Enhancing Breast Cancer Risk Prediction in Hispanic Women Through Transfer Learning

Zhengwei Song (zs2539)

Practicum Supervisor: Tian Gu, Ph.D.

Columbia University Mailman School of Public Health

March 14, 2024

Overview and Student Role

In this project, I applied advanced concepts of high-dimensional regression and transfer learning algorithms to real-world data. I used high-dimensional genotype data obtained from the NIH's database of Genotypes and Phenotypes (dbGaP), employing innovative single nucleotide polymorphism (SNP) selection methods such as clumping to select the relevant SNPs associated with breast cancer. To handle the computational demands, I used high-performance computing resources and ran the data analysis on powerful clusters. The goal is to construct and critically assess the effectiveness of improved risk prediction models for breast cancer, specifically on the Hispanic population, who are underrepresented in biomedical study and have weak performance in breast cancer risk prediction.

Background

As a leading cause of cancer death among American women, Hispanic women are 30% less likely to be diagnosed with breast cancer than non-Hispanic white women [1]. Hispanics have a higher risk of developing more aggressive tumors such as HER2 positive and triple negative breast cancer compared to non-Hispanic Whites, but without accurate prediction tools for early prevention [2]. Recent studies on White women have shed light on achieving a more effective breast cancer risk stratification by integrating genetic and clinical factors [3]. It is of great interest and needs to develop such an integrative analysis framework for Hispanic women. However, due to the limited participation of underrepresented populations in research, the breast cancer risk prediction performance of Hispanic women is often weak compared to White women [4]. Therefore, we aim to use novel transfer learning models to boost the breast cancer risk performance in Hispanic women by leveraging shared knowledge from other ancestral populations such as White and Asian.

Methods

1. SNP Selection

To reduce the redundancy of genetic markers caused by linkage disequilibrium (LD) and to focus on the most representative SNPs for association analyses, I implemented a SNP clumping procedure in the following way:

Parameters: Clumping was performed using PLINK 1.9.10. I defined clumps by setting the LD correlation threshold at an $r^2 < 0.6$. This means that the correlation between the 1 and 2 allele counts is less than 0.6, which can achieve a balance between minimizing genetic signal redundancy and preserving sufficient information for accurate genetic analysis across diverse populations. The physical distance for clumping was set to 500 kb to ensure SNPs within this window were considered in LD. Index SNPs were selected based on their p-values from a prior GWAS analysis in the Japanese Population, with a threshold of $p < 1x10^{-3}$ for genome-wide significance. Only SNPs meeting this significance level were eligible as index SNPs.

Process: The clumping process began with the most significant SNP, designating it as the index SNP for the first clump. I then identified all SNPs within the defined physical distance of the index SNP and with an LD correlation above the threshold with the index SNP. These SNPs were clumped together, and only the index SNP was retained for subsequent analysis. The process was repeated iteratively, moving to the next most significant SNP not yet clumped until all significant SNPs were evaluated.

2. Models

2.1 Transfer Learning Models and Comparisons

I have genotype data of SNPs (X) from a target Hispanic population of sample n. I have fitted a simple linear model to regression age and gender on the binary breast cancer outcome (Y_0). After obtaining the estimated outcome $\hat{Y}$ for each sample using the simple model, I computed the residual $Y = Y_0 - \hat{Y}$ as the new outcome. This is commonly used in genetic risk prediction to exclude the basic effect on the outcome. The goal is to model the relationship between X and Y. I implemented two transfer learning methods to improve breast cancer risk prediction in Hispanic populations by leveraging shared information from other populations. I also compared the performance with two benchmark methods. Specifically,

Target-Only Model (baseline model): I trained the Ridge models exclusively on the target dataset, comprising clinical and genetic data from underrepresented Hispanic populations, without incorporating external data during the learning process. This approach served as our baseline for performance comparison:

$$\hat{\beta}_{target} = argmin_{\beta} \ \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|^2,$$

where λ is the tuning parameter that can be obtained by cross validation.

Source-Only Model: Two separate breast cancer Ridge models were trained on ancestry-specific datasets: one from a European population and one from a Japanese population. We denoted the sample as $\{(X_k, Y_k)\}_{k=1,2}$. These models were then directly applied to the target population to evaluate the feasibility of cross-population predictive modeling:

$$\hat{\beta}_k = argmin_{\beta} \ \frac{1}{N_k} \|Y_k - X_k\beta\|_2^2 + \lambda_k \|\beta\|^2,$$

where λ_k are the tuning parameters.

Distance-based Transfer Learning (DistTL): The DistTL method enhanced the estimation of target model parameters by imposing a penalty on the L2-distance between the target and source model estimates [5]. By doing so, it leveraged the potentially richer information from a larger source dataset to inform the target estimation. This distance-based penalty confined the target parameters within a specified radius of the source estimates in the parameter space, promoting similarity where smaller L2 distances correspond to more informative source models for improved guidance in fitting the target model:

$$\hat{\beta}_{distTL} = argmin_{\beta} \ \frac{1}{n} \|Y_k - X_k\beta\|_2^2 + \lambda_d \left\|\beta - \hat{\beta}_k\right\|^2,$$

where λ_d is the tuning parameters.

Angle-based Transfer Learning (AngleTL): AngleTL utilized the source model parameters to guide the learning of the target model estimation [5]. This was accomplished by penalizing the angle-based distance between the target and the source model parameters. The source model parameters provided directional information in high-dimensional space, so that a smaller distance represents a more helpful source model for better guiding the target model fitting:

$$\hat{\beta}_{angleTL} = argmin_{\beta} \ \frac{1}{n} \|Y_k - X_k\beta\|_2^2 + \lambda_a \|\beta\|^2 + \eta \hat{\beta}_k^T \beta,$$

where λ_a and η are the tuning parameters. A multi-source AngleTL algorithm was also implemented following Gu et al. [5].

2.2 Data and Preprocessing

The data comes from two studies. The Multiethnic Cohort Study is a population-based prospective cohort study that was initiated between 1993 and 1996 and includes subjects from various ethnic groups including Latinos primarily from Californian (great Los Angeles area) and Japanese-Americans. State drivers' license files were the primary sources used to identify study subjects in Hawaii and California. Also, the initial stage of the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer genome-wide association study (GWAS) included genotype data in 1,145 postmenopausal women of European ancestry with invasive breast cancer and 1,142 controls from The Nurses' Health Study. In this report, all the datasets were divided into Source Data and Target data by ancestry as follows:

Target data**:** n=1,064 Hispanics (520 cases and 544 controls) descent populations from The Multiethnic Cohort Study.

Two source data: N_1=2,287 European (1,145 cases & 1,142 controls) and N_2=1,707 Japanese (885 cases & 822 controls) descent populations, from The Nurses' Health Study and The Multiethnic Cohort Study, respectively.

Both datasets were subjected to standard preprocessing steps, including normalization, missing data imputation, and feature selection.

2.3 Model Training and Validation

The binary outcome breast cancer phenotype was first residualized as mentioned before. Transforming the binary data into residual data adjusted by age and 20 genotype principal components (PCs) was to isolate the genetic contributions to disease risk, independent of age and population stratification. This adjustment increased the statistical power and accuracy of genetic association studies by removing confounding effects, ensuring findings are more directly comparable across diverse populations and studies.

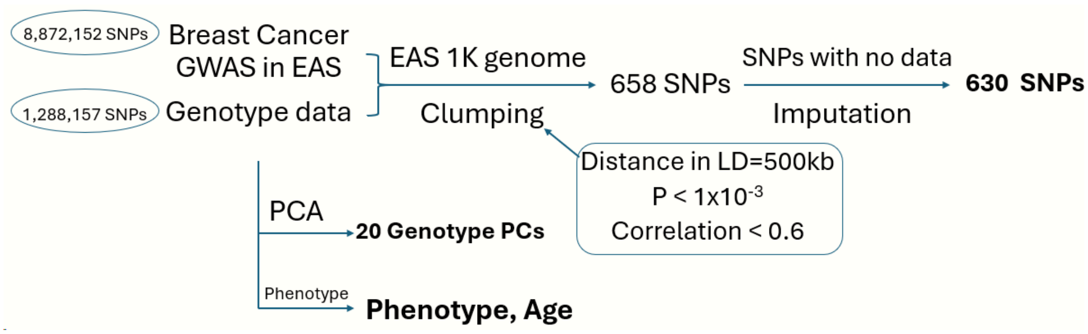
Each model was trained using 70% random samples from the target samples. The performance was validated on the remaining 30% of the target data that was not used in the model training. I repeated the process 20 times and reported the average performance using metrics such as Root Mean Square Error (RMSE) and squared correlation between the predicted and the observed outcome (R^2). In each iteration, I trained and evaluated each model on the same testing data. An improvement in prediction performance in the target population using transfer learning methods over the target-only approach was considered evidence of successful knowledge transfer. I used the GitHub code to implement the aforementioned four models (https://github.com/biostat-duan-lab/multiTL).

Results

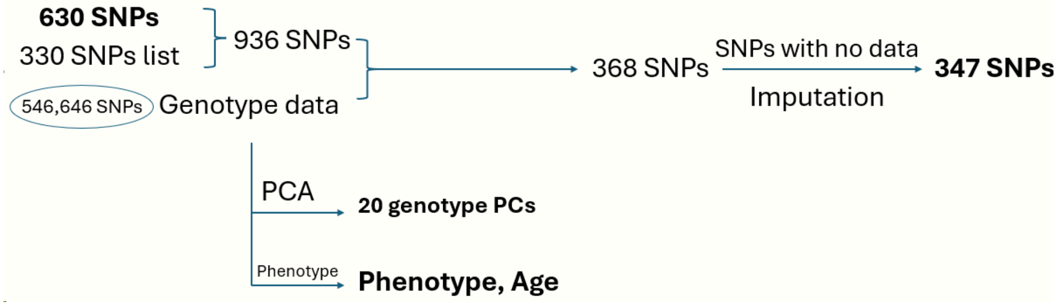
1. Clumping and Data Manipulation

For data on Latinos and Japanese individuals, I used the East Asian (EAS) GWAS data ieu-b-70 from the Integrative Epidemiology Unit (IEU) OpenGWAS project [6] and the data from The Multiethnic Cohort Study for overlap. The overlapping genotype data were then referenced against the EAS 1000 (1K) Genome GWAS data for Clumping, resulting in 658 SNPs.

After data quality processing and imputation based on the mode of each SNP, SNPs without data were excluded, ultimately leaving 630 SNPs (Figure 1A). Regarding the data source for Europeans, for model comparison, identical variable names were necessary. Therefore, by referencing the GWAS analysis results available for Europeans, which include 330 SNPs [7], these 330 SNPs were merged with the previously clumped 630 SNPs, resulting in 936 SNPs. These 936 SNPs were then matched with the European genotype data from The Nurses' Health Study used in this report, yielding 368 SNPs. Similarly, after imputation using the mode and excluding SNPs without data or those that always contained the same value, a total of 347 SNPs were obtained for subsequent Transfer learning analysis (Figure 1B).



(A) The Multiethnic Cohort Study



(B) The Nurses' Health Study

Figure 2. Clumping and Data Manipulation for (A) The Multiethnic Cohort Study and (B) The Nurses' Health Study.

2. Transfer learning model comparison

In terms of performance on the test set, all models had R^2 values ranging from -0.1 to 0.2, while RMSE values were between 0.75 and 1.15. Among them, on R^2, DistTL was clearly higher than the target-only model, reaching values close to that of the source-only model, followed by AngleTL. However, on RMSE, DistTL was clearly higher than both the target-only and source-only models, with the other models having similar values (Figure 3).

The AngleTL models, while showing a median increase in R^2 slightly above zero (Figure A1), overall performed less effectively than both the source-only and DistTL models. This suggested that while AngleTL might contribute positively to model performance, it might not be as robust as DistTL methods in this context.

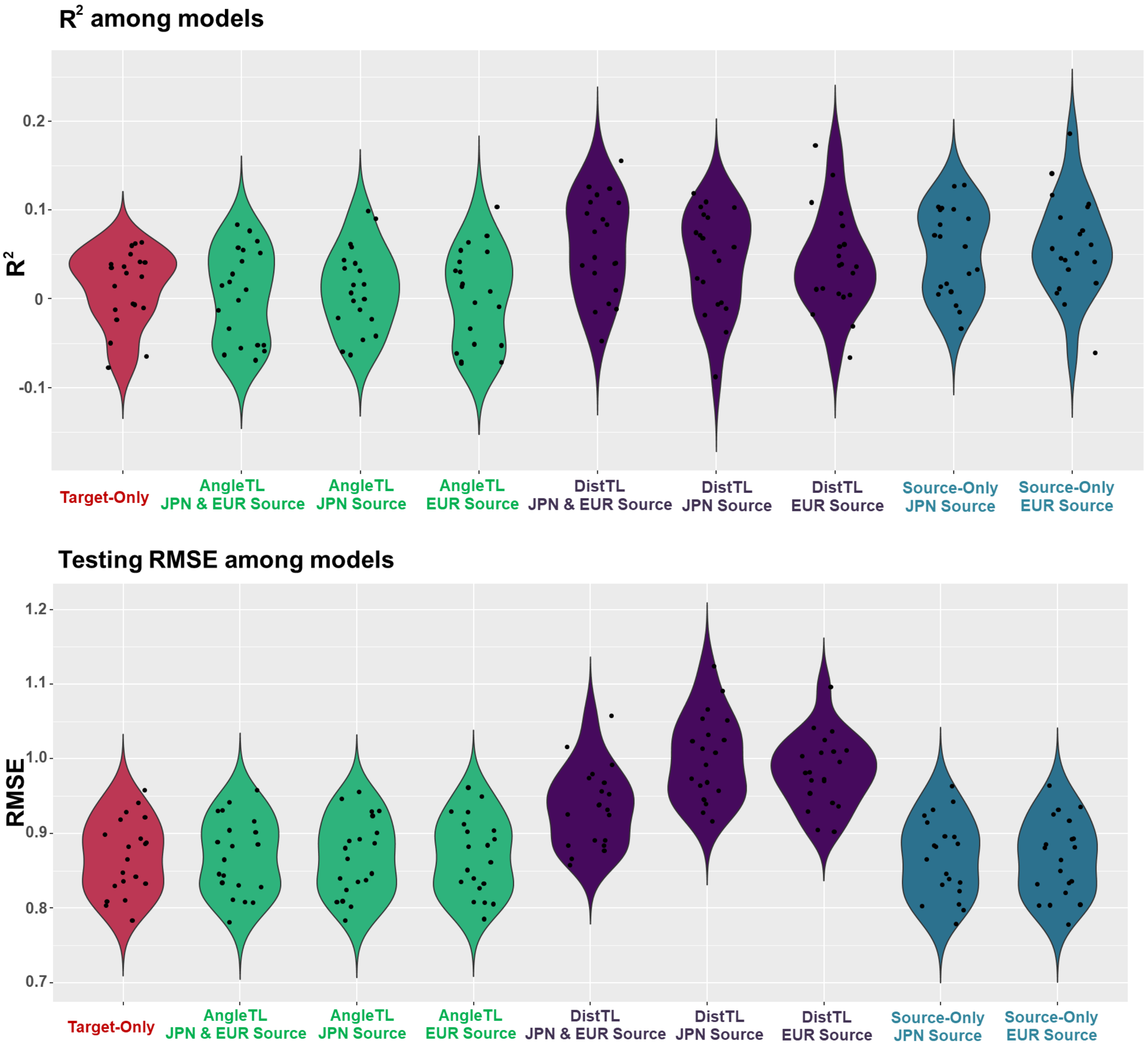


Figure 3. Testing R^2 & Testing RMSE among target-only (red), AngleTL (green), DistTL (purple), and source-only (cyan) models

Conclusions

The process of SNP clumping and the resulting data manipulation effectively eliminated the redundancy of genetic markers and focused on the most representative SNPs for association analyses. This fundamental step ensured that the transfer learning models built subsequently were based on high-quality and relevant data, which improved the reliability of the findings.

Transfer Learning Models showed good performance for addressing disparities in breast cancer risk prediction among underrepresented Hispanic populations, indicating the potential for transfer learning. The DistTL model, in particular, demonstrated a clear improvement in R^2 values compared to the target-only model, with increases ranging from 50% to 350%. This substantial improvement in predictive accuracy for Hispanic women outperformed the source-only model derived from different ancestral populations. The study showed that utilizing advanced machine learning techniques, such as transfer learning, can improve disease risk prediction in underrepresented populations in biomedical research. By utilizing genetic data from extensively researched populations, we can enhance the accuracy and fairness of predictive models for groups such as Hispanic women, who are historically underserved in the biomedical study.

References

1. Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2018, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2018/, based on November 2020 SEER data submission, posted to the SEER web site, April 2021.

2. Serrano-Gomez, S. J., Sanabria-Salas, M. C., & Fejerman, L. (2020). Breast Cancer Health Disparities in Hispanics/Latinas. Current breast cancer reports, 12(3), 175–184. https://doi.org/10.1007/s12609-020-00370-3

3. Mbuya-Bienge, C., Pashayan, N., Kazemali, C. D., Lapointe, J., Simard, J., & Nabi, H. (2023). A Systematic Review and Critical Assessment of Breast Cancer Risk Prediction Tools Incorporating a Polygenic Risk Score for the General Population. Cancers, 15(22), 5380. https://doi.org/10.3390/cancers15225380

4. Shieh, Y., Fejerman, L., Lott, P. C., Marker, K., Sawyer, S. D., Hu, D., Huntsman, S., Torres, J., Echeverry, M. E., Bohórquez, M. E., Martínez-Chéquer, J. C., Polanco-Echeverry, G.,

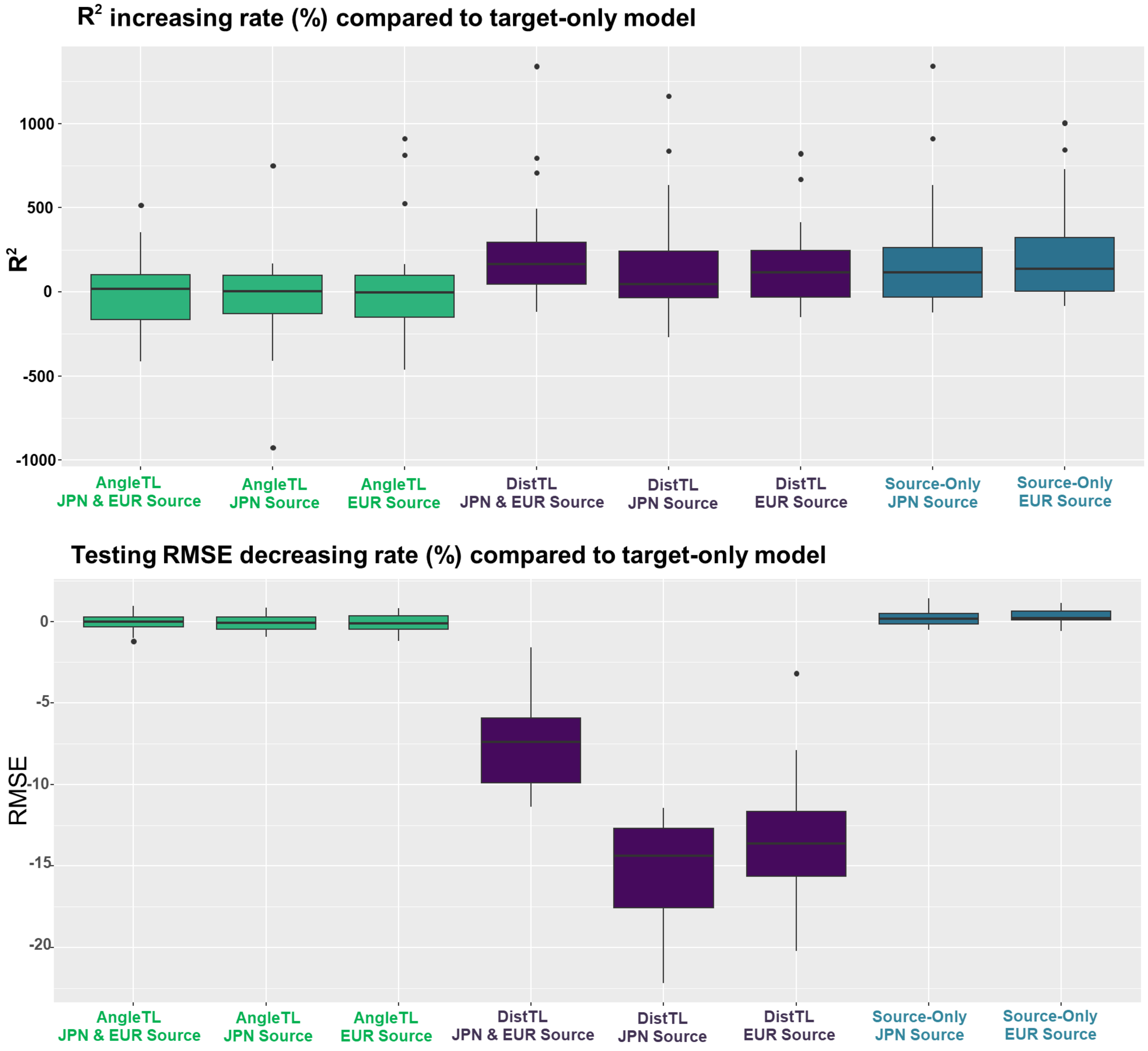
Estrada-Flórez, A. P., the COLUMBUS Consortium, Haiman, C. A., John, E. M., Kushi, L. H., Torres-Mejía, G., Vidaurre, T., Weitzel, J. N., Casavilca Zambrano, S., Carvajal-Carmona, L. G., Ziv, E., Neuhausen, S. L., &... (2020). A Polygenic Risk Score for Breast Cancer in US Latinas and Latin American Women. JNCI: Journal of the National Cancer Institute, 112(6), 590–598. https://doi.org/10.1093/jnci/djz174

5. Gu, T., Han, Y., & Duan, R. (2023). Robust angle-based transfer learning in high dimensions. arXiv:2210.12759 [stat.ME].

6. Lyon, M., Andrews, S. J., Elsworth, B., Gaunt, T. R., Hemani, G., & Marcora, E. (2020). The variant call format provides efficient and robust storage of GWAS summary statistics. bioRxiv. https://doi.org/10.1101/2020.05.29.115824

7. Zhang, H., Ahearn, T. U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., Jiang, X., O'Mara, T. A., Zhao, N., Bolla, M. K., Dunning, A. M., Dennis, J., Wang, Q., Ful, Z. A., Aittomäki, K., Andrulis, I. L., Anton-Culver, H., Arndt, V., Aronson, K. J., Arun, B. K., … García-Closas, M. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nature genetics, 52(6), 572–581. https://doi.org/10.1038/s41588-020-0609-2

Appendix



R^2 increasing rate (%) compared to target-only model

Testing RMSE decreasing rate (%) compared to target-only model

Figure A1. Increase Rate of Testing R^2 & Decrease Rate of Testing RMSE among AngleTL (green), DistTL (purple), and source-only (cyan) models compared to the target-only model