Data Inequality, Machine Learning and Health Disparity

Yan Gao [1] **Yan Cui** [1]

Abstract

Over 80% of clinical genetics and omics data were collected from individuals of European ancestry (EA), which comprise approximately 16% of the world's population. This severe data disadvantage for the non-EA populations is set to generate new health disparities as machine learning powered biomedical research and health care become increasingly common. The new health disparity arising from data inequality can potentially impact all data-disadvantaged ethnic groups in all diseases where data inequality exists. Thus, its negative impact is not limited to the diseases for which significant racial/ethnic disparities have already been evident. In a recent work (Gao & Cui, 2020), we showed that the current prevalent scheme for machine learning with multi-ethnic data, the mixture learning scheme, and its main alternative, the independent learning scheme, are prone to generating machine learning models with relatively low performance for data-disadvantaged ethnic groups due to inadequate training data and data distribution discrepancies among ethnic groups. We found that transfer learning can provide improved machine learning models for data-disadvantaged ethnic groups by leveraging knowledge learned from other groups having more abundant data. These results indicate that transfer learning can provide an effective approach to reduce health care disparities arising from data inequality among ethnic groups.

1. The impact of data inequality on machine learning and health equity

Over 80% of the GWAS (genome-wide association study) and clinical omics data ware collected from individuals of European Ancestry (EA). Other populations, which constitute about 84% of the world's population, have a severe data disadvantage. Inadequate training data may lead to non-optimal Artificial Intelligence (AI) models with low prediction accuracy, which means AI-powered precision medicine would be less precise for the data-disadvantaged minority groups (Fig 1). Thus, the data inequality among ethnic groups may lead to new health care disparities through AI models.

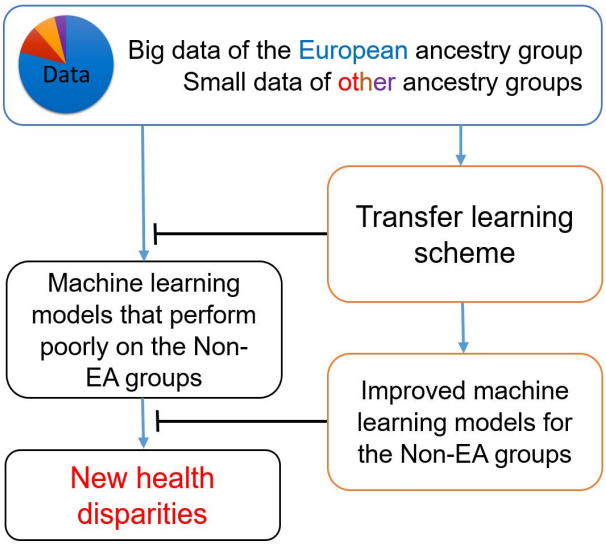

Figure 1. Transfer learning for prevention and reduction of health disparities arising from data inequality.

2. Multiethnic machine learning schemes

For machine learning with multiethnic data, the current prevalent scheme is the **mixture learning** scheme, in which data for all ethnic groups are mixed and used indistinctly for model training and testing. A major alternative approach is the **independent learning** scheme, in which data from each ethnic group is used for model training and testing. Here we propose to use **transfer learning** for multi-ethnic data. In transfer learning, a model is pre-trained on the majority group data and then the knowledge learned is transferred to assist model development for each data-disadvantaged ethnic group.

[1]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA. Correspondence to: Yan Cui <ycui2@uthsc.edu>.

The 2021 ICML Workshop on Computational Biology.

3. Results

We used a large set of machine learning experiments (224 learning tasks of predicting cancer clinical outcomes from mRNA and protein expression data) to show that the current prevalent multiethnic machine learning schemes, mixture learning and independent learning, tend to generate machine learning models with relatively low performance for data-disadvantaged ethnic groups due to inadequate training data and data distribution discrepancies among ethnic groups. We also found that transfer learning can improve machine learning model performance for data-disadvantaged ethnic groups, and thus provides an effective approach to reduce health care disparities arising from data inequality among ethnic groups (Gao & Cui, 2020) (Fig.3).

4. Methods

The 224 machine learning tasks were assembled using The Cancer Genome Atlas (TCGA) data. A total of 21 types of cancers and four clinical outcome endpoints, overall survival (OS), disease-specific survival (DSS), progression-free interval (PFI), and disease-free interval (DFI), were represented in these learning tasks. We used a six-layer neural network with the pyramid architecture as base model of machine learning, and used two fine-tuning algorithms and a domain adaptation method named Contrastive Classification Semantic Alignment (CCSA) for transfer learning.

5. Key Messages

- Biomedical data-disadvantage has become a ubiquitous health risk factor for the non-European populations (84% of the world's population) in the AI era.
- Machine learning profoundly empowers genomic medicine but in the meantime opens up a major pathway for the manifestation of this risk factor.
- There is an urgent need to block this risk factor from manifesting through the pathway of machine learning.
- Using transfer learning may reduce the negative impacts of the data inequality on health equity.

References

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

Multiethnic Machine Learning Scheme	Experiment	Training Data Composition	Testing Data Composition
Mixture Learning	Mixture 0	AA + EA	AA + EA
	Mixture 1		EA
	Mixture 2		AA
Independent Learning	Independent 1	EA	EA
	Independent 2	AA	AA
Transfer Learning	Transfer Learning	EA (source domain) AA (target domain)	AA

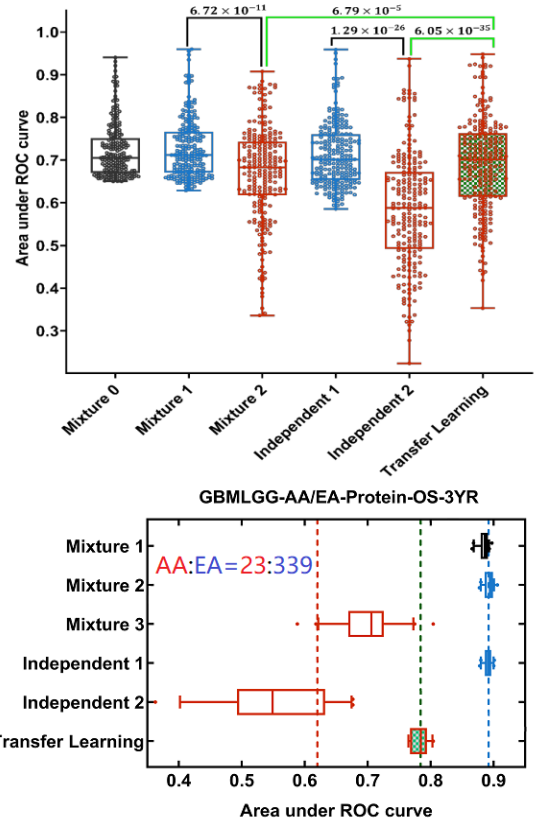

Figure 2. **Comparison of the multiethnic machine learning schemes. Top panel**: The machine learning experiments. **Middle panel**: Each boxplot shows the AUROC (Area under ROC curve) values for the 224 learning tasks for a machine learning experiment listed in top panel. **Bottom panel**: One example of the learning tasks – Prediction of 3-year overall survival of Glioma patients from protein expression data. The cohort includes 23 AA and 339 EA patients. The box plots show AUROC values for the six experiments (20 independent runs for each experiment). The grey color represents performance for the whole cohort, blue represents performance for the EA group, and red represents performance for the AA group. **GMBLGG**: Glioma; **AA**: African Ancestry; **EA**: European Ancestry; **OS**: Overall Survival.