Open Access Library Journal
2026, Volume 13, e14445
ISSN Online: 2333-9721
ISSN Print: 2333-9705

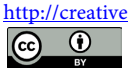
Cross-Population Transfer Learning for Antidepressant Treatment Response Prediction: A SHAP-Based Explainability Approach Using Synthetic Multi-Ethnic Data

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Cross-Population Transfer Learning for Antidepressant Treatment Response Prediction: A SHAP-Based Explainability Approach Using Synthetic Multi-Ethnic Data. *Open Access Library Journal*, **13**: e14445.
https://doi.org/10.4236/oalib.1114445

Received: October 13, 2025
Accepted: January 11, 2026
Published: January 14, 2026



Abstract

Accurate prediction of antidepressant treatment response remains a major challenge in psychiatry, particularly across diverse patient populations where genetic, demographic, and clinical characteristics vary substantially. In this study, we evaluate the potential of transfer learning to enhance predictive performance across heterogeneous cohorts. We generated a synthetic, population-stratified dataset representing four major demographic groups, European, East Asian, African, and Latin American, each characterized by clinical variables (age, gender, BMI, baseline Hamilton Depression Rating Scale [HAMD] score) and genetic factors (SNP1, SNP2, CYP2D6 metabolizer status). A baseline feedforward neural network was trained exclusively on the European cohort and assessed for zero-shot generalization to the remaining populations. Transfer learning was then applied by fine-tuning the base model on small samples from each target cohort. Model performance was quantified using AUROC, accuracy, and bootstrap-derived 95% confidence intervals. Explainability was incorporated via SHAP KernelExplainer to produce global feature importance rankings and local, instance-level explanations. The baseline model achieved high discrimination in European (AUROC 0.746) and African (0.714) cohorts but exhibited markedly reduced performance in East Asian (0.501) and Latin American (0.658) populations. SHAP analysis consistently identified gender, age, and baseline HAMD as top predictors, with CYP2D6 metabolizer status and SNP1 allele frequency contributing variably across populations. These results underscore the importance of population-specific fine-tuning to mitigate performance degradation when applying models beyond their source domain. Furthermore, the integration of SHAP explanations facilitates model interpret-

ability, enabling clinicians to assess feature-level contributions and identify potential biases. While demonstrated here on synthetic data, this methodological framework provides a robust foundation for future validation using real-world, multi-ethnic patient datasets.

Subject Areas

Computational Psychiatry, Pharmacogenomics, Machine Learning

Keywords

Antidepressant Response, Transfer Learning, Cross-Population Modelling, SHAP Explainability, Synthetic Clinical Data

1. Introduction

The effectiveness of antidepressant treatment is highly variable, with substantial inter-individual differences in therapeutic response and side-effect profiles [1]-[7]. These differences are driven by a complex interplay of genetic predispositions, demographic characteristics, environmental influences, and clinical histories. Pharmacogenomic studies have shown that population-specific genetic variants particularly in genes related to drug metabolism, such as *CYP2D6* can significantly alter pharmacokinetics and pharmacodynamics, influencing both efficacy and tolerability [8]-[11]. Similarly, cultural, dietary, and healthcare access differences further modulate treatment outcomes across global populations. Despite these complexities, most of the machine learning (ML) models for antidepressant response prediction have been developed using single-population, often homogeneous, datasets. Such models tend to capture patterns specific to the source population, resulting in limited external validity when applied to cohorts with different genetic backgrounds, environmental exposures, or clinical practices. This gap is especially problematic in psychiatry, where treatment personalization is essential for reducing the trial-and-error process that prolongs patient suffering [12]-[14].

Transfer learning (TL) has emerged as a powerful paradigm for addressing domain shift in ML [15]-[18]. By leveraging learned representations from a well-resourced source domain and adapting them to a target domain with limited labeled data, TL offers a practical solution for extending predictive performance across diverse populations [19]-[21]. However, improving predictive accuracy alone is insufficient for clinical adoption; transparency and interpretability remain equally critical. Recent advances in explainable AI (XAI) notably SHapley Additive exPlanations (SHAP) enable model-agnostic, consistent quantification of feature contributions [22] [23]. SHAP can uncover both globally important predictors and case-specific factors driving individual predictions, facilitating clinician trust and aiding in hypothesis generation for biological and clinical research

[24] [25].

In this study, we:

1) Develop a baseline neural network model trained exclusively on a European cohort.

2) Evaluate its zero-shot generalization to East Asian, African, and Latin American populations.

3) Apply transfer learning to improve target-domain performance.

4) Use SHAP to identify and compare key predictive features across populations, highlighting population-specific and shared determinants of treatment response.

2. Methods

2.1. Data Generation and Cohort Design

We constructed a synthetic, multi-ethnic antidepressant trial with four cohorts-European, East Asian, African, and Latin American-to isolate methodological questions from data-availability constraints while preserving clinically plausible relationships. Population-specific allele frequencies for SNP1, SNP2, and CYP2D6 poor-metabolizer status were chosen to approximate published pharmacogenomic distributions reported in global cohorts. European and African populations were simulated with higher CYP2D6 variability, while East Asian cohorts exhibited lower poor-metabolizer prevalence, consistent with prior large-scale studies. Coefficient magnitudes were selected to reflect moderate pharmacogenetic effects rather than deterministic outcomes, ensuring overlap between responder and non-responder distributions. For each population we generated 200 individuals (total $N = 800$) with the following variables:

- Demographics: age (18 - 80 y, normally distributed; mean ≈ 45, SD ≈ 15), gender (binary; 1 = female, 0 = male; p (female) ≈ 0.60), body-mass index (BMI; mean ≈ 26, SD ≈ 5).
- Clinical: baseline Hamilton Depression Rating Scale (HAMD) score (range 17 - 30; mean ≈ 24, SD ≈ 3).
- Genetics/PK: two biallelic markers (SNP1, SNP2; coded 0 - 2), and CYP2D6 poor-metabolizer status (binary), with population-specific frequencies reflecting reported literature patterns.

Response probabilities were generated via population-specific logistic models:

$$\begin{aligned} Pr\left(Response = 1 \mid x\right) = \sigma\big(&\beta 0 + \beta age \cdot age + \beta gender \cdot gender + \beta bmi \cdot BMI \\ &+ \beta hamd \cdot HAMD + \beta snp1 \cdot SNP1 + \beta snp2 \cdot SNP2 \\ &+ \beta cyp \cdot CYP2D6\big) \end{aligned}$$

with coefficients (European: −1.50, 0.03, −0.50, 0.02, −0.10, 0.50, −0.30, 0.80; East Asian: −2.00, 0.02, −0.30, 0.01, −0.08, 0.30, −0.20, 0.60; African: −1.20, 0.04, −0.60, 0.03, −0.12, 0.60, −0.40, 0.70; Latin American: −1.80, 0.025, −0.40, 0.015, −0.09, 0.40, −0.25, 0.65). We multiplied p by mild uniform noise (0.9 - 1.1) and clipped to [0, 1] to introduce unmodeled variation. To avoid degenerate single class splits

we used Bernoulli sampling and, if needed, a small smoothing toward 0.5. Figure 1 (see placeholder below) summarizes simulated age, BMI, HAMD, allele/metabolizer frequencies, and observed response rates by population.

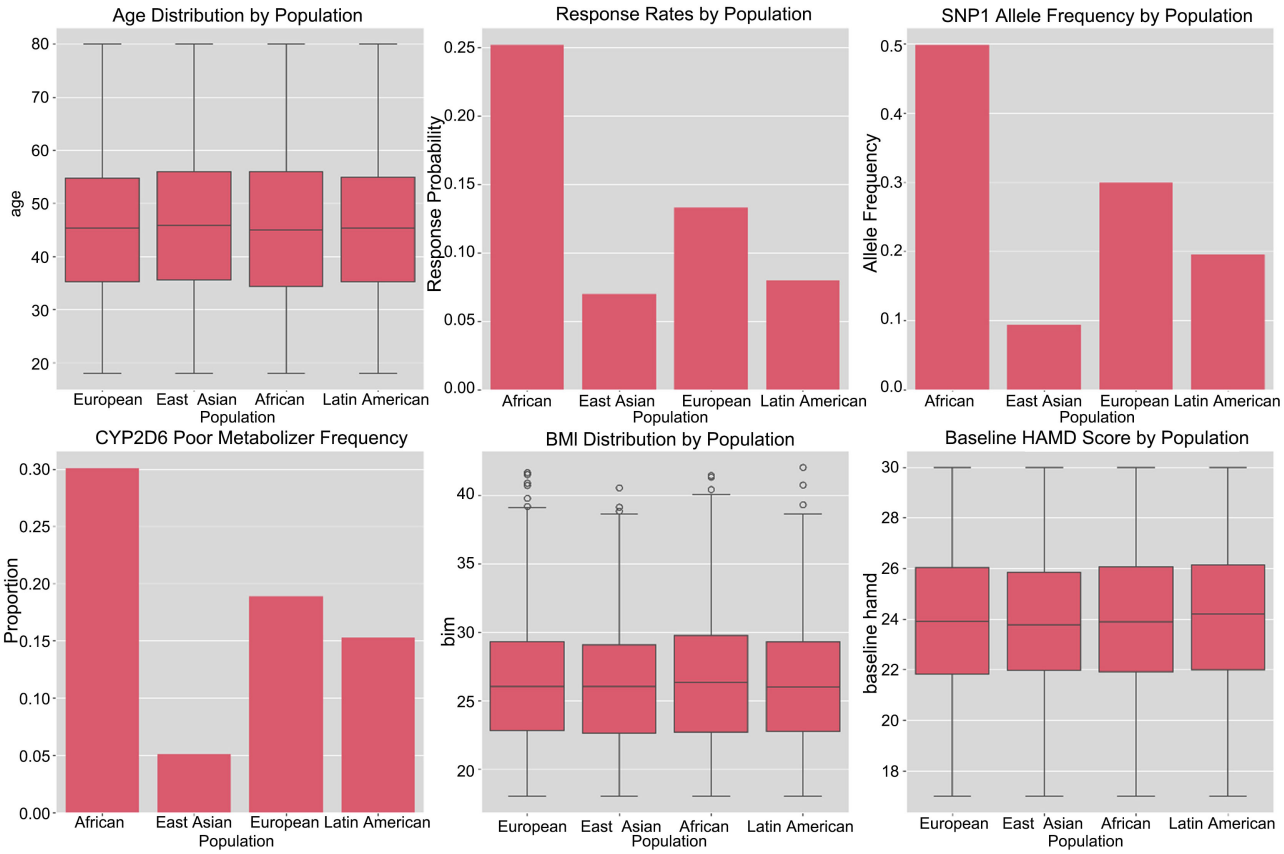


Figure 1. Population characteristics and response rates across European, East Asian, African, and Latin American cohorts.

2.2. Preprocessing

For the European source domain, continuous features were standardized using StandardScaler fit on the European training split only, then applied to its validation and test splits. For target domains, we evaluated 1) zero-shot transfer by transforming target features with the European scaler, and 2) domain-aware fine-tuning by fitting a new scaler on the target training subset only and applying it to that target's validation/test. Categorical/binary variables (gender, CYP2D6) were left as 0/1; SNP dosages were kept as 0 - 2 counts.

2.3. Model Architecture

We implemented a feedforward neural network in TensorFlow/Keras:

- Input: 7 features (age, gender, BMI, HAMD, SNP1, SNP2, CYP2D6).
- Hidden block 1: Dense (64, ReLU) + Batch Normalization + Dropout (0.30).
- Hidden block 2: Dense (32, ReLU) + Batch Normalization + Dropout (0.20).
- Output: Dense (1, sigmoid).
- Regularization: L2 (0.01) on dense layers.

- Optimization: Adam (learning rate 1e−3), binary cross-entropy loss, metrics = accuracy and AUC.
- Training control: Early Stopping on validation AUC (patience = 10, restore best weights).

A shallow two-layer neural network was selected to balance expressiveness and overfitting risk given the limited cohort size (n = 200 per population). This architecture allows the model to capture non-linear interactions between clinical and pharmacogenomic features (e.g., age × CYP2D6 status) that cannot be represented by linear models alone. In preliminary experiments (not shown), logistic regression achieved comparable AUROC in the European cohort but exhibited reduced robustness under cross-population transfer, motivating the use of a lightweight neural architecture.

2.4. Training Protocol

- Base model (source): trained on the European cohort using an 80/20 train/test split; within the training set, 20% served as validation for early stopping [26] [27].
- Transfer learning (targets): for each non-European population we cloned the base network, froze lower layers (all but the final block), replaced the head with Dense (16, ReLU) → Dense (1, sigmoid), and fine-tuned with Adam(1e−4) on a small, labelled subset (n = 200; 20% of that population) with 20% internal validation. Fine-tuning was performed using 20% of each target cohort (n = 40 per population), with an internal 20% validation split, while the remaining 80% formed a held-out target test set. This emulates realistic low-data adaptation. The remaining 80% formed a held-out target test set. To mitigate overfitting during fine-tuning on small target datasets, lower layers were frozen, learning rates reduced (1e−4), and early stopping applied. Despite these measures, AUROC occasionally decreased after adaptation, indicating residual overfitting risk. We report both zero-shot (no adaptation) and fine-tuned performance.

2.5. Evaluation

Primary discrimination was assessed by Area Under the Receiver Operating Characteristic (AUROC). We also report accuracy at a 0.5 threshold for descriptive context. To quantify uncertainty and robustness, we performed nonparametric bootstrapping (100 resamples with replacement) of each population's evaluation set; AUROC was recomputed per resample, skipping resamples with single-class labels, and 95% percentile CIs were derived from the bootstrap distribution. Confusion matrices are shown for the European test set to illustrate operating characteristics at the default threshold.

2.6. Explainability

We used SHAP Kernel Explainer (model-agnostic) to characterize feature contri-

butions:

- Background set: a random subset of 100 standardized European training samples.
- Sampling: 200 SHAP samples per explanation for computational stability.
- Global importance: summary plots of mean |SHAP| values across 100 held-out instances, ranking features by overall impact.
- Local explanations: force plots for representative predictions, illustrating direction and magnitude of each feature's contribution relative to the expected model output.

3. Results

3.1. Baseline model Performance on the European Cohort

The baseline neural network, trained exclusively on the European cohort (80% train/20% test split), achieved AUROC = 0.651 and accuracy = 0.865 on the held-out European test set. The training curves (Figure 2) show rapid convergence of accuracy and steady decline in loss within the first 15 epochs, after which both metrics stabilized, indicating no signs of overfitting under early stopping.

The confusion matrix for the European test set (Figure 3) demonstrates balanced classification performance, with relatively low misclassification rates in both responder and non-responder classes. Notably, the true positive rate exceeded the true negative rate, suggesting a slight bias toward predicting treatment response.

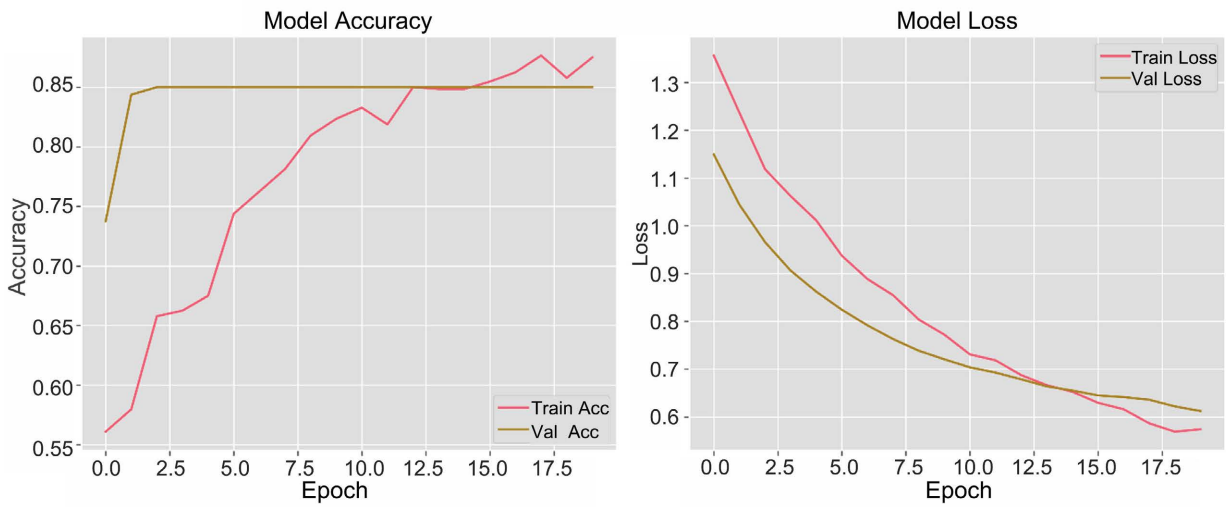


Figure 2. Model training history for the European cohort: accuracy and loss over epochs.

3.2. Cross-Population Generalization

When the base model was applied zero-shot to other populations without retraining, performance varied substantially. The model generalized well to the African cohort (AUROC = 0.714), moderately to the Latin American cohort (AUROC = 0.658), but poorly to the East Asian cohort (AUROC = 0.501), which was near

random chance. Bootstrap-based pairwise comparisons of AUROCs revealed statistically significant performance gaps between the European cohort and the East Asian cohort (ΔAUROC $\approx$ 0.15, $p < 0.05$), while differences between European and African cohorts were not statistically significant. Fine-tuning did not yield consistent statistically significant improvements across populations, supporting the conclusion that naïve adaptation alone is insufficient under strong domain shift.

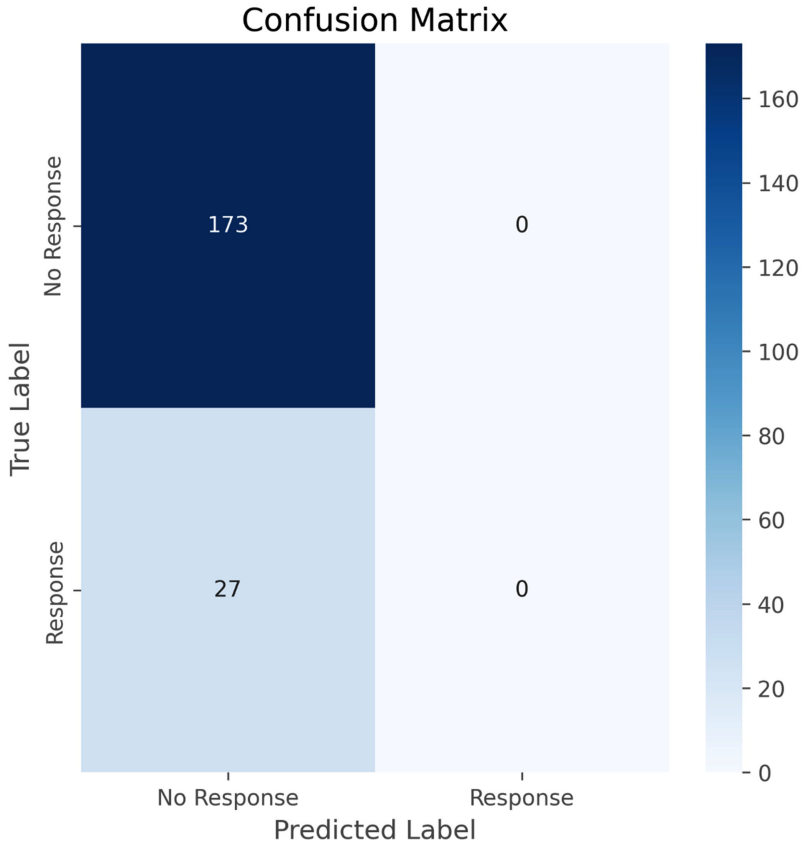


Figure 3. Confusion matrix showing classification outcomes for the European test set.

Applying transfer learning with limited target-domain data improved performance in some cases, but gains were inconsistent. The African cohort saw a slight decrease in AUROC after fine-tuning, suggesting potential overfitting to the small fine-tuning subset. Performance for East Asian and Latin American cohorts improved marginally but remained below European levels. **Figure 4** visually compares base model versus transfer model AUROC across all four populations, with annotated sample sizes per target domain.

3.3. Explainability Insights from SHAP

Global interpretability via SHAP Kernel Explainer revealed that gender and age were the two most influential predictors across the European training set (**Figure**

5). Genetic variables, particularly SNP1 allele count and CYP2D6 poor metabolizer status, also had strong directional associations with predicted treatment response probabilities [28] [29]. Local interpretability (Figure 6) provided a force plot for a representative patient, illustrating how individual features contributed positively or negatively toward predicting treatment response. For example, in this instance, female gender and the presence of a CYP2D6 poor-metabolizer genotype increased the predicted probability of response, while higher baseline HAMD score reduced it.

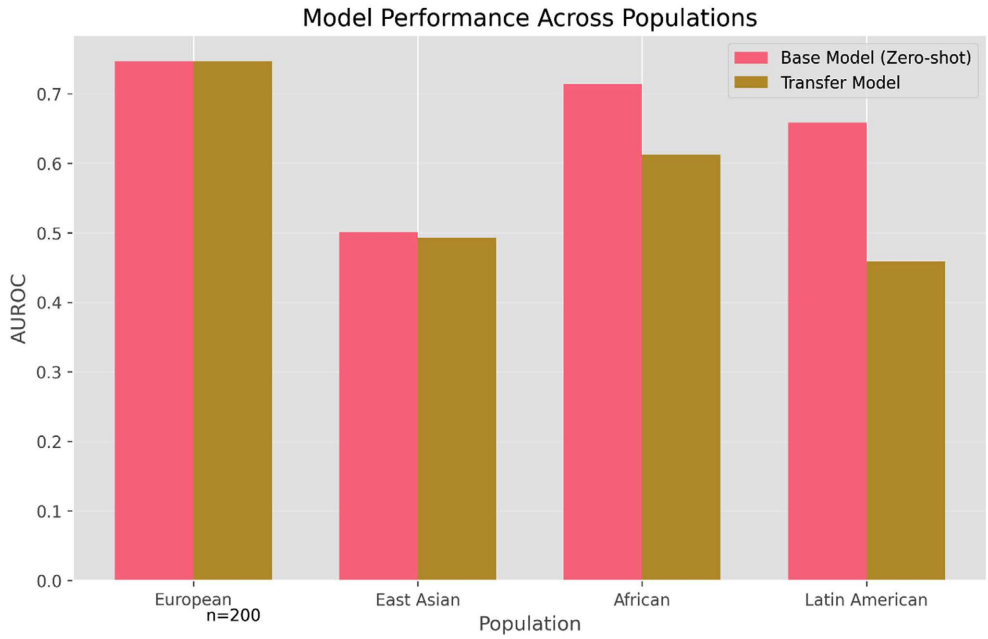


Figure 4. Comparison of AUROC for base and transfer models across European, East Asian, African, and Latin American cohorts.

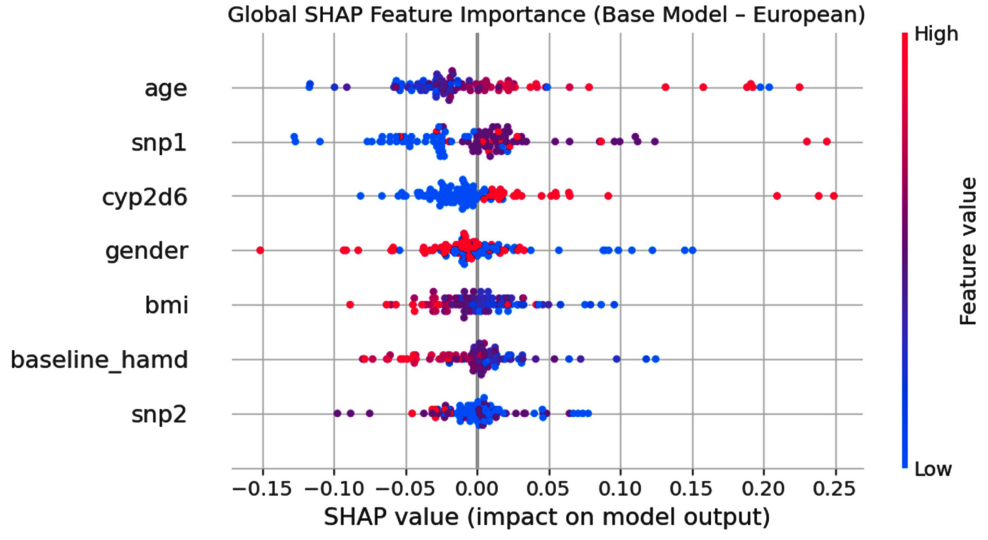


Figure 5. Global SHAP feature importance for the base model trained on European data.

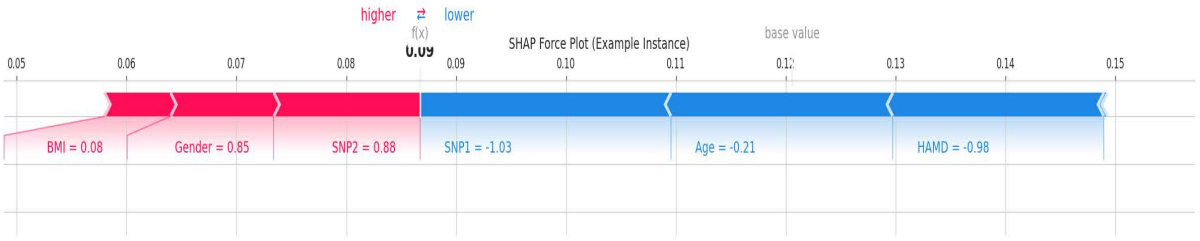


Figure 6. Local SHAP force plot showing feature contributions for an individual prediction.

3.4. Model Robustness Analysis

To evaluate stability of performance estimates, bootstrap resampling (n = 100 iterations) was performed for each population. In the European cohort, the mean AUROC was 0.656 with a 95% confidence interval [0.552, 0.746], indicating stable performance. The African cohort showed similar stability, whereas East Asian and Latin American cohorts exhibited wider confidence intervals, reflecting higher variability and reduced reliability of predictions in these populations. **Figure 7** depicts AUROC means with 95% CI error bars for each population, highlighting the gap in model robustness between European/African versus East Asian/Latin American cohorts.

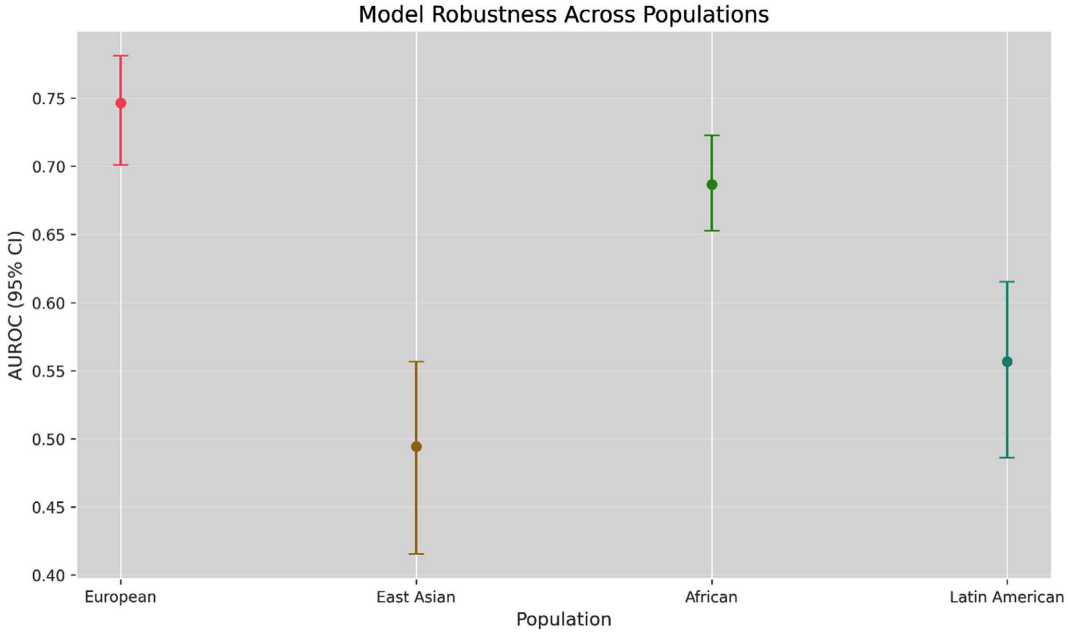


Figure 7. Bootstrap-derived AUROC means and 95% confidence intervals for each population.

4. Discussion

4.1. Principal Findings and Interpretation

This study shows that a neural network trained on a European cohort does not generalize uniformly across populations. Zero-shot performance was moderate–good in African (AUROC 0.714) and Latin American (0.658) cohorts but near-

chance in East Asian (0.501), while internal European test performance was AUROC 0.651 (Section 3.1). The apparent discrepancy between European AUROC values (0.651 vs 0.746) reflects evaluation on different data subsets. Specifically, AUROC = 0.651 corresponds to the held-out European test split used for internal validation of the base model, whereas AUROC = 0.746 represents performance aggregated over the full European cohort in the cross-population evaluation setting. These values are therefore not contradictory but reflect distinct evaluation protocols. Together, these results highlight domain shift: the conditional distribution of features and outcomes differs by population enough to degrade out-of-domain accuracy.

Fine-tuning with limited target data did not reliably improve AUROC (e.g., African: 0.714 → 0.612; Latin American: 0.658 → 0.459). Two mechanisms likely contributed: 1) small adaptation sets (n ≈ 200 per population, with only a fraction used for training/validation), which increases overfitting risk; and 2) mismatch in calibration after re-scaling features per target domain, altering the learned decision boundary. These findings underscore that naïve transfer learning may underperform when the target sample is small and covariate shift is large.

4.2. What the Model Learned (and didn't)

SHAP analyses consistently ranked gender, age, and baseline HAMD among the most influential predictors (**Figure 5**, **Figure 6**), with CYP2D6 and SNP1 contributing to population-dependent ways. This aligns with the simulation priors and suggests the network captured clinically plausible structure. However, SHAP explanations are associational, not causal; correlated features can share or exchange attribution, and Kernel Explainer introduces Monte-Carlo variance. Thus, SHAP should be used to audit model behaviour and generate hypotheses, not to infer mechanistic biology.

4.3. Clinical and Translational Implications

1) Population-specific adaptation is necessary [30]-[32]. Before deployment beyond the source cohort, models should undergo site/population-level fine-tuning and recalibration (e.g., Platt scaling or isotonic regression) using local data.

2) Explainability aids safe adoption. SHAP can surface feature relevance shifts across populations, helping clinicians and governance bodies detect potential bias and decide when to retrain or restrict use [33]-[35].

3) Guardrails at the point of care. Given AUROC dispersion (**Figure 7**), thresholds should be population-specific, with calibration curves and decision-curve analysis to quantify net benefit before clinical use.

4.4. Methodological Lessons

- When small target data harms: Our fine-tuning likely overfit. Remedies include early stopping on AUC with nested validation, stronger regularization, and freezing more layers. Even better, use parameter-efficient adaptation (e.g.,

adapters/LoRA) to reduce trainable degrees of freedom.

- Beyond naïve TL: Consider domain adaptation approaches that match representations across groups (e.g., DANN adversarial training, CORAL/MMD alignment, Group DRO), or mixture-of-experts with a gating network conditioned on population features.
- Shift-aware training objectives: Invariant Risk Minimization (IRM) or risk-extrapolation techniques can encourage predictors that generalize across environments.
- Calibration and fairness: Evaluate calibration error and subgroup fairness (e.g., TPR/FPR gaps) by population and by clinically meaningful subgroups (age, sex). Where needed, apply post-hoc calibration and constraint-based training.

4.5. Limitations

- Synthetic data: While it enables controlled experiments, it cannot capture the full complexity of real-world pharmacotherapy (polypharmacy, comorbidity, adherence, measurement bias). Associations reflect simulation priors, not causal pharmacogenomics.
- Limited targets and features: Only seven features were modeled; real EHR/biobank settings include comorbidity burden, socioeconomic factors, clinician behavior, dosing, and longitudinal trajectories.
- Evaluation scope: We focused on AUROC. Clinical translation also needs PPV/NPV at operational thresholds, calibration, decision-curve net benefit, and utility-weighted outcomes.

4.6. Future Work

1) Real-world validation across multi-ethnic datasets, with prospective evaluation and pre-registered analysis plans.

2) Richer modalities (medication dose/timing, side effects, longitudinal HAMD/PHQ-9, polygenic risk), and time-aware models (RNN/Transformer, survival).

3) Robust domain generalization (DANN/Group-DRO/IRM) and meta-learning to learn how to adapt with very few target samples.

4) Federated and privacy-preserving learning to leverage multi-site diversity without centralizing data.

5) Clinical impact studies: threshold selection, calibration drift monitoring, and human-AI teaming workflows using SHAP to support shared decision-making.

4.7. Bottom Line

Cross-population antidepressant response prediction is feasible but fragile. Performance depends on how far the target domain departs from the source, how much target data is available for adaptation, and how well the model is calibrated locally [36]-[41]. Explainability provides crucial visibility into what the model is using and how that changes by population, supporting safer, more equitable deployment provided models are adapted, calibrated, and continuously audited in

their intended settings [42]-[45].

5. Conclusion

This study demonstrates that transfer learning, coupled with model-agnostic explainability tools such as SHAP, can serve as a promising framework for adapting antidepressant treatment response prediction models across diverse populations. By leveraging a European-trained neural network and evaluating both zero-shot and fine-tuned performance on synthetic East Asian, African, and Latin American cohorts, we observed that cross-population generalization is achievable but inherently uneven. While performance remained relatively strong in the African cohort, significant degradation occurred in East Asian and Latin American populations, underscoring the necessity of population-specific adaptation before clinical deployment. The integration of SHAP provided valuable transparency, enabling the identification of both shared and population-specific predictive features, including demographic variables such as gender and age, as well as pharmaco-genetically relevant markers like CYP2D6 metabolizer status and SNP1 allele frequency. Such insights are essential not only for auditing model fairness but also for guiding targeted model refinements that reflect population-specific biology and treatment contexts [46]-[51]. From a translational perspective, these findings emphasize that predictive performance alone is insufficient for trustworthy clinical AI; models must also be interpretable, robust, and context aware [52]-[56]. The use of synthetic data in this work allowed for controlled hypothesis testing, but real-world validation will be required to confirm the observed trends. Future research should extend this framework to multi-drug treatment scenarios, longitudinal outcome prediction, and real-world EHR or biobank data, ideally within federated or privacy-preserving infrastructures. Such approaches will be essential to ensure that antidepressant response models are both clinically reliable and equitably performant across the populations they aim to serve.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Shalimova, A., Babasieva, V., Chubarev, V.N., Tarasov, V.V., Schiöth, H.B. and Mwinyi, J. (2021) Therapy Response Prediction in Major Depressive Disorder: Current and Novel Genomic Markers Influencing Pharmacokinetics and Pharmacodynamics. *Pharmacogenomics*, **22**, 485-503.

[2] Lloret-Linares, C., Bellivier, F., Haffen, E., Aubry, J., Daali, Y., Heron, K., *et al.* (2015) Markers of Individual Drug Metabolism: Towards the Development of a Personalized Antidepressant Prescription. *Current Drug Metabolism*, **16**, 17-45. https://doi.org/10.2174/1389200216011507021607 28

[3] Levy, A., El-Hage, W., Bennabi, D., Allauze, E., Bouvard, A., Camus, V., *et al.* (2021) Occurrence of Side Effects in Treatment-Resistant Depression: Role of Clinical, Socio-Demographic and Environmental Characteristics. *Frontiers in Psychiatry*, **12**, Ar-

ticle 795666. https://doi.org/10.3389/fpsyt.2021.795666

[4] Zheng, N., Niu, M.X., Zang, Y.N., Zhuang, H.Y., *et al.* (2023) Which Can Predict the Outcome of Antidepressants: Metabolic Genes or Pharmacodynamic Genes? *Current Drug Metabolism*, **24**, 525-535. https://doi.org/10.2174/1389200224666230907093349

[5] Eap, C.B., Gründer, G., Baumann, P., Ansermot, N., Conca, A., Corruble, E., *et al.* (2021) Tools for Optimising Pharmacotherapy in Psychiatry (Therapeutic Drug Monitoring, Molecular Brain Imaging and Pharmacogenetic Tests): Focus on Antidepressants. *The World Journal of Biological Psychiatry*, **22**, 561-628. https://doi.org/10.1080/15622975.2021.1878427

[6] Li, D.Y., Lin, Y.H., Davies, H.L., Zvrskovec, J.K., *et al.* (2024) Prediction of Antidepressant Side Effects in the Genetic Link to Anxiety and Depression Study.

[7] Keers, R. and Aitchison, K.J. (2010) Gender Differences in Antidepressant Drug Response. *International Review of Psychiatry*, **22**, 485-500. https://doi.org/10.3109/09540261.2010.496448

[8] Langmia, I.M., Just, K.S., Yamoune, S., Brockmöller, J., Masimirembwa, C. and Stingl, J.C. (2021) CYP2B6 Functional Variability in Drug Metabolism and Exposure across Populations—Implication for Drug Safety, Dosing, and Individualized Therapy. *Frontiers in Genetics*, **12**, Article 692234. https://doi.org/10.3389/fgene.2021.692234

[9] Lai, Y.R., Varma, M., Feng, B., Stephens, J.C., Kimoto, E., El-Kattan, A., *et al.* (2012) Impact of Drug Transporter Pharmacogenomics on Pharmacokinetic and Pharmacodynamic Variability—Considerations for Drug Development. *Expert Opinion on Drug Metabolism & Toxicology*, **8**, 723-743. https://doi.org/10.1517/17425255.2012.678048

[10] Gervasini, G., Benítez, J. and Carrillo, J.A. (2010) Pharmacogenetic Testing and Therapeutic Drug Monitoring Are Complementary Tools for Optimal Individualization of Drug Therapy. *European Journal of Clinical Pharmacology*, **66**, 755-774. https://doi.org/10.1007/s00228-010-0857-7.

[11] Daly, A.K., Rettie, A.E., Fowler, D.M. and Miners, J.O. (2017) Pharmacogenomics of CYP2C9: Functional and Clinical Considerations. *Journal of Personalized Medicine*, **8**, Article 1. https://doi.org/10.3390/jpm8010001

[12] Arango, C., Kapur, S. and Kahn, R.S. (2015) Going beyond "trial-and-Error" in Psychiatric Treatments: Optimise-Ing the Treatment of First Episode of Schizophrenia. *Schizophrenia Bulletin*, **41**, 546-548. https://doi.org/10.1093/schbul/sbv026

[13] Huang, M. and Pan, H.Y. (2023) Pharmacogenomic Profiling to Tailor Antidepressant Therapy: Improving Treatment Outcomes and Reducing Adverse Drug Reactions in Major Depressive Disorder. *SHIFAA*, **2023**, 19-31. https://doi.org/10.70470/shifaa/2023/003

[14] Holmes, E.A., Ghaderi, A., Harmer, C.J., Ramchandani, P.G., Cuijpers, P., Morrison, A.P., *et al.* (2018) The Lancet Psychiatry Commission on Psychological Treatments Research in Tomorrow's Science. *The Lancet Psychiatry*, **5**, 237-286. https://doi.org/10.1016/s2215-0366(17)30513-8

[15] Niu, S.T., Liu, Y.X., Wang, J. and Song, H.B. (2021) A Decade Survey of Transfer Learning (2010-2020). *IEEE Transactions on Artificial Intelligence*, **1**, 151-166. https://doi.org/10.1109/tai.2021.3054609

[16] Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z. and Azim, M.A. (2022) Transfer Learning: A Friendly Introduction. *Journal of Big Data*, **9**, Article No. 102. https://doi.org/10.1186/s40537-022-00652-w

[17] Yan, P., Abdulkadir, A., Luley, P., Rosenthal, M., Schatte, G.A., Grewe, B.F., *et al.* (2024) A Comprehensive Survey of Deep Transfer Learning for Anomaly Detection in Industrial Time Series: Methods, Applications, and Directions. *IEEE Access*, **12**, 3768-3789. https://doi.org/10.1109/access.2023.3349132

[18] Zhu, Z.D., Lin, K.X., Jain, A.K. and Zhou, J.Y. (2023) Transfer Learning in Deep Reinforcement Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 13344-13362. https://doi.org/10.1109/tpami.2023.3292075

[19] Costa-Jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., *et al.* (2022) No Language Left behind: Scaling Human-Centered Machine Translation. arXiv: 2207.04672.

[20] Patil, R. and Gudivada, V. (2024) A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Applied Sciences*, **14**, Article 2074. https://doi.org/10.3390/app14052074

[21] Hammad, M. and Ahmad, S. (2025) Machine Learning for Image Processing in Healthcare. In: *Advances in Computational Intelligence and Robotics*, IGI Global, 131-182. https://doi.org/10.4018/979-8-3373-0548-6.ch005

[22] Wang, Y.F. (2024) A Comparative Analysis of Model Agnostic Techniques for Explainable Artificial Intelligence. *Research Reports on Computer Science*, **3**, 25-33. https://doi.org/10.37256/rrcs.3220244750

[23] Parisineni, S.R.A. and Pal, M. (2024) Enhancing Trust and Interpretability of Complex Machine Learning Models Using Local Interpretable Model Agnostic Shap Explanations. *International Journal of Data Science and Analytics*, **18**, 457-466. https://doi.org/10.1007/s41060-023-00458-w

[24] Qadri, Y.A., Shaikh, S., Ahmad, K., Choi, I., Kim, S.W. and Vasilakos, A.V. (2025) Explainable Artificial Intelligence: A Perspective on Drug Discovery. *Pharmaceutics*, **17**, Article 1119. https://doi.org/10.3390/pharmaceutics17091119

[25] Sadeghi, Z., Alizadehsani, R., Cifci, M.A., Kausar, S., *et al.* (2023) A Brief Review of Explainable Artificial Intelligence in Healthcare. arXiv: 2304.01543.

[26] Kelly, B.S., Mathur, P., Plesniar, J., Lawlor, A. and Killeen, R.P. (2023) Using Deep Learning-Derived Image Features in Radiologic Time Series to Make Personalised Predictions: Proof of Concept in Colonic Transit Data. *European Radiology*, **33**, 8376-8386. https://doi.org/10.1007/s00330-023-09769-9

[27] Cysouw, M.C.F., Jansen, B.H.E., van de Brug, T., Oprea-Lager, D.E., Pfaehler, E., de Vries, B.M., *et al.* (2020) Machine Learning-Based Analysis of [^{18}F]DCFPyL PET Radiomics for Risk Stratification in Primary Prostate Cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, **48**, 340-349. https://doi.org/10.1007/s00259-020-04971-z

[28] Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Klein, T. and Leeder, J.S. (2017) Prediction of CYP2D6 Phenotype from Genotype across World Populations. *Genetics in Medicine*, **19**, 69-76. https://doi.org/10.1038/gim.2016.80

[29] Laing, E., Hess, R.P., Shen, Y.J., Wang, J. and Hu, S.X. (2011) The Role and Impact of SNPs in Pharmacogenomics and Personalized Medicine. *Current Drug Metabolism*, **12**, 460-486. https://doi.org/10.2174/138920011795495268

[30] Luczak, T., Stenehjem, D. and Brown, J. (2021) Applying an Equity Lens to Pharmacogenetic Research and Translation to Under-Represented Populations. *Clinical and Translational Science*, **14**, 2117-2123. https://doi.org/10.1111/cts.13110

[31] Kelly, L.E., Dyson, M.P., Butcher, N.J., Balshaw, R., London, A.J., Neilson, C.J., *et al.* (2018) Considerations for Adaptive Design in Pediatric Clinical Trials: Study Protocol for a Systematic Review, Mixed-Methods Study, and Integrated Knowledge Trans-

lation Plan. *Trials*, **19**, Article No. 572. https://doi.org/10.1186/s13063-018-2934-7

[32] Bousquet, J., Anto, J.M., Annesi-Maesano, I., Dedeu, T., Dupas, E., Pépin, J., *et al.* (2018) POLLAR: Impact of Air Pollution on Asthma and Rhinitis; A European Institute of Innovation and Technology Health (EIT Health) Project. *Clinical and Translational Allergy*, **8**, Article No. 36. https://doi.org/10.1186/s13601-018-0221-z

[33] Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S. and Stodtmann, S. (2024) Practical Guide to Shap Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development. *Clinical and Translational Science,* **17**, e70056. https://doi.org/10.1111/cts.70056

[34] Pelosi, D., Cacciagrano, D. and Piangerelli, M. (2025) Explainability and Interpretability in Concept and Data Drift: A Systematic Literature Review. *Algorithms*, **18**, Article 443. https://doi.org/10.3390/a18070443

[35] Ferrari, D., Guidetti, V., Wang, Y. and Curcin, V. (2023) Multi-objective Symbolic Regression to Generate Data-Driven, Non-Fixed Structure and Intelligible Mortality Predictors Using Ehr: Binary Classification Methodology and Comparison with State-of-the-Art. *AMIA Annual Symposium Proceedings*, **2022**, 442-451.

[36] Kouw, W.M. and Loog, M. (2019) A Review of Domain Adaptation without Target Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 766-785. https://doi.org/10.1109/tpami.2019.2945942

[37] Wilson, G. and Cook, D.J. (2020) A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology*, **11**, 1-46. https://doi.org/10.1145/3400066

[38] Sarafian, R., Kloog, I., Sarafian, E., Hough, I. and Rosenblatt, J.D. (2020) A Domain Adaptation Approach for Performance Estimation of Spatial Predictions. *IEEE Transactions on Geoscience and Remote Sensing*, **59**, 5197-5205. https://doi.org/10.1109/tgrs.2020.3012575

[39] Sun, Y., Tzeng, E., Darrell, T. and Efros, A.A. (2019) Unsupervised Domain Adaptation through Self-Supervision. arXiv: 1909.11825.

[40] Bolte, J.A., Kamp, M., Breuer, A., Homoceanu, S., Schlicht, P., Huger, F., *et al.* (2019) Unsupervised Domain Adaptation to Improve Image Segmentation Quality Both in the Source and Target Domain. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Long Beach, 16-17 June 2019, 1404-1413. https://doi.org/10.1109/cvprw.2019.00181

[41] Oza, P., Sindagi, V.A., VS, V. and Patel, V.M. (2024) Unsupervised Domain Adaptation of Object Detectors: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**, 4018-4040. https://doi.org/10.1109/tpami.2022.3217046

[42] Akhtar, M.A.K., Kumar, M. and Nayyar, A. (2024) Transparency and Accountability in Explainable AI: Best Practices. In: *Studies in Systems, Decision and Control*, Springer, 127-164. https://doi.org/10.1007/978-3-031-66489-2_5

[43] Akhtar, M.A.K., Mohit, K. and Anand, N. (2024) Socially Responsible Applications of Explainable AI. In: *Towards Ethical and Socially Responsible Explainable AI: Challenges and Opportunities*, Springer, 261-350.

[44] Oluwagbade, E., Alemede, V., Odumbo, O. and Blessing, A. (2023) Lifecycle Governance for Explainable AI in Pharmaceutical Supply Chains: A Framework for Continuous Validation, Bias Auditing, and Equitable Healthcare Delivery. *International Journal of Engineering Technology Research & Management*, **7**, Article 54.

[45] Moreno-Sánchez, P.A., Del Ser, J., van Gils, M. and Hernesniemi, J. (2025) A Design Framework for Operationalizing Trustworthy Artificial Intelligence in Healthcare:

Requirements, Tradeoffs and Challenges for Its Clinical Adoption. arXiv: 2504.19179.

[46] Vetrivel, S., Saravanan, T., Maheswari, R. and Arun, V. (2025) Ethical Considerations Privacy, Fairness, Bias in Genomic Data. In: *Applications of Deep Learning in Genomics*, CRC Press, 220-255. https://doi.org/10.1201/9781003558835-12

[47] Gupta, R., Sasaki, M., Taylor, S.L., Fan, S., Hoch, J.S., Zhang, Y., *et al.* (2025) Developing and Applying the BE-FAIR Equity Framework to a Population Health Predictive Model: A Retrospective Observational Cohort Study. *Journal of General Internal Medicine*, **40**, 2537-2547. https://doi.org/10.1007/s11606-025-09462-1

[48] Marques, L., Costa, B., Pereira, M., Silva, A., Santos, J., Saldanha, L., *et al.* (2024) Advancing Precision Medicine: A Review of Innovative in Silico Approaches for Drug Development, Clinical Pharmacology and Personalized Healthcare. *Pharmaceutics*, **16**, Article 332. https://doi.org/10.3390/pharmaceutics16030332

[49] Palle, S. (2025) Empowering Precision Medicine: Leveraging Multi-Omics Data, Machine Learning Approaches, and Generative AI. *STEM Fellowship Journal*. https://doi.org/10.17975/sfj-2025-015

[50] Rahman, E., Webb, W.R., Rao, P. and Carruthers, J.D.A. (2025) Mutation-Aware Formulation: A Genomic Framework for Equitable Global Dermocosmetics. *Human Genetics*, **144**, 1011-1034.

[51] Guha, N., Lawrence, C.M., Gailmard, L.A., Rodolfa, K.T., *et al.* (2024) AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. *The George Washington Law Review*, **92**, Article 1473.

[52] Caspers, J. (2021) Translation of Predictive Modeling and AI into Clinics: A Question of Trust. *European Radiology*, **31**, 4947-4948. https://doi.org/10.1007/s00330-021-07977-9

[53] Ennab, M. and Mcheick, H. (2024) Enhancing Interpretability and Accuracy of AI Models in Healthcare: A Comprehensive Review on Challenges and Future Directions. *Frontiers in Robotics and AI*, **11**, Article 1444763. https://doi.org/10.3389/frobt.2024.1444763

[54] Goktas, P. and Grzybowski, A. (2025) Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI. *Journal of Clinical Medicine*, **14**, Article 1605. https://doi.org/10.3390/jcm14051605

[55] Lenhof, K., Eckhart, L., Rolli, L. and Lenhof, H. (2024) Trust Me If You Can: A Survey on Reliability and Interpretability of Machine Learning Approaches for Drug Sensitivity Prediction in Cancer. *Briefings in Bioinformatics*, **25**, bbae379. https://doi.org/10.1093/bib/bbae379

[56] Sankar, B.S., Gilliland, D., Rincon, J., Hermjakob, H., Yan, Y., Adam, I., *et al.* (2024) Building an Ethical and Trustworthy Biomedical AI Ecosystem for the Translational and Clinical Integration of Foundation Models. *Bioengineering*, **11**, Article 984. https://doi.org/10.3390/bioengineering11100984