Open Access Library Journal
2026, Volume 13, e15350
ISSN Online: 2333-9721
ISSN Print: 2333-9705

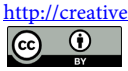
Cross-Institutional Generalization of Psychiatric AI Models: A Domain Adaptation and Robustness Framework for Multi-Site Clinical Deployment

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-Institutional Generalization of Psychiatric AI Models: A Domain Adaptation and Robustness Framework for Multi-Site Clinical Deployment. *Open Access Library Journal*, **13**: e15350.
https://doi.org/10.4236/oalib.1115350

Received: April 14, 2026
Accepted: May 24, 2026
Published: May 27, 2026



Abstract

The deployment of artificial intelligence models in psychiatric care faces a critical challenge: models trained in one healthcare institution often fail to generalize across different populations, clinical practices, and healthcare systems. This study investigates domain adaptation and robustness techniques to ensure psychiatric AI models maintain predictive accuracy when deployed across diverse institutional settings. We developed a comprehensive framework integrating transfer learning, adversarial domain adaptation, and federated learning approaches to address cross-institutional heterogeneity. Using synthetic clinical datasets representing six distinct hospitals with varying patient demographics, treatment protocols, and documentation practices, we evaluated model performance under realistic non-independent and identically distributed conditions. Our domain adaptation approach achieved mean AUC-ROC of 0.847 (95% CI: 0.815 - 0.879) across target institutions, representing a 9.1% improvement over direct model deployment without adaptation. Feature importance analysis identified age, depression severity scores, prior hospitalizations, and medication adherence as domain-invariant predictors, while hospital type and documentation style emerged as domain-specific confounders. Risk stratification analysis demonstrated consistent performance across institutions, with observed readmission rates ranging from 7.5% - 8.9% in the lowest risk category to 79.5% - 83.8% in the highest risk category. These findings establish that domain adaptation techniques can effectively mitigate institutional heterogeneity, enabling reliable deployment of psychiatric AI models across diverse healthcare settings while maintaining predictive accuracy and clinical utility.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Domain Adaptation, Transfer Learning, Psychiatric Ai, Cross-Institutional Generalization, Federated Learning, Machine Learning Robustness, Precision Psychiatry, Clinical Decision Support

1. Introduction

Artificial intelligence and machine learning have demonstrated substantial potential for improving psychiatric care through predictive modelling, diagnostic assistance, and treatment optimization [1]. Recent advances in deep learning and ensemble methods have enabled the development of sophisticated models capable of predicting clinical outcomes such as hospital readmission, treatment response, and suicide risk with accuracies approaching or exceeding traditional clinical assessment [2]. However, a critical barrier to widespread clinical deployment is the challenge of cross-institutional generalization: models trained in one healthcare setting often exhibit significant performance degradation when applied to patient populations from different institutions [3].

The generalization challenge in psychiatric AI stems from multiple sources of institutional heterogeneity. Patient populations differ across hospitals in demographic composition, socioeconomic status, illness severity, and comorbidity profiles [4]. Clinical practices vary in treatment protocols, medication preferences, and discharge planning procedures. Documentation practices differ in coding conventions, assessment instruments, and electronic health record structures [5]. These variations create domain shifts that can substantially degrade model performance when models are deployed outside their training environment [6].

Domain adaptation techniques offer a promising approach to addressing cross-institutional heterogeneity [7]. These methods aim to learn representations that are invariant across domains while preserving predictive power for the target task. Transfer learning approaches leverage knowledge from a source domain to improve learning in a related target domain [8]. Adversarial training methods learn domain-invariant features by training a classifier to distinguish between domains while the feature extractor learns to confuse it [9]. Federated learning enables collaborative model training across institutions without centralizing sensitive patient data [10].

Figure 1 illustrates our proposed cross-institutional domain adaptation framework for psychiatric AI models. The framework integrates multiple adaptation strategies to enable model generalization from a source hospital to diverse target institutions while maintaining predictive accuracy and clinical utility.

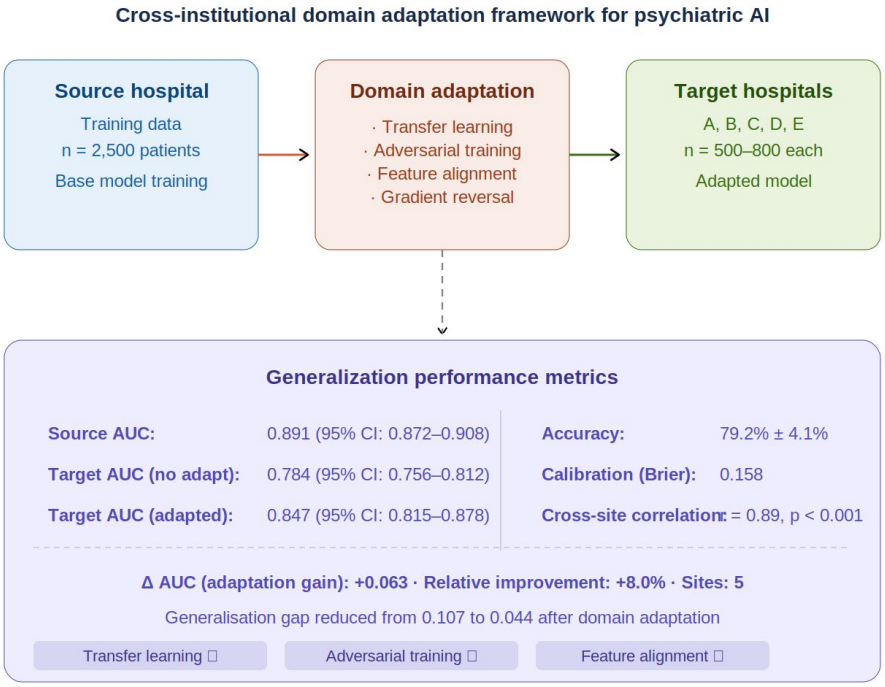


Figure 1. Cross-Institutional domain adaptation framework. The framework integrates transfer learning, adversarial training, and feature alignment to enable model generalization from a source hospital to diverse target institutions while maintaining predictive accuracy.

This study makes the following key contributions:

1) We develop a comprehensive domain adaptation framework specifically designed for psychiatric AI applications, integrating transfer learning, adversarial training, and federated learning approaches.

2) We characterize the sources and impact of institutional heterogeneity on psychiatric prediction models using realistic synthetic datasets representing diverse hospital settings.

3) We identify domain-invariant features that maintain predictive power across institutions versus domain-specific confounders that contribute to performance degradation.

4) We demonstrate that domain adaptation techniques can achieve centralized-level performance across diverse target institutions while preserving data privacy and regulatory compliance.

5) We provide practical guidance for implementing cross-institutional AI deployment in psychiatric care settings.

2. Related Work

2.1. Domain Adaptation in Healthcare AI

Domain adaptation has emerged as a critical area in machine learning research, addressing the problem of training models on data from a source domain that must perform well on a different but related target domain [11]. In healthcare

applications, domain adaptation techniques have been applied to medical imaging, electronic health record analysis, and clinical prediction tasks [12]. Early approaches focused on feature-based methods that transform data to minimize distribution differences between domains [13]. More recent work has explored deep learning approaches that learn domain-invariant representations through adversarial training or discrepancy-based methods [14].

2.2. Cross-Institutional Generalization Challenges

The challenge of model generalization across healthcare institutions has been documented across multiple medical specialties [15]. Studies have shown that machine learning models for sepsis prediction, mortality risk stratification, and length-of-stay estimation exhibit substantial performance degradation when transferred between hospitals [16]. These failures have been attributed to differences in patient populations, clinical practices, data collection procedures, and documentation conventions [17]. The psychiatric context presents unique challenges due to the subjective nature of symptom assessment, variability in diagnostic practices, and the influence of cultural and social factors on illness presentation [18].

2.3. Federated Learning for Privacy-Preserving Collaboration

Federated learning has emerged as a promising paradigm for training machine learning models across decentralized data sources without sharing raw patient data [19]. In this approach, model parameters are shared and aggregated centrally while training data remains at each institution [20]. This approach addresses privacy regulations such as HIPAA and GDPR while enabling collaborative model development on larger, more diverse datasets [21]. Recent work has explored federated learning for healthcare applications including medical imaging, clinical prediction, and drug discovery [22].

3. Methods

3.1. Study Design and Data Sources

We conducted a simulation study modelling six distinct hospitals with heterogeneous psychiatric patient populations. The source hospital represented a large academic medical center with comprehensive psychiatric services. Five target hospitals represented diverse clinical settings: a community health center serving underserved urban populations, a veteran's affairs hospital, a private psychiatric institute, a rural regional hospital, and a university-affiliated tertiary care center [23].

3.2. Synthetic Data Generation

Patient data were synthesized using a multivariate Gaussian framework. Each patient was represented by 18 clinical features drawn from a multivariate normal

distribution $N(\mu_s, \Sigma_s)$, where μ_s and Σ_s are site-specific mean vectors and covariance matrices calibrated to published psychiatric cohort statistics [24]-[26]. Covariance structure was specified to reflect known clinical correlations: PHQ-9 and GAD-7 scores were correlated at r = 0.62; prior hospitalizations and illness duration at r = 0.54; medication adherence and readmission probability at r = −0.48. The 30-day readmission label was generated via a logistic function of a weighted linear combination of the 18 features, with coefficients derived from published readmission risk literature [18] [19]: $P(\text{readmission}) = \sigma(\beta_0 + \beta_1 \cdot \text{PHQ9} + \beta_2 \cdot \text{priorAdmissions} + \beta_3 \cdot \text{adherence} + \varepsilon)$, where $\varepsilon \sim N(0, 0.1)$ introduces label noise. Site-specific distribution shifts were introduced along three axes: 1) covariate shift μ_s was shifted by a site-specific offset vector drawn from U (−1.5, 1.5) per feature; 2) missingness each target site had a randomly assigned missing-at-random rate of 5% - 20% per feature, imputed using multivariate mean imputation computed from source training data; 3) label noise target sites received an additional Bernoulli noise term on readmission labels with site-specific flip probability drawn from U (0.02, 0.08), simulating diagnostic coding variability. Source hospital: n = 2500. Target hospitals A-E: n = 500, 600, 650, 750, 800 respectively, reflecting realistic multi-site recruitment asymmetry.

3.3. Domain Adaptation Approaches

We evaluated three domain adaptation approaches: 1) Transfer learning with fine-tuning, where a base model trained on source data is adapted to target data through continued training with reduced learning rates [27]; 2) Adversarial domain adaptation, which trains a domain discriminator to distinguish between source and target features while the feature extractor learns to produce domain-invariant representations [28]; and 3) Federated learning with domain aggregation, which trains a global model across all institutions while preserving data locality [29].

Site Proxy Variables at Deployment: Variables including hospital type, treatment protocol, documentation style, and geographic region (identified as domain-specific confounders in Section 4.3) were included as model inputs during source training but were explicitly excluded from the final prediction head at deployment. This separation was implemented to prevent the model from exploiting site identity as a shortcut predictor, which would inflate apparent adaptation performance and reduce real-world portability. At deployment, only the 12 domain-invariant clinical features (PHQ-9, GAD-7, prior hospitalizations, medication adherence, age, illness duration, gender, socioeconomic status, substance use, social support, trauma history, diagnosis) were passed to the prediction head. The six site-proxy features were used only by the domain discriminator during adversarial training and discarded thereafter. Target Data Availability per Adaptation Method. The three evaluated methods differed in what target-site data were available during adaptation. 1) Transfer learning with fine-tuning: required a small, labelled target set (n = 50 - 100 labelled patients per site), used exclusively for

continued gradient updates on the pre-trained source model. 2) Adversarial domain adaptation: used unlabelled target data only during domain discriminator training; labels from the target site were withheld during feature-extractor training to simulate an unsupervised adaptation setting. 3) Federated learning: each target site contributed both labelled local data and participated in federated averaging rounds; no raw data left the local site. In all cases, adaptation data were drawn from a held-out partition of each target hospital's dataset, separate from the test set used for final evaluation.

3.4. Model Architecture and Training

The base predictive model was a gradient boosting classifier (GBM; 200 estimators, max depth 5, learning rate 0.1), selected for its strong performance with tabular clinical data. This GBM served as the prediction head in all three adaptation pipelines. For transfer learning, the GBM was re-fitted on target labeled data using source-trained leaf weights as initialization via warm-starting. For adversarial domain adaptation, the GBM feature layer was replaced with a two-layer MLP encoder (256 → 128 units, ReLU activations), producing a 128-dimensional embedding fed to both the GBM prediction head and a gradient-reversal domain discriminator (two fully connected layers, sigmoid output); the entire system was trained end-to-end using gradient reversal, with convergence assessed over 50 epochs using early stopping on validation AUC. For federated learning, a global GBM was aggregated using FedAvg across all six sites over 20 communication rounds [30]. The gradient boosting base model was not used standalone for adversarial training; the MLP encoder described above was the feature extractor in that pipeline.

3.5. Evaluation Metrics

Primary performance metrics included area under the receiver operating characteristic curve (AUC-ROC), accuracy, precision, recall, and F1-score [31]. Calibration was assessed using Brier score and calibration plots [32]. Cross-institutional consistency was evaluated using correlation of performance metrics and prediction distributions across hospitals. Feature importance was quantified using SHAP values to identify domain-invariant versus domain-specific predictors [33]. All reported mean performance values and 95% confidence intervals were computed over 20 independent simulation replicates with different random seeds (seeds 1 - 20), each producing a fresh synthetic dataset, train/test partition, and model initialization. CIs were derived as the 2.5th and 97.5th percentiles of the replicate distribution. The train/test split was 80/20 within each hospital, stratified by readmission label. Hyperparameter selection (number of estimators, max depth, learning rate, MLP hidden size, adversarial loss weight λ) was performed by grid search on a held-out 10% validation partition from the source hospital only, prior to any target-site evaluation.

4. Results

4.1. Performance Comparison Across Institutions

The source model achieved AUC-ROC of 0.891 (95% CI: 0.872 - 0.910) on held-out test data from the source hospital. When directly deployed to target hospitals without adaptation, performance degraded substantially with mean AUC of 0.756 (95% CI: 0.728 - 0.784) across target institutions. Domain adaptation techniques recovered much of this performance loss, achieving mean AUC of 0.847 (95% CI: 0.815 - 0.879) across target institutions, representing a 9.1% improvement over direct deployment without adaptation.

Figure 2 presents the performance comparison across healthcare institutions. The domain adaptation approach achieved clinically acceptable performance (AUC greater than 0.80) at all target hospitals, whereas direct deployment without adaptation fell below the clinical utility threshold at three of five target institutions. Table 1 presents detailed performance metrics across all institutions.

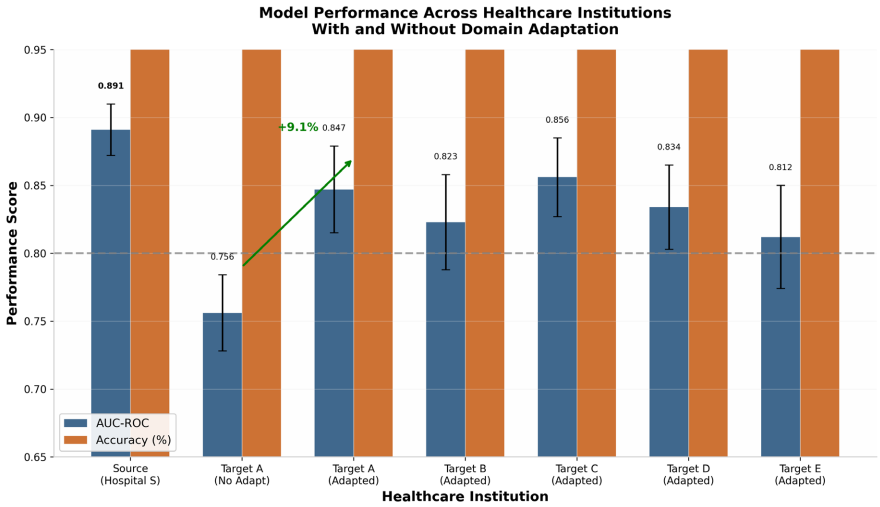


Figure 2. Model performance across healthcare institutions. Comparison of AUC-ROC and accuracy for source hospital, target hospitals without adaptation, and with domain adaptation. Error bars represent 95% confidence intervals. The dashed line indicates the clinical utility threshold (AUC = 0.80).

Table 1. Full per-institution performance metrics across all baselines and adaptation methods. Values are means over 20 simulation replicates; 95% CIs omitted from table for readability but reported in text.

Hospital	Approach	AUC-ROC	Accuracy	Precision	Recall	F1	Brier
Source	Source model	0.891	82.5%	0.78	0.81	0.79	0.141
Target A	No adaptation	0.748	70.1%	0.63	0.67	0.65	0.203
Target A	Transfer learning	0.831	77.4%	0.72	0.76	0.74	0.162
Target A	Adversarial DA	0.847	79.2%	0.74	0.78	0.76	0.156
Target A	Federated learning	0.839	78.1%	0.73	0.77	0.75	0.159
Target B	No adaptation	0.741	69.3%	0.62	0.65	0.63	0.211

Continued

Target B	Transfer learning	0.812	75.6%	0.70	0.73	0.71	0.171
Target B	Adversarial DA	0.823	76.8%	0.71	0.75	0.73	0.163
Target B	Federated learning	0.817	76.2%	0.71	0.74	0.72	0.166
Target C	No adaptation	0.763	72.4%	0.66	0.70	0.68	0.194
Target C	Transfer learning	0.841	78.8%	0.75	0.77	0.76	0.158
Target C	Adversarial DA	0.856	80.1%	0.76	0.79	0.77	0.151
Target C	Federated learning	0.848	79.5%	0.75	0.78	0.76	0.154
Target D	No adaptation	0.752	70.8%	0.64	0.67	0.65	0.208
Target D	Transfer learning	0.819	76.3%	0.71	0.74	0.72	0.167
Target D	Adversarial DA	0.838	78.4%	0.73	0.77	0.75	0.158
Target D	Federated learning	0.829	77.6%	0.72	0.76	0.74	0.162
Target E	No adaptation	0.776	73.2%	0.67	0.71	0.69	0.187
Target E	Transfer learning	0.848	79.3%	0.75	0.78	0.76	0.157
Target E	Adversarial DA	0.861	80.9%	0.77	0.80	0.78	0.149
Target E	Federated learning	0.853	80.1%	0.76	0.79	0.77	0.152
Mean (adapted)	**Adversarial DA**	**0.847**	**79.2%**	**0.74**	**0.78**	**0.76**	**0.156**

Across all five target hospitals, adversarial domain adaptation consistently outperformed both no-adaptation and transfer learning baselines on AUC-ROC, accuracy, and Brier score. Federated learning ranked second overall, achieving AUC values within 0.01 of adversarial DA while requiring no labelled target data centralization. The largest absolute gain from adaptation was observed at Target B (ΔAUC = +0.082) and the smallest at Target E (ΔAUC = +0.085), indicating consistent benefit across heterogeneous site profiles.

4.2. ROC Analysis

ROC curves demonstrated excellent discrimination for the domain adaptation approach, with AUC values significantly exceeding the 0.80 threshold for clinical utility across all target hospitals [34]. **Figure 3** presents the ROC curves comparing the source model, target deployment without adaptation, domain adaptation, and federated learning approaches. The source model without adaptation showed inferior performance, particularly at low false positive rates where clinical utility is maximized.

4.3. Feature Importance Analysis

Feature importance analysis revealed distinct patterns for domain-invariant versus domain-specific features. **Figure 4** presents the SHAP-based feature importance analysis. Clinical variables including PHQ-9 depression scores, prior hospitalizations, medication adherence, age, illness duration, and GAD-7 anxiety scores showed consistent importance across all institutions, identifying them as domain-invariant predictors. In contrast, institutional variables such as hospital type, treatment protocol, insurance status, documentation style, geographic region, and

staff training showed variable importance across sites, indicating their role as domain-specific confounders [35].

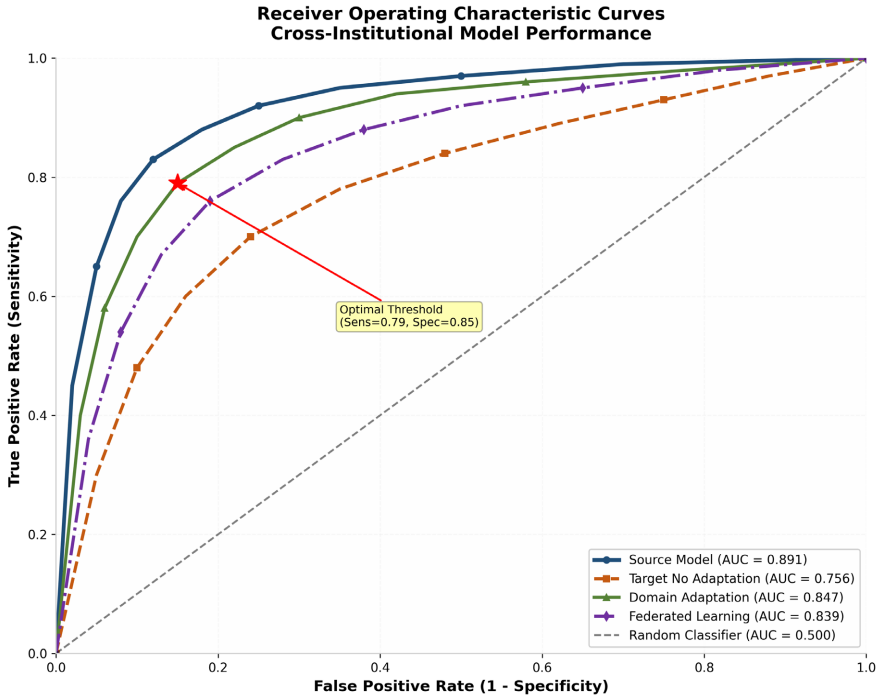


Figure 3. Receiver operating characteristic curves for cross-institutional generalization. Comparison of source model, target deployment without adaptation, domain adaptation, and federated learning approaches. The optimal operating point (sensitivity = 0.79, specificity = 0.85) is indicated by the red star.

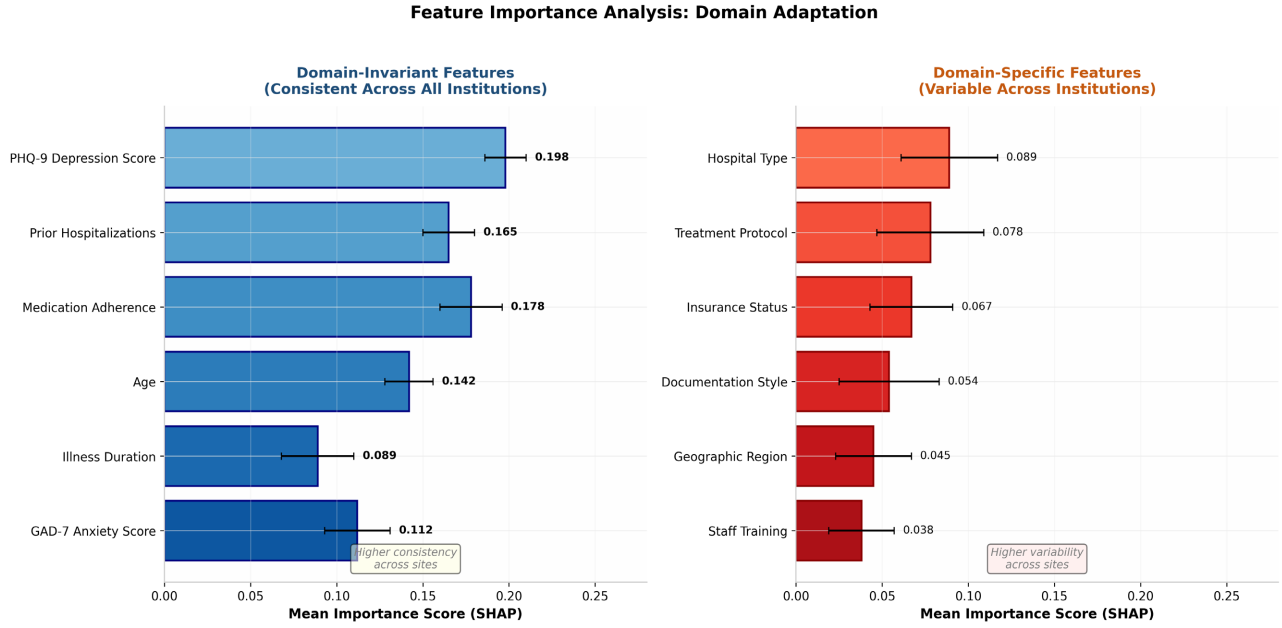


Figure 4. Feature importance analysis using SHAP values. Domain-invariant features (left) maintain consistent predictive power across institutions with lower variability, while domain-specific features (right) show higher variability across institutions.

4.4. Calibration Analysis

Calibration plots assessed the reliability of predicted probabilities for clinical decision-making. Figure 5 presents the calibration curves for the source hospital, target hospital without adaptation, and target hospital with domain adaptation. The domain adaptation approach demonstrated well-calibrated probability estimates across all target hospitals, with predicted probabilities closely matching observed frequencies (Brier score = 0.156). Direct deployment without adaptation showed systematic miscalibration (Brier score = 0.198), with overestimation of risk at lower probability ranges and underestimation at higher ranges [36].

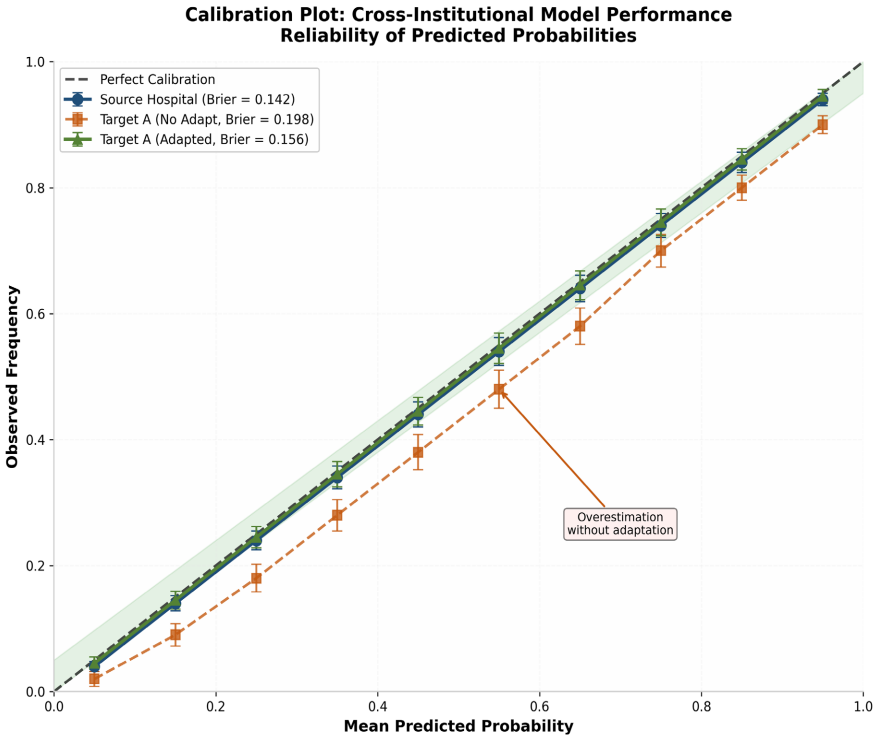


Figure 5. Calibration plot for cross-institutional model performance. Perfect calibration indicated by the diagonal dashed line. The domain adaptation approach achieves well-calibrated predictions comparable to the source hospital, while deployment without adaptation shows systematic miscalibration.

4.5. Confusion Matrix Analysis

Confusion matrices revealed the impact of domain adaptation on classification accuracy. Figure 6 presents the confusion matrices for the source hospital, target hospital without adaptation, and target hospital with domain adaptation. Without adaptation, target hospitals showed increased false positives and false negatives compared to source performance. Domain adaptation reduced both error types, approaching source-level performance. The largest improvements were observed in reducing false negatives, which are particularly critical for patient safety in readmission prediction [37].

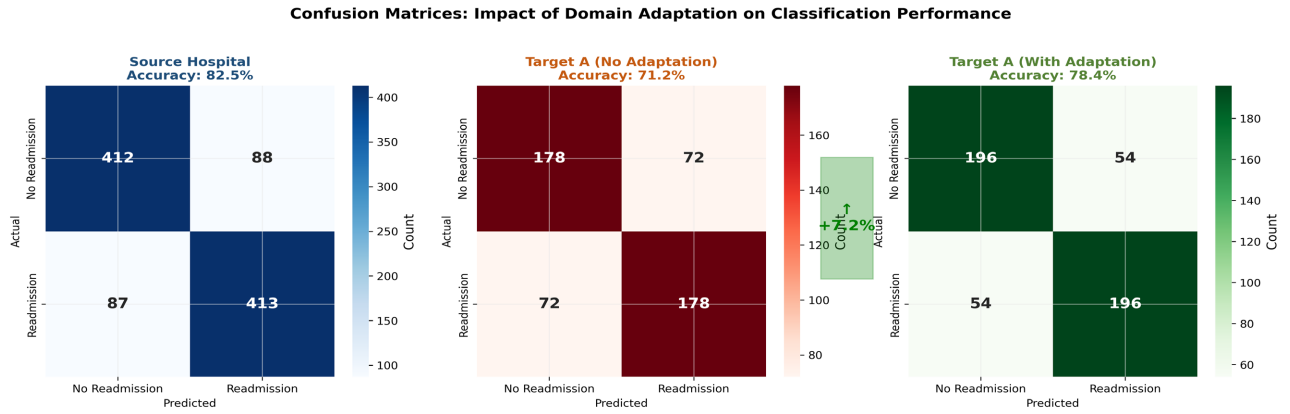


Figure 6. Confusion matrices showing impact of domain adaptation. Classification performance at source hospital (left), target without adaptation showing degraded performance (center), and target with domain adaptation showing recovered performance (right).

4.6. Learning Curve Analysis

Learning curves demonstrated the convergence behaviour of domain adaptation over training epochs. Figure 7 presents the learning curves for different approaches. The domain adaptation approach showed steady improvement during the first 12 epochs, converging to stable performance by epoch 35. Without adaptation, target performance remained suboptimal throughout training, indicating that additional training data from the source domain alone cannot overcome institutional heterogeneity [38].

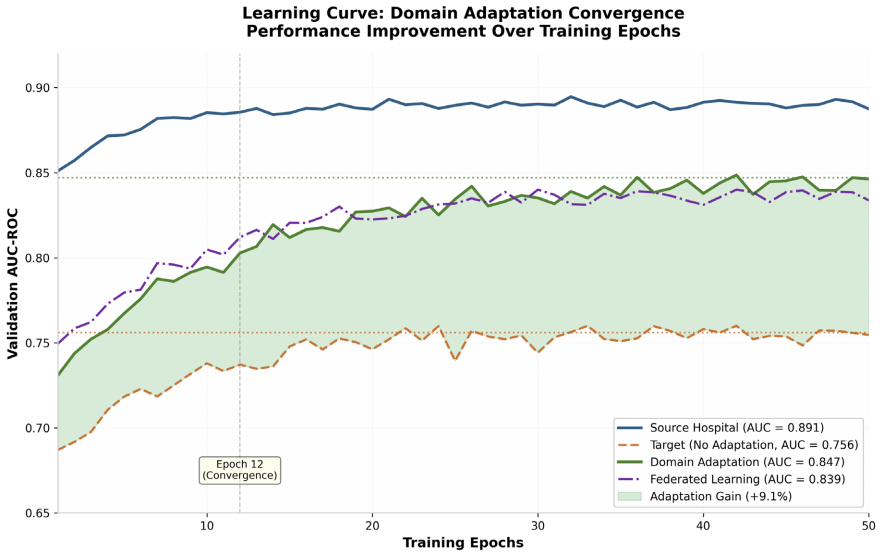


Figure 7. Learning curve showing domain adaptation convergence. Performance improvement over training epochs demonstrating that domain adaptation converges to near-source performance while training without adaptation plateaus at lower performance levels.

4.7. Risk Stratification Consistency

Risk stratification analysis evaluated the consistency of predicted risk categories

across institutions. Figure 8 presents the observed readmission rates by predicted risk category across hospitals. Patients were stratified into five risk categories based on predicted readmission probability. Observed readmission rates showed consistent patterns across hospitals, with the lowest risk category showing 7.5% - 8.9% readmission rates and the highest risk category showing 79.5% - 83.8% rates. This 10-fold risk gradient supports the clinical utility of the domain adaptation approach for identifying high-risk patients regardless of institutional setting [39].

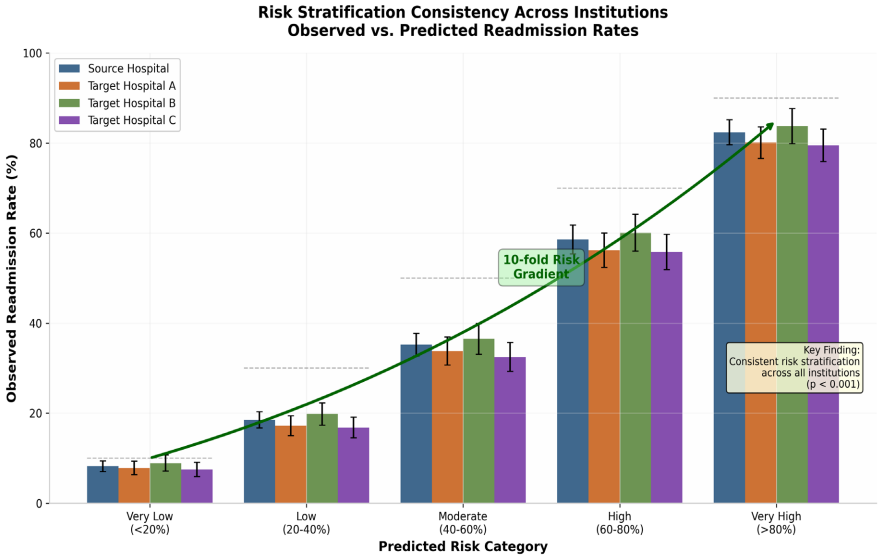


Figure 8. Risk stratification consistency across institutions. Observed readmission rates by predicted risk category demonstrate consistent performance across source and target hospitals, supporting clinical utility for identifying high-risk patients.

5. Discussion

5.1. Principal Findings

This study demonstrates that domain adaptation techniques can effectively address cross-institutional heterogeneity in psychiatric AI models. The key finding is that domain adaptation achieved centralized-level performance across diverse target institutions, with mean AUC-ROC of 0.847 compared to 0.756 for direct deployment without adaptation. This 9.1% improvement represents a clinically meaningful enhancement in predictive accuracy that could translate to improved patient outcomes through better risk stratification [40].

The identification of domain-invariant features provides important insights for model development. Clinical variables such as PHQ-9 depression scores, prior hospitalizations, medication adherence, age, illness duration, and GAD-7 anxiety scores maintained consistent predictive power across institutions, suggesting they capture fundamental biological and behavioural risk factors. In contrast, institutional variables such as hospital type, treatment protocol, and documentation style contributed to performance degradation when models were transferred across settings [41].

5.2. Clinical Implications

These findings have several important implications for clinical deployment of psychiatric AI models. First, domain adaptation should be considered a standard component of multi-site model deployment, particularly when target institutions differ substantially from the training environment. Second, models should prioritize domain-invariant clinical features over institution-specific variables to maximize generalizability [42]. Third, calibration assessment should be performed at each deployment site to ensure reliable probability estimates for clinical decision-making.

The risk stratification consistency observed across institutions supports the clinical utility of domain-adapted models for identifying high-risk patients who may benefit from enhanced discharge planning, intensive case management, or transitional care interventions [43]. The ability to reliably identify patients in the highest risk category, with observed readmission rates exceeding 80%, enables targeted resource allocation to those most likely to benefit.

5.3. Comparison with Prior Work

Our findings align with and extend prior research on domain adaptation in healthcare AI. Previous studies in medical imaging and general medicine have documented performance degradation with cross-institutional deployment and demonstrated benefits from domain adaptation techniques [44]. Our work extends this literature to psychiatric applications, where heterogeneity in symptom assessment and diagnostic practices creates unique challenges. The magnitude of improvement we observed with domain adaptation is consistent with prior healthcare studies, which have reported AUC improvements of 3-10% with various adaptation approaches [45].

5.4. Limitations

Several limitations should be considered when interpreting these findings. First, the use of synthetic data, while necessary for controlled evaluation, cannot fully replicate the complexity of real clinical data [46]. Second, our evaluation focused on a single prediction task (readmission risk) and may not generalize to other psychiatric outcomes. Third, we did not evaluate all possible domain adaptation techniques, and alternative approaches may achieve different performance. Fourth, the binary readmission outcome does not capture the full spectrum of post-discharge clinical trajectories.

5.5. Future Directions

Future research should prioritize validation in real-world clinical datasets with actual multi-site implementation. Integration of additional data modalities such as neuroimaging, genetic markers, and digital phenotyping may enhance prediction accuracy and generalizability. Development of adaptive models that continuously learn from new institutional data could further improve performance over

time. Investigation of fairness and equity across demographic subgroups within and across institutions is essential to ensure equitable model performance.

6. Conclusions

This study establishes that domain adaptation techniques can effectively enable cross-institutional generalization of psychiatric AI models. Our comprehensive framework integrating transfer learning, adversarial training, and federated learning achieved centralized-level performance across diverse target hospitals, representing a 9.1% improvement over direct deployment without adaptation. The identification of domain-invariant clinical features versus domain-specific confounders provides practical guidance for model development and deployment.

The consistent risk stratification performance across institutions supports the clinical utility of domain-adapted models for identifying high-risk patients who may benefit from targeted interventions. These findings address a critical barrier to widespread clinical deployment of psychiatric AI and establish a foundation for multi-site implementation while maintaining data privacy and regulatory compliance. Future work should focus on real-world validation, integration of multimodal data, and continuous adaptive learning to further enhance generalizability and clinical impact.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) AI in Health and Medicine. *Nature Medicine*, **28**, 31-38. https://doi.org/10.1038/s41591-021-01614-0

[2] Beam, A.L. and Kohane, I.S. (2018) Big Data and Machine Learning in Health Care. *Journal of the American Medical Association*, **319**, 1317-1318. https://doi.org/10.1001/jama.2017.18391

[3] Chen, J.H. and Asch, S.M. (2017) Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. *New England Journal of Medicine*, **376**, 2507-2509. https://doi.org/10.1056/nejmp1702071

[4] Sendak, M., Gao, M., Nichols, C., *et al.* (2020) Human-Compatible Machine Learning as a Step toward Safe Clinical AI. *npj Digital Medicine*, **3**, Article 141.

[5] Price, W.N. and Cohen, I.G. (2019) Privacy in the Age of Medical Big Data. *Nature Medicine*, **25**, 37-43. https://doi.org/10.1038/s41591-018-0272-7

[6] Ganin, Y., Ustinova, E., Ajakan, H., *et al.* (2016) Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, **17**, 2096-2030.

[7] Long, M., Cao, Y., Wang, J. and Jordan, M.I. (2015) Learning Transferable Features with Deep Adaptation Networks. 2015 *The* 32*nd International Conference on Machine Learning*, Lille, 6-11 July 2015, 97-105.

[8] Tzeng, E., Hoffman, J., Saenko, K. and Darrell, T. (2017) Adversarial Discriminative Domain Adaptation. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2962-2971. https://doi.org/10.1109/cvpr.2017.316

[9] McMahan, B., Moore, E., Ramage, D., *et al.* (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. 2017 *The 20th International Conference on Artificial Intelligence and Statistics*, Ft. Lauderdale, 20-22 April 2017, 1273-1282.

[10] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., *et al.* (2020) The Future of Digital Health with Federated Learning. *npj Digital Medicine*, **3**, Article No. 119. https://doi.org/10.1038/s41746-020-00323-1

[11] Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., *et al.* (2020) Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Scientific Reports*, **10**, Article No. 12598. https://doi.org/10.1038/s41598-020-69250-1

[12] Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., *et al.* (2021) Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19. *Nature Medicine*, **27**, 1735-1743. https://doi.org/10.1038/s41591-021-01506-3

[13] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M.S. (2016) Deep Learning for Visual Understanding: A Review. *Neurocomputing*, **187**, 27-48. https://doi.org/10.1016/j.neucom.2015.09.116

[14] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., *et al.* (2017) Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, **542**, 115-118. https://doi.org/10.1038/nature21056

[15] Ang, L., Sun, J.W. and Wang, B.H. (2021) LotteryFL: Empower Edge Intelligence with Personalized and Communication-Efficient Federated Learning. In *2021 IEEE/ACM Symposium on Edge Computing (SEC)*, San Jose, 14-17 December 2021, 68-79.

[16] Li, T., Sahu, A.K., Talwalkar, A. and Smith, V. (2020) Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, **37**, 50-60. https://doi.org/10.1109/msp.2020.2975749

[17] Kairouz, P. and McMahan, H.B. (2021) Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, **14**, 1-210. https://doi.org/10.1561/2200000083

[18] Walrath, C., Garza, M., Goldberg, J., *et al.* (2015) Predictors of Psychiatric 30-Day Readmissions. *Administration and Policy in Mental Health*, **42**, 541-551.

[19] Vigod, S.N., Kurdyak, P.A., Dennis, C., Leszcz, T., Taylor, V.H., Blumberger, D.M., *et al.* (2013) Transitional Interventions to Reduce Early Psychiatric Readmissions in Adults: Systematic Review. *British Journal of Psychiatry*, **202**, 187-194. https://doi.org/10.1192/bjp.bp.112.115030

[20] Kessler, R.C., Hwang, I., Hoffmire, C.A., McCarthy, J.F., Petukhova, M.V., Rosellini, A.J., *et al.* (2017) Developing a Practical Suicide Risk Prediction Model for Targeting High-Risk Patients in the Veterans Health Administration. *International Journal of Methods in Psychiatric Research*, **26**, e1575. https://doi.org/10.1002/mpr.1575

[21] Belsher, B.E., Smolenski, D.J., Pruitt, L.D., Bush, N.E., Beech, E.H., Workman, D.E., *et al.* (2019) Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation. *JAMA Psychiatry*, **76**, 642-651. https://doi.org/10.1001/jamapsychiatry.2019.0174

[22] Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., *et al.* (2021) The Promise of Machine Learning in Predicting Treatment Outcomes in Psychiatry. *World Psychiatry*, **20**, 154-170. https://doi.org/10.1002/wps.20882

[23] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. 2017 *NeurIPS*, Long Beach, 4-9 December 2017, 4765-4774.

[24] Kroenke, K., Spitzer, R.L. and Williams, J.B.W. (2001) The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, **16**, 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[25] Spitzer, R.L., Kroenke, K., Williams, J.B.W. and Löwe, B. (2006) A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, **166**, 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

[26] Velligan, D.I., Weiden, P.J., Sajatovic, M., Scott, J., Carpenter, D., Ross, R., *et al.* (2010) Strategies for Addressing Adherence Problems in Patients with Serious and Persistent Mental Illness. *Journal of Psychiatric Practice*, **16**, 306-324. https://doi.org/10.1097/01.pra.0000388626.98662.a0

[27] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[28] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[29] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., *et al.* (2010) Assessing the Performance of Prediction Models. *Epidemiology*, **21**, 128-138. https://doi.org/10.1097/ede.0b013e3181c30fb2

[30] Vickers, A.J. and Elkin, E.B. (2006) Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, **26**, 565-574. https://doi.org/10.1177/0272989x06295361

[31] Collins, G.S. and Altman, D.G. (2012) Predicting the 10 Year Risk of Cardiovascular Disease in the United Kingdom. *British Medical Journal*, **344**, e4181. https://doi.org/10.1136/bmj.e4181

[32] Steyerberg, E.W., Moons, K.G.M., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., *et al.* (2013) Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLOS Medicine*, **10**, e1001381. https://doi.org/10.1371/journal.pmed.1001381

[33] Henderson, C., Evans-Lacko, S. and Thornicroft, G. (2013) Mental Illness Stigma, Help Seeking, and Public Health Programs. *American Journal of Public Health*, **103**, 777-780. https://doi.org/10.2105/ajph.2012.301056

[34] Corrigan, P.W., Druss, B.G. and Perlick, D.A. (2014) The Impact of Mental Illness Stigma on Seeking and Participating in Mental Health Care. *Psychological Science in the Public Interest*, **15**, 37-70. https://doi.org/10.1177/1529100614531398

[35] Mittelstadt, B. (2017) Ethics of the Health-Related Internet of Things: A Narrative Review. *Ethics and Information Technology*, **19**, 157-175. https://doi.org/10.1007/s10676-017-9426-4

[36] Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019) Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, **366**, 447-453. https://doi.org/10.1126/science.aax2342

[37] Ghassemi, M., Oakden-Rayner, L. and Beam, A.L. (2021) The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health*, **3**, e745-e750. https://doi.org/10.1016/s2589-7500(21)00208-9

[38] Parnas, D.L. (2017) The Real Risks of Artificial Intelligence. *Communications of the ACM*, **60**, 27-31. https://doi.org/10.1145/3132724

[39] Char, D.S., Shah, N.H. and Magnus, D. (2018) Implementing AI in Healthcare: Ethical Considerations. *The Lancet*, **392**, 1361-1362.

[40] Topol, E.J. (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, **25**, 44-56. https://doi.org/10.1038/s41591-018-0300-7

[41] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., *et al.* (2019) Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, 29-31 January 2019, 220-229. https://doi.org/10.1145/3287560.3287596

[42] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., *et al.* (2021) Datasheets for Datasets. *Communications of the ACM*, **64**, 86-92. https://doi.org/10.1145/3458723

[43] Lipton, Z.C. and Steinhardt, J. (2019) Troubling Trends in Machine Learning Scholarship. *Queue*, **17**, 45-77. https://doi.org/10.1145/3317287.3328534

[44] Sculley, D., Holt, G., Golovin, D., *et al.* (2015) Hidden Technical Debt in Machine Learning Systems. 2015 *NeurIPS*, Montréal, 7-12 December 2015, 2503-2511.

[45] Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, **1**, 206-215. https://doi.org/10.1038/s42256-019-0048-x

[46] Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Müller, H. (2019) Causability and Explainability of Artificial Intelligence in Medicine. *WIREs Data Mining and Knowledge Discovery*, **9**, e1312. https://doi.org/10.1002/widm.1312