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

Anonymous Authors[1]

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).

Deep learning, has recently achieved remarkable success in learning from high dimensional images and sequences (LeCun et al., 2015), but there are several challenges in applying neural networks to genomic data (Min et al., 2017). 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 augmentation, 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

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

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 ith 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 $X = \{x_1, ..., x_N\}$ are the samples, and $Y = \{E, S\}$ represents label vectors of event or last follow-up times $S = \{s_1, ..., s_N\}$ and event status $E = \{e_1, ..., e_N\}$. 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. $Y = \{Y^1, ..., Y^T\}$ and $X = \{X^1, ..., X^T\}$ 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 models extends SurvivalNet by addition of an adversarial cohort classification loss. Let $X_{comb} = \{x_1, ..., x_M\}$ and $Y_{comb} = \{y_1, ..., y_M\}$ denote the combination of all X^t and Y^t, respectively, including M samples in total. A set of one-hot vectors $Y_D = \{d_1, ..., d_M\}$ indicate cohort membership, so that $d_{it} = 1$ means that the ith sample belongs to the tth 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:

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

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

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).

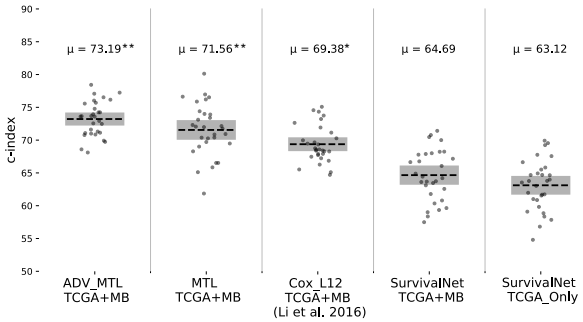

Figure 1. METABRIC and TCGA breast cancer datasets were combined to improve performance on TCGA, using the proposed and baseline methods. ⋆⋆ indicates significant improvement over both Cox–$\ell_{2,1}$ and single–task models. ⋆ indicates significant improvement over single–task models.

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).

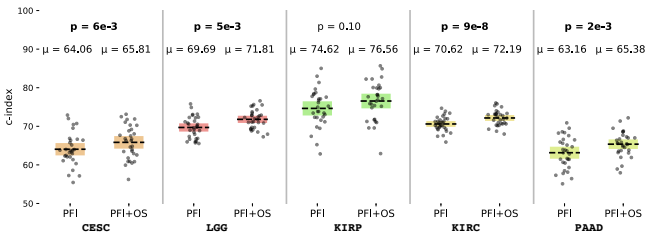

Figure 2. Progression-free interval (PFI) prediction performance with and without multi-task learning with overall survival (OS) labels.

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.

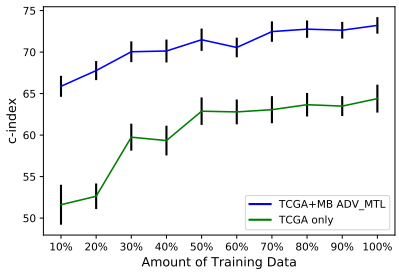

Figure 3. Learning curves of SurvivalNet and ADV–MTL (target: TCGA BRCA.

References

Abu-Mostafa, Y. S. The vapnik-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., Samarajiwa, 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.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Li, Y., Wang, L., Wang, J., Ye, J., and Reddy, C. K. Transfer learning for survival analysis via efficient l2, 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.

Min, S., Lee, B., and Yoon, S. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

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. L1-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 genomical? *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 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.

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.

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).

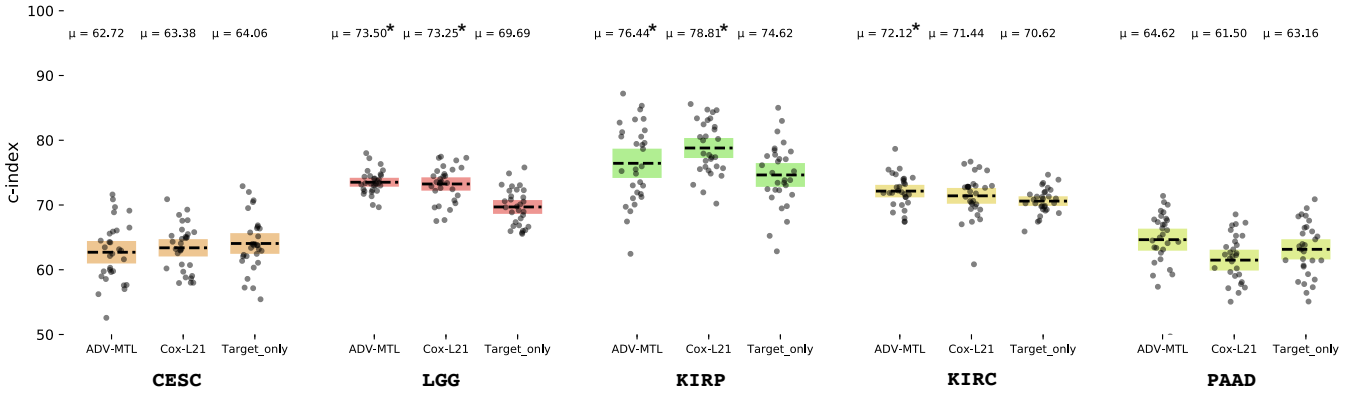

Figure S1. Survival prediction accuracy was improved by multi-task learning and adversarial representation learning on several benchmark datasets. $\star$ indicate significant improvement ($p < 0.05$) of multi-task methods over target–only setting.