nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20167-3 OPEN

A meta-learning approach for genomic survival analysis

Yeping Lina Qiu[1,2], Hong Zheng[2], Arnout Devos[3], Heather Selby[2] & Olivier Gevaert[2,4]✉

RNA sequencing has emerged as a promising approach in cancer prognosis as sequencing data becomes more easily and affordably accessible. However, it remains challenging to build good predictive models especially when the sample size is limited and the number of features is high, which is a common situation in biomedical settings. To address these limitations, we propose a meta-learning framework based on neural networks for survival analysis and evaluate it in a genomic cancer research setting. We demonstrate that, compared to regular transfer-learning, meta-learning is a significantly more effective paradigm to leverage high-dimensional data that is relevant but not directly related to the problem of interest. Specifically, meta-learning explicitly constructs a model, from abundant data of relevant tasks, to learn a new task with few samples effectively. For the application of predicting cancer survival outcome, we also show that the meta-learning framework with a few samples is able to achieve competitive performance with learning from scratch with a significantly larger number of samples. Finally, we demonstrate that the meta-learning model implicitly prioritizes genes based on their contribution to survival prediction and allows us to identify important pathways in cancer.

[1] Department of Electrical Engineering, Stanford University, Stanford, USA. [2] Department of Medicine, Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, USA. [3] School of Computer and Communication Sciences, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. [4] Department of Biomedical Data Science, Stanford University, Stanford, USA. ✉email: ogevaert@stanford.edu

Cancer is a leading cause of death in the world. Accurate prediction of its survival outcome has been an interesting and challenging problem in cancer research over the past decades. Quantitative methods have been developed to model the relationship between multiple explanatory variables and survival outcome, including fully parametric models[1,2] and semi-parametric models, such as the Cox proportional hazards model[3]. The Cox model makes a parametric assumption about how the predictors affect the hazard function, but makes no assumption about the baseline hazard function itself[4]. In most real world scenarios, the form of the true hazard function is either unknown or too complex to model, making the Cox model the most popular method in survival analysis[5].

In clinical practice, historically, survival analysis has relied on low-dimensional patient characteristics, such as age, sex, and other clinical features in combination with histopathological evaluations such as stage and grade[6]. With advances in high-throughput sequencing technology, a greater amount of high-dimensional genomic data is now available and more molecular biomarkers can be discovered to determine survival and improve treatment. With the cost of RNA sequencing coming down significantly, from an average of $100M per genome in 2001 to $1k per genome in 2015[7], it is becoming more feasible to use this technology to prognosticate. Such genomic data often has tens of thousands of variables which requires the development of new algorithms that work well with data of high dimensionality.

To address these challenges, several implementations of regularized Cox models have been proposed[8–10]. A regularized model adds a model complexity penalty to the Cox partial likelihood to reduce the chance of overfitting. More recently, the increasing modeling power of deep learning networks has aided in developing suitable survival analysis platforms for high-dimensional feature spaces. For example, autoencoder architectures have been employed to extract features from genomic data for liver cancer prognosis prediction[11]. The Cox model has also been integrated in a neural network setting to allow greater modeling flexibility[12–15].

In studying a specific rare cancer's survival outcome, one interesting problem is whether it is possible to make use of the abundant data that is available for more common relevant cancers and leverage that information to improve the survival prediction. This problem is commonly approached with transfer-learning[16], where a model which has been trained on a single task (e.g., 1 or more abundant cancers) is used to fine-tune on a related target task (rare cancer). In survival analysis, transfer-learning has shown to significantly improve prediction performance[17]. Deep neural networks used to analyze biomedical-imaging data can also take advantage of information transfer from data in other settings. For example, multiple studies show that convolutional neural networks pretrained on ImageNet data can be used to build performant survival models with histology images[18,19].

In this context, meta-learning is an area in deep learning research that has gained much attention in recent years which addresses the problem of "learning to learn"[20,21]. A meta-learning model explicitly learns to adapt to new tasks quickly and efficiently, usually with a limited exposure to the new task environment. Such a framework may potentially adapt better than the traditional transfer-learning setting where there is no explicit adaptation incorporated in the pre-training algorithm. This problem setting with limited exposure to a new task is also known as few-shot learning: learn to generalize well, given very few examples (called shots) of a new task[17]. Recent advances in meta-learning have shown that, compared to transfer-learning, it is a more effective approach to few-shot classification[22,23], regression[20], and reinforcement learning[24,20]. In this study, we propose a meta-learning framework based on neural networks for survival analysis applied in a cancer research setting. Specifically, for the application of predicting survival outcome, we demonstrate that our method is a preferable choice compared to regular transfer-learning pre-training and other competing methods on three cancer datasets when the number of training samples from the specific target cancer is very limited. Finally, we demonstrate that the meta-learning model implicitly prioritizes genes based on their contribution to survival prediction and allows us to uncover biological pathways associated with cancer survival outcome.

Methods

Datasets. We use the RNA-sequencing data from The Cancer Genome Atlas (TCGA) pan-cancer datasets[25]. We remove the genes with NA values and normalized the data by log transformation and z-score transformation. The feature dimension is 17,176 genes after preprocessing. The data contains 9707 samples from 33 cancer types. The outcome is the length of survival time in months. 78% of the patients are censored, which means that the subject leaves the study before an event occurs or the study terminates before an event occurs to the subject.

Survival prediction model. To describe the effect of categorical or quantitative variables on survival time, several approaches are commonly considered[13]. The most popular method is the Cox-PH model[3], which is a semi-parametric proportional hazards model, where the patient hazards depend linearly on the patient features and the relative risks of the patients are expressed in the hazard ratios. Survival trees and random survival forests are an attractive alternative approach to the Cox models[26]. They are an extension of classification and regression trees and random forests for time-to-event data, and are fully non-parametric and flexible. Artificial neural networks (ANNs)-based models have also been used to predict survival, but the survival time is often converted to a binary variable or discrete variables and the prediction is framed as a classification problem[27,28]. To overcome the potential loss of accuracy in the previous methods, ANNs based on proportional hazards are recently developed. It is shown that when applied to high-dimensional RNA-seq data, the neural network extension of the Cox model achieves better performance than the Cox-PH (including Ridge and LASSO regularization), random survival trees, and other ANN-based models[13]. It can directly integrate the meta-learning optimization algorithm and is therefore the most suitable choice of model structure in our framework.

Meta-learning. Our proposed survival prediction framework is based on a neural network extension of the Cox regression model that relies on semi-parametric modeling by using a Cox loss[13]. The model consists of two modules: the feature extraction network and the Cox loss module (Fig. 1). We use a neural network with two hidden layers to extract features from the RNA sequencing data input, which yields a lower dimensional feature vector for each patient. The features are then fed to the Cox loss module, which performs survival prediction by doing a Cox regression with the features as linear predictors of the hazard[3]. The parameters of the Cox loss module $\boldsymbol{\beta}$ are optimized by minimizing the negative of the partial log-likelihood:

$$\mathcal{L}(\beta) = -\sum_{y_i=\text{uncensored}} \left[\mathbf{z}_i\beta - \log\left(\sum_{y_j \geq y_i} e^{\mathbf{z}_i\beta} \right) \right], \quad (1)$$

where y_i is the survival length for patient i, $\mathbf{z}_i$ contains the extracted features for patient i, and $\boldsymbol{\beta}$ is the coefficient weight vector between the features and the output. Since $\mathbf{z}_i$ is the output of the feature extraction module, it can be further

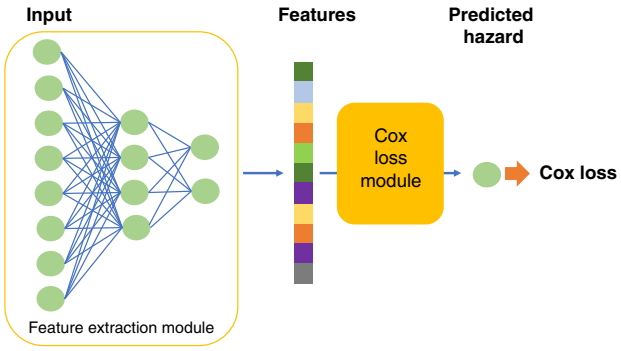


Fig. 1 Schematic showing the survival prediction model architecture. The model consists of a feature extraction module, and a Cox loss module.

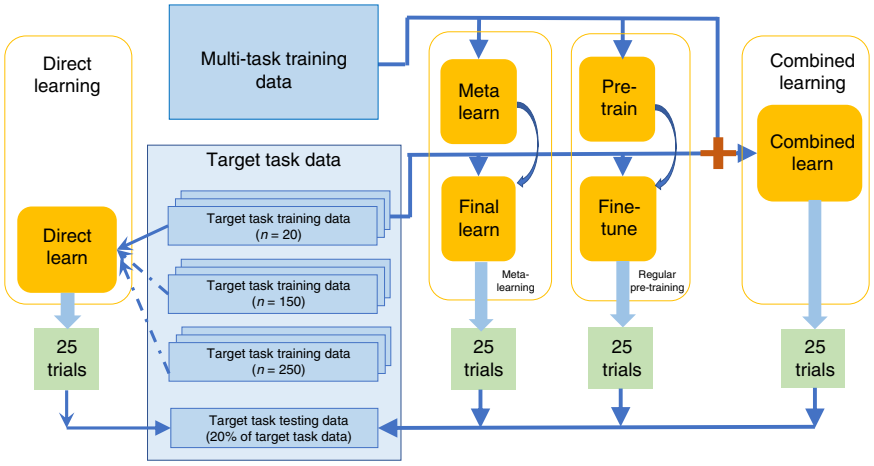


Fig. 2 Data flow schematic. 25 trials are conducted each for the meta-learning, regular pre-training, combined learning, and direct learning frameworks.

represented by

$$\mathbf{z}_i = f_{\varphi}(\mathbf{x}_i), \tag{2}$$

where $\mathbf{x}_i$ is the input predictor of patient i, f denotes a nonlinear mapping that the neural network learns to extract features form the predictor, and φ denotes the model parameters including the weights and biases of each neural network layer. The feature extraction module parameters φ and Cox loss module parameters β are jointly trained in the model. For convenience in the following discussion we denote the combined parameters as θ.

The optimization of parameters θ consists of two stages: a meta-learning stage, and a final learning stage. The meta-learning stage is the key process, where the model aims to learn a suitable parameter initialization for the final learning stage, so that during final learning the model can adapt efficiently to previously unseen tasks with a few training samples[29]. In order to reach the desired intermediate state, a first-order gradient-based meta-learning algorithm is used to train the network during the meta-learning stage[20,29].

Specifically, at the beginning of meta-learning training, the model is randomly initialized with parameter θ. Consider that the training samples for the meta-learning stage consist of n tasks T_τ, $\tau = 1, 2 \ldots n$. A task is defined as a common learning theme shared by a subgroup of samples. Concretely, these samples come from a distribution on which we want to carry out a classification task, regression task, or reinforcement learning task. The algorithm continues by sampling a task T_τ and using samples of T_τ to update the inner-learner. The inner-learner learns T_τ by taking k steps of stochastic gradient descent (SGD) and updating the parameters to θ_τ^k:

$$\begin{aligned} \theta_\tau^0 &= \theta \\ \theta_\tau^1 &\leftarrow \theta_\tau^0 - \alpha \mathcal{L}'_{\tau,0}(\theta_\tau^0) \\ \theta_\tau^2 &\leftarrow \theta_\tau^1 - \alpha \mathcal{L}'_{\tau,1}(\theta_\tau^1) \\ &\cdots \\ \theta_\tau^k &\leftarrow \theta_\tau^{k-1} - \alpha \mathcal{L}'_{\tau,k-1}(\theta_\tau^{k-1}) \end{aligned}, \tag{3}$$

where θ_τ^k is the model parameter at step k for learning task τ, $\mathcal{L}_{\tau,k-1}$ is the loss computed on the kth minibatch sampled from task τ, $\mathcal{L}_{\tau,1}$ is the loss computed on the second minibatch sampled from task τ and so on. The 'prime' symbol denotes differentiation, and α is the inner learner step size. Note that this learning process is the separate for all tasks, starting from the same initialization θ.

After an arbitrary m ($<n$) number of tasks are independently learnt by the above k-step SGD process, and obtaining θ_τ^k, $\tau = 1, 2, \ldots m$, we make one update across all these m tasks with the meta-learner to get a better initialization θ:

$$\theta \leftarrow \theta + \gamma \frac{1}{m} \sum_{\tau=1}^{m} (\theta_\tau^k - \theta), \tag{4}$$

where γ is the learning step of the meta-learner. The term $\frac{1}{m}\sum_{\tau=1}^{m}(\theta_\tau^k - \theta)$ can be considered as a gradient, so that for example a popular optimization algorithm such as Adam[30] can be used by the meta-learner to self-adjust learning rates for each parameter. The entire process of inner-learner update and meta-learner update is repeated until a chosen maximum number of meta-learning epochs is reached. This algorithm is shown to encourage the gradients of different minibatches of a given task to align in the same direction, thereby improving generalization and efficient learning later on[29].

In the final learning stage, the model is provided with a few-sample dataset of a new task. First, the model is initialized with the meta-learnt parameters θ, which are then fine-tuned with the new task-training data to $\theta_\tau^{k'}$ and finally the fine-tuned model is evaluated with testing data from the new task. The training procedure of final learning does not require a special algorithm, and can be conducted with regular mini-batch stochastic gradient descent. This final learning stage is equal to the inner-learning loop for a single task in Eq. (3) without any outer loop.

Algorithm 1 summarizes the complete procedure.

Algorithm 1

Meta-learning for few-shot survival prediction

Initialize randomly $\theta = \{\phi, \beta\}$, the feature extractor and Cox model parameters, respectively
Let the (inner) survival loss function be defined as in Eq. (1):
$\mathcal{L} = -\sum_{y_i=\text{uncensored}} \left[f_\varphi(\mathbf{x}_i)\beta - \log\left(\sum_{y_j \geq y_i} e^{f_\varphi(\mathbf{x}_j)\beta} \right) \right]$
for $i \leftarrow 0$ **to** n **do**
  for m randomly sampled tasks T_τ **do**
    Compute θ_τ^k, denoting k update steps with $\mathcal{L}$, as in Eq. (3)
    end
    Update $\theta \leftarrow \theta + \gamma \frac{1}{m} \sum_{\tau=1}^{m} (\theta_\tau^k - \theta)$
    $i \leftarrow i + m$
end
Return θ

Experimental setup. In order to assess the meta-learning method's performance, we compare it with several alternative training schemes based on the same neural network architecture: regular pre-training, combined learning, and direct learning. First, meta-learning initially learns general knowledge from a dataset containing tasks that are relevant but not directly related to the target, and then learns task-specific knowledge from a very small target task dataset. We define the first dataset as the "multi-task training data", and the second as the "target task training data". Secondly, regular pre-training also has a two-stage learning process on the same datasets, but unlike meta-learning without explicitly focusing on learning to reach an initialization that is easy to adapt to new tasks. Thirdly, combined learning does not involve a two-stage learning process, but also leverages knowledge from the relevant tasks by combining the multi-task training data and the target task training data together in one dataset to train a prediction model. Direct learning on the other hand, only uses the target task training samples. To illustrate the effectiveness of the methods with few samples, we consider three cases of direct learning: a large sample size, a medium sample size, and a small sample size which is the same size as the "target task training data" used for the other methods (i.e. regular pre-training, combined learning and meta-learning) (Fig. 2).

In our experiments, the "multi-task training data" is the pan-cancer RNA sequencing data containing samples from any cancer sites except one cancer site that we define as the target cancer site. The associated target cancer data is considered as the "target task data". This target task dataset is split into training data and testing data, stratified by disease sub-type and censoring status. For meta-learning, regular pre-training, and combined learning we will not use all of the training set for the target task, as we want to assess the performances when the algorithm is exposed to only a small number of target task training samples. Therefore, we will randomly draw 20 samples from the training dataset as one "target task training data". We choose a small sample size of 20 because it is a possible case in real life situations where the target task is the study of rare diseases[31], or where new technologies are used to produce data which only have the capacity to produce a small sample. For direct learning, we randomly draw three different sizes from the training datasets to form training data, 20 for the small size,

150 for the medium size, and 250 for the large size. All methods are evaluated on the common testing data of the target task.

Finally, as a linear baseline, we use the combined learning training sample (multi-task training data and target task training data) to train a linear cox regression model. We conduct 25 experiment trials for each method, where each trial is trained with a randomly drawn "target task training dataset as described above.

Evaluation. We evaluate the survival prediction model performances with two commonly used evaluation metrics: the concordance index (C-index)[4] and the integrated Brier score (IBS)[32]. Fistly, the C-index is a standard performance measure of a model's predictive ability in survival analysis. It is calculated by dividing the number of all pairs of subjects whose predicted survival times are correctly ordered, by the number of pairs of subjects that can possibly be ordered. A pair cannot be ordered if the earlier time in the pair is censored or both events in the pair are censored. A C-index value of 1.0 indicates perfect prediction where all the predicted pairs are correctly ordered, and a value of 0.5 indicates random prediction. Secondly, the IBS is used to evaluate the error of survival prediction and is represented by the mean squared differences between observed survival status and the predicted survival probability at a given time point. The IBS provides an overall calculation of the model performance at all available times. An IBS value of 0 indicates perfect prediction, while 1 indicates entirely inaccurate prediction.

We select target cancer sites from TCGA with the following two inclusion criteria: (1) a minimum of 450 samples, providing enough training samples for different benchmarking training schemes and (2) a minimum of 30% non-censoring samples, enabling more accurate evaluation than more heavily censored cohorts. This results in the following cancers: glioma, including glioblastoma (GBM) and low-grade-glioma (LGG); non-small cell lung cancer, including lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), and head-neck squamous cell carcinoma (HNSC). These three types of cancers are also of clinical interest, because gliomas are the most common type of malignant brain tumor, and lung cancer is the deadliest cancer in the world[33,34]. HNSC, on the other hand, is a less widely studied type of cancer, which nonetheless attracted growing attention in the recent decade since the release of the publicly available largest dataset in HNSC by TCGA[35,36].

In addition, to further validate the model in the small sample training setting, we select an additional rare cancer cohort, mesothelioma (MESO), with <90 samples in total. Due to the small sample size, we do not compare to the medium or large sample direct learning, but only compare to the small sample direct learning.

Finally, we use a fully independent testing cohort to validate the model. We use a non-small cell lung cancer cohort consisting of 129 patient samples collected from the Stanford University School of Medicine and Palo Alto Veterans Affairs Healthcare System[37]. The data is available at National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO)[38]. For the meta-learning, regular pre-training and combined learning methods, we use the same meta-learnt and pre-trained models that are trained with TCGA data for testing the non-small cell lung cancer. The final training and testing is done on the independent dataset. Due to the small sample size, we also only include small sample direct learning for comparison.

For the three large target cancer cohorts, 20% of the target data is used for testing, and we evaluate the C-index and IBS in 25 experimental trials for each method. For the small cancer cohort and independent data cohort, 50% of the data is used for testing, and we conduct 10 trials for each method due to limited training samples for sampling.

Hyper-parameter selection. To avoid overfitting, we do not conduct a separate hyper-parameter search for each of the cancer datasets. Instead, we search for hyper-parameters on one type of cancer and apply the chosen parameters to all experiments. We select the largest cancer cohort, glioma, and use 5-fold cross-validation for hyper-parameter selection. For each given set of hyper-parameters, we average the results from five validation sets (each is 20% of training data). Since there is similarity in the algorithm between methods (combined learning, direct learning, and regular pre-training), we share hyper-parameters between experiments when it makes sense, as detailed below.

All methods use the same neural network architecture with two hidden fully connected layers of size 6000 and 2000, and an output fully connected feature layer of size 200. Each layer uses the ReLU activation function[39]. Initially we experiment with 4 different structures: 1 or 2 hidden layers with feature size of 200 or 50, respectively. We chose the optimal structure detailed before and use it as the architecture for all methods in our subsequent discussion.

For the regular pre-training model, we search for hyper-parameters for the pre-training stage and fine-tuning stage separately. For both stages, we test the mini-batch gradient descent and Adam optimizers, and determine learning rates with grid search on a grid of [0.1, 0.05, 0.01, 0.005, 0.001] for SGD and a grid of [0.001, 0.0005, 0.0001, 0.00005, 0.00001] for Adam. We test batch sizes of 50, 100, 200, and 800 for pre-training. The selected parameters for the pre-train stage are: an SGD optimizer with learning rate of 0.001, L2 regularization scale of 0.1 and batch size of 800. The selected parameters for the fine-tune stage are: an SGD optimizer with learning rate of 0.001, L2 regularization scale of 0.1, and batch size of 20 which is the size of each target cancer training dataset. For the combined learning model and direct learning model, since the algorithm is very similar to the regular pre-training model's pre-train stage, we use the same parameters selected for the pre-train. The batch sizes for direct learning is half of the size of training data.

Table 1 Selected hyper-parameters for meta-learning's meta-learning stage.

Hyper-parameter	Value
Task-level optimizer	SGD
Task-level learning rate	0.01
Task-level gradient steps	5
Task-level Batch size	100
Meta-level optimizer	ADAM
Meta-level learning rate	0.0001
Meta-level tasks batch size	10
L2 regularization scale	0.1

For the meta-learning model, we search for hyper-parameters for the meta-learning only. For the final learning stage, we use the same hyper-parameters as in the fine-tune stage of the regular pre-training model, as both methods can use similar algorithms in the last stage of training. From our previous discussion, in the meta-learning stage an SGD optimizer and an Adam optimizer are suitable for the inner learner and meta-learner, respectively. For the learning rates, we perform grid search on a grid of [0.1, 0.05, 0.01, 0.005, 0.001] for SGD and a grid of [0.001, 0.0005, 0.0001, 0.00005, 0.00001] for Adam. Batch size is searched from [50, 100, 200, 800], the number of tasks for averaging one meta-learner update is searched from [5, 10, 20], and the number of gradient descent steps for the inner-learner is searched from [3, 5, 10, 20]. The selected parameters for the meta-learning stage are shown in Table 1.

Finally, in order to evaluate the effect of fluctuations of the meta-learning hyper-parameters, and ensure that our results reflect the average performance over fluctuations, we conduct a series of tests on the validation data where in each experiment we vary one of the five unique meta-learning hyper-parameters from the chosen value by tuning it up or down by one grid, obtaining 10 sets of varied hyper-parameters. We do 5-fold cross-validation for each set of varied hyper-parameters and compute the C-index from the resulting 50 experiments. We also do 50 random experiments using the selected hyper-parameters and compare the average results of varied versus selected hyper-parameters. We conduct a two-sample t-test on the two results, and conclude that the results obtained by varied parameters do not have a significant difference from the results obtained by the chosen parameters (mean C-index difference of 0.005 with p value = 0.50). Therefore, our results are robust with respect to fluctuations of the hyper-parameters and our conclusions are not based on excessive hyper-parameter tuning.

Interpretation of the genes prioritized by the meta-learning model. We apply risk score backpropagation[15] to the meta-learned models to investigate the feature importance of genes for each of the three target cancer sites. For a given sample, each input feature is assigned a risk score by taking the partial derivatives of the risk with respect to the feature. A positive risk score with high absolute value means the feature is important in poor prediction (high risk), and a negative risk score with high absolute value means the feature is important in good prediction (low risk). The features are ranked by the average of risk score across all samples.

Two approaches were adopted for annotating the genes with ranked risk scores generated by the meta-learning model. Firstly, the top 10% high-risk genes (genes with positive risk scores) and the top 10% low-risk genes (genes with negative risk scores) from each cancer type were subjected to gene set over-representation analysis, by comparing the genes against the gene sets annotated with well-defined biological functions and processes. We model the association between the genes and each gene set using a hypergeometric distribution and Fisher's exact test. Secondly, instead of arbitrary thresholding in the first approach, all the genes, together with their ranked risk scores were incorporated in the gene set enrichment analysis with the fgsea R package[40] which calculates a cumulative gene set enrichment statistic value for each gene set. The gene set databases used in this analysis include Kyoto Encyclopedia of Genes and Genomes (KEGG)[41], The Reactome Pathway Knowledgebase[42] and WikiPathways[43].

Results

Meta-learning outperforms regular pre-training and combined learning. For all of the large target cancer sites, meta-learning achieves similar or better performance than regular pre-training or combined learning (Fig. 3; Table 2). For the glioma cohort, the mean C-index for meta-learning is 0.86 (0.85–0.86 95% CI), compared to 0.84 (0.83–0.85 95% CI) for regular pre-training and

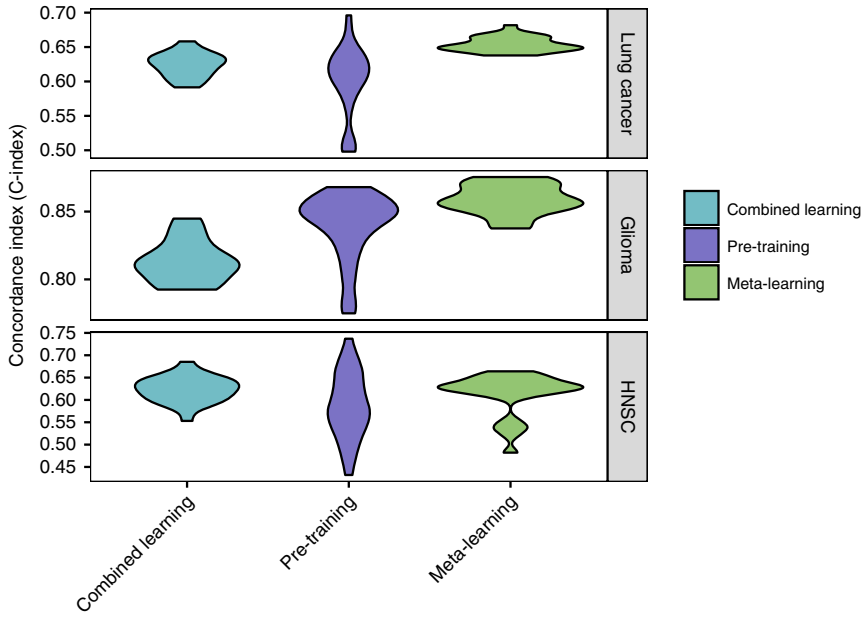


Fig. 3 C-Index for target cancer survival prediction, comparing combined learning, regular pre-training and meta-learning. The figure shows results for lung cancer (top panel), glioma (middle panel) and head and neck cancer (bottom panel).

Table 2 Integrated Brier scores (IBS) with 95% confidence intervals ($n = 25$ trials) for target cancer survival prediction with 20 samples, unless specified otherwise.

Method	Glioma	Lung cancer	HNSC
Direct (250 samples)	**0.24 ± 0.02**	0.19 ± 0.01	0.20 ± 0.01
Direct (150 samples)	0.25 ± 0.01	0.19 ± 0.01	0.21 ± 0.01
Direct	0.30 ± 0.02	0.24 ± 0.02	0.30 ± 0.02
Combined	0.29 ± 0.02	0.21 ± 0.02	0.26 ± 0.02
Pre-training	0.31 ± 0.02	0.23 ± 0.02	0.26 ± 0.02
Meta-learning	0.28 ± 0.01	**0.16 ± 0.01**	**0.16 ± 0.00**

Lower value is better. Best performing method in bold.

0.81 (0.81–0.82 95% CI) for combined training. For the lung cancer cohort, the mean C-index is 0.65 (0.65–0.66 95% CI) for meta-learning, 0.60 (0.58–0.61 95% CI) for regular pre-training, and 0.62 (0.62–0.63 95% CI) for combined training. For the HNSC cohort, the result is 0.61 (0.59–0.63 95% CI) for meta-learning, 0.59 (0.57–0.61 95% CI) for regular pre-training and 0.62 (0.61–0.64 95% CI) for combined training. Note that, the variance of the meta-learning results across 25 random trials also tends to be the smallest, which is most observable for the lung cancer and glioma cohorts. In addition, each of these multi-layer neural networks also shows better performance on average than a linear baseline model. The linear baseline model achieves a C-index of 0.61 for lung cancer (0.60–0.62 95% CI), 0.77 for glioma (0.74–0.80 95% CI), and 0.59 for HNSC (0.58–0.62 95% CI).

Meta-learning achieves competitive predictive performance compared to direct learning. Next, we compare our meta-learning approach with regular direct learning on the target task training samples with different cohort sizes. The performance of direct learning drops significantly when the number of training samples decreases from 250 to 20, which is anticipated because a great amount of information is lost and the model can hardly learn well. However, meta-learning and pre-training can compensate for such lack of information by transferring knowledge from the pan-cancer data explicitly and implicitly, respectively. We show that meta-learning achieves similar or better prediction performance than large-sample direct training in lung cancer and HNSC, and reaches comparable performance with medium-sample direct training in glioma (Fig. 4; Table 2). For the lung cancer cohort, the mean C-index is 0.57 (0.56–0.58 95% CI) for large sample direct learning, 0.54 (0.52–0.56 95% CI) for medium sample direct learning, 0.53 (0.50–0.55 95% CI) for small sample direct learning, and 0.65 (0.65–0.66 95% CI) for meta-learning. For the glioma cohort, the mean C-indices for large sample, medium sample and small sample direct learning are 0.86 (0.86–0.87 95% CI), 0.85 (0.85–0.86 95% CI), and 0.82 (0.81–0.84 95% CI), respectively, and for meta-learning the mean C-index is 0.86 (0.85–0.86 95% CI). For the HNSC cohort, the mean C-index is 0.62 (0.60–0.64 95% CI) for large sample direct learning, 0.57 (0.54–0.59 95% CI) for medium sample direct learning, 0.53 (0.49–0.56 95% CI) for small sample direct learning, and 0.61 (0.59–0.63 95% CI) for meta-learning. Thus, in all three cancer sites, meta-learning reaches competitive performances as large sample direct learning and can outperform it in certain cases.

Risk score ranked genes are enriched in key cancer pathways. Next, we investigated for each cancer site what genes are most important in the meta-learning model (Fig. 5, Supplementary Tables 1–6). In gliomas, the high-risk genes are associated with viral carcinogenesis (p value = 0.002), Herpes simplex infection (p value = 0.007), cell cycle (p value = 0.03), apoptosis (p value = 0.03), DNA damage response (p value = 0.04), all of which are also enriched in gene set enrichment analysis with positive enrichment scores (Fig. 5a). The low-risk genes are associated with HSF1 activation (p value = 0.02) which is involved in hypoxia pathway, and tryptophan metabolism (p value = 0.04), the latter is also enriched in gene set enrichment analysis with negative enrichment score. Tryptophan catabolism has been increasingly recognized as an important microenvironmental factor in anti-tumor immune responses[44] and it is a common therapeutic target in cancer and neurodegeneration diseases[45].

In head and neck cancer, the high-risk genes are associated with PTK6 signaling (p value = 0.01), which regulates cell cycle and growth, and cytokines and inflammatory response (p value = 0.009). The low-risk genes are associated with

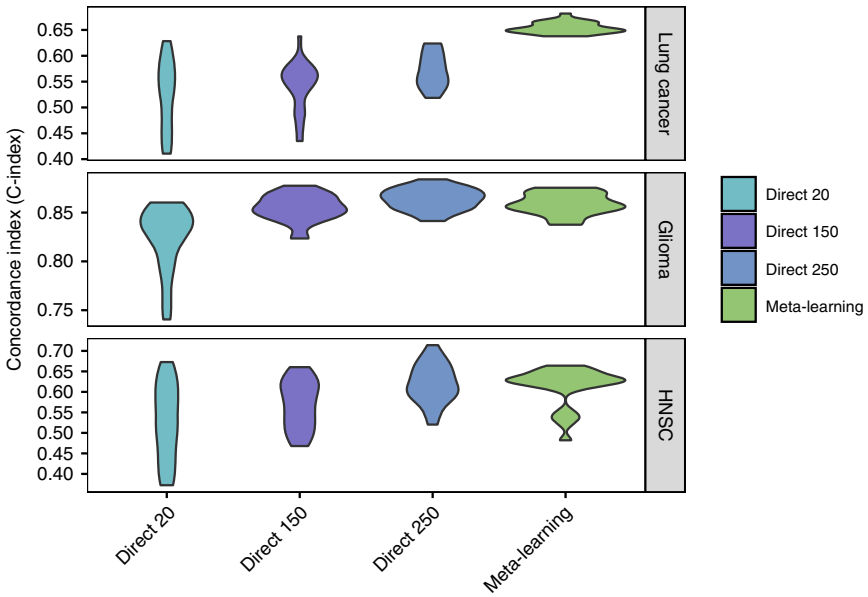
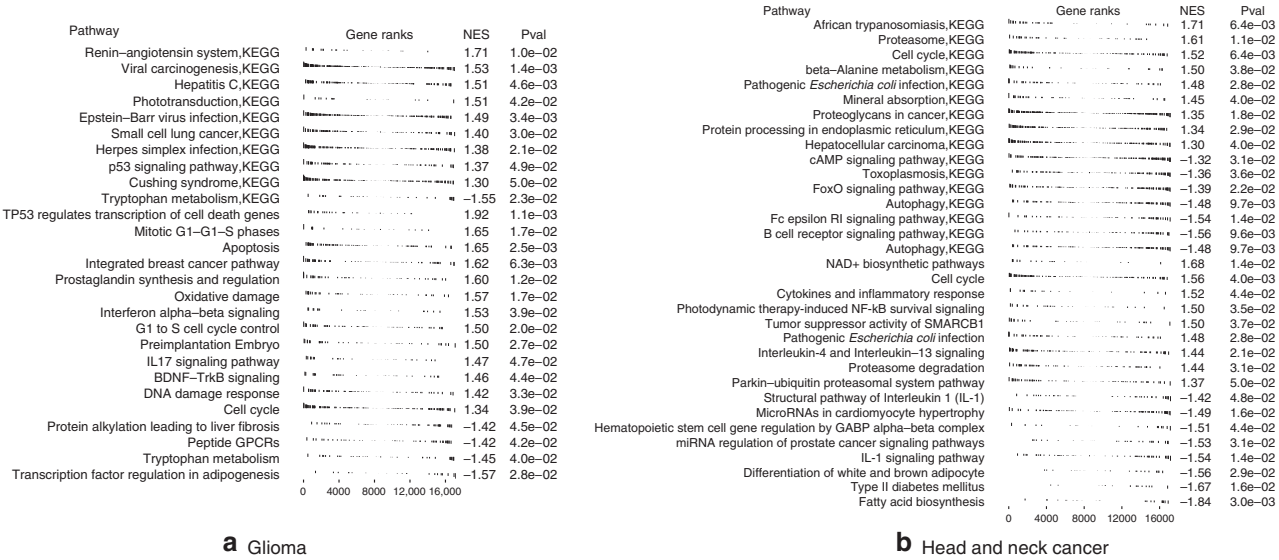


Fig. 4 C-Index for target cancer survival prediction, comparing direct learning with large (250), medium (150), and small (20) size samples and meta-learning. The figure shows results for lung cancer (top panel), glioma (middle panel), and head and neck cancer (bottom panel).



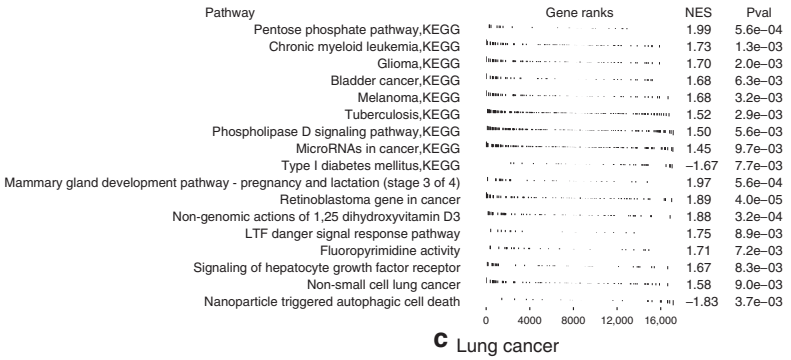


Fig. 5 Gene set enrichment analysis of the ranked gene list in (a) Glioma (b) Head and neck cancer (c) Lung cancer. The gene set databases used in this analysis included Kyoto Encyclopedia of Genes and Genomes (KEGG) and WikiPathways. Pval enrichment *p*-value, NES normalized enrichment score. In **a** and **b** the enrich pathways with *p* value below 0.05 were displayed. In **c** the enrich pathways with *p* value below 0.01 were displayed.

autophagy (p value = 0.02), which is also enriched in gene set enrichment analysis. Other enriched pathways include B cell receptor signaling pathway, cell cycle, and interleukin 1 signaling pathway (Fig. 5b). Interleukin 1 is an inflammatory cytokine which plays a key role in carcinogenesis and tumor progression[46].

In lung cancer, the top high-risk genes are associated with "non-small cell lung cancer" pathway (p value = 0.01), tuberculosis (p value = 0.008), Hepatitis B and C virus infection (p value = 0.03), and many pathways implicated previously in cancer. These pathways are also enriched in gene set enrichment analysis (Fig. 5c). Pulmonary tuberculosis has been shown to increase the risk of lung cancer[47,48]. The low-risk genes are associated with energy metabolism (p value = 0.03), ferroptosis (p value = 0.037), and AMPK signaling pathway (p value = 0.046), all related to energy metabolism, particularly lipid metabolism. AMPK signaling pathway activation by an AMPK agonist was shown to suppresses non-small cell lung cancer through inhibition of lipid metabolism[22]. AMPK signaling and energy metabolism are also enriched in gene set enrichment analysis. Other enriched pathways include Notch signaling, interleukin signaling, ErbB signaling, and signaling pathways regulating pluripotency of stem cells.

Validation on the small sample rare cancer cohort. Next we conduct validation on the small sample rare cancer cohort. It is shown that meta-learning achieves similar or better performance than regular pre-training, combined learning, or small sample direct learning (Fig. 6). The mean C-index for meta-learning is 0.66 (0.63–0.69 95% CI), compared to 0.62 (0.59–0.64 95% CI) for regular pre-training, 0.65 (0.63–0.67 95% CI) for combined training, and 0.60 (0.59–0.62 95% CI) for small sample direct learning.

Validation on the independent lung cancer cohort. Finally, we test on the independent lung cancer cohort. We use the same meta-learnt and pre-trained models that are trained with TCGA data for the lung cancer target site. On this cohort, it is shown that meta-learning has better performance than regular pre-training, combined learning, or small sample direct learning (Fig. 7). The mean C-index for meta-learning is 0.63 (0.61–0.65 95% CI), compared to 0.58 (0.55–0.61 95% CI) for regular pre-training, 0.59 (0.55–0.64 95% CI) for combined training, and 0.54 (0.50–0.58 95% CI) for small sample direct learning.

Discussion

Previous studies have shown that when analyzing high-dimensional genomic data, deep learning survival models can achieve comparable or superior performance compared to other methods (e.g., Cox elastic net regression, random survival forests)[49]. However, the performance of deep learning is often limited by the relatively small amount of available data[15]. To address this issue, our work investigates different deep learning paradigms to improve the performance of deep survival models, especially in the setting of small size training data.

In previous studies the most common way to build deep survival models is to train neural networks with a large number of target task training samples from scratch, a process we call direct training. Direct training with a large sample size (e.g. $n = 250$) can thus be considered as a baseline. As expected, the performance of direct training drops when the number of training samples decreases (e.g. from $n = 150$ to $n = 20$). On the other hand, combined learning, regular pre-training, and meta-learning all leverage additional data from other sources, thereby enabling them to achieve better performances when the training sample size is small. We use a small number of target cancer site training samples (e.g. $n = 20$) with these methods and investigate their performance.

When only small (task) sample sizes are available for meta-training, a Bayesian approach to meta-learning is an option[50,51]. Although Bayesian meta-learning with few(er) training tasks has shown less meta-overfitting (on training tasks) and performance improvements over regular meta-learning in low-dimensional settings, in high-dimensional settings the improvement is marginal[51].

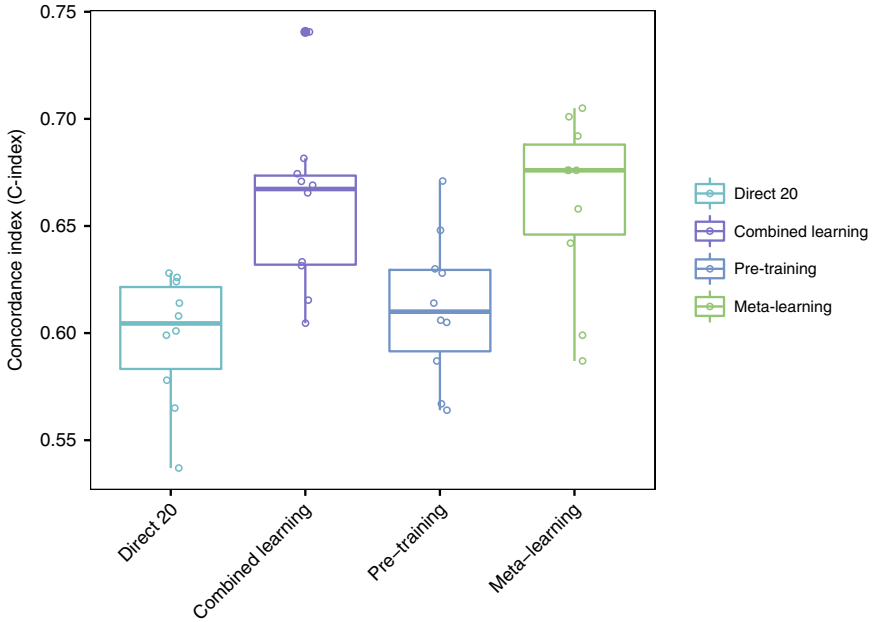


Fig. 6 C-Index for survival prediction on the mesothelioma cohort, comparing small (20) size sample direct learning, combined learning, regular pre-training, and meta-learning. $n = 45$ testing samples from mesothelioma cohort over 10 independent experiments. The upper and lower bars in the boxplots represent the largest and smallest data points excluding any outliers. The upper and lower bounds of the boxes represent the 25th percentile and 75th percentile of the data, respectively. The middle bars represent the median of the data.

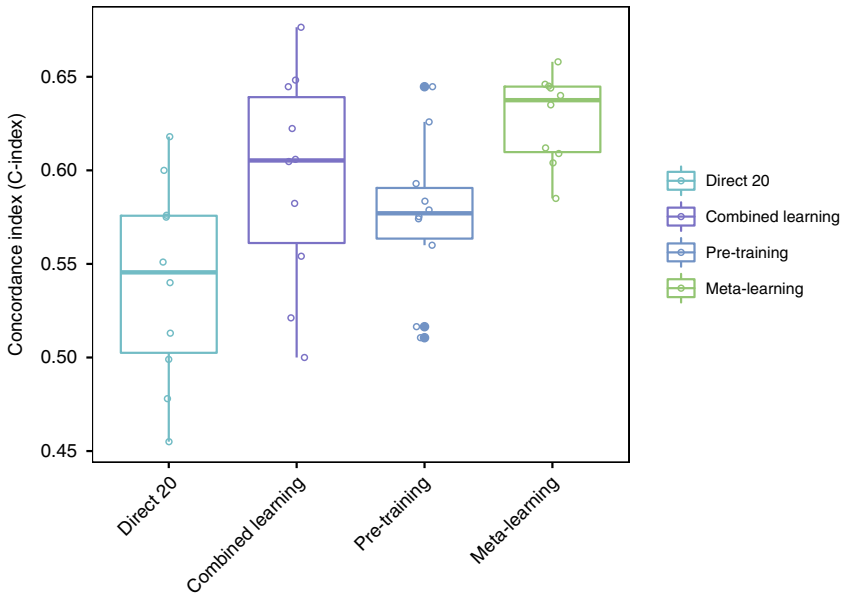


Fig. 7 C-Index for survival prediction on the independent lung cancer cohort, comparing small (20) size sample direct learning, combined learning, regular pre-training, and meta-learning. $n = 64$ testing samples from lung cancer cohort over 10 independent experiments. The upper and lower bars in the box-plots represent the largest and smallest data points excluding any outliers. The upper and lower bounds of the boxes represent the 25th percentile and 75th percentile of the data, respectively. The middle bars represent the median of the data.

It is important to note that combined learning, regular pre-training, and meta-learning are exposed to exactly the same information, but differ only in their algorithms. Combined learning is a one-stage learning process, whereas pre-training and meta-learning are two-stage learning methods. Meta-learning shows better predictive performance than combined or regular pre-training, indicating that it is able to adapt to a new task more effectively due to the improved optimization algorithm targeting the few-sample training environment.

It has been shown that methods which use only target task data (direct learning with different size samples) and methods which use additional information (combined, pre-training and meta-learning) perform differently, and one type of approach may be better than the other on different cancer sites. For example, on glioma, direct learning tends to do better overall; whereas on lung cancer, the other methods outperform direct learning. This may be due to that fact that the amount of information that can be learnt from related data versus from the target data is different for each cancer site. If there is significant information within the target cancer samples alone, then direct training will be more effective than learning from other cancer samples. On all three cancer sites we observe that meta-learning achieves similar or better performance than medium-size direct training, and outperforms large-size direct training in some cases. However, the advantage of meta-learning may not generalize to every cancer site. Certain cancers may have very unique characteristics so that transfer of information from other cancers may not help in prediction regardless of improved adaptivity. For the three cancer sites, the affinity of each target cancer to other types of cancers in the pan-cancer data aids the performance of meta-learning, which efficiently transfers the information from other cancers to the target cancers. On the other hand, some cancers are more dissimilar from other cancer sites which makes information transfer difficult. For example, for another cancer site, kidney cancer, specifically kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP), both meta-learning and pre-training do not produce good survival prediction. This can be visualized in Supplementary Fig. 1, comparing the affinity between different target cancers with the rest of the cancers on a t-distributed stochastic neighbor embedding (t-SNE) graph. Therefore, in order for meta-learning to achieve good performance, the related tasks training data need to contain a reasonable amount of transferable information to the target task.

The performance of meta-learning can be explained by the learned learning algorithm at the meta-learning stage where the model learns from related tasks. We further investigate how to optimize meta-learning performance. We examine results from two sampling approaches when forming one task, where we either draw samples only from one cancer, or draw samples from multiple types of cancer. It is a more natural choice to consider each cancer type as a separate task, but we found that the latter leads to improved performance. To explain this improvement, we examine the gradient of the meta-learning loss function. It can be shown that the gradient of the loss function contains a term that encourages the gradients from different minibatches for a given task to align in the same direction (Supplementary Note 1). If the two minibatches contain samples from the same type of cancer, their gradient might already be very similar and thus this higher order term would not have a large effect. On the other hand, if the second minibatch contains samples from a different type of cancer than the first, the algorithm will learn something that is common to both of them and thereby help to improve generalization.

From a molecular point of view, certain cancer types are related to each other and there is an inherent presence of inter-dependency. For example, colon and rectal cancers are found to have considerably similar patterns of genomic alteration[52]. In our work, the multi-task training data contains many cancer sites including cancer sites that may have inter-dependencies. We did not focus on modeling the inter-dependencies within the multi-task training data, but on transferring information from the multi-task data to the new tasks. However, handling cross-task relations in meta-learning is an interesting topic that could potentially improve generalization further. Recent work has proposed methods to accommodate the relations between tasks[53]. This would be worth investigating in future work.

The gene set enrichment analysis results validate our model for prioritizing the genes for survival predication. In the three cancer types investigated, the resulting gene lists are enriched in key pathways in cancer including cell cycle regulation, DNA damage response, cell death, interleukin signaling, NOTCH signaling pathway, etc.

Apart from the well-recognized cancer pathways, our results also reveal potential players affecting cancer development and prognosis, that are not well-studied yet. Viruses have been linked to the carcinogenesis of several cancers, including human papilloma virus in cervical cancer, hepatitis B and C viruses in liver cancer, and Epstein-Barr virus in several lymphomas and nasopharyngeal carcinoma[54]. Our results further suggest that viruses might also play a role in glioma and lung cancer, where the high-risk genes are enriched in several viral carcinogenesis pathways. In gliomas, the enriched pathways that are unfavorable for survival include Epstein-Barr virus and herpes simplex infection. In lung cancer, both hepatitis B and C virus infection pathways are enriched. This suggests that that hepatotropic viruses may affect the respiratory system, including the association with lung cancer. For example, hepatitis B virus infection has been associated with poor prognosis in patients with advanced non-small cell lung cancer[55]. The role of Epstein-Barr virus in gliomagenesis have also been studied but the results remain inconclusive[56]. Whether these viruses do play a role in carcinogenesis and further affect cancer prognosis, or the association we observed reflects an abnormal immune system that is unfavorable for the survival of cancer patients remains to be investigated.

While it would be interesting to decipher the mechanisms of viral infections in tumorigenesis in these cancers, it is difficult to establish a direct link using genomic analysis from currently available data. More well-designed experiments are needed. That being said, the field of viral infections in tumorigenesis is under active research now and there are several interesting studies that suggest a wider role of viral infections in cancer. The recent study of viral landscape in cancer by Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium[57] examined whole genome and whole transcriptome sequencing data from 2658 cancers across 38 tumor types. Apart from the well-known viral etiology in cancer (HPV in cervical cancer and head and neck cancer, HBV in liver cancer, and EBV in gastric cancer), the study also found frequent appearance of herpesviruses (EBV and HHV-6B) in brain cancers. A 2018 study in Neuron[58] found frequent presence of herpesviruses in the brain tissues. Although this study is designed to study Alzheimer's disease, the fact that herpesviruses are frequently found in brain tissues warrants further research of the role of herpesviruses in not only neurodegenerative diseases, but also cancer. A 2018 study in Cancer Research[59] found virus infection shapes the tumor immune microenvironment and genetic architecture of six virus-associated tumor types. They found that EBV infection was associated with decreased receptor diversity in multiple cancers. The altered immune profile in the tumor microenvironment may affect tumor progression and patient survival, but more study is needed to confirm it.

As for the enriched pathways that are favorable for cancer survival, we identified pathways related to metabolism, in particular, lipid metabolism, in all the three cancer types investigated. In glioma, the top enriched pathway favorable for cancer survival is adipogenesis regulation. In head and neck cancer, differentiation of adopocyte and fatty acid biosynthesis are top enriched favorable pathways. In lung cancer, ferroptosis and AMPK signaling pathway are both related to energy metabolism. Ferroptosis is a process driven by accumulated iron-dependent lipid ROS that leads to cell death. Small molecules-induced ferroptosis has a strong inhibition of tumor growth and enhances the sensitivity of chemotherapeutic drugs, especially in drug resistance[60]. AMPK plays a central role in the control of cell growth, proliferation, and autophagy through the regulation of mTOR activity and lipid metabolism[22,61]. The link between cancer and metabolism is worth investigating in future studies.

To conclude, in survival analyses one problem that researchers have encountered is the insufficient amount of training samples for machine learning algorithms to achieve good performances. We address this problem by adapting a meta-learning approach which learns effectively with only a small number of target task training samples. We show that the meta-learning framework is able to achieve similar performance as learning from a significantly larger number of samples by using an efficient knowledge transfer. Moreover, in the context of limited training sample exposure, we demonstrate that this framework achieves superior predictive performance over both regular pre-training and combined learning methods on two types of target cancer sites. Finally, we show that meta-learning models are interpretable and can be used to investigate biological phenomena associated with cancer survival outcome.

The problem of small data size may be a limiting factor in many biomedical analyses, especially when studies are conducted with data that is expensive to produce, or in the case of multi-modal data[12]. Our work shows the promise of meta-learning for biomedical applications to alleviate the problem of limited data. In future work, we intend to extend this approach to analysis with medical imaging data, such as histopathology data and radiology data, for building predictive models on multi-modal data with limited sets of patients.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data used in this manuscript are publicly available. The TCGA Gene expression data is version 2 of the adjusted pan-cancer gene expression data obtained from Synapse: https://www.synapse.org/#!Synapse:syn4976369.2. The independent lung cancer data can be obtained from: https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics. The databases used in gene set enrichment analysis are publicly available: Kyoto Encyclopedia of Genes and Genomes (KEGG) at https://www.genome.jp/kegg/; the Reactome Pathway Knowledgebase at https://reactome.org/download-data; and WikiPathways at https://www.wikipathways.org/index.php/Download_Pathways. The remaining data are available within the Article, Supplementary Information or available from the authors upon request.

Code availability

The code used to analyze the data in this manuscript is in the GitHub repository with URL: https://github.com/gevaertlab/metalearning_survival[62].

Received: 23 April 2020; Accepted: 16 November 2020;
Published online: 11 December 2020

References

1. Hosmer, D. W., Lemeshow, S. & May, S. *Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley Series in Probability and Statistics* (John Wiley & Sons, 2008).
2. Klein, J. P. & Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data* (Springer Science & Business Media, 2006).
3. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc.: Ser. B* **34**, 187–202 (1972).
4. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).
5. Kleinbaum, D. G. & Klein, M. The Cox proportional hazards model and its characteristics. In *Survival Analysis*, 97–159 (Springer, 2012).
6. 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 (2016).
7. Park, S. T. & Kim, J. Trends in next-generation sequencing and a new era for whole genome sequencing. *Int. Neurourol. J.* **20**, S76 (2016).

8. Goeman, J. J. L1 penalized estimation in the cox proportional hazards model. *Biometrical J.* **52**, 70–84 (2010).
9. Park, M. Y. & Hastie, T. L1-regularization path algorithm for generalized linear models. *J. R. Stat. Soc.: Ser. B* **69**, 659–677 (2007).
10. Wong, K. Y. et al. An integrative boosting approach for predicting survival time with multiple genomics platforms. Preprint at https://doi.org/10.1101/338145 (2018).
11. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **24**, 1248–1259 (2018).
12. Cheerla, A. & Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35**, i446–i454 (2019).
13. Ching, T., Zhu, X. & Garmire, L. X. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* **14**, e1006076 (2018).
14. Luck, M., Sylvain, T., Cardinal, H., Lodi, A. & Bengio, Y. Deep learning for patient-specific kidney graft survival analysis. Preprint at 1705.10245 (2017).
15. Yousefi, S. et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.* **7**, 1–11 (2017).
16. Pratt, L. Y. Discriminability-based transfer between neural networks. In *Advances in Neural Information Processing Systems*, 204–211 (1993).
17. Li, Y., Wang, L., Wang, J., Ye, J. & Reddy, C. K. Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 231–240 (IEEE, 2016).
18. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, 2009).
19. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
20. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, 1126–1135 (JMLR. org, 2017).
21. Vilalta, R. & Drissi, Y. A perspective view and survey of meta-learning. *Artif. Intell. Rev.* **18**, 77–95 (2002).
22. Chen, X. et al. Novel direct ampk activator suppresses non-small cell lung cancer through inhibition of lipid metabolism. *Oncotarget* **8**, 96089 (2017).
23. Devos, A. & Grossglauser, M. Regression networks for meta-learning few-shot classification. In *7th ICML Workshop on Automated Machine Learning (2020)* (2020).
24. Duan, Y. et al. Rl2: fast reinforcement learning via slow reinforcement learning. Preprint at https://arxiv.org/abs/1611.02779 (2016).
25. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68 (2015).
26. Ishwaran, H. et al. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
27. Chi, C.-L., Street, W. N. & Wolberg, W. H. Application of artificial neural network-based survival analysis on two breast cancer datasets. In *AMIA Annual Symposium Proceedings*, Vol. 2007, 130 (American Medical Informatics Association, 2007).
28. Petalidis, L. P. et al. Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Mol. Cancer Ther.* **7**, 1013–1024 (2008).
29. Nichol, A., Achiam, J. & Schulman, J. On first-order meta-learning algorithms. Preprint at 1803.02999 (2018).
30. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at 1412.6980 (2014).
31. Hee, S. W. et al. Does the low prevalence affect the sample size of interventional clinical trials of rare diseases? An analysis of data from the aggregate analysis of clinicaltrials. gov. *Orphanet J. Rare Dis.* **12**, 44 (2017).
32. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
33. Ceccarelli, M. et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).
34. Herbst, R. S. & Lippman, S. M. Molecular signatures of lung cancer—toward personalized therapy. *N. Engl. J. Med.* **356**, 76–78 (2007).
35. Brennan, K., Koenig, J. L., Gentles, A. J., Sunwoo, J. B. & Gevaert, O. Identification of an atypical etiological head and neck squamous carcinoma subtype featuring the cpg island methylator phenotype. *EBioMedicine* **17**, 223–236 (2017).
36. Tonella, L., Giannoccaro, M., Alfieri, S., Canevari, S. & De Cecco, L. Gene expression signatures for head and neck cancer patient stratification: are results ready for clinical application? *Curr. Treat. Options Oncol.* **18**, 32 (2017).
37. Bakr, S. et al. A radiogenomic dataset of non-small cell lung cancer. *Sci. data* **5**, 1–9 (2018).
38. Barrett, T. et al. Ncbi geo: archive for high-throughput functional genomic data. *Nucleic Acids Res.* **37**, D885–D890 (2009).
39. Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning* (ICML-10), 807–814 (2010).
40. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. Preprint at https://doi.org/10.1101/060012 (2016).
41. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
42. Croft, D. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–D477 (2014).
43. Slenter, D. N. et al. Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**, D661–D667 (2018).
44. Platten, M., Wick, W. & Van den Eynde, B. J. Tryptophan catabolism in cancer: beyond ido and tryptophan depletion. *Cancer Res.* **72**, 5435–5440 (2012).
45. Platten, M., Nollen, E. A. A., Röhrig, U. F., Fallarino, F. & Opitz, C. A. Tryptophan metabolism as a common therapeutic target in cancer, neurodegeneration and beyond. *Nat. Rev. Drug Discov.* **18**, 379–401 (2019).
46. Mantovani, A., Barajon, I. & Garlanda, C. Il-1 and il-1 regulatory pathways in cancer progression and therapy. *Immunol. Rev.* **281**, 57–61 (2018).
47. Wu, C.-Y. et al. Pulmonary tuberculosis increases the risk of lung cancer: a population-based cohort study. *Cancer* **117**, 618–624 (2011).
48. Yu, Y.-H. et al. Increased lung cancer risk among patients with pulmonary tuberculosis: a population cohort study. *J. Thorac. Oncol.* **6**, 32–37 (2011).
49. Kim, D. W. et al. Deep learning-based survival prediction of oral cancer patients. *Sci. Rep.* **9**, 1–10 (2019).
50. Finn, C., Xu, K. & Levine, S. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 9516–9527 (2018).
51. Yoon, J. et al. Bayesian model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 7332–7342 (2018).
52. Network, C. G. A. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330 (2012).
53. Yao, H. et al. Automated relational meta-learning. Preprint at https://doi.org/10.1101/2001.00745 (2020).
54. Martin, D. & Gutkind, J. S. Human tumor-associated viruses and new insights into the molecular mechanisms of cancer. *Oncogene* **27**, S31–S42 (2008).
55. Peng, J.-W., Liu, D.-Y., Lin, G.-N., Xiao, J. J. & Xia, Z.-J. Hepatitis b virus infection is associated with poor prognosis in patients with advanced non small cell lung cancer. *Asian Pac. J. Cancer Prev.* **16**, 5285–5288 (2015).
56. Akhtar, S., Vranic, S., Cyprian, F. S. & Al Moustafa, A.-E. Epstein–barr virus in gliomas: cause, association, or artifact? *Front. Oncol.* **8**, 123 (2018).
57. Zapatka, M. et al. The landscape of viral associations in human cancers. *Nat. Genet.* **52**, 320–330 (2020).
58. Rizzo, R. Controversial role of herpesviruses in alzheimer's disease. *PLoS Pathog.* **16**, e1008575 (2020).
59. Varn, F. S., Schaafsma, E., Wang, Y. & Cheng, C. Genomic characterization of six virus-associated cancers identifies changes in the tumor immune microenvironment and altered genetic programs. *Cancer Res.* **78**, 6413–6423 (2018).
60. Lu, B. et al. The role of ferroptosis in cancer development and treatment response. *Front. Pharmacol.* **8**, 992 (2018).
61. Han, D., Li, S.-J., Zhu, Y.-T., Liu, L. & Li, M.-X. Lkb1/ampk/mtor signaling pathway in non-small-cell lung cancer. *Asian Pac. J. Cancer Prev.* **14**, 4033–4039 (2013).
62. Qiu, Y. L., Zheng, H., Devos, A., Selby, H. & Gevaert, O. Supporting software metalearning_survival. A meta-learning approach for genomic survival analysis. https://doi.org/10.5281/zenodo.4116296 (2020).

Acknowledgements

A.D. acknowledges funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 754354.

Author contributions

Y.L.Q., A.D., and O.G. conceived of the presented idea and designed the computational framework. Y.L.Q. performed the analytic experiments. H.Z. carried out genetic analysis. H.S. contributed to the preparation of data. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-20167-3.

Correspondence and requests for materials should be addressed to O.G.

Peer review information *Nature Communications* thanks Dipak Dey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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