A transfer learning approach based on random forest with application to breast cancer prediction in underrepresented populations

Tian Gu[1], Yi Han[2] and Rui Duan[1,†]

[1] *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA*

[2] *School of Mathematical Sciences, Shanghai Jiaotong University, Shanghai, 200240, China*

[†] *E-mail: rduan@hsph.harvard.edu*

Despite the high-quality, data-rich samples collected by recent large-scale biobanks, the underrepresentation of participants from minority and disadvantaged groups has limited the use of biobank data for developing disease risk prediction models that can be generalized to diverse populations, which may exacerbate existing health disparities. This study addresses this critical challenge by proposing a transfer learning framework based on random forest models (TransRF). TransRF can incorporate risk prediction models trained in a source population to improve the prediction performance in a target underrepresented population with limited sample size. TransRF is based on an ensemble of multiple transfer learning approaches, each covering a particular type of similarity between the source and the target populations, which is shown to be robust and applicable in a broad spectrum of scenarios. Using extensive simulation studies, we demonstrate the superior performance of TransRF compared with several benchmark approaches across different data generating mechanisms. We illustrate the feasibility of TransRF by applying it to build breast cancer risk assessment models for African-ancestry women and South Asian women, respectively, with UK biobank data.

Keywords: Transfer Learning; Random Forest; Underrepresented Population; Breast Cancer.

1. Introduction

Risk prediction tools can guide disease prevention, early detection, and intervention. Some well-known examples include the Gail model for assessing breast cancer risks,[1] and the Bach model for lung cancer risk prediction,[2] which are helpful for both risk stratification and cancer screening recommendations. Over the past few decades, genome-wide association studies (GWAS) have identified significant genetic loci associated with many complex diseases, suggesting the great potential for combining genetic information with epidemiological, clinical, and other risk factors to further improve the performance of risk prediction models.[3] With the development of large-scale biobanks, such as the UK biobank (UKB),[4] the Mass General Brigham (MGB) biobank,[5] and the Million Veteran Program (MVP) mega-biobank,[6] clinical information obtained from electronic health records is linked with participants' genomic data, health survey data, and other health-related measures, providing unique opportunities to

develop enhanced risk prediction tools that integrate different types of risk factors.[7]

However, a long-standing problem is the lack of participants from minority and disadvantaged groups in biomedical studies, which may lead to underperformance of risk prediction models in these underrepresented populations, and might exacerbate health disparities.[8,9] For example, most breast cancer risk prediction models have been developed based on data from White women, resulting in underestimated risk in Black women and inaccurate estimation for other racial groups such as American Indian or Alaska Native.[10] Many large-scale biobanks also have disproportionately fewer participants from non-European ancestry than the European ancestry populations. There are less than 6% participants of non-European ancestry in UKB, while the MGB biobank only contains 6% African Americans, 5% Hispanics, and 4% Asians. Such lack of representation has raised significant challenges for developing and evaluating risk assessment tools for underrepresented populations. More inclusive data collection strategies are needed to tackle these challenges, while methodological advancements are also essential to improve the use of existing resources.

Transfer learning methods have been successfully applied in many areas, including text recognition and imaging classification,[11] due to their capability of leveraging shared information from source populations with relatively sufficient data to build prediction models in a target population with limited data. Unlike many transfer learning methods that require individual-level data from both the source and target populations,[12,13] we consider the situation where we can only obtain fitted models from a source population instead of their individual-level data. This is a common situation in biomedical studies, where data are often protected by various regularities or rules to be made publicly available, while trained models can be shared through open-source platforms such as GitHub, or more protected environments such as the Phenotype Knowledgebase website (PheKB).[14] As increasing efforts have been devoted to building collaborative environments for evaluating and validating machine learning algorithms across different health care datasets, sharing fitted models is expected to become increasingly feasible.[15] Consequently, model-based transfer learning methods that can leverage existing fitted models are needed.

Existing model-based transfer learning methods mainly involve parametric models such as regression,[16,17] which may have limited predictive power when the model is misspecified. Network-based deep transfer learning methods mostly follow the idea of fine-tuning a pre-trained neural network,[18] which often lacks clear model interpretation, practical guidance, and theoretical justification.[19] Among many risk prediction models, tree-based methods such as random forest (RF) have been widely used in biomedical research, including risk prediction,[20,21] disease diagnosis,[22,23] and digital phenotyping.[24] Tree-based methods enjoy several advantages, including the ability to handle non-linear relationships, the property to learn feature importance, and good interpretability. Importantly, recent studies have laid the theoretical foundation of RF models,[25] which further helps researchers to understand how well these methods work under different scenarios.

The development of model-based transfer learning methods built upon RF models is still an open area due to the non-parametric nature of RF. Recently, a few strategies have been proposed based on using target data to either refine each source tree's structure or adjust

the numeric threshold of each split.[26,27] Such structure-based transfer learning methods may not perform well in cases where the optimal tree structures in the two populations are highly different and each source tree performs relatively poorly in the target population. In addition, pruning and adjusting a large number of trees with limited target data may be inefficient. The lack of performance of the structure-based transfer learning methods are demonstrated in our data application.

In this paper, we propose a RF-based transfer learning framework termed TransRF. Our method is based on an ensemble of multiple transfer learning approaches covering various types of similarity between the source and target models. Unlike existing work that relies on tree structural similarities, our method is more robust and applicable to different scenarios. More importantly, with slight modifications, our approach can be extended to adapt a broader range of prediction models beyond RF. We evaluate our method using extensive simulation studies and apply it to predict breast cancer patients in African-ancestry (AFR) women and South Asian (SAS) women, respectively, using UKB data.

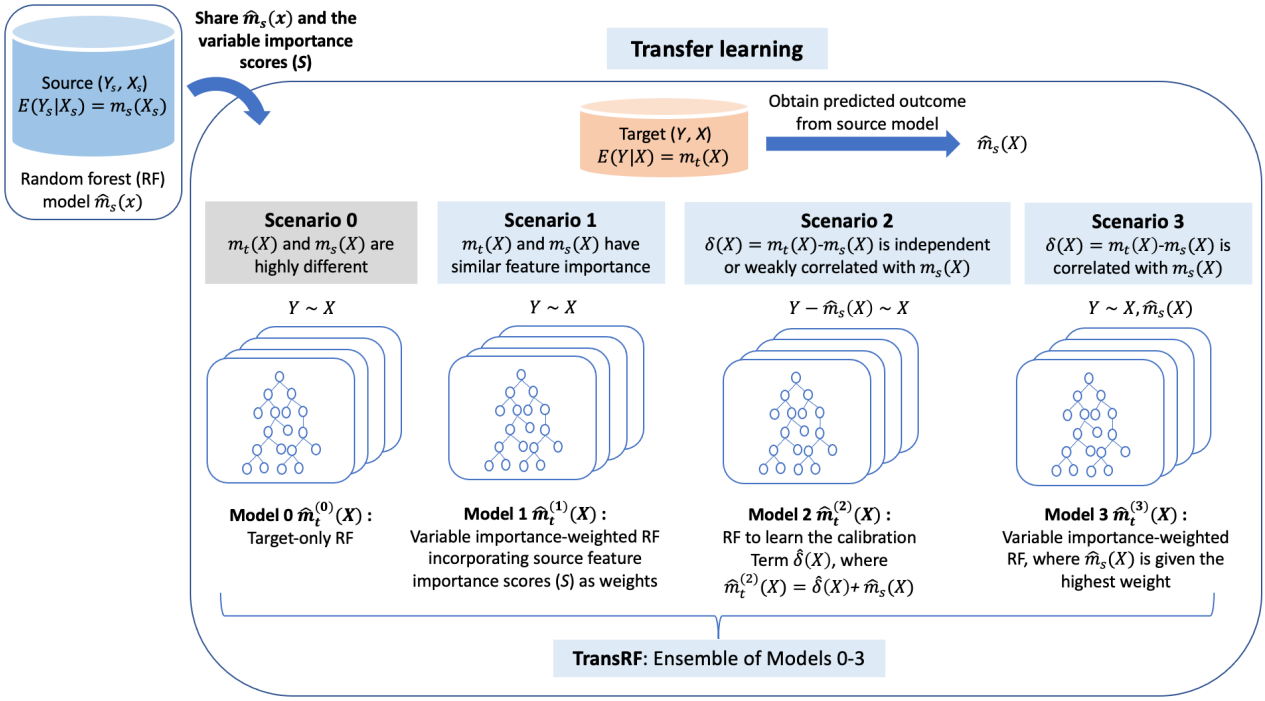

Fig. 1. The schematic illustration of TransRF, an ensemble of a forest trained using only the target data (scenario 0) and multiple forests that transferred information from a source forest (described by scenarios 1-3).

2. Method

2.1. *Overview and notation*

We start with an overview of the proposed framework. TransRF aims to improve the prediction performance in an underrepresented population with limited data by incorporating a RF model

trained in a source population with relatively more sufficient data. To leverage the information contained in the source model, we develop transfer learning models that cover several practical scenarios, in which the source model shares certain similarities with the target population. An ensemble learning strategy is used to combine multiple transfer learning models to improve the method's robustness. A schematic illustration is presented in Fig. 1.

We denote (Y, X) as the target data, where $Y \in \mathbb{R}^n$ is the outcome variable and $X \in \mathbb{R}^{n \times p}$ is the p-dimensional feature variables. Correspondingly, we denote data from the source population as (Y_s, X_s). To improve the applicability of the method, we consider the case where only a fitted source model $\hat{m}_s(x)$ is available, which is an estimator of the true conditional mean function $m_s(x) = \mathbb{E}(Y_s|X_s = x)$. The distribution of the target data can be different from the source data, i.e., either the feature distribution or the conditional distribution $m_t(x) = \mathbb{E}(Y|X = x)$ could be different from the source. Our goal is to estimate $m_t(x)$, using target data (Y, X) and the fitted source model $\hat{m}_s(x)$.

2.2. *Three ways to incorporate the source model*

Leveraging feature importance. One potential similarity between the source and the target models is that they may have similar feature importance rankings (see **Scenario 1** in Fig. 1). When training a RF model with limited target data, we can use the variable importance scores obtained from the source model, which we denoted by $S = (s_1, \ldots, s_p)$. This is especially useful when the number of features is large. The importance scores can be normalized to weights to determine the probabilities of selecting the features in each tree.[28] We denote the fitted model as $\hat{m}_t^{(1)}(x)$, and refer it to *Model 1*. Intuitively, Model 1 is expected to perform well if the source and target share similar feature importance rankings, even if the underlying $m_t(x)$ and $m_s(x)$ are highly different. When $m_s(x)$ and $m_t(x)$ are close, Model 1 might be less effective as it does not directly use the predicted values from the source model. Thus, we introduce the following two scenarios.

Calibration of the source model by learning the discrepancy. Due to population heterogeneity, the predicted values $\hat{m}_s(X)$ may not be accurate when directly applying the source model to the target data. We propose to use the target data to calibrate the source model. Denote the discrepancy between the two underlying true models as $\delta(X) = m_t(X) - m_s(X)$. One possible situation is that $\delta(X)$ is independent or weakly correlated with $m_s(X)$ (see **Scenario 2** in Fig. 1), meaning that the discrepancy term captures complementary information of the source model. In such a case, instead of fitting a model using the original outcome Y, we propose to obtain the residual term, defined as $Y - \hat{m}_s(X)$, which is the difference between the observed outcomes and the predicted values. Treating the source model as an anchor, we fit a RF model using the residual term as the outcome and X as the features. When $\delta(X)$ has some sparse or low-dimensional structure, we can benefit from such sparsity by targeting the discrepancy term.[29] Finally, we obtain $\hat{m}_t^{(2)}(X) = \hat{\delta}(X) + \hat{m}_s(X)$, which we refer to as *Model 2* hereafter.

Calibration of the source model by adding a new feature. We now consider the case where the discrepancy term $\delta(X)$ is correlated with the source model $\hat{m}_s(X)$ so that the above Model 2 might not be able to learn $\delta(X)$ accurately. In other words, $\hat{m}_s(X)$ could be an

important feature for predicting the discrepancy so as to predict Y. In this case, we propose to add $\hat{m}_s(X)$ as an additional feature for predicting Y (see **Scenario 3** in Fig. 1). Since $\hat{m}_s(X)$ is likely an important feature, we propose using weighted RF and assigning it a large weight. We can assign equal or different weights for other features, X, according to prior knowledge of whether certain features have different effects in two populations. We denote the fitted model as $m_t^{(3)}(x)$, and refer it to *Model 3*.

2.3. *Ensemble learning to boost the robustness and prevent negative transfer*

Each of the models described above relies on certain assumptions about the true underlying functions $m_t(x)$ and $m_s(x)$, where the validity of the assumptions is unverifiable in practice. As we will later show in the simulation studies, the performance of Models 1-3 varies under different settings. In addition, when the source population is highly different from the target population, the source model could not provide any useful information to the training of the target model, and the above models may even have lower performance compared to a RF model trained by using the target data alone (the *target-only model, or Model 0 shown in Fig.1*, denoted as $\hat{m}_t^{(0)}$). To prevent such "negative transfer" and to leverage the strength of each model, we propose to obtain an ensemble model which is a linear combination of $\hat{m}_t^{(0)}, \hat{m}_t^{(1)}, \hat{m}_t^{(2)}$ and $\hat{m}_t^{(3)}$. We denote the *TransRF model* as

$$\hat{m}_t(x) = \sum_{i=0}^{3} w_i \hat{m}_t^{(i)}(x)$$

where w_i is the weight of the i-th model. Many existing methods can be used to obtain the ensemble weights. For example, with a small validation dataset $(\tilde{X}, \tilde{Y})$, we can obtain the ensemble model by fitting a linear regression model treating $\hat{m}_t^{(0)}(\tilde{X}), \hat{m}_t^{(1)}(\tilde{X}), \hat{m}_t^{(2)}(\tilde{X})$ and $\hat{m}_t^{(3)}(\tilde{X})$ as features. Alternatively, we can use methods such as Q-aggregation[30] to learn the weights. The sample size of the validation dataset can be relatively small compared to the training data, and a cross-fitting strategy can be used to potentially achieve better accuracy.

As illustrated in Fig. 1, TransRF requires only the fitted RF model and the corresponding feature importance scores from the source population, especially preferable in settings where individual-level data is not shareable across sites. Our framework can be modified to incorporate other possible transfer learning models that might work better in scenarios not described above, such as the structure-based transfer learning models.[26]

3. Simulation studies

We conduct Monte Carlo simulations to assess TransRF and several comparisons under three settings. Due to space limitations, we outline the data generating procedures in this section and leave the detailed choices of parameters, transformation, and distribution functions in the online Supplementary Materials. In each setting, we generate X and X_s from a multivariate truncated normal distribution with different means to mimic the potential shifts in feature distributions. The dimension of features is set to $p = 20$. The mean function $m_s(x)$ and $m_t(x)$ are set to be some non-linear functions of X, which are different across settings. We then add random noise to the mean functions $m_s(x)$ and $m_t(x)$ to obtain the outcomes in the source and the target populations. For each simulated dataset, we generate target data of size $n = 200$ for

the training purpose and an independent testing set with $n_{\text{test}} = 100$. A source sample of size $n_{\text{src}} = 1000$ is generated to fit the source model. We evaluate the model performance using the mean squared prediction error (MSE) of the testing set over 200 simulation replications. We now describe the three simulation settings:

(i) In Setting 1, we consider that the source and the target populations share a similar variable importance ranking, where the similarity between the two populations is measured by the correlation of their variable importance rankings. To generate $m_s(x)$ and $m_t(x)$, we apply some non-linear transformations on each feature in X and obtain the transformed features Z. We then combine the transformed features through a linear combination to obtain $m_s(x)$ and $m_t(x)$, i.e., $m_s(x) = Z\beta_s$ and $m_t(x) = Z\beta_t$, where β_s and β_t are p-dimensional vectors whose magnitude determines the feature importance. By changing the correlation between β_t and β_s, we vary the similarity degree of their feature importance.

(ii) In Setting 2, we consider that the discrepancy between $m_s(X)$ and $m_t(X)$ is independent or weakly correlated with $m_s(X)$. To achieve this, we first generate $m_s(x)$ in the same way described in Setting 1. We then generate $\delta(x)$, the function of a random subset of all the features, on which we apply different feature transformations and linear combinations compared to $m_s(x)$. We obtain $m_t(x) = m_s(x) + \delta(x)$. We vary the variance explained by the source model $m_s(x)$ to control the similarity between the source and the target populations.

(iii) In Setting 3, we consider that the discrepancy term is correlated with $m_s(X)$. We generate Y_s following the same data generating mechanism in Setting 2 except that we set $m_t(X) = Cm_s(X) + \delta(X)$, where C is a constant. In this case, the true discrepancy is $m_t(X) - m_s(X) = (C-1) * m_s(X) + \delta(X)$. With $C \neq 1$, $m_s(X)$ is correlated with the discrepancy, and we vary C to alter the strength of the correlation.

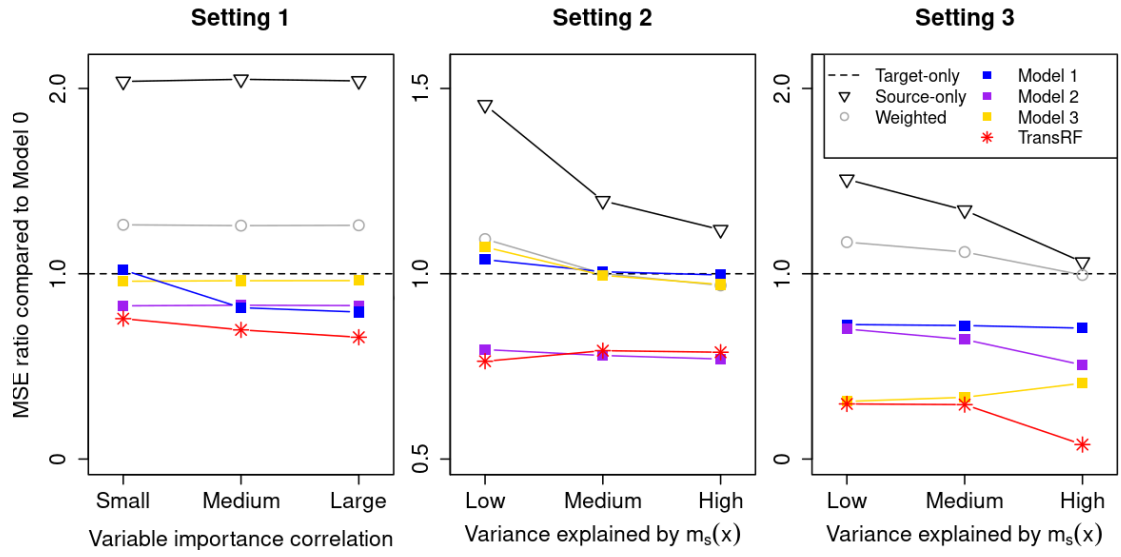

Fig. 2. MSE ratio compared to Model 0 (the target-only model) in simulation settings 1 (left), 2 (middle), and 3 (right).

In each setting, we use Model 0, i.e., the target-only model, as the reference and compare the performance of six models with it: (1) Source-only: $\hat{m}_s(x)$; (2) Weighted: a weighted average of

predictions from source-only model and target-only model, using inverse MSE of validation data as weights; (3) Model 1: $\hat{m}_t^{(1)}(x)$; (4) Model 2: $\hat{m}_t^{(2)}(x)$; (5) Model 3: $\hat{m}_t^{(3)}(x)$; and (6) TransRF: the proposed method, combining Models 0-3. Note that for methods (2) and (6), a validation dataset is needed to train the weights, where we randomly split $n_{val} = 50$ samples from the training data. For each method (k), $k \in \{1, \ldots, 6\}$ described above, we report its MSE ratio compared to the reference, denoted as MSE_k/MSE_0, where a ratio larger than 1 represents worse performance than the reference. In contrast, a ratio smaller than 1 represents improved prediction compared to the reference. We build TransRF algorithm in R software[31] on the basis of *viRandomForests* package. Code to implement TransRF along with the example data, and Supplementary Materials are available at https://github.com/gutian-tiangu/TransRF.

3.1. *Simulation results*

Results of Setting 1 (the left panel of Fig. 2) show that the performance of Model 1 improves over the increasing correlation of feature importance. When the correlation is large, Model 1 outperforms most of the compared methods, while it performs slightly worse than Model 0 when the correlation is low. Interestingly, Model 2 performs well across all settings, which might be due to the discrepancy term $m_t(x) - m_s(x)$ under this setting is weakly correlated with the source mean structure when we alter the correlations between the feature importance. TransRF has the best performance over different correlation levels and the MSE ratios to Model 0 range from 0.66 to 0.76.

In Setting 2 (the middle panel of Fig. 2), we observe that when the performance of the source model increases, Model 2 outperforms all the compared methods. Since Model 2 has much better performance than Models 0, 1, and 3, TransRF has nearly identical performance as Model 2, with a MSE of 0.78 times that of Model 0.

In Setting 3 (the right panel of Fig. 2), we alter the parameter C in $m_t(X) = Cm_s(X) + \delta(X)$ from 10 to 1 corresponding to three levels shown in the x-axis. When C is getting closer to 1, the variance explained by the source model increases (from low to high), and so as the performance of the source-only model. When C is larger than 1, Model 3 performs better than other methods and similarly to TransRF. When $C = 1$, the performance of all the other methods improves, and therefore TransRF has better performance, where its MSE ratios to Model 0 range between 0.08 and 0.30.

In summary, the performance of each transfer learning model varies in different settings, where each model has the best performance in a specific scenario. TransRF that combines Models 0-3 often outperforms its underlying constituents and is robust against negative transfer.

4. Application using UKB data

We apply TransRF to UKB breast cancer data, treating European (EUR) women as the source population, and AFR women SAS women as the target population, respectively.

4.1. *Defining breast cancer case and control, ancestry, and other variables*

We identify breast cancer cases using the ICD-10 code (C50) following a recently released UKB disease phenotyping definition.[32] When using retrospective data like UKB to build risk

prediction models, one should exclude the prevalent cases where observations already had breast cancer diagnosed before they entered the study, and only keep incident cases who developed breast cancer after entering the study. In our example, as the target sample size is minimal, we want to keep as many target samples as possible. We decide to include both incident and prevalent cases and only use time-invariant predictors (excluding variables that can potentially happen after the diagnosis). When selecting candidate controls, we identify women who have not been diagnosed with breast cancer or ovarian cancer (ICD-10 code, C56) as these cancers are closely related.[33] We select a subset of subjects to obtain controls with a case-control ratio approximately equal to 1:2.

To define the ancestry population for EUR, AFR and SAS, we use a mutual set of self-reported ancestry through UKB survey data and the principal component-based ancestry prediction proposed by Zhang, Dey, and Lee.[34] Only those whose self-claimed ancestry matched the ancestry prediction are included. We define the following clinical variables that are commonly known as breast cancer risk factors: ever smoking (yes or no), age at the start of menstruation in years, had a college degree (yes or no), ever had a live birth (yes or no).[35,36] For a small percentage of participants who had missing age at the start of menstruation (<3%), we impute the missingness with a mean age of 13. We identify 479 SAS samples (173 cases and 306 controls), 440 AFR samples (126 cases and 314 controls), and 43,576 EUR samples (14,240 cases and 29,336 controls) that contain complete data of outcomes and clinical variables. For each target population, we randomly split a validation set of size 50 (20 cases and 30 controls) and a testing set of size 90 (30 cases and 60 controls), whereas the remaining samples are used as training data (339 samples including 123 cases and 216 controls when using SAS as the target; and 300 samples including 76 cases and 224 controls when using AFR as the target).

4.2. *Genotyping, quality control and imputation*

Details on genotype calling and quality control for UKB data are described elsewhere.[4] We include 330 novel breast cancer susceptibility single-nucleotide polymorphisms (SNPs) identified in a GWAS study.[37] We perform standard quality control, including removing participants who have mismatched self-reported sex versus biological sex, those who failed UKB official genotype quality control, and all pairs of participants who are estimated to be genetically related. A total of 272 SNPs are found in the UKB data, used as genetic predictors, among which 151 contain a small percent of missingness (over 90% SNPs with missingness have a missing rate $< 5\%$). For each SNP with missingness, we impute the missingness using the value with the largest frequency.

4.3. *Results*

Fig. 3 shows the area under the operating characteristic curve (AUC) of different transfer learning methods after incorporating source model information for SAS as the target model in the left panel and AFR as the target model in the right panel. When using AFR as the target population, compared with Model 0 (dashed vertical line, AUC=0.61), Model 1 by directly sharing the variable importance score has the highest AUC, equal to 0.70. Model 2 that learns the discrepancy term has an AUC of 0.69, while Model 3 by including source predicted values as the most important feature does not show improved performance with AUC equal to 0.60.

TransRF by aggregating Models 0-3 shows an AUC of 0.70, a 10% improvement compared to the target-only model and a 5% improvement compared to the weighted model by naively aggregating Model 0 and the source-only predictions. On the contrary, the SER model by using the target data to fine-tune the source tree structure[26] shows the worst performance among others. This may result from insufficient target data to refine the tree or dissimilarity between the source and the target tree structure.

When comparing the results that each uses SAS and AFR as the target population, we observe that each transfer learning model performs differently, e.g., Model 3 has the worst performance in transferring EUR information to AFR while it has the best performance when leveraging EUR information to SAS. This might be due to different similarities of genetic architectures between EUR and AFR versus EUR and SAS.[38]

In Table 1, we present the top 20 important variables from the source and each target-only model, along with the corresponding variable importance scores. Age at the start of menarche is found in all three models, and it is the most important variable in both the EUR model and AFR Model 0. Two predictors, rs16886165 and "Ever had a college degree", overlap in EUR model and AFR Model 0, while the top one feature of SAS Model 0, rs4784227, is also found important in the EUR model. In addition, rs9315973 is identified with high importance in both AFR and SAS Model 0's, an intron variant belongs to gene EPSTI1 that is known to be associated with many traits and diseases, including breast cancer in European and East Asian.[39]

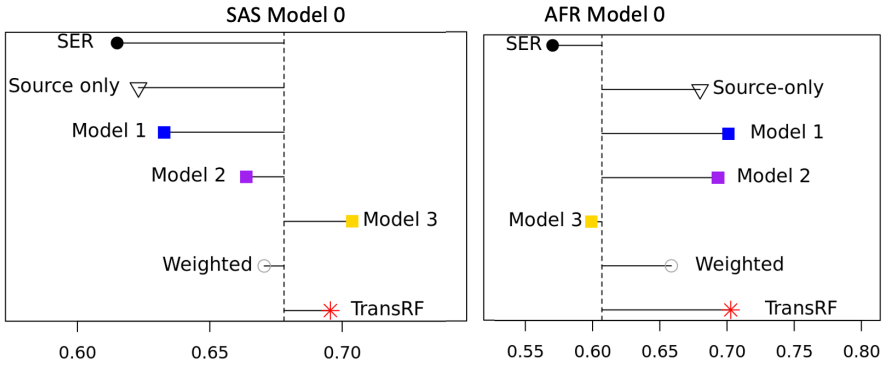

Fig. 3. AUC of transfer learning methods compared to Model 0 (the target-only model) for SAS (left) and AFR (right).

It is worth noting that rs16886165 is an intergenic variant identified as associated with breast cancer in European populations.[40,41] The known risk effect of rs4784227, an intron variant mapped to gene CASC16, associated with breast cancer has been validated in European[42,43] and East Asian ancestries.[44–46] Other than these two SNPs, the ranking of the rest of the top SNPs between the target and the source populations is not consistent, which might suggest underlying differences in genetic architectures across populations.[38] However, with a limited sample size in the target population, the estimated feature importance scores may have large variability.

5. Discussion

In this study, we propose TransRF, a RF-based transfer learning framework targeting risk prediction in underrepresented populations. By incorporating fitted models from a large source population, TransRF combines the strengths of several novel transfer learning models motivated by various practical situations. Our simulation studies reveal that the effectiveness of different

Table 1. Top 20 variables (importance score) from the fitted EUR model, Model 0 treating South Asian (SAS) as the target population, and Model 0 treating African Ancestry (AFR) as the target population. Variables identified from all three populations indicated in bold text. Variables shared by the EUR and SAS populations are indicated in blue. Variables shared by the EUR and AFR populations are indicated in red. Variables shared by the SAS and AFR populations are indicated in orange.

Rank	Fitted EUR model (score)	SAS Model 0 (score)	AFR Model 0 (score)
1	**Menarche age (0.056)**	rs4784227 (0.043)	**Menarche age** (0.148)
2	rs4442975 (0.021)	rs7848334 (0.04)	Had a college degree (0.048)
3	rs630965 (0.02)	rs12472404 (0.031)	rs2454399 (0.045)
4	rs10941679 (0.017)	rs12422552 (0.031)	rs144767203 (0.031)
5	rs16886165 (0.016)	rs332529 (0.03)	rs2181965 (0.028)
6	rs910416 (0.016)	**Menarche age (0.029)**	rs2403907 (0.02)
7	rs6913578 (0.016)	rs4866496 (0.028)	rs9693444 (0.019)
8	rs10096351 (0.015)	rs78540526 (0.027)	rs56387622 (0.018)
9	rs7072776 (0.014)	rs4868701 (0.026)	rs35542655 (0.018)
10	Had a college degree (0.014)	rs719338 (0.02)	rs7924772 (0.018)
11	rs9931038 (0.014)	rs7842619 (0.02)	rs16886165 (0.018)
12	rs35668161 (0.014)	rs3010266 (0.019)	rs3819405 (0.017)
13	rs552647 (0.014)	rs10832963 (0.019)	rs2356656 (0.016)
14	rs661204 (0.012)	rs9315973 (0.018)	rs9364472 (0.016)
15	rs4784227 (0.012)	rs55872725 (0.018)	rs9315973 (0.016)
16	rs17343002 (0.012)	rs7830152 (0.018)	rs11693806 (0.015)
17	rs11249433 (0.012)	rs28539243 (0.016)	rs7800548 (0.014)
18	rs10164323 (0.011)	rs335160 (0.015)	rs665889 (0.013)
19	rs10197246 (0.011)	rs9712235 (0.015)	rs889310 (0.012)
20	rs4602255 (0.011)	rs7121616 (0.014)	Ever had a live birth (0.012)

transfer learning models varies with the underlying relationship between the source and the target models. TransRF reaches comparable performance to the transfer learning method with the best performance across different scenarios, demonstrated by both simulation studies and the application to UKB data.

Our paper considers the practical situation where we can only obtain a fitted model from the source population, whereas the individual-level data are unavailable. A relevant problem is transfer learning in a federated setting, where summary-level statistics (not necessarily the trained model) can be shared across populations. In such a setting, Li et al.[47] proposed a federated transfer learning algorithm based on penalized generalized linear regression models, which requires sharing the gradients of likelihood functions iteratively across populations, and we refer to the relevant works discussed therein. This type of method is more applicable to research networks with specific infrastructures to facilitate timely information sharing and model updating. In contrast, the model-based transfer learning framework proposed in this paper can be helpful in a broader range of applications.

There are several limitations to this study. In the breast cancer example, both instant and prevalent cases are included. Due to the limited sample size in the target population, only

including the breast cancer incidents will result in too few target samples. Although we only use time-invariant features or features that are most likely to happen before breast cancer diagnosis, such as education level and menarche age, there is still uncertainty in terms of their actual temporal relationships. We aim to use this data example to show the feasibility of TransRF. As a future direction, we will explore the potential of TransRF for disease risk prediction by incorporating more precise temporal information based on codified and unstructured information in biobank data.

Acknowledgements

This research was supported by the Harvard Data Science Initiative Postdoctoral Fellow Research Fund.

References

1. M. H. Gail *et al.*, Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, *J. Natl. Cancer Inst.* **81**, 1879 (1989).
2. K. A. Cronin *et al.*, Validation of a model of lung cancer risk prediction among smokers, *J. Natl. Cancer Inst.* **98**, 637 (2006).
3. A. Lee *et al.*, Boadicea: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors, *Genetics in Medicine* **21**, 1708 (2019).
4. C. Bycroft *et al.*, The UK biobank resource with deep phenotyping and genomic data, *Nature* **562**, 203 (2018).
5. E. W. Karlson *et al.*, Building the partners healthcare biobank at partners personalized medicine: Informed consent, return of research results, recruitment lessons and operational considerations, *Journal of Personalized Medicine* **6** (2016).
6. J. M. Gaziano *et al.*, Million veteran program: A mega-biobank to study genetic influences on health and disease, *Journal of clinical epidemiology* **70**, 214 (2016).
7. C. J. O'Donnell, Opportunities, challenges and expectations management for translating biobank research to precision medicine, *European Journal of Epidemiology* **35**, 1 (2020).
8. P. Kim *et al.*, Minority participation in biobanks, *Biobanking* , 43 (2019).
9. W. L. Teagle *et al.*, Comorbidities and ethnic health disparities in the UK biobank, *JAMIA Open* **5**, p. ooac057 (2022).
10. A. W. Kurian *et al.*, Performance of the ibis/tyrer-cuzick model of breast cancer risk by race and ethnicity in the women's health initiative, *Cancer* **127**, 3742 (2021).
11. K. Weiss *et al.*, A survey of transfer learning, *Journal of Big data* **3**, 1 (2016).
12. Y. Yao *et al.*, Boosting for transfer learning with multiple sources, *CVPR* , 1855 (2010).
13. R. Chattopadhyay *et al.*, Multisource domain adaptation and its application to early detection of fatigue, *ACM Trans Knowl Discov Data* **6**, 1 (2012).
14. J. C. Kirby *et al.*, Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability, *JAMIA* **23**, 1046 (2016).
15. J. Shiffman *et al.*, The emergence and effectiveness of global health networks: findings and future research, health policy and planning, *Health Policy and Planning* **31** (2016).
16. T. Gu *et al.*, Commute: communication-efficient transfer learning for multi-site risk prediction, *MedRxiv* (2022).
17. P. Han, A discussion on "a selective review of statistical methods using calibration information from similar studies" by qin, liu and li, *Stat. Theory Relat. Fields* , 1 (2022).
18. C. Tan *et al.*, A survey on deep transfer learning, *ICANN* , 270 (2018).
19. J. R. Geis *et al.*, Ethics of artificial intelligence in radiology: summary of the joint european and north american multisociety statement, *CAR Journal* **70**, 329 (2019).

20. B. Dai *et al.*, Using random forest algorithm for breast cancer diagnosis, *International Symposium on Computer, Consumer and Control* , 449 (2018).
21. Q. Wang *et al.*, Random forest with self-paced bootstrap learning in lung cancer prognosis, *ACM TOMM* **16**, 1 (2020).
22. F. Yang *et al.*, Using random forest for reliable classification and cost-sensitive learning for medical diagnosis, *BMC bioinformatics* **10**, 1 (2009).
23. M. Zhu *et al.*, Class weights random forest algorithm for processing class imbalanced medical data, *IEEE Access* **6**, 4641 (2018).
24. J. Benoit *et al.*, Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses, *Harvard Review of Psychiatry* **28**, 296 (2020).
25. S. Athey *et al.*, Generalized random forests, *Ann. Stat.* **47**, 1148 (2019).
26. N. Segev *et al.*, Learn on source, refine on target: A model transfer learning framework with random forests, *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1811 (2016).
27. W. Fang *et al.*, Adapted tree boosting for transfer learning, *IEEE* , 741 (2019).
28. Y. Liu *et al.*, Variable importance-weighted random forests, *Quant. Biology* **5**, 338 (2017).
29. J. Klusowski, Sparse learning with cart, *Advances in NeurIPS* **33**, 11612 (2020).
30. G. Lecué *et al.*, Optimal learning with q-aggregation, *Ann. Stat.* **42**, 211 (2014).
31. R. C. Team *et al.*, R: A language and environment for statistical computing (2021).
32. D. J. Thompson *et al.*, UK biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits, *MedRxiv* (2022).
33. J. Schildkraut *et al.*, Evaluating genetic association among ovarian, breast, and endometrial cancer: evidence for a breast/ovarian cancer relationship., *Am. J. Hum. Genet.* **45**, p. 521 (1989).
34. D. Zhang *et al.*, Fast and robust ancestry prediction using principal component analysis, *Bioinformatics* **36** (2020).
35. K. Al-Ajmi *et al.*, Risk of breast cancer in the uk biobank female cohort and its relationship to anthropometric and reproductive factors, *PLoS One* **13**, p. e0201097 (2018).
36. S. K. Hussain *et al.*, Influence of education level on breast cancer risk and survival in sweden between 1990 and 2004, *International Journal of Cancer* **122**, 165 (2008).
37. H. Zhang *et al.*, Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses, *Nat. Genet* **52**, 572 (2020).
38. L. Shang *et al.*, Genetic architecture of gene expression in european and african americans: an eqtl mapping study in genoa, *Am. J. Hum. Genet.* **106**, 496 (2020).
39. K. Michailidou *et al.*, Association analysis identifies 65 new breast cancer risk loci, *Nature* **551**, 92 (2017).
40. G. Thomas *et al.*, A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (rad51l1), *Nat. Genet* **41**, 579 (2009).
41. N. Brandes *et al.*, Genetic association studies of alterations in protein function expose recessive effects on cancer predisposition, *Scientific reports* **11**, 1 (2021).
42. J. Hu *et al.*, Supervariants identification for breast cancer, *Genetic epidemiology* **44**, 934 (2020).
43. F. J. Couch *et al.*, Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer, *Nat. Commun.* **7**, 1 (2016).
44. K. Ishigaki *et al.*, Large-scale genome-wide association study in a japanese population identifies novel susceptibility loci across different diseases, *Nat. Genet* **52**, 669 (2020).
45. J. Long *et al.*, Identification of a functional genetic variant at 16q12. 1 for breast cancer risk: results from the asia breast cancer consortium, *PLoS genetics* **6**, p. e1001002 (2010).
46. S. Sakaue *et al.*, A cross-population atlas of genetic associations for 220 human phenotypes, *Nat. Genet* **53**, 1415 (2021).
47. S. Li *et al.*, Targeting underrepresented populations in precision medicine: A federated transfer learning approach, *ArXiv* (2021).