Proceedings of Machine Learning Research 1:1–6, 2021 Survival Prediction: Algorithms, Challenges and Applications

Multi-ethnic Survival Prediction: Transfer Learning with Cox Neural Networks

Yan Gao YGAO45@UTHSC.EDU
Yan Cui YCUI2@UTHSC.EDU
Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA

Abstract

Extensive collections of personal omics data from large clinical cohorts provide an unprecedented opportunity to develop high-performance machine learning systems for precision medicine. However, most clinical omics data were collected from individuals of European ancestry. Such ancestrally imbalanced data may lead to inaccurate machine learning models for the data-disadvantaged ethnic groups and thus generate new health care disparities. In this work, we develop a transfer learning scheme for survival analysis with multi-ethnic data. We perform machine learning experiments on real and synthetic clinical omics datasets to show that transfer learning can improve the prognostic accuracy of Cox neural network models for data-disadvantaged ethnic groups.

Keywords: Survival analysis, Transfer learning, Cox neural network

1. Introduction

Clinical omics studies hold great promise to elucidate complex disease mechanisms and generate critical data for developing predictive analytics that is essential to precision medicine. However, recent statistics show that the vast majority of participants of clinical genetics and omics studies are of European descent Sirugo et al. (2019); Gurdasani et al. (2019); Guerrero et al. (2018); Martin et al. (2019). Such data inequality presents a great challenge for training machine learning models that work equally well for all ethnic groups. In a recent study Gao and Cui (2020), we found that current prevalent multi-ethnic machine learning schemes, mixture learning and independent learning, tend to generate low performance models for data-disadvantaged ethnic groups, and that transfer learning can reduce the model performance disparities. However, it is not clear how the clinical omics data inequality would affect the performance of Cox neural network Yousefi et al. (2017); Ching et al. (2018); Katzman et al. (2018); Luck et al. (2017); Kvamme et al. (2019); Wang et al. (2020) models for time-to-event prediction of clinical outcomes and whether transfer learning would be effective to address this challenge in time-to-event prediction tasks. Here, we extend the machine learning experiments supporting these conclusions from classification tasks to time-to-event prediction tasks, which are widely used in clinical studies. We investigate the impacts of data inequality on caner prognosis using Cox neural networks. We use machine learning experiments on both real and simulated cancer clinical omics data to show that Cox neural networks generate significant performance gaps between ethnic groups due to data inequality and data distribution discrepancy between ethnic groups, and that transfer learning can improve the prognostic prediction for data-disadvantaged ethnic groups.

2. Methods

We used protein expression data from TCGA (Genome Data Commons, `https://gdc.cancer.gov/about-data/publications/pancanatlas`) for overall survival (OS) prognosis for patients with kidney renal clear cell carcinoma (KIRC) or Glioma (GBMLGG). The KIRC-Protein-OS dataset contains expression data of 195 proteins and censored survival time data from 406 European American (EA) and 44 African American (AA) patients. The GBMLGG-Protein-OS dataset contains expression data of 176 proteins and censored survival time data from 582 EA and 48 AA patients.

Here we created a Cox neural network model with six layers. The input layer has 195 nodes for KIRC and 176 nodes for GBMLGG, a fully connected (FC) layer with 128 nodes followed by a dropout layer (drop out rate = 0.5), then another FC layer (with 64 nodes) also followed by a dropout layer (p = 0.5), and finally a Cox regression layer. We used the ReLU activation function for each FC layer to avoid the gradient vanish problem Goodfellow et al. (2016). In model fitting, we optimized the object function, $l(\beta) = -\sum_{i \epsilon U} L(\beta) + \lambda_1 |W| + \lambda_2 ||W||_2$, where $L(\beta) = log[\prod_{i=1}^{m} \frac{e^{\beta X_i}}{\sum_{j \epsilon R(T_i)} e^{\beta X_j}}]^{\delta_i} = \sum_{i=1}^{m} \delta_i [\beta X_i - log\left\{\sum_{j \epsilon R(T_i)} e^{\beta X_j}\right\}]$, β represents the weights of the Cox layer, $\sum_{i \epsilon U} L(\beta)$ is the partial likelihood, U is the set of uncensored patients, λ_1 and λ_2 are regularization parameters, W represents the weights in the network, δ_i is the event status of patient i.

In transfer learning, knowledge learned from the source domain where training data is abundant can be transferred to assist machine learning in the target domain where training data is inadequate Pan and Yang (2009); Tan et al. (2018); Weiss et al. (2016). Here we set EA as the source domain and AA as the target domain. We used two fine-tuning methods for transfer learning: (1) We pre-trained the Cox neural network model $M = f(T_{EA}, E_{EA}|X_{EA})$ using the EA group data, and then fine-tuned it with the AA group data: $M^{'} = finetuning(M|T_{AA}, E_{AA}, X_{AA})$, where X_k,T_k, and E_k denote the protein expression, event time, and the event status of group k. In the fine-tuning step, we used a smaller learning rate (0.002) since the model had been partially fitted. (2) The second fine-tuning method is based on stacked auto-encoder Sevakula et al. (2018); Singh et al. (2016); Vincent et al. (2010). We used the unlabeled data from the EA group to pre-train a stacked denoising auto-encoder with 5 layers: an input layer, a FC layer with 128 nodes, a bottleneck layer with 64 nodes, a FC layer with 128 nodes and an output layer with same nodes as the input layer.

For each experiment, we applied a 10-fold stratified cross-validation for training and testing split, in which samples are stratified on ethnicity and event status. We performed 20 runs for each experiment with different random partitions. We evaluated model performance using the concordance index Harrell et al. (1982) (C-index). We used the one-sided Wilcoxon Signed rank test to calculate the p-values for the statistical significance of the performance differences Fig. 1.

We developed a statistical model to generate synthetic datasets such that each generated dataset will contain two ethnic groups, EA and AA. The model comprises three types of parameters, N_1 and N_2, n_{de}, and $N = \{n_{u,v}|u, v \in [-1, 0, 1]\}$. The data inequality is controlled by N_1 and N_2, which denotes the total number of samples in the EA and AA groups. We generated the simulated feature matrix X_{ij} using the ssizeRNA package and used n_{de} to specify the number of differentially expressed features which controls the

marginal distribution difference between the two ethnic groups. For the i^{th} sample in group k, the survival time T_i^k was generated using the exponential Cox model Austin (2012) $T_i^k = \frac{U}{exp(\sum_{j=1}^{m} \beta_j^k x_{ij})}$, where x_{ij} is the j^{th} feature of individual i, $\beta_j^k \in [-1, 0, 1]$ represents the effect of feature j to survival time of group k, and U is an exponential distribution with a mean of 3000. We used a cut-off time threshold Thr^k to simulate the event status $E_i^k = \left\{ \begin{matrix} 0 \ if \ T_i^k < Thr^k \\ 1 \ otherwise \end{matrix} \right\}$ and set $T_i^k = Thr^k$ if $E_i^k = 0$. With two ethnic groups, a pair of β_j^1 and β_j^2 would have nine possible combinations. The number of features associated with each of the nine possible combinations is denoted as $N = \{n_{u,v} | u, v \in [-1, 0, 1]\}$, which controls the conditional distribution discrepancy between two ethnical groups.

3. Results

We compared the performance of different multi-ethnic machine learning schemes with seven experiments (Table 1). Fig. 1 shows the box plots of C-index from seven experiments on the multi-ethnic survival analysis for two types of cancer: kidney renal clear cell carcinoma (KIRC) or Glioma (GBMLGG). Each box plot represents 20 independent runs with different random training and testing data partition.

In mixture learning, data from both ethnic groups were used to train the models, and then the model was tested on the data of both ethnic groups (Mixture0), on EA data only (Mixture1) and on AA data only (Mixture2). We observed significant model performance gaps between EA (Mixture1) and AA (Mixture2) groups with p-values of 3.38×10^{-8} and 3.39×10^{-8} for KIRC and GBMLGG, respectively.

In independent learning, data of each ethnic group were used separately to train and test independent models for each ethnic group (Table 1). We also observed significant model performance gaps between EA (Independent1) and AA (Independent2) groups for both cancer types Fig. 1).

In naive transfer learning, the model trained using source domain (EA) data was directly applied to the target domain without any adaptation. The naive transfer learning also showed low performance on the data-disadvantaged AA group (Fig. 1).

Table 1: Multi-ethnic Machine Learning Scheme Comparison (*SD: Synthetic data)

Multi-ethnic ML Scheme	**Experiment**	**Ethnic Composition**		**C-index**			
		Training Data	**Testing Data**	**KIRC**	**GBMLGG**	**SD1***	**SD2***
Mixture Learning	Mixture 0	AA + EA	AA + EA	0.68	0.74	0.81	0.86
	Mixture 1		EA	0.69	0.75	0.83	0.87
	Mixture 2		AA	0.52	0.59	0.56	0.65
Independent Learning	Independent 1	EA	EA	0.69	0.75	0.85	0.89
	Independent 2	AA	AA	0.43	0.64	0.55	0.54
Naive Transfer	Naive Transfer	EA	AA	0.45	0.63	0.51	0.64
Transfer Learning	Transfer Learning	EA (source) AA (target)	AA	0.66	0.69	0.73	0.69

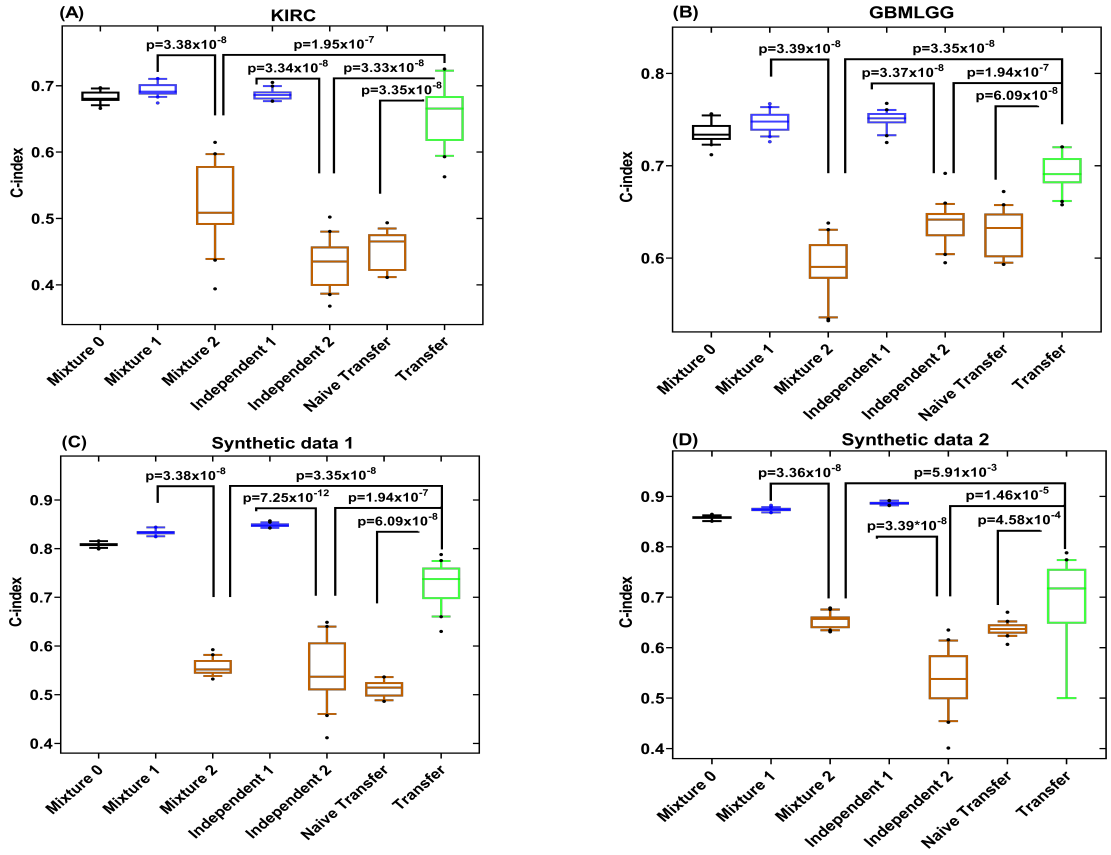

Figure 1: **Multi-ethnic machine learning scheme comparison on real and synthetic datasets**. Prediction of overall survival for (A) KIRC and (B) GBMLGG patients from protein expression data, (C) Synthetic Data 1, generated using parameters estimated from the KIRC dataset, and (D) Synthetic data 2, generated using parameters estimated from the GBMLGG dataset.

Using transfer learning, we achieved significant performance improvements for the data-disadvantaged AA group in both cancers (Fig. 1). The fine-tuning methods 1 and 2 showed the best performance for KIRC and GBMLGG respectively.

We generated two synthetic datasets using a statistical model described in the Method section. Synthetic datasets 1 and 2 was generated using parameters estimated from the KIRC and GBMLGG datasets respectively. For each synthetic dataset, the machine learning experiments showed a performance pattern similar to that of the real data (Fig. 1), which was characterized by performance gaps from the mixture and independent learning schemes and by transfer learning reduction of the performance gaps.

4. Discussion

Our results show that ancestrally imbalanced data may lead to significant model performance disparity in survival analysis with Cox neural networks and transfer learning can reduce the performance gaps by improving the model performance for data disadvantaged ethnic groups. We developed a new synthetic data generator to simulated multi-ethnic omics data associated with time-to-event clinical outcome endpoints. The experiments on the synthetic data show that the performance patterns of the multi-ethnic machine learning schemes can be reproduced from the synthetic data generated using the statistical model incorporating two key factors: data inequality and distribution mismatch between ethnic groups. This is consistent with our previous observations from machine learning experiments using binary classification tasks for omics-based cancer prognosis Gao and Cui (2020). We also performed new machine experiments on naïve transfer learning, which showed that the direct application of machine learning model learning from one ethnic groups to another ethnic group may lead to low performance.

References

Peter C Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in medicine*, 31(29):3946–3958, 2012.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

Yan Gao and Yan Cui. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nature communications*, 11(1):1–8, 2020.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Santiago Guerrero, Andrés López-Cortés, Alberto Indacochea, Jennyfer M García-Cárdenas, Ana Karina Zambrano, Alejandro Cabrera-Andrade, Patricia Guevara-Ramírez, Diana Abigail González, Paola E Leone, and César Paz-y Miño. Analysis of racial/ethnic representation in select basic and applied cancer research studies. *Scientific reports*, 8(1): 1–8, 2018.

Deepti Gurdasani, Inês Barroso, Eleftheria Zeggini, and Manjinder S Sandhu. Genomics of disease risk in globally diverse populations. *Nature Reviews Genetics*, 20(9):520–535, 2019.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1): 1–12, 2018.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *arXiv preprint arXiv:1907.00825*, 2019.

Margaux Luck, Tristan Sylvain, Héloïse Cardinal, Andrea Lodi, and Yoshua Bengio. Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*, 2017.

Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Rahul K Sevakula, Vikas Singh, Nishchal K Verma, Chandan Kumar, and Yan Cui. Transfer learning for molecular cancer classification using deep neural networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(6):2089–2100, 2018.

Vikas Singh, Nikhil Baranwal, Rahul K Sevakula, Nishchal K Verma, and Yan Cui. Layerwise feature selection in stacked sparse auto-encoder for tumor type prediction. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1542–1548. IEEE, 2016.

Giorgio Sirugo, Scott M Williams, and Sarah A Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11 (12), 2010.

Di Wang, Kevin He, and Lana X Garmire. Cox-nnet v2. 0: improved neural-network based survival prediction extended to large-scale emr dataset. *arXiv preprint arXiv:2009.04412*, 2020.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E Lewis, Congzheng Song, David A Gutman, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):1–11, 2017.