹ Chromosome 7 contains the gene encoding lunatic fringe (LFNG), a key Notch signaling repressor, such that its overexpression could suppress Notch signaling.²⁹ The identification of *RBPJ* mutations as a Notch pathway member associated with a poor prognosis, our link between gene expression of Notch pathway members to patient outcome, and the finding of downstream effectors of the Notch pathway, such as *Hes* and *Hey* family members, being downregulated in progressed oligodendrogliomas collectively point in the direction of uncovering other inactivating Notch family members, likely within amplified or deleted loci and providing a platform for assessing Notch pathway for predicting clinical behavior.

Mutations of *PIK3* subunits were highly weighted negative prognostic markers in our neural network analysis; were enriched in a subset of our endpoints of advanced disease; and were markers of shorter survival on multivariable analysis. Mutations of *PIK3CA* are activating, while those of *PIK3R* are inactivating, and both result in enhanced PI3K activity, with downstream activation of Akt and mammalian target of rapamycin, which are associated with aggressive clinical behavior in many cancers.³⁰ Our neural network identified INPP4B, a known suppressor of PI3K signaling,³¹ as a protein whose increased expression was strongly associated with improved outcome. The PI3K pathway also strongly converges with SMAD2/3 and P38/MKK3 pathways, which were identified as among the most enriched with negative prognostic transcripts in our neural network.^{15,16} Lastly, our IHC

analysis indicated pAkt expression was associated with higher-grade features and may have utility as a prognostic marker.

Importantly, our identification of Notch and PI3K pathways' association with survival risk and disease progression does not demonstrate a causal or temporal relationship, and represents an inherent limitation of our study. The use of a machine-learning method does not resolve the issues of feature covariance that also limit the interpretation of models generated by more conventional approaches. We cannot prove *NOTCH1* or *PIK3* subunit mutations evolved temporally from a lower grade tumor, causing its progression. It is entirely possible oligodendrogliomas with Notch inactivation and PI3K activation are in fact distinct genetic subsets at their initiation and these tumors are more rapidly progressive. Longitudinal investigation of patient cohorts with primary and recurrent tumors is needed to identify temporal evolution.^{32,33} Future investigation will also require the elucidation of downstream targets of Notch and PI3K pathways that may drive glioma progression.

METHODS

Study design

We used clinical and genomic data from the Open Access Data Tier of the TCGA LGG dataset for 169 oligodendroglioma (IDH mutant, 1p/19q co-deleted) (<http://cancergenome.nih.gov/>; last accessed September 7th, 2016). Clinical variables consisted of age, gender, extent of resection, overall survival time, survival status, progression-free survival time, and progression status; tumor characteristics included location and histologic grade based on the 2007 WHO brain tumor classification.¹³

Deep learning survival model

We trained a Cox proportional hazards deep learning neural networks to model OS.³⁴ Two models were constructed: (1) a genetic-protein model based on clinical factors (radiation therapy, histologic grade), age, gender, mutations, focal and arm-level copy number events (CNAs), and reverse phase protein array profiles, and 2) a transcriptional model based on mRNA sequencing factors alone. Mutations and CNAs were filtered using MutSig *P*-value threshold of 0.1, and Genomic Identification of Significant Targets in Cancer (GISTIC) *P*-value threshold of 0.25.^{35,36} The prognostic significance of each feature was assessed using mathematical derivatives to evaluate the sensitivity of risk to changes in feature values. Prognostic significance weights in the mRNA model were further used to perform pathway analysis to identify pathways enriched with either good or poor prognosis transcripts. Pathway analysis was performed with GSEA using the Canonical Pathways gene set from the MSigDB curated gene sets.

The accuracy of these modeling approaches in the oligodendroglioma cohort was evaluated using Monte-Carlo cross validation. We first randomly assigned 80% of samples to a training set, and the remaining 20% of samples to a testing set. A predictive model was trained using the training sample, and the accuracy of this model was evaluated using Harrell's concordance-index (*c*-index) on the testing samples. This process was repeated for 20 randomized partitions of the dataset. For the genetic-proteomic model, a three layer network consisting of 100 neurons per layer was used. For the transcriptional model, a three layer network consisting of 500 neurons per layer was used. In both cases, these models were trained for 25 epochs using RMSprop optimization with a learning rate of 1e-3 and a dropout rate of 10%. Further details of this modeling approach are available in our previous work.³⁴

Clinical data was obtained from the TCGA data portal (last accessed 22 January 2016). OS was defined as months from initial diagnosis to death. Survival curves were estimated using the Kaplan-Meier method; log-rank tests were used to compare curves between groups. Progression free survival (PFS) was defined as months from initial diagnosis to disease progression or death. PFS curves were estimated using the Kaplan-Meier method; log-rank tests were used to compare curves between groups. Single and multi-variable models (non machine-learning) were also fit using Cox regression under the proportional hazards assumption for OS and PFS.

Genomic data

Gene expression, mutation, and CNA data were obtained from the TCGA portal (<https://tcga-data.nci.nih.gov>). Genetic alterations with at least 5% frequency were included in the analysis (Table S1A). Variants were considered as mutants if there was an amino acid change and genes were filtered using $q \leq 0.05$ in MutSig analysis. Mutations were then converted into dichotomous variables (mutation and wild-type). Arm level copy number data was obtained from GDAC GISTIC hosted analysis results (<https://gdac.broadinstitute.org/>). Values of chromosomal arm gain or loss were listed as a fraction of the chromosomal arm, where gains were positive values and losses were negative values. A threshold absolute value of 0.10 of the fraction of the chromosomal arm was used to signify chromosomal gain or loss. Frequency of chromosomal gains and losses are summarized in Table S1B.

Radiographic imaging review

Preoperative MR imaging studies for TCGA patients were obtained from TCIA (<http://www.cancerimagingarchive.net/>; last accessed 8 February 2016) for 55 untreated patients. Institutional neuroradiologists and neurosurgeons reviewed MR images for the presence of unequivocal contrast-enhancement.

Quantification of cellular density and nearest-neighbor analysis

Whole-slide digital pathology images ($n = 142$) were obtained from the CDSA (<http://cancer.digitalslidearchive.net/>; last accessed 11 August 2016). Images (20x) were analyzed using an image analysis algorithm to identify cell nuclei and to quantify cellular density in areas of tumor infiltration.³⁷ The spacing between neighboring nuclei was calculated using KD-trees, and these distances were modeled using a Poisson point process. The densities of tumor and normal regions were deconvolved using a mixture Poisson model to identify the density parameter in tumor regions, $\lambda_{\text{tumor}}^{-1}$. The median tumor density across patients was used to define "less dense" and "more dense" categories. Cell density was also analyzed visually by a neuropathologist (JV), blinded to nearest neighbor analysis, and scored as: "low", "intermediate", or "high". Algorithm and human assessments of density were highly concordant (Wilcoxon-rank sum < 0.05 level).

Gene expression of MKI67 as a marker for cellular proliferation

A "high" category for *MKI67* expression was defined (≥ 700) to correspond to 15% MIB-1/Ki-67 labeling index using a linear regression model.¹¹ Samples with *MKI67* < 700 were designated 'low'.

Statistics

Associations between contrast-enhancement and mutational status were calculated using the χ^2 test for independence; for expected counts less than 5, Fisher's exact test was used. Statistical associations between 2 groups of continuous or ordinal variables, such as the cellular density calls, were calculated using Wilcoxon rank-sum tests. The Pearson correlation coefficient was used to measure the linear dependence between continuous variables. All *P*-values reported are two-sided and regarded as statistically significant if $P < 0.05$. The software used for statistical analysis and graphical representations include: SPSS v23 (SPSS Statistics, IBM Corp., NY) and R Studio v0.99. All boxplots have the median marked as the center line, and whisker lines indicate the lower and upper quartiles (25 and 75%, respectively).

Validation set

Fifty-one patients with primary oligodendroglioma (*IDH* mutant, 1p/19q co-deleted) were identified at Emory University Hospital with approval from the institution's IRB committee and with a waiver of consent (IRB 00088647). MRIs were reviewed by a neuroradiologist (CAH) for contrast enhancement. Histologic slides were reviewed by two neuropathologists (DJB and JV). IHC staining was performed for Ki-67; a proliferation index was calculated using digital image analysis (Aperio Positive Pixel Count). Cell density was calculated by dividing cell count by area in regions of interest (mm^2). IHC for Notch signaling was assessed using anti-HEY2 rabbit polyclonal antibody (catalog #AB5716, Millipore, 1:100) and for PI3K using anti-pAkt (S473) rabbit monoclonal antibody (#EP2109Y, Abcam, 1:100). HEY2 and pAkt IHC slides were reviewed and scored based on staining intensity. Selected samples underwent DNA isolation and focused sequencing of the *NOTCH1* gene using Sanger sequencing, included the epidermal-growth-factor-like domain

(EGF-like) spanning amino acids 300 to 500. Targeted sequencing was performed using a glioma gene panel on the Fluidigm platform.

DATA AVAILABILITY

All data used in this investigation is accessible in Supplementary Data File S1.

ACKNOWLEDGEMENTS

The authors thank the Tissue Procurement Service and the Research Pathology Laboratory of the Cancer Tissue and Pathology Shared Resource, as well as the Proteomics Shared Resource, at the Winship Cancer Institute, supported by the NCI Cancer Center Support Grant (P30CA138292). This work was generously supported by the National Brain Tumor Society (NBTS), the loglio research project, and Oligo Nation. The U.S. Public Health Service supported this work through National Institutes of Health grants R01CA176659 (D.J. Brat), U24CA194362 (L.A.D. Cooper), K22LM011576 (L.A.D. Cooper), the National Center for Advancing Translational Sciences of the National Institutes of Health grants UL1TR000454 (S.H. Halani) and TL1TR000456 (S.H. Halani), and the Winship Cancer Institute NCI Cancer Center Support Grant (P30CA138292).

AUTHOR CONTRIBUTIONS

The following authors contributed to: Conception and design of study (S.H.H., S.Y., L.A.D.C., D.J.B.); acquisition of data (S.H.H., S.Y., J.V.V., F.A., A.B.S., Z.Z., T.J., C.G.E., C.A.H., J.E., B.G., J.J.O., L.P., L.A.D.C., D.J.B.); analysis and interpretation of data (S.H.H., S.Y., F.A., L.P., L.A.D.C., D.J.B.); drafting and revising the manuscript (S.H.H., S.Y., J.V.V., F.A., A.B.S., Z.Z., T.J., C.G.E., C.A.H., J.E., B.G., J.J.O., L.P., L.A.D.C., D.J.B.); final approval of the submitted manuscript (S.H.H., S.Y., J.V.V., F.A., A.B.S., Z.Z., T.J., C.G.E., C.A.H., J.E., B.G., J.J.O., L.P., L.A.D.C., D.J.B.).

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Precision Oncology* website (<https://doi.org/10.1038/s41698-018-0067-9>).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

- Cancer Genome Atlas Research, N. et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).
- Parsons, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
- Yan, H. et al. IDH1 and IDH2 mutations in gliomas. *New Engl. J. Med.* **360**, 765–773 (2009).
- Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K. *World Health Organization Histological Classification of Tumours of the Central Nervous System*. International Agency for Research on Cancer, France (2016).
- Eckel-Passow, J. E. et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *New Engl. J. Med.* **372**, 2499–2508 (2015).
- Bettgowda, C. et al. Mutations in CIC and FUBP1 contribute to human oligodendroglioma. *Science* **333**, 1453–1455 (2011).
- Ceccarelli, M. et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).
- Aoki, K. et al. Prognostic relevance of genetic alterations in diffuse lower-grade gliomas. *Neuro-Oncology* **20**, 66–77 (2017).
- Obermeyer, Z. & Emanuel, E. J. Predicting the future—big data, machine learning, and clinical medicine. *New Engl. J. Med.* **375**, 1216–1219 (2016).
- Reyes-Botero, G. et al. Contrast enhancement in 1p/19q-codeleted anaplastic oligodendrogliomas is associated with 9p loss, genomic instability, and angiogenic gene expression. *Neuro Oncol.* **16**, 662–670 (2014).
- Trembath, D., Miller, C. R. & Perry, A. Gray zones in brain tumor classification: Evolving concepts. *Adv. Anat. Pathol.* **15**, 287–297 (2008).
- Wesseling, P., van den Bent, M. & Perry, A. Oligodendroglioma: Pathology, molecular mechanisms and markers. *Acta Neuropathol.* **129**, 809–827 (2015).
- Louis, D. N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K. *WHO Classification of Tumours of the Central Nervous System*. 4th edn, Intl. Agency for Research, Lyon (2007).

- Olar, A. et al. IDH mutation status and role of WHO grade and mitotic index in overall survival in grade II-III diffuse gliomas. *Acta Neuropathol.* **129**, 585–596 (2015).
- Singh, A. M. et al. Signaling network crosstalk in human pluripotent cells: A Smad2/3-regulated switch that controls the balance between self-renewal and differentiation. *Cell Stem Cell* **10**, 312–326 (2012).
- Locatelli, S. L. et al. Dual PI3K/ERK inhibition induces necroptotic cell death of Hodgkin Lymphoma cells through IER3 downregulation. *Sci. Rep.* **6**, 35745 (2016).
- Wang, N. J. et al. Loss-of-function mutations in Notch receptors in cutaneous and lung squamous cell carcinoma. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 17761–17766 (2011).
- Cerami, E. et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
- Kopan, R. & Ilagan, M. X. The canonical Notch signaling pathway: Unfolding the activation mechanism. *Cell* **137**, 216–233 (2009).
- Yap, L. F. et al. The opposing roles of NOTCH signalling in head and neck cancer: A mini review. *Oral Dis.* **21**, 850–857 (2015).
- Rampias, T. et al. A new tumor suppressor role for the Notch pathway in bladder cancer. *Nat. Med.* **20**, 1199–1205s (2014).
- Agrawal, N. et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**, 1154–1157 (2011).
- Radtke, F. & Raj, K. The role of Notch in tumorigenesis: oncogene or tumour suppressor? *Nat. Rev. Cancer* **3**, 756–767 (2003).
- Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
- Stransky, N. et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
- Rebay, I. et al. Specific EGF repeats of Notch mediate interactions with Delta and Serrate: Implications for Notch as a multifunctional receptor. *Cell* **67**, 687–699 (1991).
- Xie, Q. et al. RBPJ maintains brain tumor-initiating cells through CDK9-mediated transcriptional elongation. *RBPJ maintains brain tumor-initiating cells through CDK9-mediated transcriptional elongation* **126**, 2757–2772 (2016).
- Kulic, I. et al. Loss of the Notch effector RBPJ promotes tumorigenesis. *J. Exp. Med.* **212**, 37–52 (2015).
- UniProt: A hub for protein information. *Nucleic Acids Res.* **43** D158–D169 (2015).
- Thorpe, L. M., Yuzugullu, H. & Zhao, J. J. PI3K in cancer: Divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat. Rev. Cancer* **15**, 7–24 (2015).
- Gewinner, C. et al. Evidence that Inositol polyphosphate 4-phosphatase type II is a tumor suppressor that inhibits PI3K signaling. *Cancer Cell* **16**, 115–125 (2009).
- Kim, H. et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.* **25**, 316–327 (2015).
- Johnson, B. E. et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).
- Yousefi, S. et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.* **7**, 11707 (2017).
- Beroukhim, R. et al. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20007–20012 (2007).
- Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
- Cooper, L. A. et al. Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images. *Lab Invest.* **95**, 366–376 (2015).



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

Chapter 4

Learning Clinical Outcomes from Heterogeneous Data Sources

This chapter is a collection of articles on learning from combinations of heterogeneous data sources. All articles included in this chapter are open access and available online.

Heterogeneity is a common challenge is combining clinical and genomic data from different studies, hospitals, etc. due to demographical and technological differences and batch effects. In this chapter, we introduce multi-task learning models that can learn similarities between heterogeneous cohorts without being misled by such differences and consequently achieve better performance in predicting outcomes. We use an adversarial classification in the multi-task learning framework to further encourage the learning of a shared representation among all data sources. Our proposed methods are shown to outperform existing multi-task learning algorithms and single-task baselines. Chapter 4.1 includes a short conference paper on this topic, and chapter 4.2 expands on the former by more extensive experiments and offering a comparative interpretation of some of the models.

4.1 Learning Clinical Outcomes from Heterogeneous Genomic Data Sources: An Adversarial Multi-task Learning Approach, ICML, 2019 Adaptive and Multitask Learning Workshop

This section is an exact copy of the following open-access conference paper:

Safoora Yousefi, Amirreza Shaban, Mohamed Amgad, and Lee AD Cooper. *Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models*. International Conference on Machine Learning, Adaptive and Multitask Learning Workshop 2019.

Abstract. Translating the high-dimensional data generated by genomic platforms into reliable predictions of clinical outcomes remains a critical challenge in realizing the promise of genomic medicine largely due to the small number of independent samples. We show that neural networks can be trained to predict clinical outcomes using heterogeneous genomic data sources via multi-task learning and adversarial representation learning, allowing one to combine multiple cohorts and outcomes in training. Experiments demonstrate that the proposed method helps mitigate data scarcity and outcome censorship in cancer genomics learning problems.

Learning Cancer Outcomes from Heterogeneous Genomic Data Sources: An Adversarial Multi-task Learning Approach

Safoora Yousefi¹ Amirreza Shaban² Mohamed Amgad³ Lee Cooper³

Abstract

Translating the high-dimensional data generated by genomic platforms into reliable predictions of clinical outcomes remains a critical challenge in realizing the promise of genomic medicine largely due to small number of independent samples. We show that neural networks can be trained to predict clinical outcomes using heterogeneous genomic data sources via multi-task learning and adversarial representation learning, allowing one to combine multiple cohorts and outcomes in training. Experiments demonstrate that the proposed method helps mitigate data scarcity and outcome censorship in cancer genomics learning problems.

1. Introduction

Since the emergence of high throughput experiments such as Next Generation Sequencing, the volume of genomic data produced has been increasing exponentially (Stephens et al., 2015). A single biopsy can generate tens of thousands of transcriptomic, proteomic, or epigenetic features. The ability to generate genomic data has far outpaced the ability to translate these data into clinically-actionable information, as typically only a handful of molecular features are used in diagnostics or in determining prognosis (Bailey et al., 2018; Van De Vijver et al., 2002; Network, 2015).

Cancer genomic datasets often have small sample size (hundreds of samples), and much larger dimensionality (tens of thousands of features), making it difficult to train complex models such as neural networks (Abu-Mostafa, 1989). Furthermore, of those available samples, often large proportions have censored outcomes. Several approaches have been employed to alleviate this data insufficiency including dimensionality reduction, feature selection, data augmenta-

tion, and transfer learning (Ching et al., 2018).

An alternative approach is to integrate genomic data from multiple studies and hospitals to increase training set size. Heterogeneity of available genomic datasets due to technical and sample biases poses challenges to this approach. Cohorts from different sources typically have difference demographic or disease stage distributions, may be subject to different signal capture calibration and post-processing artifacts. This means that naively combining heterogeneous cohorts is both difficult and may degrade model accuracy due to batch effects (Tom et al., 2017).

Building upon SurvivalNet (Yousefi et al., 2016; 2017) - a neural network model for survival prediction- we propose a multi-task learning approach that enables: a) training SurvivalNet on multiple heterogeneous data sources while avoiding the issues that arise from naively combining datasets, and b) training on multiple clinical outcomes from the same cohort, thus helping to address the issue of censorship often encountered in clinical datasets. We further enhance our proposed method by introducing an adversarial cohort classification loss that prevents the model from learning cohort-specific noise, thus enabling task-invariant representation learning. Experiments demonstrate that our proposed methods can be used to alleviate data scarcity and outcome censorship in several cancer genomics learning problems, leading to superior performance on target cohorts and outperforming previous multi-task survival analysis methods.

2. Background and Related Work

Survival analysis with Cox proportional hazards model:

Survival analysis refers to any problem where the variable of interest is time to some event, which in cancer is often death or progression of disease. Time-to-event modelling is different from ordinary regression due to a specific type of missing data problem known as censoring. Incomplete or censored observations are important to incorporate into the model since they could provide critical information about long-term survivors (Harrell Jr, 2015). The most widely used approach to survival analysis is the semi-parametric Cox proportional hazards model (Cox, 1972). It models the

¹Department of Computer Science, Emory University, Atlanta, GA, USA ²College of Computing, Georgia Institute of Technology, Atlanta, GA, USA ³Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA. Correspondence to: Safoora Yousefi <safoora.yousefi@emory.edu>.

hazard function at time s given the predictors x_i of the i th sample as:

$$h(s|x_i, \beta) = h_0(s)e^{\beta^\top x_i} \quad (1)$$

The model parameters β are estimated by minimizing Cox’s negative partial log-likelihood:

$$L_{cox}(X, Y, \beta) = - \sum_{x_i \in U} \left(\beta^\top x_i - \log \sum_{j \in R_i} e^{\beta^\top x_j} \right) \quad (2)$$

where $X = \{x_1, \dots, x_N\}$ are the samples, and $Y = \{E, S\}$ represents label vectors of event or last follow-up times $S = \{s_1, \dots, s_N\}$ and event status $E = \{e_1, \dots, e_N\}$. For censored samples ($e = 0$), s represents time of last follow-up while for observed samples ($e = 1$), it represents event time. The outer sum is over the set of uncensored samples U and R_i is the set of *at-risk* samples with $s_j \geq s_i$. The baseline hazard $h_0(t)$ is cancelled out of the likelihood and can remain unspecified.

A non-linear alternative to Cox regression is SurvivalNet (Yousefi et al., 2016; 2017), a fully connected artificial neural network f_W with parameters W that replaces X in Equation 2 with its non-linear transformation $f_W(X)$. SurvivalNet has been shown to outperform other common survival analysis techniques such as random survival forests (Ishwaran et al., 2008) and Cox-ElasticNet (Park & Hastie, 2007) in learning from high-dimensional genomic data.

Multi-task learning for survival analysis: Both theoretical and empirical studies show that learning multiple related tasks simultaneously often significantly improves performance relative to learning each task independently (Baxter, 2000; Ben-David & Schuller, 2003; Caruana, 1997). This is particularly the case when only a few samples per task are available, since with multi-task learning, each task has more data to learn from.

The general form of the loss function when learning T tasks simultaneously is:

$$L(Y, X, W) = \sum_{t=1}^T L_t(y^t, g^t(W^t, X^t)) + \gamma \lambda(Y, X, W) \quad (3)$$

L_t and W^t , respectively, are the loss function and the parameters of task t . $Y = \{Y^1, \dots, Y^T\}$ and $X = \{X^1, \dots, X^T\}$ are the combined input data of all t tasks. g^t indicates the prediction function corresponding to task t , and λ is a regularization or auxiliary function that captures task relatedness assumptions, examples of which include $\ell_{2,1}$ norm (Argyriou et al., 2007), and cluster norm (Jacob et al., 2009). γ is a weight parameter controlling the importance of the auxiliary function.

Previous work has applied multi-task learning under different task relatedness assumptions to train Cox’s proportional hazards model using multiple genomic data sources (Wang et al., 2017; Li et al., 2016).

In this paper, our main assumption is that gene expression data lies on a lower dimensional subspace that can be utilized in several prognostic tasks. We will enforce this assumption via hard parameter sharing among tasks and the bottleneck architecture of our models. Moreover, In section 3 we describe how an adversarial classification objective can be used as auxiliary function λ to encourage task-invariant representation learning. We compare our proposed method to multi-task Cox model with $\ell_{2,1}$ regularization (Li et al., 2016).

Adversarial representation learning: The idea of using adversarial learning to match two distributions was first proposed by (Goodfellow et al., 2014) for training generative models. This idea has been applied to unsupervised domain adaptation for natural language processing and computer vision, with varying design choices including parameter sharing, type of adversarial loss, and discriminative vs. generative base model (Ganin & Lempitsky, 2015; Ganin et al., 2016; Tzeng et al., 2015; Liu & Tuzel, 2016; Tzeng et al., 2017).

We adapt this idea to multi-task learning to encourage our proposed model to learn task-invariant genomic representations. A cohort discriminator is trained to assign samples to their cohort. Simultaneously, a SurvivalNet is adversarially trained to confuse the discriminator by learning a representation of data where samples from different cohorts are indistinguishable (in addition to learning to predict survival).

3. Methods

In cases where all tasks are similar and their corresponding samples come from similar distributions, a natural approach is to simply combine the datasets and train a single-task model on the combined training data, as done in (Yousefi et al., 2017). We implement this approach using SurvivalNet to provide a performance baseline.

But the assumption that the datasets come from the same distribution rarely holds and this could be problematic in training a Cox-based model. Comparisons of survival time between pairs of samples are integral to the Cox log-likelihood loss function. When one naively combines datasets to train a model with a single Cox loss, in addition to comparisons within each cohort, comparisons between these cohorts contribute to the loss. Since the difference between distributions of these cohorts could be due to clinically insignificant factors such as batch effects, these between-cohort comparisons could be misleading in training. Our first proposed model aims to eliminate this potentially misleading signal from the training process via multi-task learning:

Multi-task learning (MTL): This proposed extension of SurvivalNet model comprises one Cox loss node per each task, so that only within-cohort comparisons contribute to

the loss. The objective function of the MTL model is the sum of all Cox losses:

$$L_{MTL} = \sum_{t=1}^T L_{Cox}(f_W(X^t), Y^t, \beta) \quad (4)$$

where f_W is the SurvivalNet model. All parameters of MTL, β and W , are shared among tasks.

Although we are encouraging sparse representation learning via the bottleneck architecture of the MTL model, that does not force the model to learn a task invariant representation. The model may learn a sparse representation, but still have enough parameters to be able to discriminate between samples from different cohorts and process them differently. The adversarial model described below addresses this limitation.

Adversarial multi-task model (ADV-MTL): This model extends SurvivalNet by addition of an adversarial cohort classification loss. Let $X_{comb} = \{x_1, \dots, x_M\}$ and $Y_{comb} = \{y_1, \dots, y_M\}$ denote the combination of all X^t and Y^t , respectively, including M samples in total. A set of one-hot vectors $Y_D = \{d_1, \dots, d_M\}$ indicate cohort membership, so that $d_{it} = 1$ means that the i th sample belongs to the t th cohort. A cohort discriminator is trained to assign the transformed samples $z_i = f_W(x_i)$ to the cohort they belong to. This component of the model is a multi-class logistic regression with a softmax cross-entropy loss. It comprises a simple neural network g_θ mapping z_i to a T-dimensional vector, where T is the number of tasks, and a softmax function that transforms the result to a T-dimensional vector of probabilities. The predicted probability that sample i belongs to cohort t is given by:

$$\hat{d}_{it} = \frac{e^{g_\theta(z_i)_t}}{\sum_{k=1}^T (e^{g_\theta(z_i)_k})},$$

and the objective function of the discriminator L_D is the cross-entropy between predicted probabilities and cohort labels:

$$L_D(f_W(X_{comb}), Y_D, \theta) = \gamma \sum_{i=1}^M \sum_{t=1}^T -d_{it} \log \hat{d}_{it} \quad (5)$$

This loss function only trains the parameters of the discriminator, namely θ the parameters of the function g_θ .

Simultaneously, a multi-task risk predictor component is adversarially trained with the following objective:

$$L_R = \sum_{t=1}^T L_{Cox}(f_W(X^t), Y^t, \beta) - \gamma L_D(f_W(X_C), Y_D, \theta) \quad (6)$$

L_R trains the parameters of the risk predictor β as well as W . By updating W with an objective function that is the opposite of that of the discriminator, we encourage learning a representation of data in which samples from different cohorts are indistinguishable. γ controls the contribution of the adversarial loss to representation learning.

4. Results

Datasets: We use several publicly available benchmark survival analysis datasets from The Cancer Genome Atlas (TCGA) and METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) (Curtis et al., 2012). Both of these sources provide gene expression data (over 20K features) and clinical outcome labels. TCGA clinical data contains overall survival (OS) and progression free interval (PFI) outcome labels (Liu et al., 2018) while METABRIC only contains OS labels. For details about datasets and preprocessing, refer to supplementary materials.

Model selection and training: In each experiment, we pick a target task and use auxiliary tasks to improve performance on the target task. We use random stratified sampling to sample 60% of target data as training and use the remaining 40% as hold-out testing data. Stratified sampling ensures similar event rates in training and testing sets. Training set is augmented with any auxiliary data at this stage if the experiment calls for it. For model selection, grid search with 5-fold cross validation is performed on the training set and the selected model is then evaluated on the hold-out testing data. We repeat this procedure on 30 randomly sampled training and testing sets and use re-sampled t-test and paired re-sampled t-test (Dietterich, 1998) to provide confidence intervals and significance analysis. In visualizing the results, we use shaded areas or error bars to depict the 95% confidence intervals of the mean c-index.

A single hidden layer with 50 ReLU hidden units was used in all risk prediction neural networks. Discriminators were fixed to a single-layer design with 20 ReLU hidden units. Learning rate, drop-out regularization rate, and L2 regularization rate of neural network parameters W , and the weight of the discriminator loss γ were tuned via grid search.

The same sampling, training, model selection and evaluation procedures was used in all experiments with all methods. All software to reproduce the results presented in this section is available at [GITHUB LINK]. For Cox- $\ell_{2,1}$, we used the authors' open-source implementation (Li et al., 2016).

Evaluation Metric: We measured model performance using *concordance index* (c-index) that captures the rank correlation of predicted and actual survival (Harrell Jr et al., 1982), and is given by:

$$CI(\beta, X) = \sum \frac{I(i,j)}{|P|} \quad (7)$$

$$I(i,j) = \begin{cases} 1, & \text{if } r_j \stackrel{P}{>} r_i \text{ and } t_j > t_i \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Where P is the set of orderable pairs. A pair of samples (x_i, x_j) is orderable if either the event is observed for both x_i and x_j , or x_j is censored and $t_j > t_i$. Optimizing Cox's partial likelihood (Equation 2) has been shown to be equivalent to optimizing c-index (Steck et al., 2008).

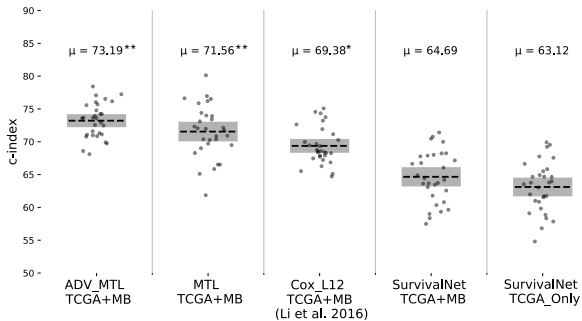


Figure 1. METABRIC and TCGA breast cancer datasets were combined to improve performance on TCGA, using the proposed and baseline methods. ** indicates significant improvement over both Cox- $\ell_{2,1}$ and single-task models. * indicates significant improvement over single-task models.

4.1. Combining two breast cancer cohorts

This section investigates the integration of two breast cancer cohorts from independent studies. We use TCGA BRCA as target, and METABRIC as auxiliary cohort. Both cohorts are diagnosed with breast cancer. In such cases where similar biological processes determine the outcomes, one would expect naively pooling cohorts together to lead to better predictions on each of the cohorts. This is the expectation particularly in this case where the auxiliary cohort has twice the number of samples as almost the target cohort (1903 vs. 1094) and three times the event rate (33% vs. 13%).

Surprisingly, we observe that simply adding METABRIC to training data (SurvivalNet TCGA+MB) does not improve prediction of c-index on TCGA ($p=0.1$). See Figure 1. MTL model achieves a significant improvement ($p=3e-4$) over SurvivalNet trained on target data only (SurvivalNet TCGA-only), and ADV-MTL significantly outperforms all other methods. Cox- $\ell_{2,1}$ achieves a significant improvement over single-task SurvivalNet methods, but is significantly outperformed by ADV-MTL ($p=1e-6$) and MTL ($p=0.01$).

4.2. Combining multiple outcome labels

As shown in Table 1, for some patients, a progression event is never observed (or recorded) during the study (censored PFI), while their overall survival outcome is observed (deceased by end of study). In such cases, overall survival could provide an extra supervision signal in training a predictive model that originally targets PFI prediction.

We use the MTL model to simultaneously use PFI and OS outcomes in training. In our experiments with five different TCGA cancer types, multi-task learning with PFI and OS always leads to improved PFI prediction performance compared to single-task SurvivalNet trained with PFI labels only (see Table 1 and Figure 2).

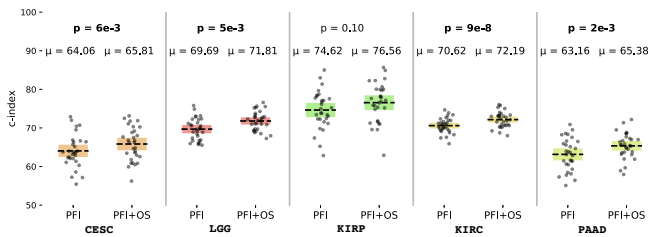


Figure 2. Progression-free interval (PFI) prediction performance with and without multi-task learning with overall survival (OS) labels.

Cancer type	Number of Samples	PFI+OS c-index	Improvement on PFI-only	Censored PFI Observed OS
CESC	304	65.83	1.69%	5.26%
KIRC	533	76.55	2.12%	11.81%
KIRP	514	76.55	1.35%	5.19%
LGG	288	72.15	1.75%	1.75%
PAAD	178	65.12	1.34%	9.55%

Table 1. Progression-free survival (PFI) prediction performance with multi-task learning with overall survival (OS). Percent of samples in each cohort with censored PFI and observed OS is given in the last column.

4.3. Discussion

To provide an insight into the significance of the improvement achieved by our models, we look at the learning curves of SurvivalNet and ADV-MTL evaluated on TCGA-BRCA. Learning curves were obtained by training the models on incrementally more training samples from the target task, and testing on a fixed sized test set (40% of target data, consistent with the rest of experiments). As shown in Figure 3, the performance improvement achieved by ADV-MTL over SurvivalNet (a 10% improvement, see Fig. 1) exceeds the improvement resulting from tripling the size of target training data from 30% to 100% in SurvivalNet. This shows that the integration of heterogeneous datasets using the proposed method is a reasonable alternative to acquisition of new training data from the target distribution which may be expensive or impossible. The ideal solution to any data insufficiency issue is enhanced data collection and standardization efforts. However, in settings where this is impractical, employing techniques like ADV-MTL and MTL can help address this at no extra cost.

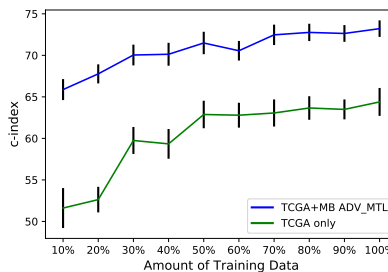


Figure 3. Learning curves of SurvivalNet and ADV-MTL (target: TCGA BRCA).

References

- Abu-Mostafa, Y. S. The vovnik-chervonenkis dimension: Information versus complexity in learning. *Neural Computation*, 1(3):312–317, 1989.
- Argyriou, A., Evgeniou, T., and Pontil, M. Multi-task feature learning. In *Advances in neural information processing systems*, pp. 41–48, 2007.
- Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.
- Baxter, J. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.
- Ben-David, S. and Schuller, R. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pp. 567–580. Springer, 2003.
- Caruana, R. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.
- Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15 (141):20170387, 2018.
- Cox, D. R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34: 187–220, 1972.
- Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiva, S., Yuan, Y., et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346, 2012.
- Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.
- Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd ICML*, 2015.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Hanahan, D. and Weinberg, R. A. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.
- Harrell Jr, F. E. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
- Harrell Jr, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., Rosati, R. A., et al. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.
- Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The annals of applied statistics*, pp. 841–860, 2008.
- Jacob, L., Vert, J.-p., and Bach, F. R. Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, pp. 745–752, 2009.
- Li, Y., Wang, L., Wang, J., Ye, J., and Reddy, C. K. Transfer learning for survival analysis via efficient $l_2, 1$ -norm regularized cox regression. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pp. 231–240. IEEE, 2016.
- Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.
- Liu, M.-Y. and Tuzel, O. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pp. 469–477, 2016.
- Network, C. G. A. R. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.
- Park, M. Y. and Hastie, T. L_1 -regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (4):659–677, 2007.
- Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., and Raykar, V. C. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pp. 1209–1216, 2008.

- Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. Big data: astronomical or genetical? *PLoS biology*, 13(7):e1002195, 2015.
- Tom, J. A., Reeder, J., Forrest, W. F., Graham, R. R., Hunkapiller, J., Behrens, T. W., and Bhangale, T. R. Identifying and mitigating batch effects in whole genome sequencing data. *BMC bioinformatics*, 18(1):351, 2017.
- Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4068–4076, 2015.
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.
- Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.
- Wang, L., Li, Y., Zhou, J., Zhu, D., and Ye, J. Multi-task survival analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 485–494. IEEE, 2017.
- Yousefi, S., Song, C., Nauata, N., and Cooper, L. Learning genomic representations to predict clinical outcomes in cancer. In *International Conference on Learning Representations (ICLR)*, 2016.
- Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., Song, C., Gutman, D. A., Halani, S. H., Vega, J. E. V., Brat, D. J., et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 7(1):11707, 2017.

Supplementary Materials

1. Data Description

The Cancer Genome Atlas (TCGA) provides publicly available clinical and molecular data for 33 cancer types. TCGA gene expression features were taken from the Illumina HiSeq 2000 RNA Sequencing V2 platform. TCGA clinical data contains overall survival (OS) and progression free interval (PFI) labels, with varying degrees of availability for different primary cancer sites (Liu et al., 2018). This data has been obtained from multiple hospitals and health-care centers, so a considerable degree of heterogeneity exists within the TCGA.

PFI is defined as the period from the date of diagnosis until the date of the first occurrence of a new tumor-related event, which includes progression of the disease, locoregional recurrence, distant metastasis, new primary tumor, or death with tumor. OS is the period from the date of diagnosis until the date of death from any cause. Since patients generally suffer from disease progression or recurrence before dying, PFI requires shorter follow-up times and has higher event rate. Additionally, OS is a noisy signal due to deaths from non-cancer causes. Therefore, wherever possible, PFI is used as the outcome variable.

We used METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) (Curtis et al., 2012) gene expression and clinical data in section 4.1. Since METABRIC comes with OS labels only, OS was used as the outcome variable in this section. TCGA breast invasive carcinoma (BRCA) was used in this section as target cohort.

In section 4.2 of the main paper and section 2 of supplementary materials, we perform experiments on a subset of TCGA cancer types. Out of the 33 TCGA cancer types, we selected those with PFI event rate higher than 20%. We used the performance of Cox-ElasticNet (Park & Hastie, 2007) on each of these cancer types as a measure of outcome label quality, and used only those cancer types where Cox-ElasticNet achieved a c-index of 60% and higher, leaving us with adrenocortical carcinoma (ACC), cervical squamous cell carcinoma (CESC), lower-grade glioma (LGG), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), mesothelioma (MESO), and pancreatic adenocarcinoma (PAAD). ACC and MESO could not be used as target cohorts since their small sample sizes did not allow for reliable model evaluation. All of the mentioned cancer types were used as auxiliary cohorts in section 2 of supplemental materials.

We discarded samples that did not have gene expression data or outcome labels. A summary of sample sizes and event rates of datasets after this preprocessing step is given in Table S1. Z-score normalization and 3-NN missing data

Dataset Name	Number of Samples	Number of Features	Event Rate	Event Type
ACC	79	20531	52%	PFI
CESC	304	20531	23%	PFI
KIRC	533	20531	30%	PFI
KIRP	514	20531	37%	PFI
LGG	288	20531	20%	PFI
MESO	84	20531	70%	PFI
PAAD	178	20531	58%	PFI
BRCA	1094	20531	13%	OS
METABRIC	1903	24368	33%	OS

Table S1. Summary of datasets.

imputation were performed on gene expression data. No further feature selection or dimensionality reduction was performed. In section 4.1, we found the intersection of Hugo IDs present in both BRCA and METABRIC datasets (17272 genes), and discarded the genes that were absent in either dataset.

2. Additional Experiments

In addition to integrating data from studies involving the same primary cancer site as in section 4.1, we may benefit from pooling cohorts diagnosed with different cancer types together to increase training size. Cancers that originate from different primary sites are known to have large differences in genetic markup, although there are some remarkable similarities that seem to play a fundamental role in carcinogenesis (Hoadley et al., 2018; Bailey et al., 2018; Hanahan & Weinberg, 2011). The idea of combining multiple cancer types relies on the premise that models of sufficient complexity and constraints can exploit these similarities to improve outcome prediction.

We repeat the experiments of section 4.1 this time using TCGA cohorts diagnosed with different cancer types. In each experiment, one cancer type is chosen as target and all others are used as auxiliary data. Results of these experiments are shown in Figure S1 in terms of c-index achieved on target test set. In 3 out of five 5, training on the combination of heterogeneous TCGA datasets with ADV-MTL model leads to significant improvement over single-task training of SurvivalNet with target training data only. Cox- $\ell_{2,1}$ achieves the same in 2 out of 5 cases. We did not observe any significant difference between ADV-MTL and Cox- $\ell_{2,1}$ in this set of experiments, except in experiments with PAAD where ADV-MTL significantly outperforms Cox- $\ell_{2,1}$ ($p=7e-3$).

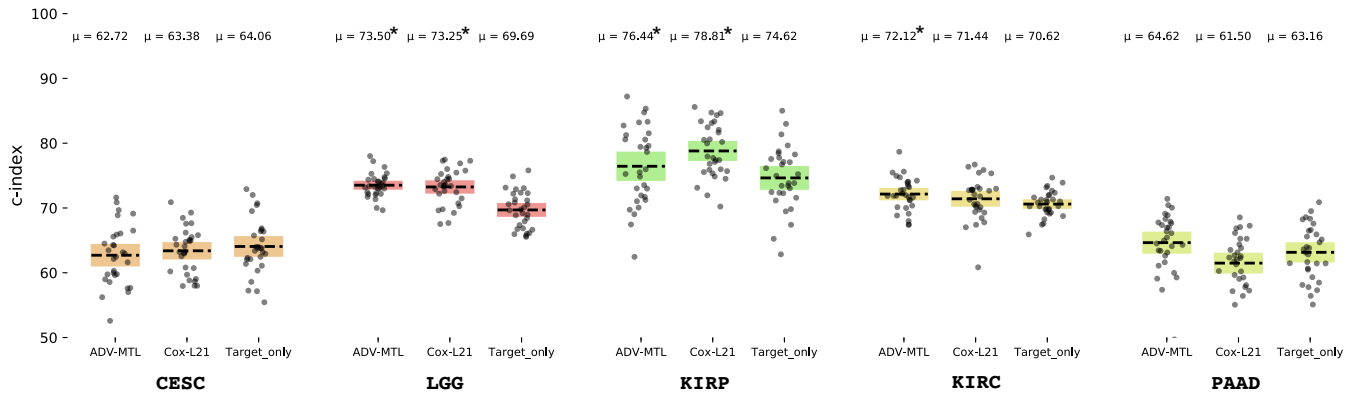


Figure S1. Survival prediction accuracy was improved by multi-task learning and adversarial representation learning on several benchmark datasets. * indicate significant improvement ($p < 0.05$) of multi-task methods over target-only setting.

4.2 Learning clinical outcomes from heterogeneous genomic datasets using adversarial and multi-task learning, Manuscript in Progress

This section is an exact copy of the following manuscript in progress:

Safoora Yousefi, Amirreza Shaban, Mohamed Amgad, Ramraj Chandradevan, and Lee AD Cooper. *Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models.* manuscript in progress.

Abstract. Neural networks have shown promise in predicting clinical outcomes from high-dimensional genomic data, but their potential may be limited due to insufficient training data. This is particularly true for predicting clinical outcomes, where a large portion of the samples may have incomplete clinical followups. In this paper, we show how multi-task and adversarial learning approaches can be used to overcome data shortages and improve accuracy in predicting clinical outcomes. We first show how multi-task learning can significantly improve the prediction of overall survival in breast cancer by combining microarray and RNA-sequencing gene expression profiles from the METABRIC and TCGA datasets. We improve prediction accuracy further using an adversarial penalty to encourage the network to learn a consistent representation across datasets. We also demonstrate the utility of this approach in leveraging samples with multiple clinical outcomes like time to progression and overall survival, and in combining data from distinct but related cancer types. We compare our proposed method to two baseline approaches and describe how this framework can help to mitigate data shortages in cancer genomics learning problems.

Learning clinical outcomes from heterogeneous genomic datasets using adversarial and multi-task learning

Safoora Yousefi

*Department of Computer Science, Emory University
Atlanta, GA, USA*

SAFOORA.YOUSEFI@EMORY.EDU

Amirreza Shaban

*College of Computing, Georgia Institute of Technology
Atlanta, GA, USA*

AMIRREZA@GATECH.EDU

Mohamed Amgad

*Department of Biomedical Informatics, Emory University School of Medicine
Atlanta, GA, USA*

MTAGELD@EMORY.EDU

Ramraj Chandradevan

*Department of Computer Science, Emory University
Atlanta, GA, USA*

CRAMRAJ8@GMAIL.COM

Lee A. D. Cooper

*Department of Pathology, Northwestern University Feinberg School of Medicine
Chicago, IL, USA*

LEE.COOPER@NORTHWESTERN.EDU

Abstract

Neural networks have shown promise in predicting clinical outcomes from high-dimensional genomic data, but their potential may be limited due to insufficient training data. This is particularly true for predicting clinical outcomes, where a large portion of the samples may have incomplete clinical followup. In this paper, we show how multi-task and adversarial learning approaches can be used to overcome data shortages and improve accuracy in predicting clinical outcomes. We first show how multi-task learning can significantly improve prediction of overall survival in breast cancer by combining microarray and RNA-sequencing gene expression profiles from the METABRIC and TCGA datasets. We improve prediction accuracy further using an adversarial penalty to encourage the network to learn a consistent representation across datasets. We also demonstrate the utility of this approach in leveraging samples with multiple clinical outcomes like time to progression and overall survival, and in combining data from distinct but related cancer types. We compare our proposed method to two baseline approaches and describe how this framework can help to mitigate data shortages in cancer genomics learning problems.

1. Introduction

Since the emergence of high-throughput proteomic and genomic platforms, the volume of molecular data produced has been increasing exponentially (Stephens et al., 2015). These platforms can generate tens of thousands of genetic, epigenetic, transcriptomic, or proteomic features from a single biopsy, and the ability to generate this data has so far outpaced the ability to translate it into clinically-actionable information, as typically only a handful of

molecular features are still used in diagnosis or prognostication (Bailey et al., 2018; Van De Vijver et al., 2002; Network, 2015).

Machine-learning has emerged as a powerful tool for analyzing high-dimensional data, with open software tools that enable scalable and distributed data analysis. A sub-field of machine learning, known as deep learning, has recently achieved remarkable success in learning from high dimensional images and sequences (LeCun et al., 2015). It involves artificial neural networks with several processing layers that learn to transform data into highly-predictive representations for specific prediction tasks. There are several challenges in applying neural networks to genomic data (Min et al., 2017). The more parameters a machine learning model has, the more independent samples it requires for training (Abu-Mostafa, 1989), and neural networks often have many thousands or millions of parameters due to their many interconnected layers. Cancer genomic datasets often have small sample sizes (typically hundreds of samples), and much larger dimensionality (tens of thousands of features), presenting a challenging scenario for machine learning algorithms. This data insufficiency issue is further pronounced in survival analysis, where large fractions (e.g. 87% in TCGA breast cancer data) of available samples may have incomplete labels. Several approaches have been employed to mitigate data insufficiency including dimensionality reduction, feature selection, data augmentation, and transfer learning (Ching et al., 2018).

An alternative approach to mitigate data insufficiency is to integrate multiple datasets to increase the training set size (e.g. two gene expression studies of breast cancer). Challenges in integrating datasets range from dealing with normal technical biases on the same platform used at different sites, to integrating data from entirely different platforms (sequence-based versus array-based), and dealing with different cohort compositions. Cohorts from multiple sites or studies often have different demographic or clinical characteristics like stage. This means that naively combining heterogeneous cohorts is both difficult and may degrade model accuracy due to batch effects (Tom et al., 2017). A significant amount of work has been done in the area of normalizing datasets for integration and removing batch effects. Many of these methods are based on linear regression and singular value decomposition, and make numerous assumptions such as orthogonality of the batch effect and biological variation, the ability of humans to distinguish between batch effects and biological effects, and assumptions on the batch structure (see for example Leek et al. (2010); Haghverdi et al. (2018)). Another limitation of such methods is that they do not take a learning objective into account to distinguish between relevant variations and batch effects.

In addition to integrating data from studies involving the same primary site, data-hungry algorithms may benefit from combining data from different cancer types to increase training set size Yousefi et al. (2017). Cancers originating from different primary sites are known to have significant differences in transcriptional profiles, although many share similar genetic alterations and altered pathways (Hoadley et al., 2018; Bailey et al., 2018; Hanahan and Weinberg, 2011). The idea of combining datasets from different primary sites relies on the premise that models of sufficient complexity and constraints can identify and exploit these similarities among the surrounding noise to improve outcome predictions.

Finally, one could tackle the problem of censoring to some extent by combining multiple outcomes for the same samples in training. Although clinical outcomes like disease-progression and overall survival are typically highly correlated, including both may provide additional observations for training machine learning models.

In this paper we describe a multi-task and adversarial representation learning approach for mitigating data insufficiency in clinical outcomes prediction. This flexible approach enables a variety of beneficial data integration tasks, from integrating multiple datasets derived from the same disease, to leveraging multiple clinical endpoints, and integrating datasets from different diseases. Where the multi-task learning component of this approach enables learning of a representation that is predictive for multiple learning objectives, the adversarial component encourages the learning of a representation that is robust to cohort-specific noise that would harm prediction accuracy. Since these approaches are based on neural networks, integration is learned in an unbiased manner and optimally for the prediction problem at hand. We show how these approaches provide significant improvements in prediction accuracy in several scenarios, including integrating early and late stage breast cancer datasets across array and sequencing based platforms, leveraging both overall survival and time-to-progression outcomes in breast cancer, and in combining datasets from different primary cancers. We analyze these results by looking inside both single task and multi-task adversarial models and using gene set enrichment analysis to see what pathways and biological themes they emphasize. We also show how these themes explain gains in prediction accuracy.

2. Materials and Methods

2.1. Survival analysis with Cox proportional hazards model

Survival analysis refers to any problem where the variable of interest is time to some event, which in cancer is often death or progression of disease. Time-to-event modelling is different from ordinary regression due to a specific type of missing data problem known as censoring. Incomplete or censored observations are important to incorporate into the model since they could provide critical information about long-term survivors (Harrell Jr, 2015). The most widely used approach to survival analysis is the semi-parametric Cox proportional hazards model (Cox, 1972). It models the hazard function at time s given the predictors x_i of the i th sample as:

$$h(s|x_i, \beta) = h_0(s)e^{\beta^\top x_i} \quad (1)$$

The model parameters β are estimated by minimizing Cox’s negative partial log-likelihood:

$$L_{cox}(X, Y, \beta) = - \sum_{x_i \in U} \left(\beta^\top x_i - \log \sum_{j \in R_i} e^{\beta^\top x_j} \right) \quad (2)$$

where $X = \{x_1, \dots, x_N\}$ are the samples, and $Y = \{E, S\}$ represents label vectors of event or last follow-up times $S = \{s_1, \dots, s_N\}$ and event status $E = \{e_1, \dots, e_N\}$. For censored samples ($e = 0$), s represents time of last follow-up while for observed samples ($e = 1$), it represents event time. The outer sum is over the set of uncensored samples U and R_i is the set of *at-risk* samples with $s_j \geq s_i$. The baseline hazard $h_0(t)$ is cancelled out of the likelihood and can remain unspecified.

A non-linear alternative to Cox regression is SurvivalNet (Yousefi et al., 2016, 2017), a fully connected artificial neural network f_W with parameters W that replaces X in Equation 2 with its non-linear transformation $f_W(X)$. SurvivalNet has been shown to outperform

other common survival analysis techniques such as random survival forests (Ishwaran et al., 2008) and Cox-ElasticNet.

This paper proposes two multi-task learning models built upon SurvivalNet to learn task-invariant representations from heterogeneous data sources. These models will be discussed in detail in section 2.9.

2.2. Multi-task learning for survival analysis

Both theoretical and empirical studies show that learning multiple related tasks simultaneously often significantly improves performance relative to learning each task independently (Baxter, 2000; Ben-David and Schuller, 2003; Caruana, 1997). This is particularly the case when only a few samples per task are available, since with multi-task learning, each task has more data to learn from. Multi-task learning has been applied to many areas of machine learning including computer vision (Zhang et al., 2014), natural language processing (Collobert and Weston, 2008), and survival analysis (Wang et al., 2017; Li et al., 2016).

Following Pan and Yang (2010), we provide a classification of multi-task learning problem settings in cancer survival analysis. Let us first define the terms *domain* and *task*. A domain is a pair $\{\mathcal{X}, P(X)\}$ which includes a feature space and a marginal probability distribution where $X = \{x_1, \dots, x_n\} \in \mathcal{X}$. A task $\{\mathcal{Y}, P(Y|X)\}$ consists of a label space and a conditional probability distribution function. $P(Y|X)$ is the ultimate predictive function that is not observed but can be learned from training data. Multi-task learning, by definition, involves different tasks, that is different $P(Y|X)$, or even different label spaces. With that in mind, we focus on the following multi-task learning scenarios in this paper (Sections 3.1, 3.2, and 3.3). :

1. Different $P(X)$: Data for the tasks come from different distributions. Examples include:
 - Standard gene expression data and progression-free survival labels are available for all cohorts, but the cohorts are diagnosed with different cancer types.
 - Standard gene expression data and progression-free survival labels are available for all cohorts, and the cohorts are diagnosed with the same cancer types but belong to different studies/hospitals.
2. Different $P(Y|X)$: All tasks are the same in nature, but the conditional distribution of labels are different. For example, learning overall survival and progression-free survival simultaneously for the same cohort of patients falls under this category.

We do not consider scenarios with different feature spaces \mathcal{X} or label spaces \mathcal{Y} , although they provide interesting directions for future work.

The general form of the loss function when learning T tasks simultaneously is:

$$L(Y, X, W) = \sum_{t=1}^T L_t(y^t, g^t(W^t, X^t)) + \gamma\lambda(Y, X, W) \quad (3)$$

l_t and W^t , respectively, are the loss function and the parameters of task t . $Y = \{Y^1, \dots, Y^T\}$ and $X = \{X^1, \dots, X^T\}$ are the combined input data of all t tasks. g^t indicates the prediction

function corresponding to task t , and λ is a regularization or auxiliary function that captures task relatedness assumptions, examples of which include cluster norm (Jacob et al., 2009), and $\ell_{2,1}$ norm (Argyriou et al., 2007). γ is a weight parameter controlling the importance of the auxiliary function.

Previous work has applied multi-task learning under different task relatedness assumptions to train Cox’s proportional hazards model using multiple genomic data sources (Wang et al., 2017; Li et al., 2016). In this paper, our main assumption is that gene expression data lies on a lower dimensional subspace that can be utilized in several prognostic tasks. We will enforce this assumption via parameter sharing and the bottleneck architecture of our models as shown in Figure 1. Moreover, In section 2.9 we describe how an adversarial classification objective can be used as auxiliary function λ to encourage task-invariant representation learning.

For simplicity, we consider settings with one target task and one auxiliary task (T=2). Ground truth labels are available for both tasks, and the goal is to make better predictions on the target task by learning relevant information from the auxiliary task.

2.3. Adversarial representation learning

The idea of using adversarial learning to match two distributions was first proposed by Goodfellow et al. (2014) for training generative models. In generative adversarial models, a generator aims to generate realistic data to mislead a discriminator that is simultaneously trained to distinguish between real and generated data. This competition drives the two components of the model to improve, until the generated data distribution is indistinguishable from the real data distribution.

This idea has been applied to unsupervised domain adaptation in several applications including natural language processing and computer vision, with varying design choices including parameter sharing, type of adversarial loss, and discriminative vs. generative base model (Ganin and Lempitsky, 2015; Ganin et al., 2016; Tzeng et al., 2015; Liu and Tuzel, 2016; Tzeng et al., 2017).

We adapt this idea to multi-task learning to encourage our proposed model to learn task-invariant representations. A cohort discriminator is trained to assign samples to their cohort. Simultaneously, a SurvivalNet is adversarially trained to confuse the discriminator by learning a representation of data where samples from different cohorts are indistinguishable (in addition to learning to predict survival). This competition will teach SurvivalNet to avoid learning cohort-specific noise. See section 2.9 for a formal definition of the proposed models.

2.4. Software

All software and parameters to reproduce the results presented in this section are publicly available at [GITHUB LINK].

2.5. Training, Model Selection, and Validation

Multitask learning experiments involve a target dataset (the dataset where performance will be evaluated) and an auxiliary dataset used to supplement the training samples from

the target dataset. The target dataset samples are split between the testing set (%40) and the training and model selection sets (%60). Training and model selection are performed by combining target samples with samples from the auxiliary dataset (the integrated training/selection set). This integrated set is randomly split into five folds, and cross validation is performed to identify the best model hyper-parameters. The best-performing model was then applied to the held-out target testing samples to measure prediction accuracy using Harrell’s concordance index (c-index).

When integrating multiple clinical outcomes like OS and PFI in a single dataset, one of these outcomes is the target and the other is the auxiliary. We exclude the target testing and validation samples from the auxiliary data in each cross validation experiment, maintaining non-overlapping training, validation and testing sets.

The above validation procedure was repeated for 30 random splits of the target dataset to generate 30 c-index performance measurements. The performance of models was compared using the re-sampled t-test and re-sampled paired t-test (Dietterich, 1998) for significance analysis and to derive confidence intervals.

All neural networks shared a 50-unit single hidden layer architecture. A grid search was performed to select the optimal learning rate (0.00001-0.001), dropout regularization rate (0-0.9), and ℓ_2 -norm regularization rate(0.001, 1.0) of the neural network weights W , and the weight of adversarial discriminator loss γ (0.1-100). Experiments involving Cox elastic net regression tuned the regularization weight λ (0.001-100) and the mixture coefficient α (0-1) using the same procedure.

2.6. Evaluation Metric

We measured model performance using *concordance index* (CI) that captures the rank correlation of predicted and actual survival (Harrell Jr et al., 1982), and is given by:

$$CI(\beta, X) = \sum_P \frac{I(i,j)}{|P|} \quad (4)$$

$$I(i,j) = \begin{cases} 1, & \text{if } r_j > r_i \text{ and } t_j > t_i \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Where P is the set of orderable pairs. A pair of samples (x_i, x_j) is orderable if either the event is observed for both x_i and x_j , or x_j is censored and $t_j > t_i$. Intuitively, CI measures the pairwise agreement of the prognostic scores r_i, r_j predicted by the model and the actual time of event for all orderable pairs. Optimizing the cox partial likelihood (Equation 2) has been shown to be equivalent to optimizing CI (Steck et al., 2008).

2.7. Data

The Cancer Genome Atlas (TCGA) provides publicly available clinical and molecular data for 33 cancer types. Experiments in this paper use TCGA gene expression features from the Illumina HiSeq 2000 RNA Sequencing V2 platform. The significant challenge that high-dimensional gene expression profiles present for machine learning algorithms was the motivation in choosing this platform to explore multitask adversarial learning. The clinical data from the TCGA PanCancer Atlas project includes both overall survival (OS) and progression free interval (PFI) labels, with varying degrees of availability and extent of

ensorship for different primary cancer sites (Liu et al., 2018). PFI is defined as the period from the date of diagnosis until the date of the first occurrence of a new tumor-related event, which includes progression of the disease, distant metastasis, or death. OS is the period from the date of diagnosis until the date of death from any cause. Since patients generally suffer from disease progression or recurrence before dying, PFI requires shorter follow-up times and has higher event rate. Additionally, OS is a noisy signal due to deaths from non-cancer causes. Therefore, wherever possible, PFI is used as the outcome variable.

For experiments involving multiple diseases, we selected TCGA cancer types that have a PFI event rate higher than 20%. We further filtered these datasets using Cox elastic net regression to measure outcome label quality, keeping those datasets with a 60% or higher concordance index, selecting adrenocortical carcinoma (ACC), cervical squamous cell carcinoma (CESC), lower-grade glioma (LGG), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), mesothelioma (MESO), and pancreatic adenocarcinoma (PAAD). ACC and MESO could not be used as target cohorts since their small sample sizes did not allow for reliable model evaluation. All of the mentioned cancer types were used as auxiliary cohorts.

For breast cancer outcome prediction experiments, we combine TCGA used METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) (Curtis et al., 2012) gene expression and clinical data in section 3.1. This data was obtained from the cBioPortal website Gao et al. (2013). Since METABRIC comes with OS labels only, OS was used as the outcome variable in this section.

2.8. Cohort Selection

TCGA breast invasive carcinoma (BRCA) was used in section 3.1 as target cohort. In section 3.2 we perform multi-task learning experiments on every possible pair of target and auxiliary cohorts chosen from a subset of cancer types. Out of the 33 TCGA cancer types, we selected those with PFI event rate higher than 20%. We used the performance of Cox-ElasticNet (Park and Hastie, 2007) on each of these cancer types as a measure of outcome label quality, and used only those cancer types where Cox-ElasticNet achieved a c-index of 60% and higher, leaving us with adrenocortical carcinoma (ACC), cervical squamous cell carcinoma (CESC), lower-grade glioma (LGG), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), mesothelioma (MESO), and pancreatic adenocarcinoma (PAAD). ACC and MESO could not be used as target cohorts since their small sample sizes did not allow for reliable model evaluation. All of the mentioned cancer types were used as auxiliary cohorts.

We discarded samples that did not have gene expression data or outcome labels. A summary of sample sizes and event rates of datasets after this preprocessing step is given in Table 1. Z-score normalization and 3-NN missing data imputation were performed on gene expression data. No further feature selection or dimensionality reduction was performed. In section 3.1, we found the intersection of Hugo IDs present in both BRCA and METABRIC datasets (17272 genes), and discarded the genes that were absent in either dataset. We used the log of BRCA gene expression data since METABRIC provides log-expression values. BRCA survival times were converted from days to months to be comparable with METABRIC.

Dataset Name	Number of Samples	Numbr of Features	Event Rate	Event Type
ACC	79	20531	52%	PFI
CESC	304	20531	23%	PFI
KIRC	533	20531	30%	PFI
KIRP	514	20531	37%	PFI
LGG	288	20531	20%	PFI
MESO	84	20531	70%	PFI
PAAD	178	20531	58%	PFI
BRCA	1094	20531	13%	OS
METABRIC	1903	24368	33%	OS

Table 1: Summary of datasets.

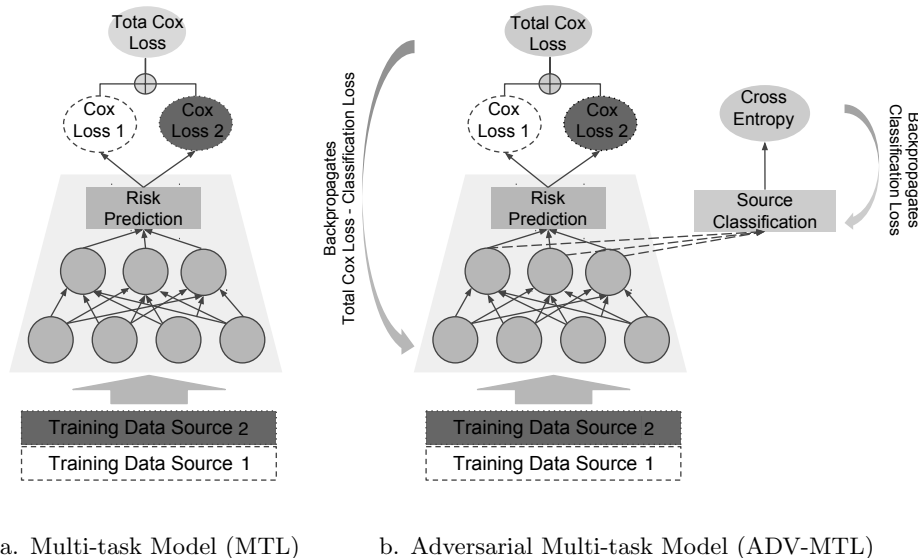


Figure 1: Model architectures used.

2.9. Proposed Methods

In cases where the target and auxiliary tasks are similar and their corresponding samples come from similar distributions, a natural approach is to simply combine (i.e. concatenate) the target and auxiliary datasets and train a single-task model on the combined training data, as done in [Yousefi et al. \(2017\)](#). We implement this approach as a baseline using both Cox-ElasticNet and SurvivalNet to provide two performance baselines.

But the assumption that the two cohorts come from the same distribution rarely holds. Comparisons of survival time between pairs of samples are integral to the Cox log-likelihood loss function. When one naively combines datasets to train a model with a single Cox loss, in addition to comparisons within target cohort and within auxiliary cohort, comparisons

between these cohorts contribute to the loss. Since the difference between distributions of these cohorts could be due to clinically insignificant factors (such as batch effects), these between-cohort comparisons could be misleading in training. Our first proposed model aims to eliminate this potentially misleading signal from the training process via multi-task learning:

Multi-task learning (MTL): This proposed extension of SurvivalNet model comprises one Cox loss node per each task. Each Cox loss node is responsible for one cohort only, so that only within-cohort comparisons contribute to the loss (See Figure 1a). The objective function of the MTL model is the sum of all Cox losses:

$$L_{MTL} = \sum_{t=1}^T L_{Cox}(f_W(X^t), Y^t, \beta) \tag{6}$$

where f_W is the SurvivalNet model. All parameters of MTL, β and W , are shared among tasks.

Although we are encouraging sparse representation learning via the bottleneck architecture of the MTL model, that does not adequately force the model to learn a task invariant representation. The model may learn a sparse representation, but still have enough parameters to be able to discriminate between samples from different cohorts and process them differently. The adversarial model described below addresses this limitation.

Adversarial model (ADV): This models extends SurvivalNet by addition of an adversarial cohort classification loss. Let $X_{comb} = \{x_1, \dots, x_M\}$ and $Y_{comb} = \{y_1, \dots, y_M\}$ denote the combination of all X^t and Y^t , respectively, including M samples in total. A set of one-hot vectors $Y_D = \{d_1, \dots, d_M\}$ indicate cohort membership, so that $d_{it} = 1$ means that the i th sample belongs to the t th cohort. A cohort discriminator is trained to assign the transformed samples $z_i = f_W(x_i)$ to the cohort they belong to. This component of the model is a multi-class logistic regression with a softmax cross-entropy loss. It comprises a linear transformation g_θ mapping z_i to a T -dimensional vector, where T is the number of tasks, and a softmax function that transforms the result to a T -dimensional vector of probabilities. The predicted probability that sample i belongs to cohort t is given by:

$$\hat{d}_{it} = \frac{e^{g_\theta(z_i)_t}}{\sum_{k=1}^T (e^{g_\theta(z_i)_k})},$$

and the objective function of the discriminator L_D is the cross-entropy between predicted probabilities and cohort labels:

$$L_D(f_W(X_{comb}), Y_D, \theta) = \gamma \sum_{i=1}^M \sum_{t=1}^T -d_{it} \log \hat{d}_{it} \tag{7}$$

This loss function only trains the parameters of the discriminator, namely θ the parameters of the linear function g_θ .

Simultaneously, a risk predictor is adversarially trained to learn a cohort-invariant representation that misleads the cohort classifier, in addition to learning to predict risk of event. The objective function of the risk predictor component of the model is:

$$L_R = L_{Cox}(f_W(X_{comb}), Y_{comb}, \beta) - \gamma L_D(f_W(X_{comb}), Y_D, \theta) \tag{8}$$

L_R trains the parameters of the risk predictor β as well as W . By updating W with an objective function that is the opposite of that of the discriminator, we encourage learning a representation of data in which samples from different cohorts are indistinguishable. γ controls the contribution of the adversarial loss to representation learning.

Adversarial multi-task model (ADV-MTL): Combining the MTL and ADV model described above, we allocate one Cox loss node for each cohort and additionally employ an adversarial cohort classification (See Figure 1b). The discriminator loss function is the same as given by Equation 7 while the rest of the model is trained with the following objective:

$$L_R = \sum_{t=1}^T L_{Cox}(f_W(X^t), Y^t, \beta) - \gamma L_D(f_W(X_{comb}), Y_D, \theta) \quad (9)$$

3. Results

3.1. Combining two breast cancer cohorts

This section investigates the integration of two breast cancer cohorts from independent studies. We use BRCA as target and METABRIC as auxiliary cohort. Both of these cohorts are diagnosed with breast cancer, and have overall survival labels. In such cases where similar biological processes determine the outcomes, one would expect naively pooling cohorts together to lead to better predictions on each of the cohorts. This is the expectation particularly in this case where the auxiliary cohort has twice the number of samples as the target cohort and three times the event rate. We train SurvivalNet and Cox-ElasticNet on the naive combination of BRCA and METABRIC as a two baselines, and compare the results to MTL, ADV, and ADV-MTL models as shown in Figure 2.

Surprisingly, we observe that simply adding METABRIC to training data has no effect on prediction c-index on BRCA ($p=0.83$). This implies that the distributions of the two datasets are so different that comparisons made between them are not providing useful insights to the model. Therefore, we eliminate these between-cohort comparisons by training a MTL model and observe a significant improvement ($p=3e-4$). Addition of adversarial classification loss to SurvivalNet lead to a slight improvement over naive SurvivalNet, while combining the MTL and ADV approaches into ADV-MTL significantly outperforms all other methods. The significance of the improvement achieved by ADV-MTL from a machine learning standpoint is discussed in section 3.5.

We measured some of the clinical and histological differences between METABRIC and TCGA-BRCA cohorts (see Fig. 3). Kaplan-Meier analysis and log-rank testing of progression-free survival revealed a significant difference between the two cohorts ($p=0.01$). Clinical variable bar charts show remarkable difference in the distribution of disease stage. Additionally, slight differences in age and histological subtypes can be observed. To measure differences in distribution of features directly used by the machine learning models, we visualized the log expression values for a subset of the Oncotype DXTM assay [Sparano et al. \(2015\)](#) in both cohorts.

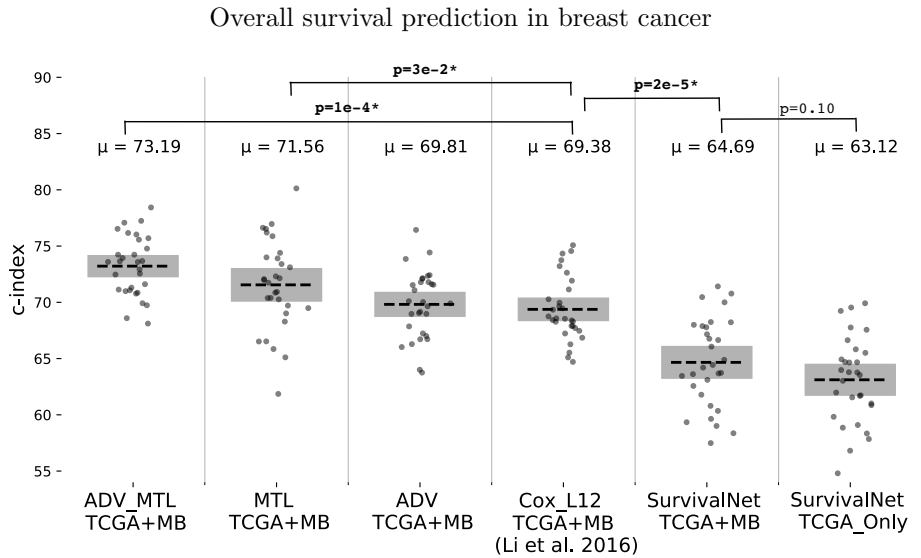


Figure 2: One solution to data insufficiency is to obtain similar cohorts and combine them with target training data. METABRIC gene expression and clinical data were combined with BRCA training data to improve performance on BRCA using models described in section 2.9. TCGA_only refers to performance of SurvivalNet when trained on TCGA only. The horizontal line indicates the performance of Cox-Elasticnet trained on naive combination of cohorts. Shaded areas indicate 95% confidence intervals. Paired t-test p-values are provided for some comparisons.

3.2. Combining cohorts with different cancer types

We repeat the experiments of section 3.1 this time using cohorts diagnosed with different cancer types. As explained in section 2.8, we use five cancer types as target and seven cancer types as auxiliary cohorts, performing experiments on each possible pair of target and auxiliary cohorts. Results of these experiments are summarized in Tables 2 and 3 in terms of average c-index achieved on target test set.

Table 2 summarizes the result of training SurvivalNet on the naive combination of target and auxiliary cohorts (NAIVE). The last column provides c-index of SurvivalNet after training on target cohort only. This naive approach leads to either significant ($p < 0.01$) deterioration or no significant difference in performance compared to SurvivalNet trained on target only, achieving significant improvement only in one case. ADV-MTL, on the other hand, achieves significant improvement over target-only setting in 10 cohort combinations (See Table 3).

In Fig. 4 we show the results of comparison between ADV-MTL and Cox- $\ell_{2,1}$ in terms of c-index achieved on target test set. In each set of experiments, one cancer type is chosen as target and all others are used as auxiliary data. In 3 out of 5, training on the combination of heterogeneous TCGA datasets with ADV-MTL model leads to significant improvement over single-task training of SurvivalNet with target training data only. Cox- $\ell_{2,1}$ achieves the same in 2 out of 5 cases. We did not observe any significant difference between ADV-MTL and Cox- $\ell_{2,1}$ in this set of experiments, except in experiments with PAAD where ADV-MTL significantly outperforms Cox- $\ell_{2,1}$ ($p = 7e-3$).

Comparison of TCGA and METABRIC breast cancer datasets

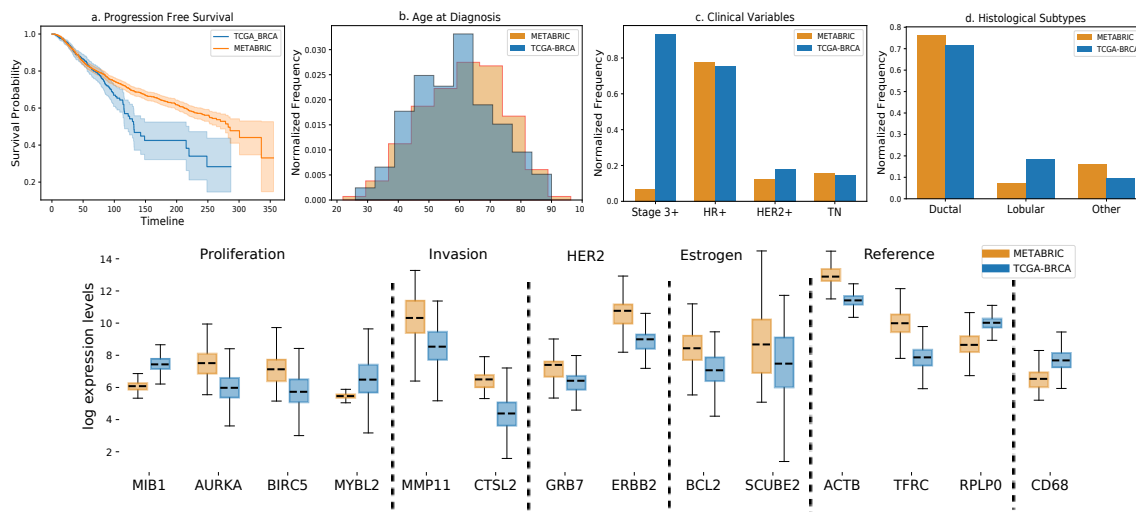


Figure 3: Differences between the TCGA and METABRIC datasets in terms of outcomes (A), cohort demographics (B, C), clinical variables (C), histologic subtypes (D), and gene expression levels (E). Progression-Free survival of the TCGA cohort is significantly worse than that of the METABRIC cohort, which is explained in part by the stark difference in proportion of advanced cases (stages 3 and 4) in the TCGA cohort. Abbreviations used in panels C and D: HR+, hormone-receptor positive cases (ER+ and/or PR+); TN, triple-negative cases (ER-, PR-, Her2-); Ductal, Infiltrating ductal carcinomas or carcinomas of no specified type (NST); Lobular, infiltrating lobular carcinomas. Genes shown in panel E are part of the OncotypeDx breast cancer recurrence panel (Sparano et al., 2015).

	+ACC	+CESC	+LGG	+KIRP	+KIRC	+PAAD	+MESO	Target-only
CESC	63.82 (±1.79)	-	62.28* (±1.91)	63.08 (±1.87)	61.49* (±1.33)	60.50* (±1.63)	65.43 (±1.44)	64.06 (±1.60)
LGG	71.12 (±0.84)	70.87 (±0.96)	-	69.45 (±1.39)	67.99* (±1.05)	71.27 (±0.86)	70.76 (±0.70)	70.62 (±0.74)
KIRP	75.14 (±1.98)	74.17 (±1.93)	73.48 (±1.84)	-	74.38 (±1.90)	74.58 (±1.36)	74.89 (±1.77)	74.62 (±1.83)
KIRC	71.20 (±1.14)	70.31 (±0.81)	70.62 (±1.08)	71.03* (±1.16)	-	70.28 (±1.13)	67.32* (±1.36)	69.69 (±1.06)
PAAD	63.77 (±1.43)	59.25* (±1.54)	63.22 (±1.43)	62.24 (±1.81)	58.80* (±2.13)	-	63.49 (±1.50)	63.16 (±1.84)

Table 2: Performance of SurvivalNet (c-index) when trained on naive combination of target data (rows) and auxiliary data (columns). Performance after training on target data only has been provided for reference. Numbers in parentheses are 95% confidence intervals. All improvements are boldened. * marks significant differences (p-values less than 0.01).

	+ACC	+CESC	+LGG	+KIRP	+KIRC	+PAAD	+MESO	Target-only
CESC	64.32 (±1.75)	-	63.41 (±1.57)	61.54 (±1.63)	62.92 (±1.84)	60.64 (±1.58)	65.26 (±1.57)	64.06 (±1.60)
LGG	71.04 (±0.93)	71.80* (±1.05)	-	71.23* (±0.71)	71.00 (±0.97)	72.24* (±0.78)	70.42 (±0.74)	70.62 (±0.74)
KIRP	76.49* (±1.89)	72.24* (±1.90)	74.91 (±1.68)	-	75.60 (±1.78)	75.0 (±1.64)	75.45 (±1.9)	74.62 (±1.83)
KIRC	72.59* (±0.92)	72.18* (±0.85)	72.50* (±0.74)	73.14* (±0.75)	-	71.83* (±1.0)	73.53* (±0.94)	69.69 (±1.06)
PAAD	64.48 (±1.66)	62.66* (±1.58)	64.17 (±1.33)	63.99 (±1.42)	64.20 (±1.49)	-	64.14 (±1.66)	63.16 (±1.84)

Table 3: Performance of ADV-MTL model when trained on combination of target data (rows) and auxiliary data (columns). Performance of SurvivalNet after training on target data only has been provided for reference. Numbers in parentheses are 95% confidence intervals. All improvements are bolded. * marks significant differences (p-values less than 0.01).

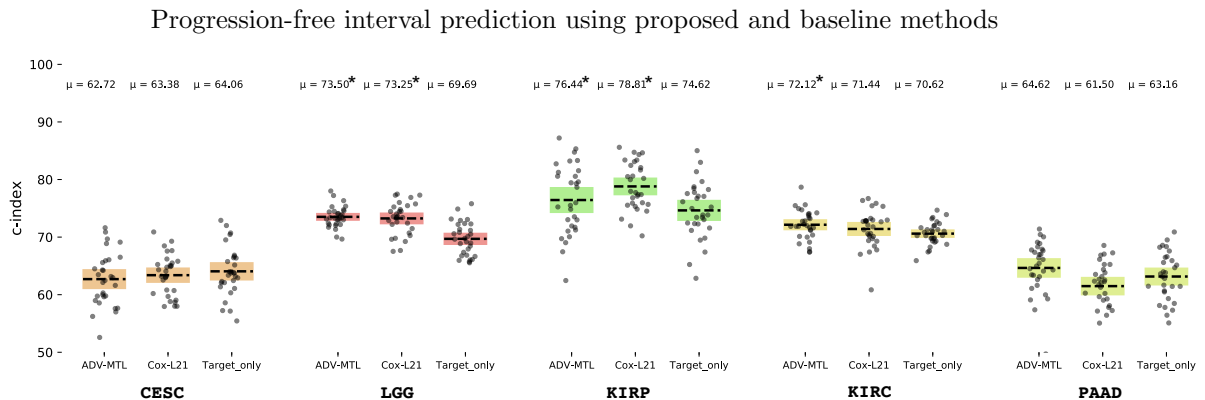


Figure 4: Survival prediction accuracy was improved by multi-task learning and adversarial representation learning on several benchmark datasets. * indicate significant improvement (p < 0.05) of multi-task methods over target-only setting.

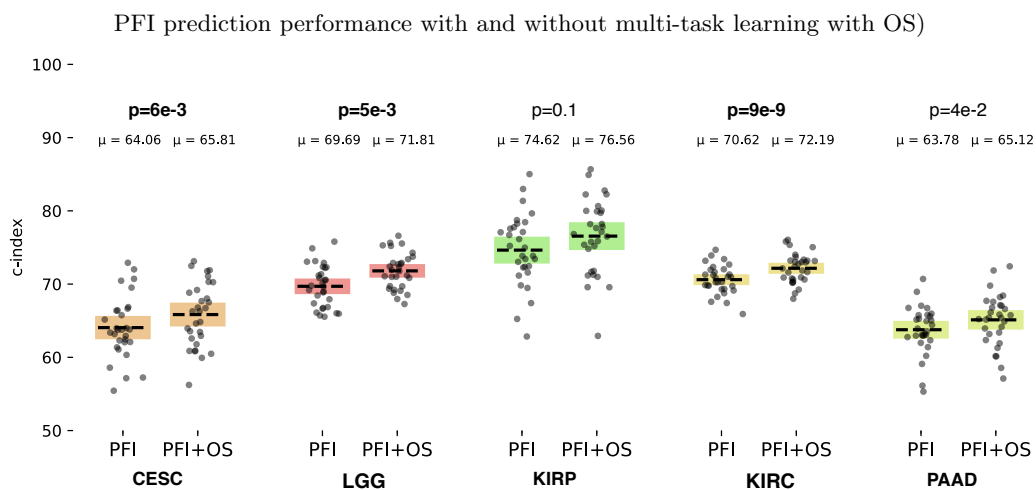


Figure 5: Progression-free interval (PFI) prediction performance with and without multi-task learning with overall survival (OS). Comparison performed on five different cancer types. Significance levels are shown on the plot for each comparison.

3.3. Combining multiple outcome labels for the same cohort

As discussed in section 2.8, TCGA samples may have multiple outcome labels. Overall survival (OS) labels are noisier, but simpler to obtain since the patients are either deceased or alive at the end of the study. As shown in Table 4, for some patients, a new tumor event is never observed (or recorded) during the study (censored PFI), while their overall survival outcome is observed (deceased by end of study). In such cases, overall survival could provide an extra supervision signal in training a predictive model that originally targets PFI prediction.

We use the MTL model to simultaneously use PFI and OS outcomes in training. Target and auxiliary domains are the same, so there is no need for adversarial domain-invariant representation learning. What differentiates the tasks from each other is the the predictive function $P(Y|X)$. Results are summarized in table 4. Multi-task learning with PFI and OS always leads to improved PFI prediction performance compared to single-task SurvivalNet trained with PFI labels only.

3.4. Model Interpretation

To gain insights into the internal mechanisms of neural network models, we used risk gradient backpropagation, described in (Yousefi et al., 2017), to interpret two breast cancer models: the baseline model (TCGA-only) and our best model (TCGA+METABRIC, ADV-MTL). Risk back-propagation uses partial derivatives of the risk prediction with respect to the input data to obtain importance scores for the various input features. In order to motivate the use of partial derivatives as a measure of feature importance, let us start by looking at a linear model of risk:

Cancer type	PFI+OS c-index	Improvement over PFI	Censored PFI and observed OS
CESC	65.83	1.69%	5.26%
KIRC	76.55	2.12%	11.81%
KIRP	76.55	1.35%	5.19%
LGG	72.15	1.75%	1.75%
PAAD	65.12	1.34%	9.55%

Table 4: Progression-free survival (PFI) prediction performance with and without multi-task learning with overall survival (OS). Comparison performed on five different cancer types. Percent of samples in each cohort with censored PFI and observed OS is given in the last column.

$$f_{\beta}(x) = x^{\top} \beta \quad (10)$$

In the case of the above linear model, it is easy to take elements of the parameter vector β corresponding to each feature in x as a measure of importance of that feature. In a neural network, however, we are dealing with a highly non-linear risk function, which we can approximate using the first-order Taylor expansion in the neighborhood of a given sample x_0 :

$$g_W(x) = g_W(x_0) + \left. \frac{\partial g_W(x)}{\partial x} \right|_{x=x_0} (x - x_0)$$

$$g_W(x) = \left. \frac{\partial g_w(x)}{\partial x} \right|_{x=x_0} x + c$$

Where c is a constant with respect to x . The partial derivative of the neural network risk predictions with respect to the input act as the feature coefficients in the linear function f_{β} , providing a measure of importance for each input feature.

After ranking genes using the above feature importance scoring method, we performed Gene Set Enrichment Analysis (GSEA) using rankings from both models to investigate gene set-level/pathway-level associations (Subramanian et al., 2005) (Figure 6). Both models learn to assign favorable prognosis to genes that are part of immune response pathways, including Interferon Gamma response pathway, Allograft rejection gene set and IL-2-STAT5 signalling. This is to be expected, since the gene expression profiles are not 100% pure, and come from an admixture of tumor cells and the surrounding microenvironment, including stromal cells and tumor-infiltrating lymphocytes (Morris and Kopetz, 2016; Aran et al., 2015). Immune response is a known indicator of good prognosis in breast cancer, and reflects host response to the tumor (Savas et al., 2016). Genes like GCNT1, MYD88, CD40, FURIN, XBP1, HDAC9, IL1RL1, and EIF4E3 were emphasized. Both models also learn to assign low risk to genes from the early and late Estrogen response pathways, including CYP4F11 and GREB1. This is also expected, given that patients with positive hormone status (i.e. having carcinomas that express ER or PR receptors) are treated with hormone therapy. In contrast, hormone negative patients, especially those who also lack HER2

receptor, have much more limited therapeutic options, and subsequently suffer from worse prognoses (Plasilova et al., 2016).

Both models also correctly learn to assign high risk to genes that are part of proliferation pathways, including G2M checkpoint, E2F targets and Myc targets, as well as pathways related to hypoxia response, glycolysis, and angiogenesis. The ADV-MTL model seems to strongly emphasize proliferation pathways, enough so to reach FDR-adjusted statistical significance (unlike the TCGA-only model). Proliferation genes like the cell cycle regulators CCNA1 and CCNB2 (cyclin A1 and B2), ATF5, and KIF4A were picked by both models, although the ADV-MTL model had a slightly higher emphasis on their adverse prognostic role. More notably, the growth factor TGF β 1 and the tumor suppressor gene TP53 were much more emphasized by the ADV-MTL model, correctly so, as being associated with high and low risk, respectively.

Other than the obvious benefit from increased sample size, there appears to be a qualitative benefit from the inclusion of cases from the METABRIC dataset during the training procedure. From figure 3 it can be seen that TCGA cases are much more advanced than METABRIC; gene expression profiles from TCGA tumors can be thought to represent a later "snapshot" of the tumor landscape than equivalent/matched profiles from METABRIC. This contrast between the early cases from METABRIC and advanced cases from TCGA may have provided the ADV-MTL model with enough variability to learn to predict risk from subtle differences in gene expression, especially using proliferation-related genes.

3.5. Significance of Results

To provide an insight into the significance of the improvement achieved by our models, we look at the learning curve of SurvivalNet on two target datasets. Learning curves were obtained by training SurvivalNet on incrementally more training samples (using the same procedure described in section 3) and testing on a fixed sized test set (40% of data, consistent with the rest of experiments). As shown in Figure 7a, the performance improvement achieved by ADV-MTL over SurvivalNet trained on the full training dataset exceeds the improvement resulting from doubling the size of target training data (from 50% to 100% of training set). In case of breast cancer, ADV-MTL trained on METABRIC and only 10% of TCGA outperforms SurvivalNet trained on all of TCGA when tested on TCGA. Considering the cost of obtaining labeled samples, we believe the ADV-MTL approach can save researchers resources by enabling the integration of heterogeneous cohorts in training.

3.6. Future Work

Data insufficiency is known to hinder successful application of machine learning models to high dimensional genomic data. In this paper, we study two neural network models for learning from combinations of heterogeneous datasets to tackle this issue. Significant improvement was achieved by identifying and integrating independent cohorts diagnosed with the same cancer type, and training the proposed models with the integrated data. We show how even genomic data obtained from different tumor sites can be used to augment training data and improve performance. Moreover, different outcome labels of the same cohort

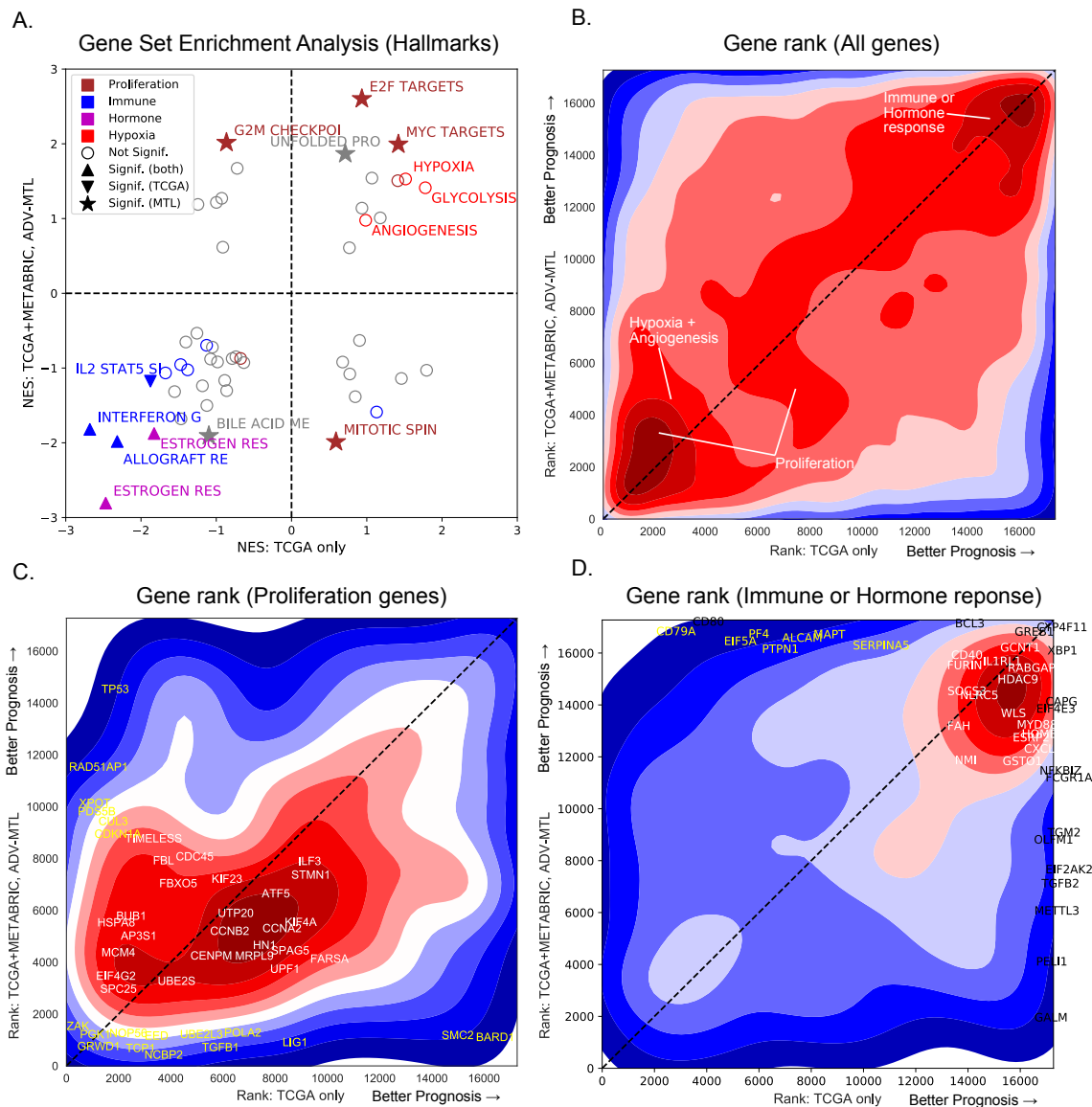


Figure 6: Model interpretation using risk backpropagation and Gene Set Enrichment Analysis (GSEA). A. GSEA results using the Hallmarks MSigDB collection. The Normalized Enrichment Score (NES) is plotted for the same gene set for the TCGA-only model (x-axis) and the final TCGA+METABRIC ADV-MTL model (y-axis). Higher enrichment scores indicate higher risk (worse prognosis). Color is used to indicate the broad category to which a particular gene set belongs, while shape is used to indicate significance of NES scores at a FDR threshold of 0.1. B-D. Comparing gene ranks in the TCGA-only versus the ADV-MTL models. Higher rank indicates lower risk (better prognosis). Density plots were used (red is higher density), and show regions of concordant ranks as densities at the lower-left and upper-right corners. A random sample of genes is also individually plotted in high- and low model concordance zones.

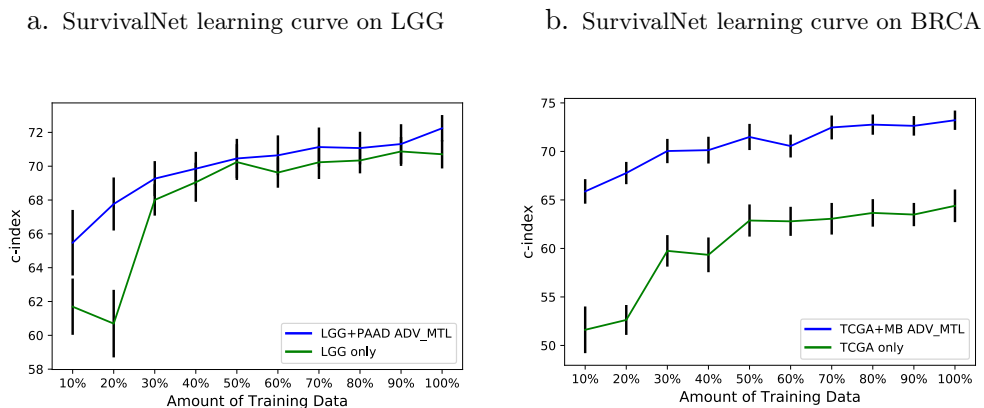


Figure 7: Learning curves of SurvivalNet on two datasets. Size of training data is gradually increased from 10% of the original size to 100%. The horizontal line depicts the performance of our ADV-MTL approach. Error bars correspond to confidence intervals of the mean c-index.

were used in multi-task learning to alleviate the outcome censoring issue and significant improvement was observed in most cases.

We show that the integration of heterogeneous datasets using our proposed method is a reasonable alternative to acquisition of new training data from the target distribution which may be expensive or impossible due to practical constraints (funding, availability, clinical setting, etc). The ideal solution to any data insufficiency issue is enhanced data collection and standardization efforts. However, in settings where this is impractical, employing techniques like ADV-MTL and MTL can help address this at no extra cost. While our work focused on combining datasets from the same feature space, future work may apply or extend the proposed models to scenarios 2 and 4 introduced in Section 2.2, namely multi-task learning using datasets with different feature spaces and/or label spaces. Studying different cancer subtypes (eg. breast cancer histologic subtypes) under a multi-task learning setting could also lead to improved prediction.

Acknowledgments

This work was funded by National Institutes of Health National Cancer Institute U01CA220401.

References

- Yaser S Abu-Mostafa. The vapnik-chervonenkis dimension: Information versus complexity in learning. *Neural Computation*, 1(3):312–317, 1989.
- Dvir Aran, Marina Sirota, and Atul J Butte. Systematic pan-cancer analysis of tumour purity. *Nature communications*, 6:8971, 2015.
- Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in neural information processing systems*, pages 41–48, 2007.

- Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.
- Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.
- Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.
- Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.
- Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.
- David R Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220, 1972.
- Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346, 2012.
- Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.
- Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd ICML*, 2015.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbiportal. *Sci. Signal.*, 6(269):pl1–pl1, 2013.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

- Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421, 2018.
- Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.
- Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
- Frank E Harrell Jr, Robert M Califf, David B Pryor, Kerry L Lee, Robert A Rosati, et al. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.
- Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.
- Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.
- Laurent Jacob, Jean-philippe Vert, and Francis R Bach. Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, pages 745–752, 2009.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733, 2010.
- Yan Li, Lu Wang, Jie Wang, Jieping Ye, and Chandan K Reddy. Transfer learning for survival analysis via efficient l2, l1-norm regularized cox regression. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 231–240. IEEE, 2016.
- Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.
- Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.
- Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.
- Jeffrey S Morris and Scott Kopetz. Tumor microenvironment in gene signatures: critical biology or confounding noise? *Clinical Cancer Research*, 22(16):3989–3991, 2016.

- Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.
- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
- Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.
- Magdalena L Plasilova, Brandon Hayse, Brigid K Killelea, Nina R Horowitz, Anees B Chagpar, and Donald R Lannin. Features of triple-negative breast cancer: Analysis of 38,813 cases from the national cancer database. *Medicine*, 95(35), 2016.
- Peter Savas, Roberto Salgado, Carsten Denkert, Christos Sotiriou, Phillip K Darcy, Mark J Smyth, and Sherene Loi. Clinical relevance of host immunity in breast cancer: from tils to the clinic. *Nature reviews Clinical oncology*, 13(4):228, 2016.
- Joseph A Sparano, Robert J Gray, Della F Makower, Kathleen I Pritchard, Kathy S Albain, Daniel F Hayes, Charles E Geyer Jr, Elizabeth C Dees, Edith A Perez, John A Olson Jr, et al. Prospective validation of a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, 373(21):2005–2014, 2015.
- Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pages 1209–1216, 2008.
- Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genetical? *PLoS biology*, 13(7):e1002195, 2015.
- Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.
- Jennifer A Tom, Jens Reeder, William F Forrest, Robert R Graham, Julie Hunkapiller, Timothy W Behrens, and Tushar R Bhangale. Identifying and mitigating batch effects in whole genome sequencing data. *BMC bioinformatics*, 18(1):351, 2017.
- Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.
- Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

- Marc J Van De Vijver, Yudong D He, Laura J Van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.
- Lu Wang, Yan Li, Jiayu Zhou, Dongxiao Zhu, and Jieping Ye. Multi-task survival analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 485–494. IEEE, 2017.
- Safoora Yousefi, Congzheng Song, Nelson Nauata, and Lee Cooper. Learning genomic representations to predict clinical outcomes in cancer. In *International Conference on Learning Representations (ICLR)*, 2016.
- Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E Lewis, Congzheng Song, David A Gutman, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 7(1):11707, 2017.
- Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014.

Chapter 5

Transfer Learning From Nucleus Detection To Classification In Histopathology Images, ISBI 2019

This chapter, mostly independent of the rest of the dissertation, describes an application of convolutional neural networks to solve a different clinical problem in cancer care: annotating histopathology images on the cell level. This chapter presents the following conference paper, not copied in this dissertation due to copyright considerations:

Safoora Yousefi and Yao Nie, *Transfer learning from nucleus detection to classification in histopathology images*. In Proceedings of International Symposium on Biomedical Imaging (ISBI) 2019 ©2019 IEEE.

Abstract. Despite significant recent success, modern computer vision techniques such as Convolutional Neural Networks (CNNs) are expensive to apply to cell-level prediction problems in histopathology images due to difficulties in providing cell-level supervision. This work explores the transferability of features learned by an object

detection CNN (Faster R-CNN) to nucleus classification in histopathology images. We detect nuclei in these images using class-agnostic models trained on small annotated patches, and use the CNN representations of detected nuclei to cluster and classify them. We show that with a small training dataset, the proposed pipeline can achieve superior nucleus detection and classification performance, and generalizes well to unseen stain types.

Bibliography

- Abadi, M. and Andersen, D. G. (2016). Learning to protect communications with adversarial neural cryptography. *arXiv preprint arXiv:1610.06918*.
- Abadi, M. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Abu-Mostafa, Y. S. (1989). The vapnik-chervonenkis dimension: Information versus complexity in learning. *Neural Computation*, 1(3):312–317.
- Abu-Mostafa, Y. S. (1990). Learning from hints in neural networks. *Journal of complexity*, 6(2):192–198.
- Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning. In *Advances in neural information processing systems*, pages 41–48.
- Baxter, J. (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39.
- Bellman, R. (1961). Curse of dimensionality. *Adaptive control processes: a guided tour*. Princeton, NJ.
- Bengio, Y. (2013). Deep learning of representations: Looking forward. In *International Conference on Statistical Language and Speech Processing*, pages 1–37. Springer.

- Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.
- Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186.
- Blum, A. and Rivest, R. L. (1988). Training a 3-node neural network is np-complete. In *Proceedings of the 1st International Conference on Neural Information Processing Systems*, pages 494–501. MIT Press.
- Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2017). Un-supervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, pages 89–99.
- Caruana, R. and De Sa, V. R. (1997). Promoting poor features to supervisors: Some inputs work better as outputs. In *Advances in Neural Information Processing Systems*, pages 389–395.
- Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk,

- H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220.
- Cruz-Roa, A., Basavanahally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., and Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Medical Imaging 2014: Digital Pathology*, volume 9041, page 904103. International Society for Optics and Photonics.
- De Laurentiis, M. and Ravdin, P. M. (1994). Survival analysis of censored data: neural network analysis detection of complex interactions between variables. *Breast cancer research and treatment*, 32(1):113–118.
- Dhungel, N., Carneiro, G., and Bradley, A. P. (2015). Deep learning and structured prediction for the segmentation of mass in mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–612. Springer.
- Dimopoulos, Y., Bourret, P., and Lek, S. (1995). Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters*, 2(6):1–4.

- Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.
- Du, M., Liu, N., and Hu, X. (2018). Techniques for interpretable machine learning. *arXiv preprint arXiv:1808.00033*.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115.
- Evgeniou, T., Michelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637.
- Faraggi, D. and Simon, R. (1995). A neural network model for survival data. *Statistics in medicine*, 14(1):73–82.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.
- Gao, S., Tibiche, C., Zou, J., Zaman, N., Trifiro, M., OConnor-McCourt, M., and Wang, E. (2016). Identification and construction of combinatorial cancer hallmark-based gene signature sets to predict recurrence and chemotherapy benefit in stage ii colorectal cancer. *JAMA oncology*, 2(1):37–45.
- Ginsbourger, D., Le Riche, R., and Carraro, L. (2007). A multi-points criterion for deterministic parallel global optimization based on kriging. In *NCP07*.
- Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256.

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.
- Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer treatment reports*, 69(10):1065–1069.
- Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421.
- Harrell Jr, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., Rosati, R. A., et al. (1982). Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546.
- Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.
- Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. (2017). Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*.
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, pages 841–860.
- Jacob, L., Vert, J.-p., and Bach, F. R. (2009). Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, pages 745–752.
- Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492.
- Kahn, S. D. (2011). On the future of genomic data. *science*, 331(6018):728–729.
- Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al. (2017). Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging*, pages 597–609. Springer.
- Khan, F. M. and Zubek, V. B. (2008). Support vector regression for censored data (svrc): a novel tool for survival analysis. In *2008 Eighth IEEE International Conference on Data Mining*, pages 863–868. IEEE.
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary

- multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106.
- Lawrence, S., Giles, C. L., Tsoi, A. C., and Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113.
- Le Cun, Y. (1989). Handwritten digit recognition with a back-propagation network. *Proceedings of the Annual Conference on Neural Information Processing Systems*.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.
- Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733.
- Leung, M. K., DeLong, A., Alipanahi, B., and Frey, B. J. (2015). Machine learning in genomic medicine: a review of computational problems and data sets. *Proceedings of the IEEE*, 104(1):176–197.
- Li, J., Lenferink, A. E., Deng, Y., Collins, C., Cui, Q., Purisima, E. O., O’Connor-McCourt, M. D., and Wang, E. (2010). Identification of high-quality cancer prognostic markers and metastasis network modules. *Nature communications*, 1:34.
- Li, Y., Wang, L., Wang, J., Ye, J., and Reddy, C. K. (2016). Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 231–240. IEEE.
- Lisboa, P. J., Wong, H., Harris, P., and Swindell, R. (2003). A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial intelligence in medicine*, 28(1):1–25.

- Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., and Van Der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6:26286.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.
- Martinez-Cantin, R. (2014). Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits. *The Journal of Machine Learning Research*, 15(1):3735–3739.
- Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. (2017). Unrolled generative adversarial networks. *International Conference on Learning Representations*.
- Močkus, J. (1975). On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer.
- Montavon, G., Orr, G., and Müller, K.-R. (2012). *Neural networks: tricks of the trade*, volume 7700. springer.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.
- Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., et al. (2018). A universal snp

- and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
- Rios, A., Kavuluru, R., and Lu, Z. (2018). Generalizing biomedical relation classification with neural adversarial domain adaptation. *Bioinformatics*, 34(17):2973–2981.
- Royston, P. and Altman, D. G. (2013). External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13(1):33.
- Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.
- Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, pages 35–47.
- Shivaswamy, P. K., Chu, W., and Jansche, M. (2007). A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 655–660. IEEE.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths

- for coxs proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1.
- Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.
- Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.
- Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., and Raykar, V. C. (2008). On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pages 1209–1216.
- Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In

- Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Tom, J. A., Reeder, J., Forrest, W. F., Graham, R. R., Hunkapiller, J., Behrens, T. W., and Bhangale, T. R. (2017). Identifying and mitigating batch effects in whole genome sequencing data. *BMC bioinformatics*, 18(1):351.
- Ture, M., Tokatli, F., and Kurt, I. (2009). Using kaplan–meier analysis together with decision tree methods (c&rt, chaid, quest, c4. 5 and id3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications*, 36(2):2017–2026.
- Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076.
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176.
- Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. A. (2011). Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial intelligence in medicine*, 53(2):107–118.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

- Wan, L., Zeiler, M., Zhang, S., Cun, Y. L., and Fergus, R. (2013). Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1058–1066.
- Wang, L., Li, Y., Zhou, J., Zhu, D., and Ye, J. (2017a). Multi-task survival analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 485–494. IEEE.
- Wang, P., Li, Y., and Reddy, C. K. (2017b). Machine learning for survival analysis: A survey. *arXiv preprint arXiv:1708.04649*.
- Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J., and Azen, S. (2000). Comparison of the performance of neural network methods and cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257.
- Xie, W., Noble, J. A., and Zisserman, A. (2018). Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, 6(3):283–292.
- Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, pages 2048–2057. JMLR.org.
- Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.
- Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer.