# Transfer learning for reducing health disparity arising from data inequality

Yan Gao[1,2] and Yan Cui[1,2,3,*]

[1]Department of Genetics, Genomics and Informatics, [2]Center for Integrative and Translational Genomics, [3]Center for Cancer Research, University of Tennessee Health Science Center, Memphis, TN 38163, USA

*Corresponding author. Email: ycui2@uthsc.edu

## Abstract

As artificial intelligence (AI) is increasingly involved in biomedical research and clinical decisions, developing unbiased AI models that work equally well for all racial and ethnic groups is of crucial importance to health disparity prevention and reduction. However, the long-standing data inequality between different racial/ethnic groups is generating new health disparities through data-driven, algorithm-based biomedical research, and clinical decisions. Here we found from a large set of machine learning experiments on cancer omics data that the current prevalent schemes of multiethnic machine learning are prone to generate significant model performance disparities between racial groups. We showed that this performance disparity is caused by data inequality and data distribution discrepancy between racial groups. We also found that transfer learning can improve the machine learning model performance on the data-disadvantaged racial groups, and thus, provides an approach to reduce the health disparities arising from the long-standing data inequality among racial groups.

Artificial intelligence (AI) is fundamentally transforming biomedical research and health care systems are increasingly relying on AI-based predictive analytics to make better diagnosis, prognosis, and therapeutic decisions[1-3]. Since data is the most important resource for developing high-quality AI models, the data inequality among racial/ethnic groups is becoming a global health problem in the AI era. Recent statistics showed that the samples of the cancer genomics research projects, including the TCGA[4], TARGET[5], OncoArray[6], and 416 cancer-related genome-wide association studies, were collected from Caucasians (91.1%), Asians (5.6%), African Americans (1.7%), Hispanics (0.5%) and other populations (0.5%)[7]. Most clinical genetics and genomics data have been collected from individuals of European ancestry and the cohort racial diversity has largely remained the same or even declined in recent years[8,9]. As a result, the non-Caucasian racial groups, which constitute about 84% of the world's population, have a long-term cumulative data disadvantage. Inadequate training data may lead to non-optimal AI models with low prediction accuracy and robustness, which will have profound negative impacts on health care for the data-disadvantaged racial/ethnic groups. Thus, the data inequality between racial groups is set to generate new health disparities.

The current prevalent scheme of machine learning with multiethnic data is the mixture learning scheme in which the data of all racial/ethnic groups are mixed and used indistinctly in model training and testing (**Fig. 1**). Under this scheme, it is unclear whether the machine model works well for all the racial/ethnic groups involved. We found from a large set of machine learning experiments that the mixture learning scheme tends to produce models with relatively low prediction accuracy for data-disadvantaged minority groups due to data distribution mismatch between the racial groups. Therefore, the mixture learning scheme often leads to unintentional and even unnoticed model performance gaps between racial groups. An alternative approach is the independent learning scheme in which data of different racial/ethnic groups are used separately to train independent models for each racial/ethnic group (**Fig. 1**). This learning scheme also tends to produce models with low prediction accuracy for data-disadvantaged minority groups due to inadequate training data. We found that the transfer learning scheme (**Fig. 1**), in many cases, can provide more accurate and robust machine learning models for data-disadvantaged racial groups. We anticipate this work to be a starting point for a new multiethnic machine learning paradigm which implements regular tests for the performance of machine learning models on all racial/ethnic groups to identify the model performance disparities between racial groups and uses transfer learning or other techniques to reduce the performance disparities. Such a new paradigm is essential for reducing health disparities arising from the long-standing data inequality among racial groups.

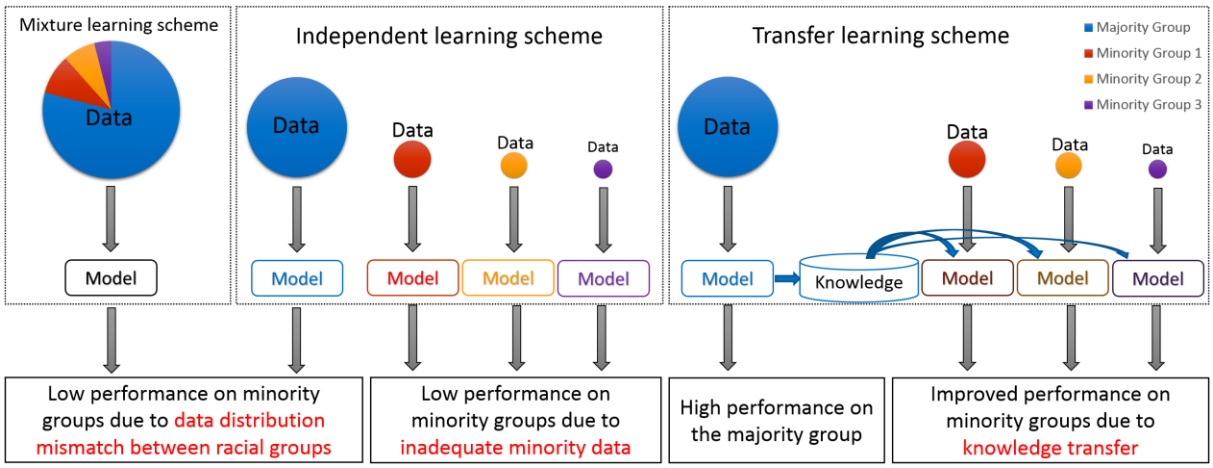



**Figure 1 | Multiethnic machine learning schemes.** In the mixture learning scheme, a model is trained and tested on the data of all racial/ethnic groups. In the independent learning scheme, a model is trained for each racial/ethnic group using its own data. In the transfer learning scheme, a model is trained on the majority group data, then the knowledge learned is transferred to assist the development of a model for each minority group.

## Genomic data inequalities among racial groups

Interrelated multi-omics factors including genetic polymorphisms, somatic mutations, epigenetic modifications, and expression alterations of RNAs and proteins collectively contribute to cancer pathogenesis and progression. The clinical omics data of large cancer cohorts provide an unprecedented opportunity to elucidate the complex molecular basis of cancers[10,11] and develop machine learning based predictive analytics for precision oncology[12-16]. However, the data inequality among racial groups continues to be conspicuous in recent large-scale genomics-focused biomedical research programs[7,17,18]. The TCGA cohort consists of 80.5% European Americans, 9.2% African Americans, 6.1% East Asian Americans and 0.3% Others[19]. The TARGET[5] and MMRF CoMMpass[20] cohorts have similar racial compositions[21], which are typical for current clinical omics datasets[7]. The data inequality among racial and ethnic groups is ubiquitous across almost all the cancer types in the TCGA and MMRF CoMMpass cohorts (see **Supplementary Fig. 1**); therefore, its negative impacts would be broad and not limited to the cancer types or subtypes for which racial disparities have already been investigated.

## Racial disparity in machine learning model performance

We assembled machine learning tasks using the omics data and clinical outcome endpoints[19] from the TCGA data of two racial groups: African American (AA) and European American (EA). A total of 1600 learning tasks were assembled using the combinations of four factors: 1) 40 types of cancers and pan-cancers[11]; 2) two types of features: mRNA and protein expression; 3) four clinical outcome endpoints: overall survival (OS), progression-free interval (PFI), disease-free interval (DFI), and disease-specific survival (DSS)[19]; and 4) five thresholds for the event time associated with the clinical outcome endpoints (**Supplementary Fig. 2**). For each learning task, each patient is assigned to the positive (or the negative) prognosis category based on whether the patient's event time for the clinical outcome endpoint of the learning task is no less than (or less than) a certain threshold.

Since the AA patients consist of less than 10% of the TCGA cohort, there are only very small numbers of AA cases in many learning tasks. We filtered out the learning tasks with too few cases to permit meaningful machine learning experiments. We then performed machine learning experiments on the remaining 447 learning tasks that have at least five AA cases and five EA cases in the positive and negative prognosis categories. For each machine learning task, we trained a deep neural network model for classification between the two prognosis categories using the mixture learning scheme. The mixture learning models achieved reasonably good baseline performance ($AUC > 0.65$) for 224 learning tasks. A total of 21 types of cancers and pan-cancers and all the four clinical outcome endpoints are represented in these learning tasks. The proportion of AA patients ranges from 0.06 to 0.25 in these learning tasks with a median of 0.12 (**Supplementary Fig. 3A**). For each of the 224 learning tasks, we performed six machine learning experiments (**Table 1**) to compare the performance of the three multiethnic machine learning schemes on the AA and EA groups (**Fig. 2**).

In the machine learning experiments, we observed that the mixture learning scheme is prone to produce biased models with a lower prediction performance for the data-disadvantaged AA group. The model performance differences between the EA and AA groups are statistically significant with a p-value of $1.19 \times 10^{-10}$ (**Fig. 2**, Mixture 1 & 2). The average EA-AA model performance gap over the 224 learning tasks is 0.06 (AUROC). Without testing the model performance of the machine learning models on each racial group separately, the performance disparity differences would be concealed by the over good performance on the entire multiethnic cohort (**Fig. 2**, Mixture 0). The independent learning scheme produced even larger EA-AA performance differences with a p-value of $4.37 \times 10^{-31}$ and the average performance gap is 0.12 (**Fig. 2**, Independent 1 & 2).

**Table 1. The machine learning experiments**

| Multiethnic Machine Learning Scheme | Experiment | Training Data Racial Composition | Testing Data Racial Composition | AUROC* | |
|---|---|---|---|---|---|
| | | | | Median | Mean |
| Mixture Learning | Mixture 0 | AA + EA | AA + EA | 0.71 | 0.72 |
| | Mixture 1 | | EA | 0.71 | 0.73 |
| | Mixture 2 | | AA | 0.68 | 0.67 |
| Independent Learning | Independent 1 | EA | EA | 0.70 | 0.71 |
| | Independent 2 | AA | AA | 0.59 | 0.58 |
| Transfer Learning | Transfer Learning | EA (source domain)<br>AA (target domain) | AA | 0.70 | 0.69 |

*Median and mean AUROC (Area under ROC curve) for each machine learning experiments on the 224 tasks.

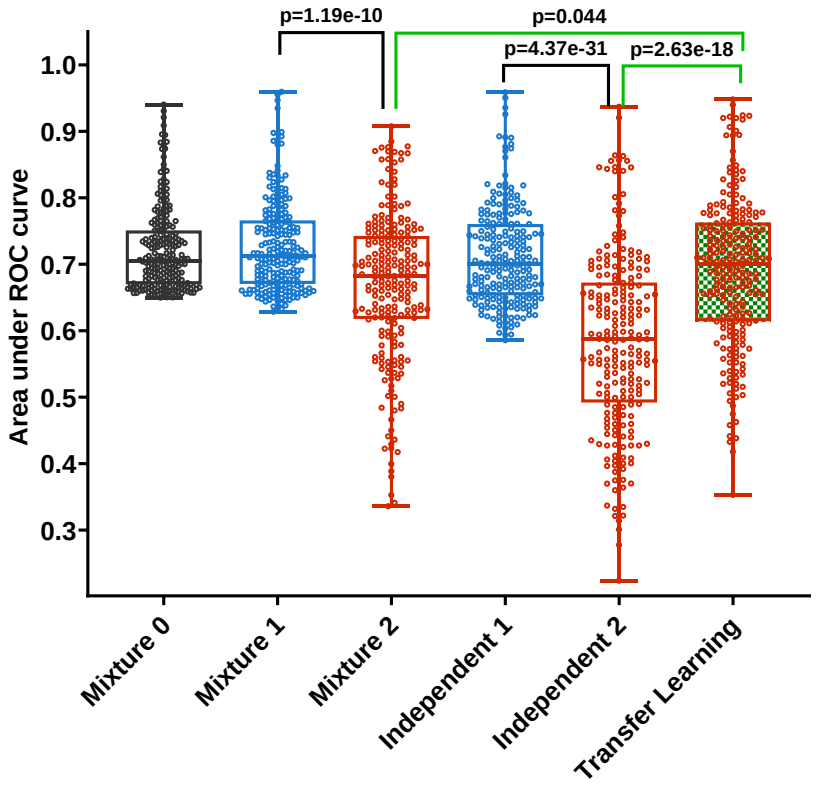



**Figure 2: The performance index for the multiethnic machine learning experiments.** Each boxplot shows the AUROC (Area under ROC curve) values of the 224 learning tasks for a machine learning experiment listed in Table 1. The Grey color represents performance on the whole cohort, the blue represents performance on the EA group, and the red color represents performance on the AA group.

## Transfer learning for improving machine learning model performance on data-disadvantaged racial groups

We compared the machine learning schemes for performance on the data-disadvantaged AA group and found that transfer learning produced models with significantly better performance on the AA group compared with the models from mixture learning ($p = 0.044$) and independent learning ($p = 2.63 \times 10^{-18}$) (**Fig. 2**). The machine learning experiment results for four learning tasks with different clinical endpoints as target variables are shown in **Fig. 3** (More results in **Supplementary Fig. 4**). We used 3-fold cross-validation and preformed 20 independent runs for each experiment with different random partitions of training and testing data to assess the machine learning model performance. The median AUROC of the six experiments are denoted as $A_{Mixture0}$, $A_{Mixture1}$, $A_{Mixture1}$, $A_{Independent1}$, $A_{Independent2}$, and $A_{Transfer}$. Results of these experiments show a consistent pattern characterized by:

1) Both mixture learning and independent learning schemes produced models with relatively high and stable performance on the EA group but low and unstable performance on the data-disadvantaged racial group (AA). We define the performance disparity gap $G = \overline{AUROC}_{EA} - \overline{AUROC}_{AA}$, where $\overline{AUROC}_{EA} = (A_{Mixture1} + A_{Independent1})/2$, and $\overline{AUROC}_{AA} = (A_{Mixture2} + A_{Independent2})/2$. $G$ is represented by the distance between the blue and red dash lines in **Fig. 3** and **Supplementary Fig. 4**.

2) The transfer learning provided the improved performance on the data-disadvantaged AA group, and thus reduced the model performance gap. The reduced model performance disparity gap is $\tilde{G} = \overline{AUROC}_{EA} - A_{Transfer}$, which is represented by the distance between the blue and green dash lines in **Fig. 3** and **Supplementary Fig. 4**.

Among the 224 learning tasks, 142 had a performance gap $G > 0.05$ and 88.7% (125/142) of the performance gaps were reduced by transfer learning.

We also performed the machine learning experiments on two additional learning tasks that involve either another race or dataset: 1) STAD-EAA/EA-2YR-PFI assembled using the TCGA Stomach Adenocarcinoma (STAD) data of EAA (East Asian American) and EA patients; and 2) MM-AA/EA-mRNA-3YR-OS assembled using the MMRF CoMMpass data of AA and EA patients. For both learning tasks show, machine learning experiments showed the same pattern of performance as described above (**Supplementary Fig. 4 A & B**).

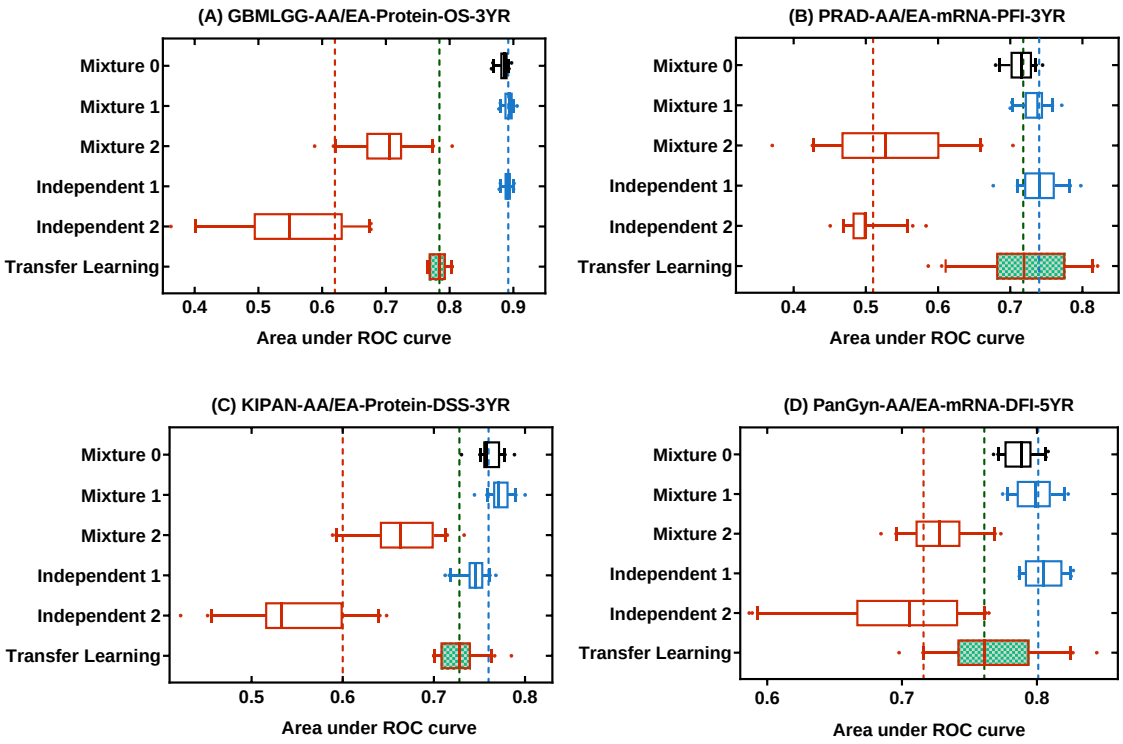



**Figure 3 | Comparison of the multiethnic machine learning schemes.** The machine learning tasks are: **(A)** GBMLGG-AA/EA-Protein-OS-3YR, **(B)** PRAD-AA/EA-mRNA-PFI-3YR, **(C)** KIPAN-AA/EA-Protein-DSS-3YR, **(D)** PanGyn-AA/EA-mRNA-DFI-5YR. In each panel, the box plots show the AUROC values for the six experiments (20 independent runs for each experiment). The red, blue and green vertical dash line represents $\overline{AUROC}_{AA}$, $\overline{AUROC}_{EA}$ and $A_{Transfer}$ respectively. The abbreviations for cancer types are explained in Supplementary Table 1.

## Key factors underlying the racial disparity in machine learning model performance

A machine learning task $\mathcal{T} = \{\mathcal{X},\ \mathcal{Y},\ f : \mathcal{X} \rightarrow \mathcal{Y}\}$ consists of a feature space $\mathcal{X}$, a label space $\mathcal{Y}$, and a predictive function $f$ learned from feature-label pairs. From a probabilistic perspective, $f$ can be written as $P(Y|X)$, where $X \in \mathcal{X}$, and $Y \in \mathcal{Y}$. It is generally assumed that each feature-label pair is drawn from a single distribution[22] $P(X,Y)$. However, this assumption needs to be tested for multiethnic omics data. Given $P(X,Y) = P(Y|X)P(X)$, both marginal distribution $P(X)$ and the conditional distribution $P(Y|X)$ may contribute to the data distribution discrepancy among racial groups. We used t-test to identify differentially expressed mRNAs or proteins between the AA and EA groups. The median percentage of differentially expressed mRNA or protein features

in the 224 learning tasks is 10%, and 70% of the learning tasks have at least 5% differentially expressed mRNA or protein features (**Supplementary Fig. 3B**). We used logistic regression to model the marginal distribution $f = P(Y|X)$, and calculated the Pearson correlation coefficient between the logistic regression parameters for the AA and EA groups. The Pearson correlation coefficient ranges from -0.14 to 0.26 in the learning tasks, with a median of 0.04 (**Supplementary Fig. 3C**). These results indicate that various degrees of marginal and conditional distribution discrepancies between the AA and EA groups exist in most of the 224 learning tasks.

We hypothesized that the data inequality represented by cohort racial composition and data distribution discrepancy between racial groups are the key factors underlying the racial disparity in machine learning model performance and that both factors can be addressed by transfer learning. To test this hypothesis, we performed the six machine learning experiments (Table 1) on synthetic data generated using a statistical model whose parameters represent the key factors underlying the machine learning model performance disparity between racial groups (Methods). Synthetic Data 1 was generated using the parameters estimated from the data for the learning task PanGyn-AA/EA-mRNA-DFI-5YR (**Fig. 3D**), which simulates the data inequality and distribution discrepancy between the racial groups in the real data (**Table 2**). On this synthetic dataset, the six machine learning experiments showed a performance pattern (**Fig. 4A**) similar to that of the real data (**Fig. 3**), which is characterized by the performance gaps from the Mixture and Independent learning schemes and by transfer learning reducing the performance gaps. Synthetic Data 2 has no distribution difference between the two racial groups (**Table 2**). For this dataset, there is no performance gap from the Mixture learning scheme, however, the performance gap from the Independent learning scheme remains (**Fig. 4B**). Synthetic Data 3 has equal numbers of cases from the two racial groups but has a distribution discrepancy between the two racial groups. Synthetic Data 4 has equal numbers of cases from the two racial groups and does not have a distribution difference between the two racial groups. For these two datasets, there is no significant performance gap from any learning scheme (**Fig. 4C&D**). These results confirm that the performance gap from the Mixture learning scheme is caused by the data distribution discrepancy between racial groups while the performance gap from the Independent learning scheme is caused by inadequate data for the disadvantaged racial group, and the transfer learning can reduce the performance gaps (**Fig. 1**).

**Table 2**. Multiethnic machine learning experiments on synthetic data

| Synthetic Data | Data Inequality | Distribution Discrepancy | Machine Learning Model Performance Gap | |
|---|---|---|---|---|
| | | | Mixture Learning | Independent Learning |
| 1 | Yes | Yes | Yes* | Yes* |
| 2 | Yes | No | No | Yes* |
| 3 | No | Yes | No | No |
| 4 | No | No | No | No |

*Performance gap > 0.05

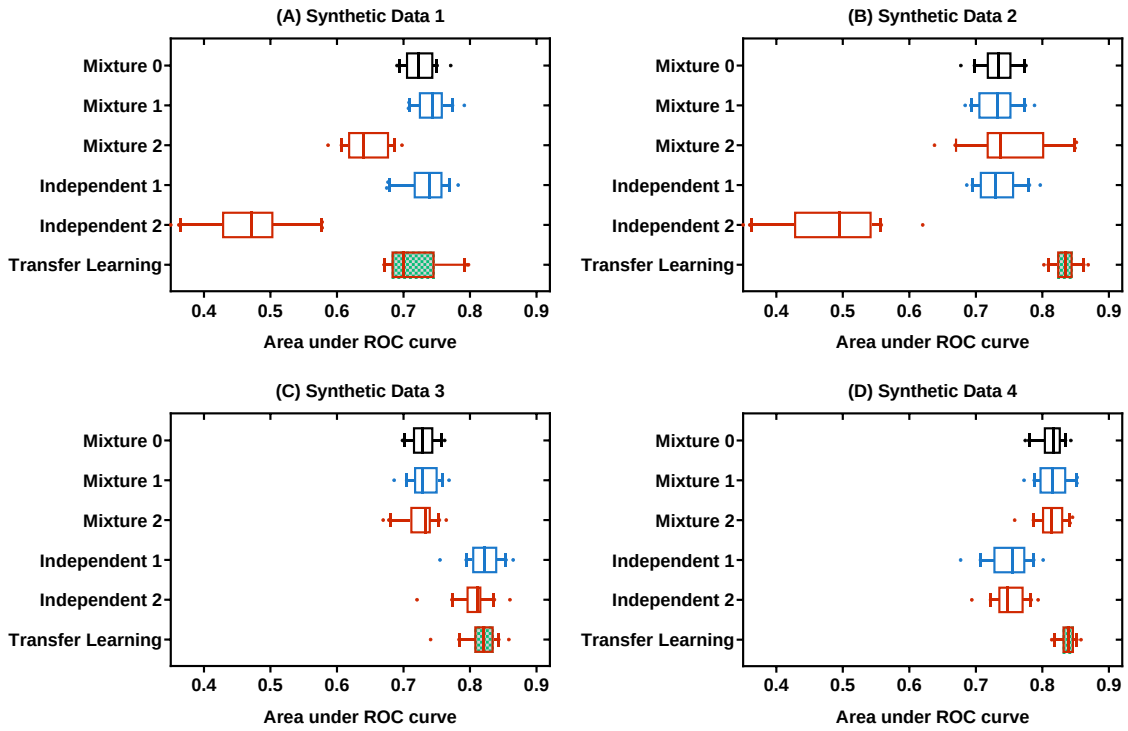



**Figure 4 | Comparison of the multiethnic machine learning schemes on synthetic data**. **(A)** Synthetic Data 1, **(B)** Synthetic Data 2, **(C)** Synthetic Data 3, **(D)** Synthetic Data 4. We used 3-fold cross validation and preformed 20 independent runs for each experiment with different random partitions of training and testing data to assess the machine learning model performance. In each panel, the box plots show the AUROC values for the six experiments (20 independent runs for each experiment).

# Conclusion

In this work, we showed that the current prevalent scheme for machine learning with multiethnic data, the mixture learning scheme, and its main alternative, the independent learning scheme, tend to generate machine learning models with relatively low performance on data-disadvantaged racial groups due to inadequate training data and data distribution discrepancy among racial groups. We also found that transfer learning can provide improved machine learning models for the data-disadvantaged racial groups by leveraging knowledge learned from the EA group that has abundant data. These results indicate that transfer learning can provide an approach to reduce the health disparity arising from the data inequality among racial groups.

# Methods

## Data source and data preprocessing

All the TCGA and MMRF CoMMpass data used in this work were downloaded from the NCI Genome Data Commons (GDC, https://gdc.cancer.gov/about-data/publications/pancanatlas). The racial groups of TCGA patients were determined based on the genetic ancestry data downloaded from The Cancer Genetic Ancestry Atlas (TCGAA, http://52.25.87.215/TCGAA). The racial groups of MMRF CoMMpass patients were determined based on the race information in the clinical data file downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov).

For the TCGA data, we used all the 189 protein expression features and the 17176 mRNA features without missing values. We further removed the samples with over 20% missing values. We also filtered out samples missing genetic ancestry or clinical endpoint data. The data matrix was standardized such that each feature has a zero mean and unit standard deviation. The ANOVA F-value for each mRNA was calculated for the training samples to select 200 mRNAs as the input features for machine learning. The feature mask, ANOVA F-value, and p values were calculated using the SelectKBest function (with the f_classif score function and k=200) of the python sklearn package[23]. For the MMRF CoMMpass data, we selected 600 mRNA features with the highest mean absolute deviation as the input features for machine learning.

## Deep neural network modeling

We used the Lasagne (https://lasagne.readthedocs.io/en/latest/) and Theano python packages (http://deeplearning.net/software/theano/) to train the deep neural networks (DNN). We used the

pyramid architecture[24] with 6 layers: an input layer with 200 nodes for mRNA features or 189 nodes for protein features, 4 hidden layers including a fully connected layer with 128 nodes followed by a dropout layer[25] and a fully connected layer with 64 nodes followed by a dropout layer, and a logistic regression output layer. To fit a DNN model, we used the stochastic gradient descent method with a learning rate of 0.01 (lr=0.01) to find the weights that minimize a loss function consisting of a cross-entropy and two regularization terms: $l(W) = -\sum_{i=1}^{m}(y_i\, log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) + \lambda_1|W| + \lambda_2\|W\|_2$ , where $y_i$ is the observed label of patient $i$, $\hat{y}_i$ is the predicted label for patient $i$, and $W$ represents the weights in the network. Traditional activation functions such as the sigmoid and hyperbolic tangent functions have a gradient vanish problem in training a deep learning model, which may lead to gradient decreasing quickly and training error propagating to forward layers. Here, we use the ReLU function $f(x) = max(0, x)$, which is commonly used in deep learning to avoid the gradient vanish problem. For each drop out layer, we set the dropout probability $p = 0.5$ to randomly omit half of the weights during the training to reduce the collinearity between feature detectors. To speed up the computation and improve the accuracy, we split the data into multiple mini-batches during the training. We used a batch size of 20 (batch_size=20) for two basic learning schemas (mixture learning and independent learning for the EA group) as there are relatively large numbers of cases available for training. For the independent learning for the AA group, we set the batch size to 4 because the number of cases available for training is limited. We set the maximum number of iteration to be 100 (max_iter=100) and applied the Nesterov momemtum[26] method (with momentum=0.9 for each DNN model) to avoid the premature stopping. We set the learning rate decay factor to be 0.03 (lr_decay=0.03) to avoid non-convergence during the training. The two regularization terms $\lambda_1$ and $\lambda_2$ were set to be 0.001.

## Transfer learning

In transfer learning[27-30], we set the EA group as the source domain and the AA or EAA group as the target domain. We applied three transfer learning methods to each learning task and selected the best AUROC as the performance index for the transfer learning scheme. The three transfer learning methods include two fine-tuning algorithms and a domain adaptation algorithm.

(1) Fine-tuning algorithm 1

Recent studies showed that fine-turning of deep neural networks (DNN) often leads to better performance and generalization in transfer learning[31]. We first pretrained a DNN model using source domain data: $M \sim f(Y_{Source}|X_{Source})$, which has the same architecture as described in the

previous section. The training parameters were set as lr=0.01, batch_size=20, $p = 0.5$, max_iter=100, and momentum=0.9. After the initial training, the DNN model was then fine-tuned using backpropagation on the target domain data: $M' = fine_tuning(M \mid Y_{Target}, X_{Target})$, where $M'$ is the final model. In the fine-tuning, the learning rate was set to be 0.002 and the batch size was 10 as the model has been partially fitted and the target dataset is small.

(2) Fine-tuning algorithm 2

In the second fine-tuning algorithm, the source domain data were used as unlabeled data to pretrain a stacked denoising autoencoder[32-34]. The stacked denoising auto-encoder has 5 layers: the input layer, a coding layer with 128 nodes, a bottleneck layer with 64 nodes, a decoding layer with 128 nodes, and an output layer has the same number of nodes with the input layer to reconstruct the input data. We used the source and target domain data to train the stacked auto-encoder with the parameters: learning rate=0.01, corruption level=0.3, batch size=32, and the maximum iteration=500. After pretraining the autoencoder, we removed the decoder and added a drop out layer (with p=0.5) after each hidden layer, and then we added a fine-tune (logistic regression) layer on the top of the last encoder layer. The network was then fine-tuned on the target domain data with a decreased learning rate (lr=0.002) as the model has already been partially fitted. The final DNN model has the same architecture as described in the previous section and was fine-tuned on target domain data with training parameters batch_size=10 and max_iter=100.

(3) Domain Adaptation

Domain adaptation is a class of transfer learning methods that improve machine learning performance on the target domain by adjusting the distribution discrepancy across domains[35,36]. We adopted the Contrastive Classification Semantic Alignment (CCSA) method[37] for domain adaptation. The CCSA method is particularly suitable for our transfer learning tasks because: 1) this method can significantly improve target domain prediction accuracy by using very few labeled target samples in training; 2) this method includes semantic alignment in training and therefore can handle the domain discrepancy in both marginal and conditional distributions. To use the CCSA method which calculates the pairwise Euclidean distance between samples in the embedding space, we applied a $L2$ norm transformation to the features of each patient such that for each patient $i$, ($\sum_{j=1}^{n} x_j^2 = 1$). The CCSA minimizes the loss function $L_{CCSA}(f) = (1 - \gamma)\, L_C(h\, o\, g) + \gamma(L_{SA}(h) + L_S(g))$, where $f = h\, o\, g$ is the target function, $g$ is an embeding function that maps the input $X$ to an embedding space $Z$, and $h$ is a function to predict the output

labels from $Z$, $L_C(f)$ denotes the classification loss (binary cross entropy) of function $f$, $L_{SA}(h)$ refers to the semantic alignment loss of function $h$, $L_S(g)$ is the separation loss of function $g$, $\gamma$ is the weight used to balance the classification loss versus the contrastive semantic alignment loss $L_{SA}(h) + L_S(g)$, $L_{SA}(h) = \frac{1}{n}\sum_{y_i^s = y_j^t} \frac{1}{2}\left\|g(x_i^s), g(x_j^t)\right\|^2$ and $L_S(g) = \frac{1}{n}\sum_{y_i^s \neq y_j^t} \frac{1}{2}\max(0, m - \left\|g(x_i^s), g(x_j^t)\right\|)^2$, $\|.\|$ is the Euclidean distance, $m$ is the margin that specifies the separability of the two domain features in the embedding space. During the training, we set m to 0.3, momentum to 0.9, batch_size to 20, learning_rate to 0.01, and max_iter=100. We used one hidden layer with 100 nodes for semantic alignment and 2 hidden layers (1 fully connected and 1 dropout layer) for classification.

## Differential expression analysis

For each learning task, we performed the permutation-based t-test on the input features to select the proteins or miRNAs that differentially expressed between the AA and EA groups. The miRNAs and proteins with a feature-wise p-value < 0.05 were selected as differentially expressed features between the two racial groups.

## Logistic regression

For each learning task, we split the data of the AA and EA groups, and then fit two multivariate logistic regression models: $Y^{AA} = 1/(1 + e^{-\boldsymbol{\beta}^{AA} * X^{AA}})$, $Y^{EA} = 1/(1 + e^{-\boldsymbol{\beta}^{EA} * X^{EA}})$, to calculate the regression parameters for each racial group.

## Synthetic data generator

We developed a statistical model to generate synthetic data for the multiethnic machine learning experiments. The simulated cohort consists of two racial groups. The degree of data inequality is controlled by the numbers of individuals of the two racial groups $n_1$ and $n_2$. We used the ssizeRNA package[38] to generate the mRNA feature matrix $x_{ij}$. The number of differentially expressed features ($n_{de}$) is the parameter controlling marginal distribution ($P(X)$) discrepancy between the two racial groups. For individual $i$ in racial group $k$, the label $y_i^k$ was generated using the logistic regression function: $y_i^k = \begin{cases} 1 & if\ z_i^k > c^k \\ -1 & otherwise \end{cases}$, where $z_i^k = \frac{1}{1 + e^{-\sum_{j=1}^{n} \beta_j^k x_{ij}}}$, $x_{ij}$ is the $j^{th}$ feature of individual $i$, and $\beta_j^k \epsilon \{-1, 1\}$ denotes the effect of feature $j$ on the label of racial group $k$, $c^k$ is the threshold for assigning a sample to the positive or negative category. A pair of $\beta_j^1$ and $\beta_j^2$ have four possible combinations representing the difference and similarity of the effect of

feature $j$ on the clinical outcome for the patients in the two racial groups. The number of the features associated with each of the four combinations is denoted as $n_{-1,-1}$, $n_{-1,1}$, $n_{1,-1}$, and $n_{1,1}$ respectively. These parameters control the conditional distribution ($P(Y|X)$) discrepancy between the two racial groups. We can generate synthetic datasets with or without data inequality and/or distribution discrepancy between two racial groups by setting the parameter values. These parameters can also be estimated from a real dataset. For example, we generated Synthetic Data 1 using the parameters estimated from the data for the learning task PanGyn-AA/EA-mRNA-DFI-5YR. We set $n_1$ and $n_2$ to equal to the number of EA and AA patients in the real data, respectively. We estimated the parameters $n_{de}$ using permutation-based t-test (feature-wise p-value < 0.05). The total number of features for the learning task PanGyn-AA/EA-mRNA-DFI-5YR is 200. We used multivariate logistic regression to calculate the regression parameters $\boldsymbol{\beta^{AA}}$ and $\boldsymbol{\beta^{EA}}$. Then we let $\beta_j^1 = \begin{cases} 1 & if\ \beta_j^{\mathrm{EA}} > median(\boldsymbol{\beta^{EA}}) \\ -1 & otherwise \end{cases}$ and $\beta_j^2 = \begin{cases} 1 & if\ \beta_j^{\mathrm{AA}} > median(\boldsymbol{\beta^{AA}}) \\ -1 & otherwise \end{cases}$, and calculated $n_{-1,-1}$, $n_{-1,1}$, $n_{1,-1}$, and $n_{1,1}$. The parameters used to generate Synthetic Data 1-4 are shown in the **Table 3.**

**Table 3**. The parameters used to generate the synthetic data

| Synthetic Data | Data inequality Parameters | | Distribution Discrepancy Parameters | | | | |
|---|---|---|---|---|---|---|---|
| | $\boldsymbol{n_1}$ | $\boldsymbol{n_2}$ | $\boldsymbol{n_{de}}$ | $\boldsymbol{n_{-1,-1}}$ | $\boldsymbol{n_{-1,1}}$ | $\boldsymbol{n_{1,-1}}$ | $\boldsymbol{n_{1,1}}$ |
| 1 | 218 | 43 | 20 | 64 | 37 | 37 | 62 |
| 2 | 218 | 43 | 0 | 100 | 0 | 0 | 100 |
| 3 | 260 | 260 | 20 | 64 | 37 | 37 | 62 |
| 4 | 260 | 260 | 0 | 100 | 0 | 0 | 100 |

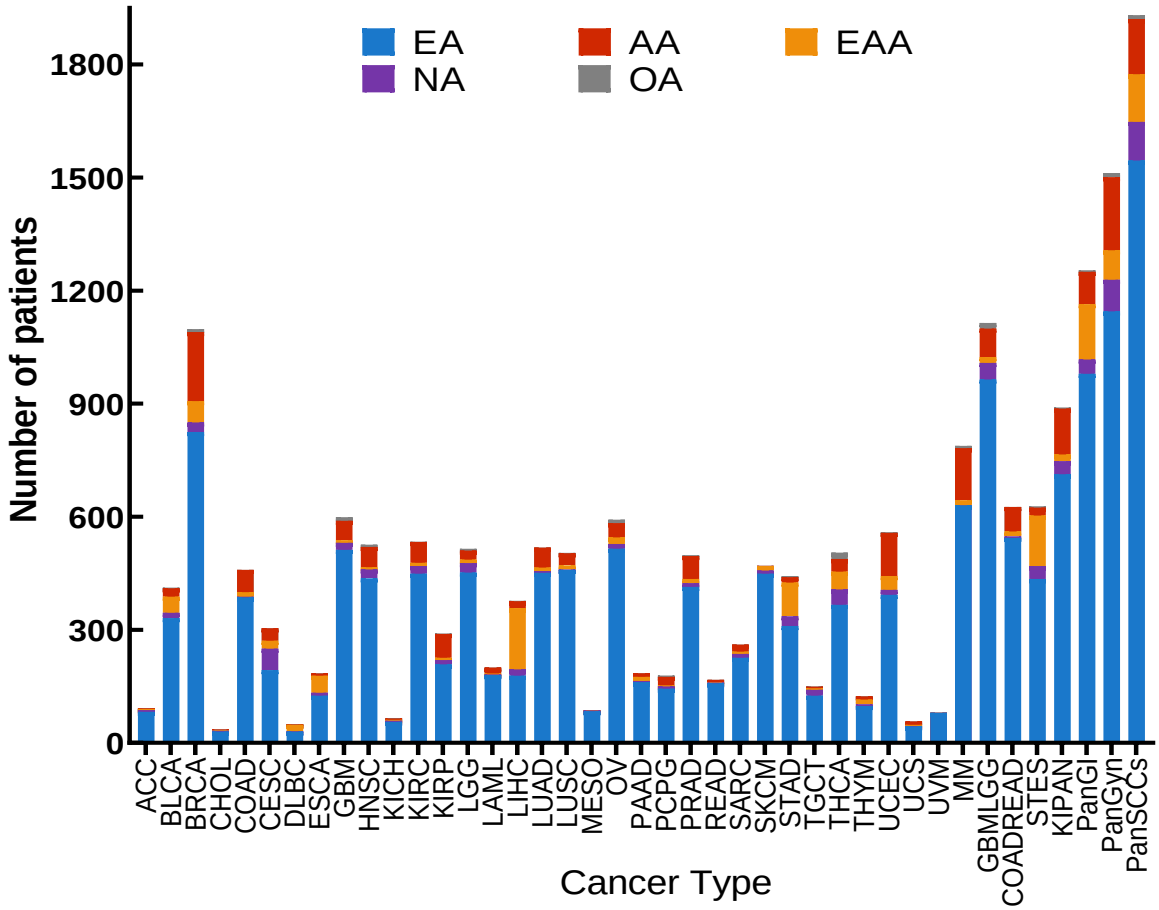
# Reference


1. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **25**, 44-56 (2019).
2. Azuaje, F. Artificial intelligence for precision oncology: beyond patient stratification. *NPJ precision oncology* **3**, 6 (2019).
3. Rajkomar, A., Dean, J. & Kohane, I. Machine Learning in Medicine. *New England Journal of Medicine* **380**, 1347-1358 (2019).
4. TCGA Research Network. (https://portal.gdc.cancer.gov/).
5. The Therapeutically Applicable Research to Generate Effective Treatments initiative. (https://ocg.cancer.gov/programs/target).

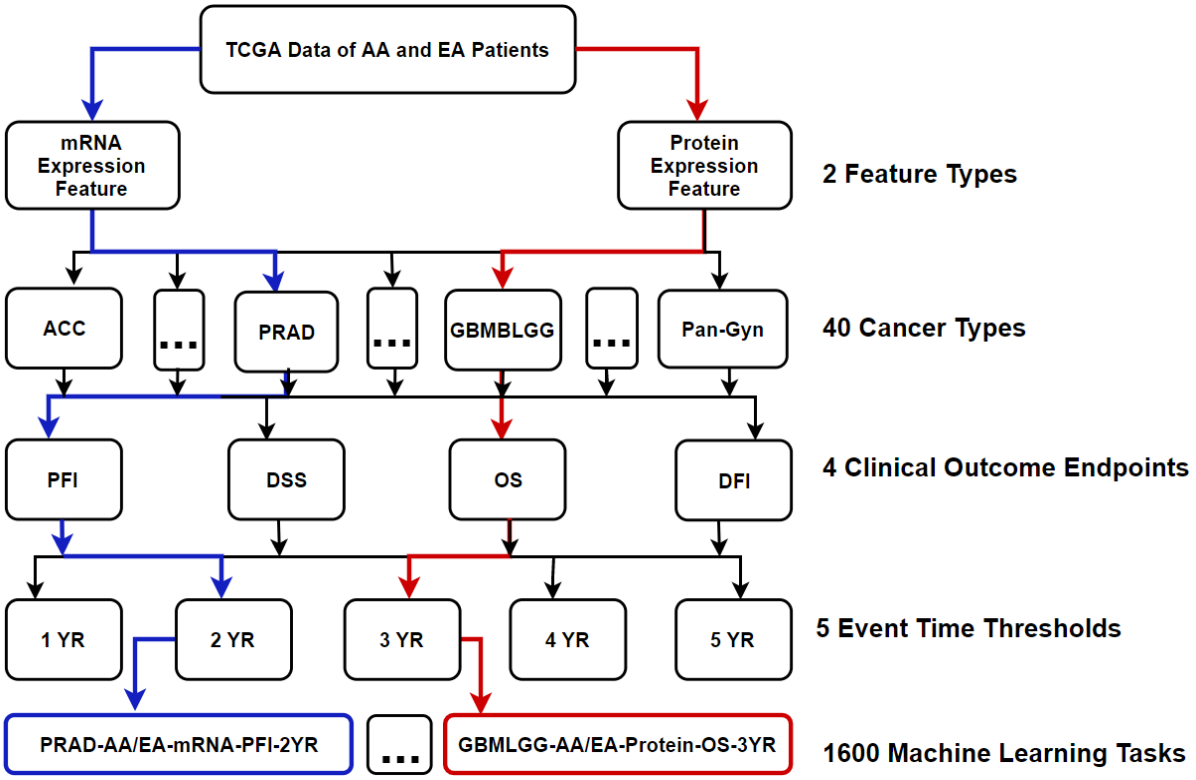
6. Amos, C.I. *et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. **26**, 126-135 (2017).
7. Guerrero, S. *et al.* Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. *Scientific Reports* **8**, 13978 (2018).
8. Genetics for all. *Nature Genetics* **51**, 579-579 (2019).
9. Martin, A.R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics* **51**, 584 (2019).
10. Hutter, C. & Zenklusen, J.C. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* **173**, 283-285 (2018).
11. Hoadley, K.A. *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291-304 (2018).
12. Malta, T.M. *et al.* Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* **173**, 338-354 (2018).
13. Way, G.P. *et al.* Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas. *Cell reports* **23**, 172-180 (2018).
14. Yousefi, S. *et al.* Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports* **7**, 11707 (2017).
15. Ching, T., Zhu, X. & Garmire, L.X. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology* **14**, e1006076 (2018).
16. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469 (2018).
17. Kim, J.I.E. & Sarkar, I.N. Racial Representation Disparity of Population-Level Genomic Sequencing Efforts. *Studies in health technology and informatics* **264**, 974-978 (2019).
18. Lyles, C.R., Lunn, M.R., Obedin-Maliver, J. & Bibbins-Domingo, K. The new era of precision population health: insights for the All of Us Research Program and beyond. *Journal of Translational Medicine* **16**, 211 (2018).
19. Liu, J. *et al.* An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400-416 (2018).
20. The Relating Clinical Outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile. (https://themmrf.org/we-are-curing-multiple-myeloma/mmrf-commpass-study/).
21. Genomic Data Commons Data Portal. (https://portal.gdc.cancer.gov/).
22. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N.D. *Dataset shift in machine learning*, (The MIT Press, 2009).
23. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825-2830 (2011).
24. Phung, S.L. & Bouzerdoum, A. A pyramidal neural network for visual pattern recognition. *IEEE transactions on neural networks* **18**, 329-343 (2007).
25. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929-1958 (2014).
26. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. On the importance of initialization and momentum in deep learning. in *International conference on machine learning* 1139-1147 (2013).
27. Pan, S.J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**, 1345-1359 (2010).
28. Tan, C. *et al.* A survey on deep transfer learning. in *International Conference on Artificial Neural Networks* 270-279 (Springer, 2018).
29. Weiss, K., Khoshgoftaar, T.M. & Wang, D. A survey of transfer learning. *Journal of Big data* **3**, 9 (2016).

30. Taroni, J.N. *et al.* MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell systems* **8**, 380-394 (2019).
31. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? in *Advances in neural information processing systems* 3320-3328 (2014).
32. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* **11**, 3371-3408 (2010).
33. Sevakula, R.K., Singh, V., Verma, N.K., Kumar, C. & Cui, Y. Transfer Learning for Molecular Cancer Classification Using Deep Neural Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **16**, 2089-2100 (2019).
34. Singh, V., Baranwal, N., Sevakula, R.K., Verma, N.K. & Cui, Y. Layerwise feature selection in Stacked Sparse Auto-Encoder for tumor type prediction. in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1542-1548 (2016).
35. Tzeng, E., Hoffman, J., Saenko, K. & Darrell, T. Adversarial discriminative domain adaptation. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7167-7176 (2017).
36. Daume Iii, H. & Marcu, D. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research* **26**, 101-126 (2006).
37. Motiian, S., Piccirilli, M., Adjeroh, D.A. & Doretto, G. Unified deep supervised domain adaptation and generalization. in *Proceedings of the IEEE International Conference on Computer Vision* 5715-5725 (2017).
38. Bi, R. & Liu, P. Sample Size Calculation for RNA-Seq Experimental Design–the ssizeRNA package. *BMC Bioinformatics* **17**, 146 (2016).

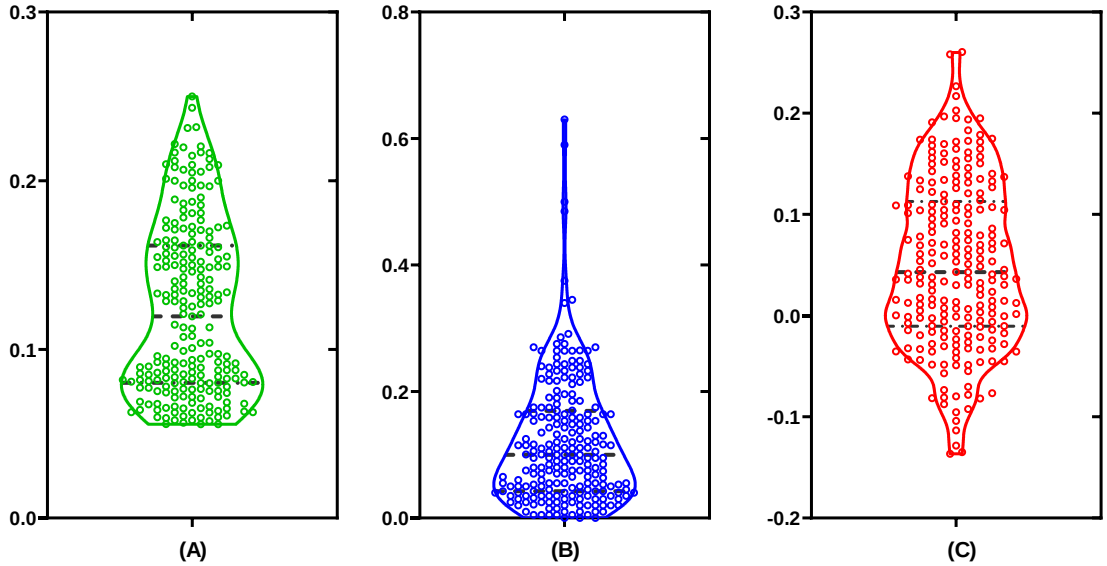
## Supplementary Data



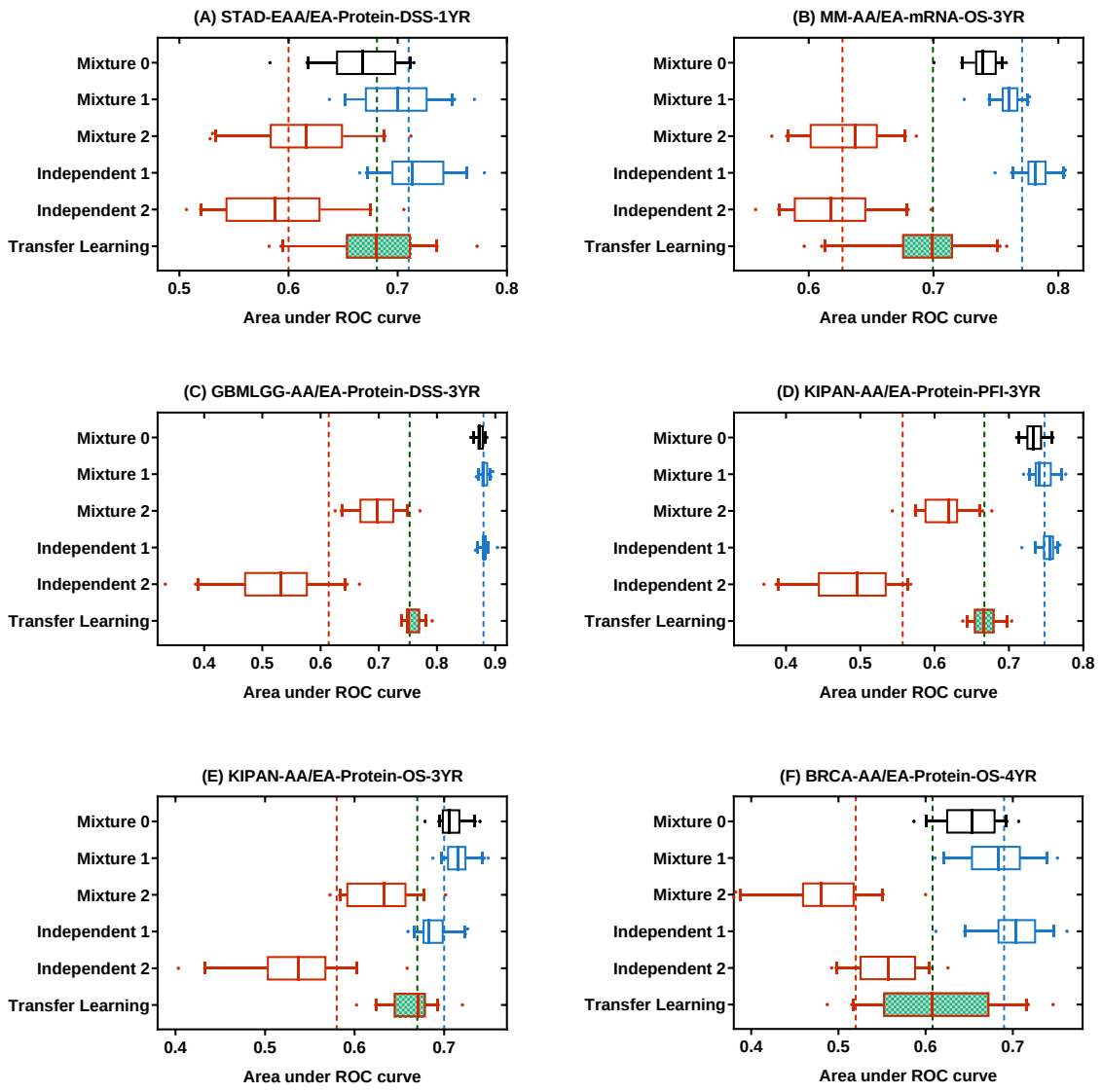

**Supplementary Figure 1 | The racial compositions of the TCGA and MMRF CoMMpass cohorts for each cancer type.** The abbreviations used in the figure: TCGA: The cancer genome atlas; AA, African American; EA, European American; EAA, East Asian American; NA, Native American; OA, Other; the abbreviations for cancer types are explained in Supplementary Table 1.




**Supplementary Figure 2 | The assembly of machine learning tasks.** The machine learning tasks were assembled using the combinations of four factors: feature type, cancer type, clinical outcome endpoint and event time for the clinical outcome endpoints. Each path (e.g. the red path and blue path) represents the assembly of a machine learning task. The abbreviations used in the figure: TCGA: The cancer genome atlas; PFI, progression-free interval; DSS, disease-specific survival; OS, overall survival; DFI, disease-free interval; the abbreviations for cancer types are explained in Supplementary Table 1.




**Supplementary Figure 3 | Assessment of the key factors underlying the racial disparity in machine learning model performance. (A)** The proportions of AA patients. **(B)** The proportions of differentially expressed mRNA or protein features between the AA and EA groups. **(C)** The Pearson correlation coefficients between the logistic regression parameters for the AA and EA groups in the 223 learning tasks (Supplementary Table 1).




**Supplementary Figure 4 | Comparison of the multiethnic machine learning schemes on cancer omics data.** The machine learning tasks are: **(A)** STAD-EAA/EA-Protein-DSS-1YR, **(B)** MM-AA/EA-mRMA-OS-3YR, **(C)** GBMLGG-AA/EA-Protein-DSS-3YR, **(D)** KIPAN-AA/EA-Protein-PFI-3YR, **(E)** KIPAN-AA/EA-Protein-OS-3YR, **(F)** BRCA-AA/EA-Protein-OS-4YR. In each panel, the box plots show the AUROC values for the six experiments (20 independent runs for each experiment). The red, blue and green vertical dash line represents $\overline{AUROC}_{AA}$, $\overline{AUROC}_{EA}$ and $A_{Transfer}$ respectively. The machine learning experiments are described in Table 1. The abbreviations for cancer types are explained in Supplementary Table 1.