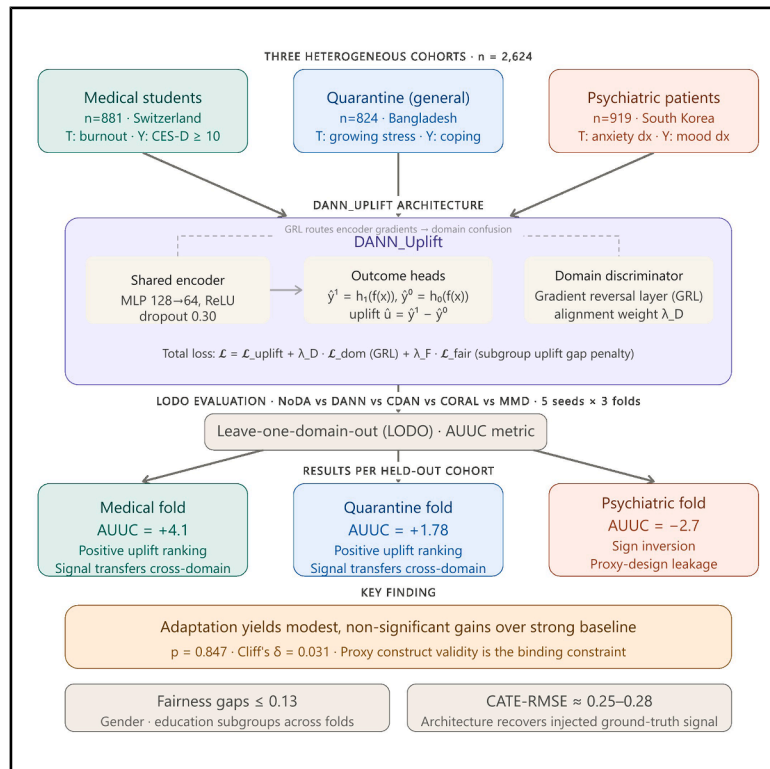


Domain adaptive uplift modeling across heterogeneous mental health cohorts

Graphical abstract



Authors

Abdullah Nayem Wasi Emran,
A. B. M. Alim Al Islam

Correspondence

abdullah.nayem@bracu.ac.bd (A.N.W.E.),
alim_razi@cse.buet.ac.bd (A.B.M.A.A.I.)

In brief

Neuroscience; Machine learning;
Psychology

Highlights

- Domain-adaptive uplift modeling transfers stress-risk rankings across cohorts
- No domain-adaptation method significantly outperforms a strong shared baseline
- Psychiatric fold failure traced to structural proxy-design leakage empirically
- Sigmoid adversarial weight schedules improve stability over fixed settings



Article

Domain adaptive uplift modeling across heterogeneous mental health cohorts

Abdullah Nayem Wasi Emran^{1,2,3,*} and A. B. M. Alim Al Islam^{1,*}¹Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh²Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh³Lead contact*Correspondence: abdullah.nayem@bracu.ac.bd (A.N.W.E.), alim_razi@cse.buet.ac.bd (A.B.M.A.I.)<https://doi.org/10.1016/j.isci.2026.116027>

SUMMARY

Predicting who will deteriorate under stress is important for targeting mental-health support; yet, treatment-effect models are rarely tested across populations. We evaluate a domain-adaptive neural uplift model on three heterogeneous cohorts—medical students, members of the general public under quarantine, and psychiatric patients ($n = 2,624$). The model combines a shared encoder, two potential-outcome heads, a domain discriminator, and an optional fairness penalty. We compare no-adaptation training with four domain-adaptation mechanisms under a leave-one-domain-out protocol, using AUUC and a semi-synthetic benchmark with known treatment effects. The model achieves positive uplift ranking in two cohorts, while the psychiatric cohort shows sign inversion of the effect proxy. Adaptation yields modest, tuning-sensitive gains over a strong baseline. These results clarify when domain adaptation helps treatment-effect ranking under distribution shift and inform cautious deployment across cohorts.

INTRODUCTION

Mental-health deterioration under stress is clinically heterogeneous. Individuals exposed to similar stressors do not worsen to the same degree, and this variation reflects differences in vulnerability, prior symptoms, coping resources, and social context.^{1,2} This heterogeneity matters in practice because preventive mental-health resources are limited: clinicians and public-health systems often need to decide whom to monitor more closely, whom to refer, and who may benefit from low-intensity psychosocial or stress-management support.³ The deployment of machine learning in mental health holds immense promise for personalized care by enabling exactly this kind of individualized prioritization—particularly through uplift modeling, which seeks to identify which individuals show the greatest differential vulnerability under specific exposure conditions. In psychiatry and occupational health this capacity is foundational to the logic of *selective prevention*: the principle that limited clinical resources are most effectively directed toward individuals whose mental-health trajectory is genuinely modifiable by an available intervention.^{4,5}

This clinical need motivates individualized prediction, but it also requires conceptual care. In the present study, we do not observe a single randomized psychiatric intervention shared across all datasets. Rather, we study a cross-dataset ranking problem using dataset-specific binary exposure or treatment proxies together with binary deterioration proxies. In the medical-student dataset, the exposure proxy is high burnout exhaustion (MBI exhaustion score at or above the 75th percentile⁶), and deterioration is operationalized as a CES-D score of ten or

above.⁷ In the quarantine dataset, the exposure proxy is heightened self-reported perceived stress (binarized from yes/maybe vs. no), and deterioration is operationalized as self-reported coping failure—an established consequence of enforced social isolation.^{8,9} In the psychiatric dataset, the exposure proxy is an anxiety-spectrum diagnosis, and the adverse outcome is a comorbid mood disorder (MDD) or major depressive disorder.¹⁰ Accordingly, throughout this study, “treatment” refers to the dataset-specific exposure/intervention proxy and “deterioration under stress” refers to worsening under that proxy relative to its comparison condition. This framing supports a hypothesis about prioritized screening or support allocation; it should not be read as a direct claim that the estimated effects are already ready for clinical treatment recommendation.

This distinction is important because mental-health prediction models often fail to generalize cleanly across populations. Differences in measurement instruments, sampling frames, base rates, and contextual mechanisms can substantially change what a model learns and what its output means in a new setting.^{11,12} The same concern is relevant for our problem setting, where the three study populations differ not only in demographics and context, but also in how stress-related exposure and deterioration are operationalized.¹³ Furthermore, ensuring that these models remain fair across diverse demographic groups is essential to prevent the perpetuation of health disparities.¹⁴

Within this broader challenge, uplift modeling is attractive because it asks a more targeted question than standard risk prediction: rather than estimating who is generally likely to experience a poor outcome, it asks who appears differentially



vulnerable under one condition relative to another—which is closer to the logic of targeted support and individualized intervention prioritization. Classical approaches to treatment-effect estimation—meta-learners such as the S-Learner, T-Learner, and X-Learner,¹⁵ and non-parametric methods such as causal forests¹⁶—offer flexible individual treatment-effect estimation within a single dataset. However, they do not natively support transfer across datasets with heterogeneous feature spaces and differing outcome distributions. Extending them to the multi-dataset setting requires either fitting independent models per dataset (discarding cross-population signal) or naively pooling samples with no mechanism to manage the resulting covariate shift—both unsatisfactory when the goal is generalization to a previously unseen study population. The neural architecture we adopt addresses both limitations simultaneously: a shared encoder learns representations across study populations, and the adversarial domain-confusion mechanism explicitly discourages encoding of dataset-specific artifacts that would impair out-of-dataset uplift ranking. Our no-adaptation (NoDA) variant, which disables the adversarial component, is functionally equivalent to a two-head neural T-Learner and thus serves as the direct neural counterpart to classical meta-learner baselines.

Mental-health datasets are rarely interchangeable, and this challenge is both empirical and clinical. Our recent cross-dataset mental-health study, which analyzed the same three study populations considered here from a descriptive and comparative perspective, showed that mental-health patterns vary meaningfully across medical, quarantine, and psychiatric settings.¹³ More broadly, Xu et al.'s GLOBEM benchmark demonstrated that longitudinal behavior models for depression detection often degrade markedly when deployed across datasets rather than within a single source population.¹² This broader translational concern is consistent with psychiatric prediction research more generally, where systematic review evidence shows that generalizability and clinical utility remain major bottlenecks even when predictive performance within a development dataset appears promising.¹¹ Clinically, long-standing diathesis-stress frameworks likewise suggest that stress exposure does not translate into deterioration uniformly across individuals; rather, outcomes depend on the interaction between stressors and pre-existing vulnerability, coping resources, and social context.^{1,2} A complementary strand of clinical research highlights that vulnerability is not exhausted by average risk alone: studies of depressive relapse and recurrence show that subsequent worsening depends on factors such as prior episode burden, residual symptoms, chronic stress exposure, and limited social support.¹⁷ Related work on personalized treatment selection has argued that clinically useful prediction should move beyond average effects toward estimating which patients are more likely to benefit from one strategy than another, as illustrated by the personalized advantage index framework.¹⁸ In practice, stepped-care models reflect the same underlying principle: that monitoring intensity and support level should be matched to predicted individual need rather than applied uniformly.¹⁹ Our use of uplift modeling is conceptually aligned with this individualized decision logic—rather than estimating only average risk, it seeks to rank differential vulnerability under alternative exposure conditions at the individual level, while re-

maintaining appropriately cautious about the observational and proxy-based nature of the treatment and outcome variables in our data.

Building a single model that can generalize treatment-effect rankings across diverse mental-health cohorts remains challenging. Recent work by Bica and van der Schaar²⁰ illustrates progress in this direction by proposing a heterogeneous transfer-learning framework for treatment-effect estimation across datasets. Adaptive machine-learning approaches have also shown promise for capturing longitudinal changes in depression and anxiety that static models cannot,²¹ and adaptive data-driven architectures more broadly have been argued to be necessary for addressing the dataset differences inherent in mental-health care.²² Theoretical foundations further support this direction: generalization bounds have been established for individualized causal-effect estimators and invariant representation learning has been shown to improve cross-domain causal generalization²³; neural network-based transfer learning under structured constraints can similarly aid treatment-effect estimation in patient populations not seen during training.²⁴ At the same time, cross-dataset generalization remains a well-documented challenge in mental-health applications.^{11,12} One promising family of methods for learning invariant representations are domain-adversarial neural networks (DANNs), which have shown success in removing domain bias and improving model generality across several health-related settings.^{25–27} For mental health-related EEG signals, domain adaptation has demonstrated strong cross-subject generalization: Wu et al.²⁸ achieved >90% cross-subject accuracy on the SEED dataset by aligning marginal and conditional distributions, and Ju et al.²⁹ showed that pseudo-label-based DANN narrows the performance gap between source and target subjects. In neuroimaging, Wang et al.³⁰ successfully corrected scanner/site differences across 20 sites for schizophrenia and brain-age prediction using a weighted ERM framework with small amounts of target data, illustrating that adaptation can also be effective in partially supervised settings. However, effectiveness varies across populations and shift mechanisms; for example, unsupervised domain adaptation methods did not outperform simple re-training under temporal dataset shift in a multi-hospital intensive care unit (ICU) study.³¹

Alongside robustness to distribution shift, ensuring fair performance across demographic groups is an important goal in mental-health prediction, since unequal model behavior can reinforce existing disparities in access, diagnosis, and treatment.^{14,32} Kim et al.³³ proposed domain-adversarial training to remove gender information from speech features in a PTSD/depression detector, improving both accuracy and fairness across genders. Causal fairness frameworks have similarly revealed that women and Black patients were less likely to receive cardiac therapy than equally eligible peers—a pattern analogous to disparities that can arise in mental-health model outputs.^{34,35} Cheong et al.³⁶ further showed that standard mitigation methods in depression detection reduced but did not fully eliminate gender disparities, suggesting more rigid constraints are needed during training. Applying fairness-aware ideas to uplift modeling can help ensure that differential-vulnerability rankings and downstream support prioritization are not systematically skewed

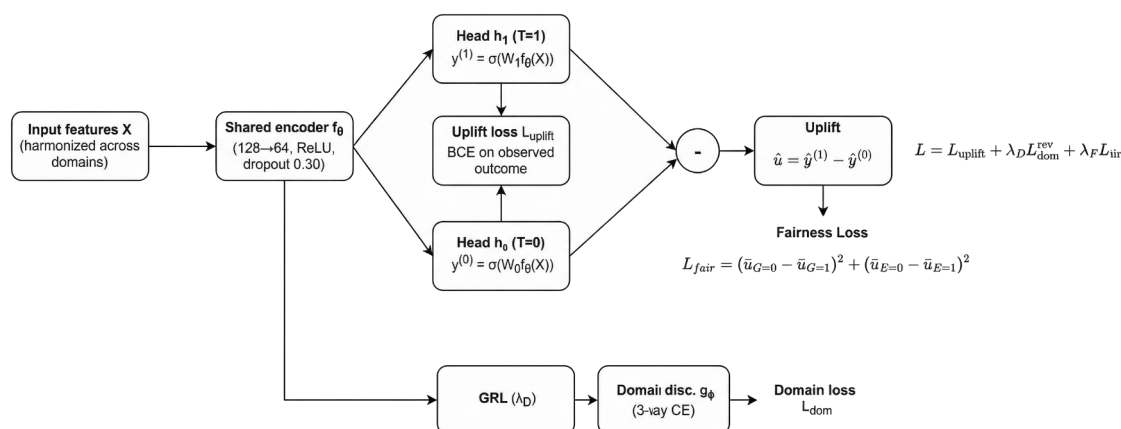


Figure 1. DANN_Uplift architecture and clinical interpretation

A shared encoder feeds two potential-outcome heads ($T = 1/T = 0$) to produce uplift $\hat{u}$. A gradient-reversal layer routes features to a domain discriminator that encourages domain-invariant representations. An optional fairness term penalizes mean uplift gaps across gender/education subgroups (Equation 10). *Clinical interpretation.* A positive $\hat{u}$ for an individual indicates that, based on the training data, this person is predicted to be more likely to exhibit the adverse outcome if exposed to the treatment proxy than if not. In a prevention context, individuals with high $\hat{u}$ are priority targets for early clinical review; $\hat{u}$ scores should be treated as flags for manual clinical assessment rather than as standalone decision tools. The LODO evaluation protocol approximates the deployment scenario in which a model trained on available study populations is applied to a new population with no labeled outcome data.

toward majority groups³⁷; this integration has seen recent progress in other domains, where causal graphs and rankability constraints have improved treatment-effect estimation in large-scale uplift settings.^{38,39} Motivated by this literature, we incorporate a simple fairness penalty and domain-adversarial machinery into our uplift model and, given that both fairness interventions and domain adaptation have shown population- and mechanism-dependent effects, we treat them as empirical hypotheses to be evaluated with diagnostics rather than as assumptions of uniform benefit.

The objective of this study is to evaluate whether a single neural uplift model, augmented with domain-adaptation (DA) and fairness-aware components, can produce more reliable treatment-effect rankings across heterogeneous mental-health datasets than a strong non-adapted baseline. Concretely, we ask three questions. First, can a single uplift model accurately rank “who will deteriorate under stress” across three fundamentally different mental-health datasets under leave-one-dataset-out evaluation, and where does this model fail? We use the term *cohort* informally throughout to refer to a distinct study population or dataset, rather than in the technical epidemiological sense of a group followed longitudinally. All three datasets used in this study are cross-sectional. Second, do DA mechanisms such as DANN, conditional adversarial (CDAN), covariance alignment (CORAL), and kernel embedding (MMD) measurably improve out-of-dataset uplift performance relative to a strong shared-encoder baseline without adaptation? Third, how sensitive is cross-dataset uplift to the adversarial alignment weight, and can subgroup uplift gaps remain small while the model still recovers known treatment effects on a semi-synthetic benchmark?

This study makes three contributions. First, we evaluate a unified neural uplift framework with a shared encoder, treatment-specific potential-outcome heads, adversarial domain

alignment, and an optional fairness penalty on subgroup uplift gaps. Second, we provide a controlled comparison of several established DA mechanisms under the same encoder and the same leave-one-dataset-out protocol, which lets us assess not only whether adaptation helps, but also where it does not; on average across seeds and folds, none of the evaluated mechanisms achieves a statistically significant uplift gain over the non-adapted baseline, a result that is informative in its own right. And third, we complement performance reporting with a tuning and diagnostic protocol based on adversarial-weight sweeps, scheduling strategies, subgroup-gap audits, and semi-synthetic validation, so that the study contributes practical guidance rather than only point estimates.

Empirically, we study cross-dataset uplift across three study populations: medical students, members of the general public under quarantine, and psychiatric patients. We compare a NoDA against several DA mechanisms under the same backbone, and we additionally examine the effect of a fairness penalty on subgroup uplift gaps. Performance is evaluated primarily with area under uplift curve (AUUC) under leave-one-dataset-out testing, alongside a semi-synthetic benchmark with known treatment effects. This design allows us to distinguish between absolute performance, transfer performance, fairness behavior, and sensitivity to the alignment hyperparameter; the full DANN_Uplift architecture is summarized in Figure 1.

RESULTS

Setup and evaluation protocol

We evaluate all models in a leave-one-domain-out (LODO) setting across three study populations: medical students (domain 0, $n = 881$), general population under quarantine (domain 1, $n = 824$), and psychiatric patients (domain 2, $n =$

Table 1. Summary of the three study populations

Study population	<i>n</i>	Prev(<i>T</i>)	Prev(<i>Y</i>)	Treatment proxy <i>T</i>	Outcome proxy <i>Y</i>
Medical students ⁴⁰	881	0.300	0.732	MBI-EX ≥ P75	CES-D ≥ 10
Quarantine (general) ⁴¹	824	0.689	0.498	Growing stress (yes/maybe vs. no)	Coping struggles (yes)
Psychiatric patients ⁴²	919	0.171	0.285	Anxiety-spectrum dx	Mood disorder/MDD dx

T = treatment proxy; *Y* = outcome proxy. Prev(*T*) and Prev(*Y*) denote treatment and outcome prevalence in the final analysis sample. *n* is the final analysis sample after preprocessing and exclusion of non-binary entries in the medical dataset for binary-gender fairness diagnostics.

919); total *n* = 2,624. Dataset-level metadata and treatment/outcome proxy definitions are summarized in Table 1, and the harmonized feature alignment is shown in Table 2. For uplift quality we report area under the uplift curve (AUUC; higher is better), computed on the target (held-out) domain. Unless noted, we use the encoder (128 → 64) with dropout 0.30, Adam, gradient clipping (1.0), and the fairness penalty weight λ_F as indicated per experiment.

Architecture selection (encoder depth/width)

We first ablated the encoder to choose a stable backbone before any adaptation; the tested configurations and their LODO-AUUC scores are summarized in Table 3.

DANN vs. NoDA

With the fixed backbone, we compared DANN training to a NoDA model. The left plot in Figure 2 shows the mean LODO-AUUC comparison between both models using the same seed and preprocessing. In this early comparison DANN and NoDA were close, with a slight edge to NoDA.

Beyond the mean comparison, the right plot in Figure 2 shows the full AUUC distributions across seed × fold samples; the strong overlap anticipates the non-significant head-to-head tests reported later.

Systematic λ_D analysis and diagnostics

Dense λ_D sweep

We ran a grid $\lambda_D \in \{0.1, 0.2, \dots, 0.8\}$ with the fixed backbone. Best fold-wise AUUCs were:

$$\begin{aligned} \text{fold 1 (psychiatric)} &= -2.684, & \text{fold 2 (medical)} &= +4.103, \\ \text{fold 3 (quarantine)} &= +1.788; & \text{mean} &= +1.069 \pm 2.817. \end{aligned}$$

The sweep revealed a non-monotonic response (Figure 3): performance typically decreased-increased-decreased-increased as λ_D grew, with peaks around 0.7–0.8 for folds 2 and 3, whereas fold 1 stayed negative throughout. This is consistent with trading off removal of harmful domain signal against over-alignment when source/target mechanisms diverge.

Fold-wise vs. shared λ_D

Selecting λ_D separately per fold improved stability, per-fold AUUC (−2.366, +2.386, +1.883) (mean + 0.634). Using a single shared λ_D (chosen by maximizing mean AUUC in the sweep) gave (−4.485, +2.626, +1.972) (mean + 0.038). The shared setting under-performed due to the psychiatric fold's sensitivity; this motivated reporting both fold-wise bests and shared settings in later comparisons.

Fairness diagnostics

To quantify subgroup equity, we computed absolute uplift gaps by gender and by education among treated samples (Figure 4).

Table 2. Cross-dataset variable alignment after harmonization

Unified feature	Medical	Quarantine	Psychiatric
Demographics			
Age	Age (years)	Age (years)	Age (years)
Gender (binary)	Sex (1 = M, 2 = F)	Gender (string)	Sex (string)
Education (years)	12 + MBBS year	— (Imputed)	Education col
Domain-specific predictors (selected)			
Trait anxiety	STAI-T	— (Imputed)	— (Imputed)
Burnout (MBI)	MBI-CY, MBI-EA	— (Imputed)	— (Imputed)
Empathy (QCAE)	QCAE-COG, QCAE-AFF	— (Imputed)	— (Imputed)
Stress/coping	— (Imputed)	Frustration, habits	— (Imputed)
Social weakness	— (Imputed)	Social_weakness	— (Imputed)
IQ/cognitive	— (Imputed)	— (Imputed)	IQ score
EEG features	— (Imputed)	— (Imputed)	10 EEG cols
Treatment and outcome			
<i>T</i>	MBI-EX ≥ P75	Growing stress	Anxiety dx
<i>Y</i>	CES-D ≥ 10	Coping struggles	Mood/MDD dx

“—” indicates missing-by-design; such columns are imputed with the training-set median.

Table 3. Encoder ablation (LODO AUUC)

Encoder	Dropout	Mean AUUC	Per-fold AUUC
128 → 64	0.30	+1.127	(−3.346, +3.309, +3.419)
128 → 64	0.00	+0.666	(−3.125, +2.178, +2.946)
64 → 32	0.20	+0.028	(−2.883, +1.273, +1.693)

Means are across the three held-out domains; parentheses show per-fold AUUC in order (psychiatric, medical, and quarantine).

The (128 → 64) architecture with dropout 0.30 delivered the best average AUUC and was therefore used in subsequent experiments.

Absolute gaps ranged from ≈ 0.00 to 0.13 across folds, with the largest gap (≈ 0.13) appearing in the psychiatric cohort's gender groups. These values are small in absolute terms; however, given the simplicity of the group-mean parity metric, the limited subgroup sample sizes within each fold (particularly in the quarantine cohort where education is missing by design), and the observational nature of the data, we caution against strong conclusions about equity. The fairness figures represent a preliminary audit rather than a validated fairness guarantee; more comprehensive evaluation including intersectional analysis remains future work.

Finally, the psychiatric target cohort exhibits a characteristic failure mode: the uplift curve lies consistently below the random baseline, indicating an *inverted* ranking of treatment benefit (Figure 5). This inversion has a precise structural explanation grounded in the proxy design. The treatment proxy ($T = 1$, anxiety-spectrum diagnosis) and the outcome proxy ($Y = 1$, MDD) are derived from the same diagnostic-label columns that are also retained as predictor features after one-hot encoding. In the processed psychiatric dataset, the two proxy categories are mutually exclusive under the present proxy construction: $\Pr(Y = 1|T = 1) = 0.000$ compared with $\Pr(Y = 1|T = 0) = 0.344$ (Table 4). The model can learn this negative association directly from the feature space, producing negative uplift rankings regardless of domain-alignment strength. The observed negative AUUC therefore reflects *proxy-design leakage* rather than a failure of domain adaptation: when treatment and outcome proxies are derived from overlapping diagnostic categories that are also present as input features, the ranking problem is structurally ill-posed for causal inference. We retain this fold in the evaluation as an instructive case and return to the limitation in limitations and future work subsection.

Schedules, transfers, and semi-synthetic validation

Scheduled λ_D

We compared fixed λ_D to two schedules (linear ramp and sigmoid ramp). Mean AUUC, *fixed* = +0.557, *linear* = −0.187, and *sigmoid* = +1.327. The sigmoid schedule outperformed fixed and linear, supporting the classical DANN heuristic of delaying stronger adversarial pressure until features stabilize.

Single-domain transfers

To isolate cross-domain generalization, we trained DANN on a single source and tested on a single target:

train 0→test 1 : + 2.815, train 0→test 2 : −2.757,
train 1→test 0 : + 3.291, train 1→test 2 : −2.981,
train 2→test 0 : + 2.745, train 2→test 1 : + 2.003.

Transfers into the medical domain (0) were consistently positive, while transfers into the psychiatric domain (2) were consistently negative. In light of the proxy audit above, these negative transfers should be interpreted cautiously: they reflect not only cross-domain difficulty but also the structural dependence between psychiatric treatment/outcome proxies and diagnostic input features.

Semi-synthetic benchmark (CATE RMSE)

We injected a known treatment effect atop real features to create a sanity-check benchmark. Across two independent runs, fold-wise conditional average treatment effect (CATE) root-mean-square errors (RMSEs) were

[0.264, 0.369, 0.219] (mean 0.284) and
[0.238, 0.312, 0.203] (mean 0.251).

These results indicate that, when the data-generating process is well-specified and the injected signal is well-aligned with the model assumptions, the architecture can recover the known effect with low error. Critically, this should be interpreted as a *model behavior check*—a sanity test demonstrating that the architecture does not systematically distort treatment-effect estimates—rather than as evidence of real-world causal identification, which would require assumptions (ignorability, overlap, and stable unit treatment value assumption) that cannot be verified in the observational datasets.

Head-to-head: DANN vs. CDAN vs. CORAL vs. MMD vs. NoDA

We now report the central comparative evaluation schemes under identical preprocessing, backbone architecture, and LODO

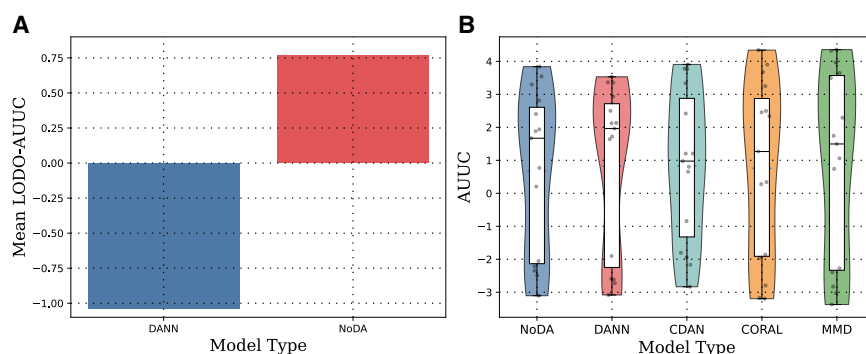


Figure 2. Comparison of uplift performance across modeling strategies

(A) Mean LODO-AUUC comparison between DANN and NoDA using a 128 → 64 encoder with dropout 0.30; individual fold scores are overlaid. (B) Distribution of LODO-AUUC scores across five modeling methods (NoDA, DANN, CDAN, CORAL, and MMD); boxplots show IQR and medians, violins show density across 15 seed × fold runs.

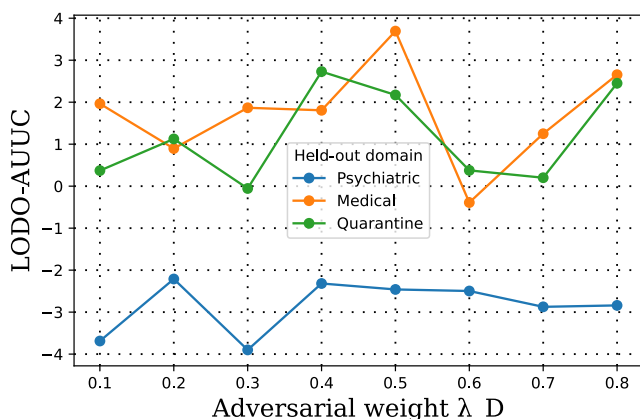


Figure 3. Dense sweep of adversarial weight λ_D for DANN models using a 128 → 64 encoder with dropout 0.30

The x axis shows λ_D values from 0.1 to 0.8, and the y axis reports LODO-AUUC scores for each fold. Curves represent individual held-out domains (fold 1, psychiatric; fold 2, medical; and fold 3, quarantine). The sweep reveals non-monotonic behavior, with performance gains in folds 2 and 3 and consistent degradation in fold 1, highlighting cohort-specific sensitivity to domain alignment strength.

evaluation. The compared methods are NoDA, DANN, CDAN, CORAL, and MMD. Each method was subjected to a light grid/Optuna search over $(l_r, \lambda_F, \lambda_D)$, with the final search ranges summarized in Table 5, and Table 6 reports the best observed mean AUUC together with uncertainty summaries across 5 seeds $\times$ 3-folds.

The wide standard deviations ($\sigma = 2.3$ – 2.9 AUUC units) and confidence-interval half-widths (± 1.2 – 1.5 AUUC) in Table 6 reflect genuine instability in cross-cohort uplift estimation under

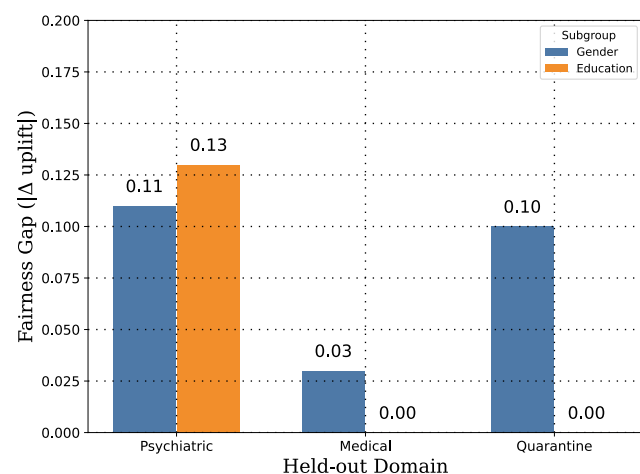


Figure 4. Fairness gaps in uplift predictions across 3-folds (psychiatric, medical, and quarantine)

Bars show the absolute difference in mean uplift between gender subgroups and education subgroups within each fold. The y axis represents $|\bar{u}_{\text{group 1}} - \bar{u}_{\text{group 2}}|$, a measure of equity in treatment effect estimation. Gaps are small in absolute terms (≈ 0.00 – 0.13), but given the group-mean parity metric and limited subgroup sizes these figures should be read as a preliminary audit rather than a validated fairness guarantee.

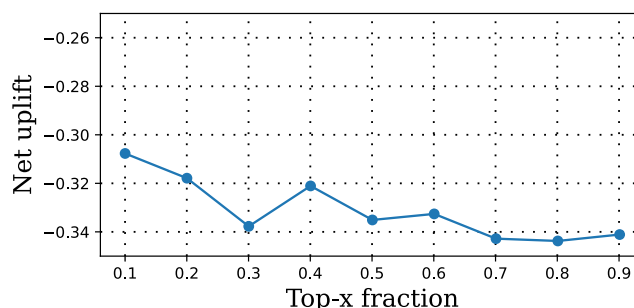


Figure 5. Uplift curve on the psychiatric test fold

The curve lies below the random baseline ($\text{AUUC} \approx -2.66$), indicating an inverted treatment-effect ranking. In this fold $\Pr(Y = 1 | T = 1) = 0.000$ under the present proxy construction: the anxiety-spectrum treatment proxy and the mood-disorder outcome proxy are mutually exclusive diagnostic categories, and both are partially encoded in the predictor features via one-hot encoding of diagnostic-label fields (`main.disorder` and `specific.disorder`). The model can learn this negative association from the feature space, producing inverted uplift rankings that reflect proxy-design leakage rather than a DA failure.

this sample size and three-cohort configuration. The overlapping distributions visible in Figure 2 reinforce that method choice matters substantially less than dataset characteristics and proxy quality in this setting. These observations should temper any interpretation of point-estimate differences between methods, including the numerical advantage of MMD over NoDA ($+0.204$ AUUC), which falls well within one standard deviation of either method.

Across 5 seeds $\times$ 3-folds ($n = 15$ paired samples), the paired median improvement of DANN over NoDA was $+0.064$ (bootstrap 95% CI $[-0.402, 0.511]$), with Cliff's $\Delta = 0.031$ (negligible effect size) and a Wilcoxon signed-rank $p = 0.847$. CDAN vs. DANN and MMD/CORAL vs. DANN were likewise non-significant (all $p > 0.75$). The difference histogram in Figure 6 visualizes the lack of consistent gain.

DISCUSSION

Cross-domain uplift performance (RQ1)

Our first question asked whether a *single* causal model can meaningfully rank “who will deteriorate under stress” across multiple cohorts. The LODO experiments indicate that this is feasible but imperfect. In two out of the three unseen test domains (medical students and quarantine respondents), the unified model achieved positive uplift performance (AUUC above zero) on the held-out cohort. For example, when trained on two cohorts and tested on the third, the model obtained relative AUUC gains on the order of $+2$ to $+3$ in the medical-student and quarantine cohorts compared with a random-uplift baseline, showing that cross-domain uplift-ranking signal can transfer under some forms of distribution shift.

In practical terms, the model can rank individuals by proxy-defined differential vulnerability in new cohorts better than chance, indicating partial cross-domain generalization at the level of uplift ranking. This addresses RQ1 by showing that a single uplift model can generalize across fundamentally different

Table 4. Cross-tabulation of treatment proxy T and outcome proxy Y in the psychiatric dataset ($n = 919$ after harmonization)

	$Y = 0$	$Y = 1$	Total
$T = 0$	500	262	762
$T = 1$	157	0	157
Total	657	262	919
$\Pr(Y = 1)$	$T = 1: 0.000 \quad T = 0: 0.344$		–

The zero cell ($T = 1, Y = 1$) confirms perfect mutual exclusivity between the anxiety-spectrum treatment proxy and the mood-disorder outcome proxy under the present proxy construction.

mental-health datasets, albeit with varying degrees of success. The unified approach also tended to outperform single-source transfers: training on all available source domains generally yielded better uplift ranking in the target domain than training on just one source domain at a time. This suggests that pooling data from multiple cohorts helps the model capture a broader spectrum of stress-outcome relationships, making it more robust than any individual source model. From a practical perspective, the most defensible implication is modest: in a research or service-planning setting, the model output could be used as a *hypothetical triage signal* to prioritize which individuals should receive additional screening or follow-up for stress-related deterioration. It should not be interpreted as a standalone treatment recommendation or as proof that the ranked individuals would truly benefit from a specific intervention in routine clinical practice.

At the same time, the psychiatric cohort exhibited a characteristic failure mode, with the uplift curve lying consistently below the random baseline. As established in systematic λ_D analysis and diagnostics section, this inversion reflects a structural proxy-design issue rather than a DA failure: the treatment proxy ($T = 1$, anxiety-spectrum diagnosis) and the outcome proxy ($Y = 1$, MDD) are derived from diagnostic-label columns that are also retained as input features, and the two proxy categories are empirically mutually exclusive in the processed sample under the current proxy definitions ($\Pr[Y = 1|T = 1] = 0.000$; Table 4). The model has access to features that encode this negative association directly, producing inverted uplift rankings regardless of domain-alignment strength. Thus, while RQ1 is answered positively overall, the psychiatric fold illustrates that proxy construct validity is a prerequisite for meaningful uplift estimation: when treatment and outcome proxies are structurally entangled with the input features, the ranking problem is ill-posed for causal inference independent of model choice.

Table 5. Final search ranges for hyperparameters across uplift models

Hyperparameter	Search range	Notes
Learning rate (lr)	1×10^{-4} – 5×10^{-4}	Log-uniform
Adversarial weight (λ_D)	0.1–0.8	Dense grid sweeps; also scheduled ramps
Fairness weight (λ_F)	1.0–15.0	Small-to-moderate regularization
Weight decay	10^{-5} – 10^{-4}	Adam

Table 6. Summary statistics of LODO-AUUC scores across five modeling methods: NoDA, DANN, CDAN, CORAL, and MMD

Method	Mean	SE	$\pm(95\% \text{ CI})$	sd	n
NoDA	+0.678	0.641	± 1.255	2.480	15
DANN ⁴³	+0.823	0.664	± 1.300	2.569	15
CDAN ⁴⁴	+0.832	0.601	± 1.177	2.326	15
CORAL ⁴⁵	+0.758	0.705	± 1.382	2.731	15
MMD ⁴⁶	+0.882	0.752	± 1.474	2.913	15

Each method was evaluated over $n = 15$ seed $\times$ fold runs. Reported metrics include mean AUUC, standard error (SE), 95% confidence interval half-width, and sample standard deviation. The central finding is that no adaptation method achieves a statistically significant improvement over the non-adapted baseline: all pairwise differences are within sampling noise, and the wide standard deviations ($\sigma = 2.3 - 2.9$ AUUC units) reflect genuine instability under this three-cohort configuration.

Value and limits of domain adaptation (RQ2)

RQ2 asked whether adding a DA layer measurably improves out-of-domain uplift performance over a strong non-adapted shared-encoder baseline. The answer is qualified. Across one-shot tuned configurations, the best adversarial methods (DANN and CDAN) achieved numerically higher mean LODO-AUUC than the NoDA baseline (e.g., mean AUUC $\approx +0.832$ vs. $\approx +0.678$ on the held-out domains). However, the size of this improvement was small—roughly $+0.15$ to $+0.20$ AUUC units—and, given the observed standard errors (≈ 0.6 – 0.75) and $n = 15$ paired observations, remained well within sampling noise.

When performance was aggregated across five seeds and 3-folds, the same pattern persisted. DANN and CDAN tended to achieve slightly better point estimates than NoDA, but the paired median improvement of DANN over NoDA was only about $+0.06$ AUUC, with a 95% bootstrap confidence interval spanning zero and a non-significant Wilcoxon signed-rank test ($p > 0.8$). Comparisons among the adaptation methods were similarly inconclusive: CDAN vs. DANN and CORAL/MMD vs. DANN were also non-significant. Thus, the present evidence does not support a statistically robust advantage of domain adaptation over a strong shared-encoder baseline in this three-dataset setting.

Among the adaptation methods, the adversarial approaches (DANN and CDAN) were numerically strongest, while CORAL and MMD were close behind. This ranking is consistent with the intuition that flexible adversarial aligners may better capture complex cross-domain discrepancies than moment-matching penalties alone. At the same time, the differences among methods were small relative to fold-to-fold and run-to-run variability, so these numerical ordering differences should not be over-interpreted.

The most important conclusion is therefore about magnitude relative to instability. Our results should not be read as showing that domain adaptation is ineffective; rather, they show that any incremental benefit is of comparable size to the noise level under the present sample size and proxy definitions. Similar findings have been reported in other medical domain-shift settings. For example, Guo et al.³¹ found that several domain generalization

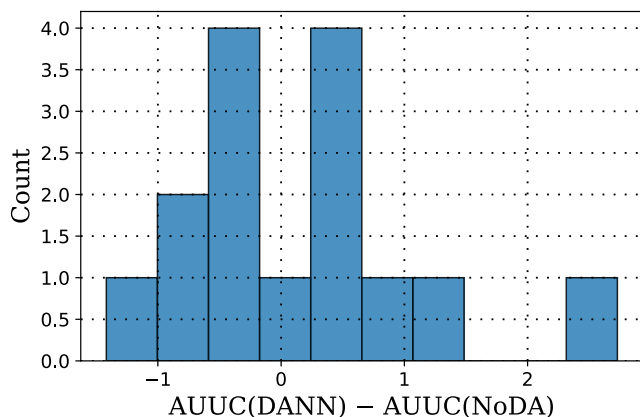


Figure 6. Distribution of paired AUUC differences between DANN and NoDA across seed $\times$ fold runs

The x axis shows $AUUC_{DANN} - AUUC_{NoDA}$ and the y axis shows frequency. The distribution is centered near zero (median = 0.064), and the 95% bootstrap confidence interval spans zero, indicating no statistically significant uplift gain from DANN.

and adaptation methods did not substantially outperform a strong ERM baseline under temporal dataset shift in ICU data. Our results echo that pattern: when source domains are already diverse and the baseline model is strong, standard DA methods may offer only incremental gains.

Overall, RQ2 is answered cautiously. Domain adaptation may provide modest numerical improvements in cross-dataset uplift ranking, but in the present three-cohort configuration these gains are unstable and not statistically significant. Accordingly, domain adaptation should be treated as an empirical option worth testing, not as a guaranteed improvement.

Robustness, fairness, and causal validity (RQ3)

RQ3 focused on robustness: how sensitive are our conclusions to the choice and schedule of the adversarial weight λ_D , and can subgroup uplift gaps remain modest while the model still recovers known treatment-effect structure on the semi-synthetic benchmark?

Tuning λ_D for stable alignment

A key methodological insight is that performance is highly sensitive to λ_D . Dense sweeps revealed a non-monotonic relationship between λ_D and AUUC: very small values produced almost no alignment (essentially reducing to NoDA), whereas large values caused over-alignment that degraded performance, likely by forcing the model to ignore meaningful cohort-specific signal. Between these extremes, an intermediate range of λ_D maximized validation AUUC, trading off bias reduction against over-smoothing.

Scheduling λ_D over training further improved stability. In particular, a smoothly increasing sigmoid schedule often matched or exceeded the best fixed setting (e.g., mean AUUC around 1.3 with a ramp vs. roughly 0.6 with a fixed moderate value) and reduced run-to-run variance. This supports the heuristic, seen in original DANN work,⁴³ that the adversarial loss should be introduced gradually so that the encoder first learns predictive features and only later enforces domain invariance.

Arbitrary static choices of λ_D risk either under-utilizing or over-driving the adversarial signal. Our sweep-plus-schedule protocol turns what is often a trial-and-error process into a reproducible strategy for future studies.

Semi-synthetic validation

To probe model behavior under a controlled treatment-effect signal, we injected a known treatment-effect function on top of real features and evaluated CATE error. Across two independent runs, fold-wise CATE RMSEs were (0.264, 0.369, 0.219) (mean 0.284) and (0.238, 0.312, 0.203) (mean 0.251), respectively. These results indicate that when the underlying signal is well aligned with the model assumptions, our architecture can recover a known injected signal with low error under controlled conditions. This makes it more plausible that failures on real datasets reflect proxy-design or construct mismatch rather than a purely architectural problem, although it does not by itself establish causal validity in the observational setting. As noted in uplift target and semi-synthetic ground truth section, this experiment should be read as a model behavior check, not as evidence of real-world causal identification, since the assumptions required for causal inference cannot be verified in the observational datasets.

Fairness diagnostics

We quantified subgroup equity by computing absolute uplift gaps between demographic groups among treated individuals (gender and education; Figure 4). Across folds, the gaps were small overall (~ 0.00 – 0.13) with mild variation by domain. Aggregating across seeds and folds using nonparametric summaries (paired bootstrap for the median difference, Cliff's δ), we did not observe a consistent reduction of these gaps attributable to adversarial adaptation: the bootstrap 95% confidence intervals for the gap difference (DANN – NoDA) overlapped zero for both gender and education, and effect sizes were negligible. In other words, on these cohorts the fairness metric remained broadly comparable between DANN and the non-adapted baseline, likely because (i) baseline gaps were already small and (ii) the dominant challenge was label/proxy mismatch, which domain alignment does not directly address.

Overall, RQ3 is answered with a mixed message: careful tuning and scheduling of λ_D are crucial for stable performance; our architecture can recover a known injected signal in a semi-synthetic setting; and fairness gaps are modest and not substantially altered by DA. Robustness and subgroup parity can be partially supported, but they do not automatically follow from adversarial alignment alone.

Cohort-specific insight: The psychiatric domain

The most striking cohort-specific result was the consistently negative uplift in the Psychiatric test fold ($AUUC \approx -2.66$; Figure 5). This inversion is not best explained by DA failure alone and is most plausibly explained by the structure of the proxy definitions. As documented in systematic λ_D analysis and diagnostics section and Table 4, the treatment proxy ($T = 1$, anxiety-spectrum diagnosis) and the outcome proxy ($Y = 1$, MDD) are both derived from the diagnostic-label columns `main.disorder` and `specific.disorder`, which are also retained as predictor features after one-hot encoding. In the processed psychiatric dataset, these two proxy categories are empirically

Table 7. Qualitative comparison of related work on treatment-effect/uplift modeling, domain adaptation, and fairness

Study	Mental-health outcome	ITE/uplift TE	Multi-dataset/domain	DA for TE/uplift	Fairness in uplift/TE	DA diagnostics
Bica et al. ²⁰	x	✓	✓	✓	x	x
Kyono et al. ⁴⁷	x	✓	✓	✓	x	x
Sun et al. ⁴⁸	x	✓	✓	✓	x	x
Shang et al. ⁴⁹	x	✓	x	x	x	x
Zahid et al. ³⁷	x	✓	x	x	✓	x
Kadioglu et al. ⁵⁰	x	✓	x	x	✓	x
Meegahapola et al. ⁵¹	✓	x	✓	x	x	✓
Fang et al. ⁵²	✓	x	✓	✓	x	x
Wang et al. ³⁸	x	✓	x	x	x	x
He et al. ³⁹	x	✓	x	x	x	x
Our study	✓	✓	✓	✓	✓	✓

A tick (✓) indicates that the aspect is an explicit, primary focus of the study. Note: Meegahapola et al.,⁵¹ apply domain adaptation to risk prediction tasks, not to treatment-effect or uplift estimation. The last column (“DA diagnostics”) is marked when the study provides an explicit analysis of domain-adaptation behavior (e.g., sensitivity to alignment strength, failure modes, or ablation of DA components).

mutually exclusive in the processed sample under the current proxy definitions: $\Pr(Y = 1|T = 1) = 0.000$ versus $\Pr(Y = 1|T = 0) = 0.344$ (Table 4). Because the model has access to features that directly encode the proxy relationship between T and Y , it can learn this negative association from the feature space regardless of alignment strength, producing inverted uplift rankings as a consequence of proxy-design leakage rather than genuine domain-shift failure.

This case illustrates why adversarial domain adaptation cannot rescue a structurally entangled proxy design. Domain-adversarial training targets the marginal feature distribution $P(X)$, encouraging the encoder to produce representations that are invariant across study populations. However, when the treatment-outcome relationship is already partially inferable from the input features—because both proxies are derived from the same categorical variable—aligning $P(X)$ across domains changes nothing about the shortcut the model exploits. This is consistent with the broader observation, reported by Guo et al.,³¹ that adversarial methods, CORAL, and MMD all fail to improve generalization when the feature-outcome association in the target domain is fundamentally different from the source: aligning feature distributions cannot correct a misspecified proxy relationship. The psychiatric fold is a concrete clinical instance of this failure mode: the structural dependence between the diagnostic treatment and the diagnostic outcome means that the uplift ranking problem is ill-posed for causal inference in this fold, independent of what alignment method is applied.

The practical lesson is specific and actionable. Before deploying an uplift model in a clinical setting where treatment and outcome are both operationalized from diagnostic categories, practitioners must verify that the treatment proxy is derived from a variable independent of the outcome classification. In the psychiatric context, this would mean using a validated continuous anxiety-symptom scale rather than the primary diagnostic category as the treatment proxy, ensuring that treatment assignment is not structurally entangled with the outcome. More broadly, the psychiatric fold provides a worked diagnostic case: rather than dismissing the negative AUUC as a modeling failure,

examining the proxy design revealed the source of the inversion precisely. This kind of structural post-hoc audit—checking $\Pr(Y = 1|T = 1)$ before interpreting uplift results—should be a routine step in any multi-dataset causal uplift pipeline, particularly when proxies are derived from categorical clinical variables.

Comparison with existing literature

Qualitative comparison with prior work is summarized in Table 7. We use “TEs” to denote treatment effects, “ITEs” for individual treatment effects, and “DA” for domain adaptation in the comparison table. Transfer CATE methods such as HTCE-learners,²⁰ Kyono et al.’s domain-adapted ITE model selection,⁴⁷ and Sun et al.’s cross-domain treatment-effect estimation framework⁴⁸ explicitly target individual treatment effects under distribution shift, but they are not set in mental-health cohorts and do not study fairness or uplift gaps. Conversely, recent fairness-aware uplift approaches (FairUDT,³⁷ and fairness evaluation without ground truth⁵⁰) focus on equity in marketing uplift campaigns, without cross-dataset transfer or domain adaptation. Complementary work in mental-health analytics applies domain adaptation to *risk prediction* rather than treatment effects, for example M3BAT for multimodal mobile health sensing⁵¹ and unsupervised cross-site fMRI adaptation for MDD identification.⁵² These models operate on single outcomes and do not estimate uplift or CATE.

Beyond the above categories, a few very recent studies have proposed novel techniques for uplift modeling within a single domain. Wang et al. developed an uplift model that integrates causal inference with graph neural networks to exploit feature relationships, achieving lower error on simulated and real marketing data.³⁸ He et al. presented a framework to improve uplift ranking by introducing loss functions that directly penalize mis-ordered treatment effect predictions, demonstrating superior performance in e-commerce targeting scenarios.³⁹ While effective in their respective contexts, neither of these approaches incorporates domain adaptation or multi-dataset generalization, and they do not consider fairness constraints. This highlights that the literature has largely treated uplift modeling, domain shift, and fairness as separate problems.

To our knowledge, our study is among the first to jointly combine (i) uplift modeling across three heterogeneous real-world mental-health cohorts, (ii) adversarial representation learning for cross-cohort transfer, and (iii) an explicit fairness regularizer on subgroup uplift gaps together with a systematic analysis of when domain adaptation helps and when its gains over a strong non-adapted baseline are modest. In Table 7, our study is the only one in this comparison to address all six dimensions simultaneously—spanning mental-health outcomes, individualized treatment-effect estimation, multi-dataset generalization, fairness considerations, and diagnostics of DA behavior. Whether this combination yields reliable clinical benefit remains an open empirical question; our findings suggest it is a necessary but not sufficient condition.

Clinical and public-health implications

The findings carry three practical messages for researchers and practitioners seeking to apply uplift modeling in mental-health settings.

Cross-context transferability is feasible but limited

The model achieved positive uplift ranking in two of three held-out populations (medical students and quarantine respondents), suggesting that features predictive of stress-related deterioration in one population may carry signal for another. As noted in cross-domain uplift performance (RQ1) subsection, this should be interpreted modestly: in a research or service-planning context, model output could serve as a *hypothetical triage signal* to prioritize individuals for additional screening or follow-up, not as a standalone clinical recommendation. In deployment, predictions from a model trained on one population and applied to another should be treated as preliminary flags for human clinical review—not as proof that ranked individuals would benefit from a specific intervention in routine practice.

Proxy construct validity is a prerequisite, not an afterthought

The psychiatric fold provides the clearest practical warning in this study: when treatment and outcome proxies are derived from overlapping categorical variables that also appear as input features, the uplift ranking problem is structurally ill-posed ($\Pr(Y = 1|T = 1) = 0.000$ in the psychiatric fold; Table 4). No unsupervised DA method evaluated here can by itself rescue this configuration. Before deploying an uplift model in a new mental-health context, we recommend verifying three conditions: (i) the treatment proxy is derived from a variable that is *independent* of the outcome classification (e.g., a validated continuous severity scale rather than a diagnostic category); (ii) the direction of the expected treatment-outcome relationship is *a priori* plausible and does not contradict the dataset's coding structure; and (iii) the treated group has adequate prevalence—our results suggest that very low treatment prevalence (as in the psychiatric fold, $\text{Prev}[T] = 0.171$) can substantially increase instability.

Fairness monitoring is necessary, not optional

Demographic uplift gaps in our experiments were small (≈ 0.00 – 0.13) and were not substantially altered by adversarial adaptation, suggesting no large subgroup disparities under this group-mean parity metric. However, small gaps under our group-mean parity metric do not constitute clearance for deployment: intersectional disparities (e.g., women with low education

vs. men with high education), within-group variance, and allocation-level inequities are not captured.⁵³ Clinical AI models should undergo ongoing fairness auditing across all protected attributes available in the target population.¹⁴ Our fairness penalty is a first step; it does not substitute for regulatory-grade bias evaluation, and any deployment should involve clinicians in reviewing whether predicted uplift gaps align with their ethical and institutional standards.

Limitations and future work

While our study provides encouraging evidence for domain-adaptive uplift modeling, it also has several limitations that suggest avenues for future work.⁵⁴

First, the magnitude of improvement due to adaptation was small, and we lacked statistical power to definitively assert superiority of our approach over a well-tuned baseline. This is partly due to sample size and the inherent noise in estimating individualized treatment effects. In the future, conducting experiments on larger or additional cohorts would help confirm the trends observed here (e.g., whether adversarial adaptation consistently yields a few percentage points of AUUC gain) or reveal stronger effects. A larger scale study might include more diverse mental health datasets—for example, data from different countries or contexts—which would also test the limits of generalization and possibly amplify the benefits of adaptation (especially if some new domains are substantially different from those seen in training).

Second, the architecture of the model used was relatively simple, as it was a fully connected network with two hidden layers. However, it is possible that using more expressive or deeper encoders could improve performance across domains. Indeed, the recent literature on the topic supports the claim that using more capable or pre-training the model on large datasets can improve the learning of invariant features. For instance, using a transformer-based model pre-trained on large behavioral datasets can improve the learning of universal patterns of stress and coping, thus improving the robustness to domain differences. Exploring such foundation models or multi-modal deep architectures (e.g., incorporating text or wearable sensor data if available) is an exciting direction. However, using deeper models also makes the problem of overfitting more likely and might require careful regularization when data are limited. A related point is that we trained from scratch on these cohorts; using transfer learning from a large corpus (not necessarily mental health-specific, even something such as a general health dataset) could initialize the model in a more domain-invariant space. Future work could evaluate fine-tuning such pre-trained models for uplift predictions, which might yield a stronger baseline and potentially amplify uplift gains or fairness.

Third, our domain adaptation was purely unsupervised, with no target outcome data used during training. In practice, if a small size of labeled target-domain data are available, semi-supervised or few-shot adaptation may yield larger gains than unsupervised alignment alone. A two-stage strategy—first learning a shared representation through adversarial alignment, then fine-tuning the uplift heads on limited target-domain supervision—may be especially useful when the target domain differs in proxy semantics or base rates.

Fourth, although we addressed fairness in terms of parity across two attributes, the ethical and practical aspects of deployment warrant ongoing attention. Our fairness penalty was a simple linear regularizer; future work could investigate richer fairness criteria, such as allocation-level equity or user-defined utility constraints. We also acknowledge that we focused on only two attributes (gender and education); other demographics and their interactions should be considered in a more comprehensive fairness analysis.^{55,56}

Future work should also consider establishing a more systematic benchmark for cross-domain uplift modeling in mental health. Our study is among the first to evaluate multiple domain adaptation techniques (DANN, CDAN, CORAL, and MMD) side-by-side on multiple mental-health cohorts with uplift metrics. Expanding this to more datasets and perhaps standardized tasks could create a benchmark analogous in spirit to WILDS (a benchmark of in-the-wild distribution shifts spanning diverse data modalities and applications, from tumor identification to wildlife monitoring to poverty mapping), but for causal uplift scenarios.

Construct validity

The fundamental limitation of this study is that the three datasets operationalize “treatment” and “outcome” in qualitatively different ways (Construct validity and comparability section). Even after binarization, AUUC summarizes performance on *proxy-defined* treatment effects, not on true clinical intervention effects. The psychiatric fold provides the sharpest illustration: when treatment and outcome are both derived from a shared categorical variable, the uplift ranking problem is structurally ill-posed regardless of model sophistication. More broadly, a positive AUUC on any fold confirms that the model ranks individuals by their proxy-defined differential vulnerability better than chance—not that those individuals would genuinely benefit from a clinical intervention. Future work should prioritize datasets where the treatment is a real-world intervention measured prospectively with validated instruments, and should also explore domain-specific proxy design, multi-task formulations, or meta-learned sharing mechanisms when a single proxy cannot be meaningfully imposed across all datasets.

Gap between semi-synthetic validation and real-world causal inference

The semi-synthetic experiment demonstrates that the architecture can recover a known artificially injected signal with low error (CATE-RMSE ≈ 0.25 – 0.28). This does not imply it recovers the true CATE in the observational datasets, where unmeasured confounders—including pre-existing comorbidities, socioeconomic factors, and help-seeking behavior—may bias uplift estimates in unknown directions. Bridging this gap requires either randomized or quasi-randomized data in which treatment assignment is independent of confounders, or a formal sensitivity analysis for unmeasured confounding.⁵⁷ Neither was feasible here given the cross-sectional observational design of all three datasets; we flag this as a prerequisite for any follow-up study that aims to make stronger causal claims.

Scalability and adversarial training stability

All experiments were run on a single graphics processing unit (T4, 16 GB VRAM). With larger or more diverse cohorts, adversarial training instability may worsen: the gradient-reversal mechanism can oscillate when the domain discriminator and

encoder are unevenly matched in capacity or when batch sizes are small relative to the number of domains. More sophisticated stabilization strategies—such as spectral normalization, Wasserstein-based divergence objectives,⁵⁸ or gradient penalty regularization—may be warranted in larger-scale settings. These extensions were outside the scope of the present study but represent an important direction for making domain-adaptive uplift modeling practically deployable at clinical scale.

Overall, this study provides a careful empirical assessment of cross-dataset uplift modeling in mental-health settings. The results show that uplift-style ranking can transfer across some populations, that domain adaptation may help modestly but not reliably, and that proxy design can dominate model behavior when treatment and outcome are not cleanly separated from the input space. By addressing RQ1–RQ3, the study clarifies both the promise and the current limits of domain-adaptive uplift modeling under heterogeneous proxy definitions. The study also highlights hyperparameter tuning and fairness considerations as important parts of any practical cross-dataset uplift pipeline rather than auxiliary add-ons. For the scientific community, the study’s findings suggest the importance of validating mental health intervention models developed for different cohorts and the possible need for adjustments through domain adaptation, proxy re-evaluation, or fairness constraints—to ensure they remain both effective and equitable.^{59,60} Going forward, combining adversarial domain-invariant learning with techniques like meta-learning, deeper architectures, and domain-specific knowledge integration holds promise for achieving more robust, scalable, and potentially clinically useful uplift modeling across diverse populations. We hope this study serves as a valuable stepping stone toward that goal, helping to chart a path for future research in generalizable and fair personalized mental health interventions.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Abdullah Nayem Wasi Emran (abdullah.nayem@bracu.ac.bd).

Materials availability

This study did not generate new materials.

Data and code availability

- All datasets analyzed in this study are publicly available from third-party repositories and were not generated by the authors. The Identification of Major Psychiatric Disorders Dataset is available at <https://osf.io/8bsvr/>. The Medical Students Mental Health Dataset is available at <https://zenodo.org/records/5702895>. The Mental Health Depression During Quarantine Life Dataset is available at <https://data.mendeley.com/datasets/pxjmjyfdh2/1>. All datasets are openly accessible subject to the terms of their respective repositories.
- The code generated during this study is publicly available at <https://github.com/Nayem9274/domain-adaptive-uplift-mh>. The repository contains the analysis notebook and a requirements.txt specifying all software dependencies needed to reproduce the reported experiments.
- No new datasets were generated or deposited as part of this study. This study does not report original data.

ACKNOWLEDGMENTS

This work was conducted at and supported by the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

AUTHOR CONTRIBUTIONS

A.B.M.A.A.I. and A.N.W.E. jointly developed the research idea. A.N.W.E. was responsible for data collection, coding, and analysis. A.B.M.A.A.I. supervised the work. A.N.W.E. prepared the manuscript, and A.B.M.A.A.I. reviewed and refined it.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **METHOD DETAILS**
 - Notation and formal definitions
 - Datasets and problem setup
 - Construct validity and comparability
 - Uplift target and semi-synthetic ground truth
 - Model: DANN_Uplift (domain-adversarial uplift network)
 - Baselines and variants
 - Evaluation protocol
 - Experimental details
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

Received: February 18, 2026

Revised: April 6, 2026

Accepted: May 4, 2026

Published: May 22, 2026

REFERENCES

1. Monroe, S.M., and Simons, A.D. (1991). Diathesis-stress theories in the context of life stress research: implications for the depressive disorders. *Psychol. Bull.* 110, 406–425.
2. Hammen, C. (2005). Stress and depression. *Annu. Rev. Clin. Psychol.* 1, 293–319.
3. World Health Organization (2022). *World Mental Health Report: Transforming Mental Health for All* (World Health Organization).
4. Muñoz, R.F., Beardslee, W.R., and Leykin, Y. (2012). Major depression can be prevented. *Am. Psychol.* 67, 285–295. <https://doi.org/10.1037/a0027666>.
5. Cuijpers, P., Beekman, A.T.F., and Reynolds, C.F. (2012). Preventing depression: a global priority. *JAMA* 307, 1033–1034. <https://doi.org/10.1001/jama.2012.271>.
6. West, C.P., Dyrbye, L.N., and Shanafelt, T.D. (2018). Physician burnout: contributors, consequences and solutions. *J. Intern. Med.* 283, 516–529. <https://doi.org/10.1111/joim.12752>.
7. Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Appl. Psychol. Meas.* 1, 385–401. <https://doi.org/10.1177/014662167700100306>.
8. Brooks, S.K., Webster, R.K., Smith, L.E., Woodland, L., Wessely, S., Greenberg, N., and Rubin, G.J. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *Lancet* 395, 912–920. [https://doi.org/10.1016/S0140-6736\(20\)30460-8](https://doi.org/10.1016/S0140-6736(20)30460-8).
9. Gunnell, D., Appleby, L., Arensman, E., Hawton, K., John, A., Kapur, N., Khan, M., O'Connor, R.C., and Pirkis, J.; COVID-19 Suicide Prevention Research Collaboration (2020). Suicide risk and prevention during the COVID-19 pandemic. *Lancet Psychiatry* 7, 468–471. [https://doi.org/10.1016/S2215-0366\(20\)30171-1](https://doi.org/10.1016/S2215-0366(20)30171-1).
10. American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. (American Psychiatric Publishing). <https://doi.org/10.1176/appi.books.9780890425596>.
11. Meehan, A.J., Lewis, S.J., Fazel, S., Fusar-Poli, P., Steyerberg, E.W., Stahl, D., and Danese, A. (2022). Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Mol. Psychiatry* 27, 2700–2708.
12. Xu, X., Liu, X., Zhang, H., Wang, W., Nepal, S., Sefidgar, Y., Seo, W., Kuehn, K.S., Huckins, J.F., Morris, M.E., et al. (2023). Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1–34.
13. Emran, A.N.W., and Islam, A.B.M.A.A. (2025). Mental health across contexts: A cross-dataset study covering medical students, quarantined individuals, and psychiatric disordered subjects. *Hum. Soc. Sci. Commun.* 12, 1157. <https://doi.org/10.1057/s41599-025-05053-x>.
14. Chen, R.J., Wang, J.J., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Lu, M. Y., Sahai, S., and Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* 7, 719–742.
15. Künzel, S.R., Sekhon, J.S., Bickel, P.J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* 116, 4156–4165. <https://doi.org/10.1073/pnas.1804597116>.
16. Wager, S., and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113, 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>.
17. Hardevel, F., Spijker, J., De Graaf, R., Nolen, W.A., and Beekman, A.T.F. (2010). Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatr. Scand.* 122, 184–191. <https://doi.org/10.1111/j.1600-0447.2009.01519.x>.
18. DeRubeis, R.J., Cohen, Z.D., Forand, N.R., Fournier, J.C., Gelfand, L.A., and Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: translating research on prediction into individualised treatment recommendations. A demonstration. *PLoS One* 9, e83875. <https://doi.org/10.1371/journal.pone.0083875>.
19. Bower, P., and Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency. *Br. J. Psychiatry* 186, 11–17. <https://doi.org/10.1192/bjp.186.1.11>.
20. Bica, I., and van der Schaar, M. (2022). Transfer learning on heterogeneous feature spaces for treatment effects estimation. *Adv. Neural Inf. Process. Syst.* 35, 37184–37198.
21. Colledani, D., Barbaranelli, C., and Anselmi, P. (2025). Fast, smart, and adaptive: using machine learning to optimize mental health assessment and monitor change over time. *Sci. Rep.* 15, 6492. <https://doi.org/10.1038/s41598-025-91086-w>.
22. Sundaram, A., Subramaniam, H., Ab Hamid, S.H., and Mohamad Nor, A. (2024). An adaptive data-driven architecture for mental health care applications. *PeerJ* 12, e17133. <https://doi.org/10.7717/peerj.17133>.
23. Johansson, F.D., Shalit, U., Kallus, N., and Sontag, D. (2022). Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *J. Mach. Learn. Res.* 23, 1–50.
24. Künzel, S.R., Stadie, B.C., Vemuri, N., Ramakrishnan, V., Sekhon, J.S., and Abbeel, P. (2018). Transfer learning for estimating causal effects using neural networks. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1808.07804>.
25. Wang, T., Ke, Y., Huang, Y., He, F., Zhong, W., Liu, S., and Ming, D. (2024). Using Semi-Supervised Domain Adaptation to Enhance EEG-Based Cross-Task Mental Workload Classification Performance. *IEEE J. Biomed. Health Inform.* 28, 7032–7039. <https://doi.org/10.1109/JBHI.2024.3452410>.
26. Zhang, Y., Gao, Y., Xu, J., Zhao, G., Shi, L., and Kong, L. (2024). Unsupervised Joint Domain Adaptation for Decoding Brain Cognitive States from fMRI Images. *IEEE J. Biomed. Health Inform.* 28, 1494–1503. <https://doi.org/10.1109/JBHI.2023.3348130>.

27. Chaddad, A., Lu, Q., Li, J., Katib, Y., Kateb, R., Tanougast, C., Bouridane, A., and Abdulkadir, A. (2023). Explainable, domain-adaptive, and federated artificial intelligence in medicine. *IEEE/CAA J. Autom. Sin.* **10**, 859–876.
28. Wu, X., Ju, X., Dai, S., Li, X., and Li, M. (2024). Multi-source domain adaptation for eeg emotion recognition based on inter-domain sample hybridization. *Front. Hum. Neurosci.* **18**, 1464431.
29. Ju, X., Su, J., Dai, S., Wu, X., Li, M., and Hu, D. (2025). Domain adversarial neural network with reliable pseudo-labels iteration for cross-subject eeg emotion recognition. *Knowl. Base Syst.* **316**, 113368.
30. Wang, R., Erus, G., Chaudhari, P., and Davatzikos, C. (2024). Adapting machine learning diagnostic models to new populations using a small amount of data: Results from clinical neuroscience. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2308.03175>.
31. Guo, L.L., Pfohl, S.R., Fries, J., Johnson, A.E.W., Posada, J., Aftandilian, C., Shah, N., and Sung, L. (2022). Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci. Rep.* **12**, 2726.
32. Dang, V.N., Cascarano, A., Mulder, R.H., Cecil, C., Zuluaga, M.A., Hernández-González, J., and Lekadir, K. (2024). Fairness and bias correction in machine learning for depression prediction across four study populations. *Sci. Rep.* **14**, 7848.
33. Kim, J.W., Yoon, H., Oh, W., Jung, D., Yoon, S.H., Kim, D.J., Lee, D.H., Lee, S.Y., and Yang, C.M. (2025). Domain adversarial training for mitigating gender bias in speech-based mental health detection. 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (IEEE), pp. 1–7.
34. Zhang, L., Richter, L.R., Wang, Y., Ostropelets, A., Elhadad, N., Blei, D.M., and Hripcsak, G. (2024). Causal fairness assessment of treatment allocation with electronic health records. *J. Biomed. Inform.* **155**, 104656.
35. Abi, R., and Joseph, J.E. (2024). Developing causal machine learning models in health informatics to assess social determinants driving regional health inequities and intervention outcomes. *Magna Sci. Adv. Biol. Pharm.* **13**, 113–129.
36. Cheong, J., Kuzucu, S., Kalkan, S., and Gunes, H. (2023). Towards Gender Fairness for Mental Health Prediction (International Joint Conferences on Artificial Intelligence Organization).
37. Zahid, A., Ali, A.R., Raza, S., Shah Nawaz, R., Kamiran, F., and Karim, A. (2025). Fairudt: Fairness-aware uplift decision trees. *Knowl. Base Syst.* **317**, 113068.
38. Wang, H., Ye, X., Wang, Y., Zhou, Y., Zhang, Z., Zhang, L., Jiang, J., and Zhai, Y. (2024). Uplift modeling based on graph neural network combined with causal knowledge. In 2024 IEEE Conference on Artificial Intelligence (CAI) (IEEE), pp. 1487–1492.
39. He, B., Weng, Y., Tang, X., Cui, Z., Sun, Z., Chen, L., He, X., and Ma, C. (2024). Rankability-enhanced revenue uplift modeling framework for on-line marketing. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5093–5104.
40. Carrard, V., Bourquin, C., Berney, S., Schlegel, K., Gaume, J., Bart, P.A., Preisig, M., Schmid Mast, M., and Berney, A. (2022). The relationship between medical students' empathy, mental health, and burnout: A cross-sectional study. *Med. Teach.* **44**, 1392–1399.
41. Amin, N., Salehin, I., Baten, M.A., and Noman, R.A. (2024). Rhmcd-20 dataset: Identify rapid human mental health depression during quarantine life using machine learning. *Data Brief* **54**, 110376.
42. Park, S.M., Jeong, B., Oh, D.Y., Choi, C.H., Jung, H.Y., Lee, J.Y., Lee, D., and Choi, J.S. (2021). Identification of major psychiatric disorders from resting-state electroencephalography using a machine learning approach. *Front. Psychiatry* **12**, 707581.
43. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**, 1–35.
44. Long, M., Cao, Z., Wang, J., and Jordan, M.I. (2018). Conditional adversarial domain adaptation. *Adv. Neural Inf. Process. Syst.* **31**.
45. Sun, B., and Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision* (Springer), pp. 443–450.
46. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–773.
47. Kyono, T., Bica, I., Qian, Z., and van der Schaar, M. (2023). Selecting treatment effects models for domain adaptation using causal knowledge. *ACM Trans. Comput. Healthc.* **4**, 1–29.
48. Sun, Y.X., Zhang, Y.L., Wang, W., Li, L., and Zhou, J. (2023). Treatment effect estimation across domains. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 2352–2361.
49. Shang, W., Li, Q., Qin, Z., Yu, Y., Meng, Y., and Ye, J. (2021). Partially observable environment estimation with uplift inference for reinforcement learning based recommendation. *Mach. Learn.* **110**, 2603–2640.
50. Kadioglu, S., and Michalsky, F. (2024). Fairness evaluation for uplift modeling in the absence of ground truth. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2403.12069>.
51. Meegahapola, L., Hassoune, H., and Gatica-Perez, D. (2024). M3bat: Unsupervised domain adaptation for multimodal mobile sensing with multi-branch adversarial training. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **8**, 1–30. <https://doi.org/10.1145/3659591>.
52. Fang, Y., Wang, M., Potter, G.G., and Liu, M. (2023). Unsupervised cross-domain functional mri adaptation for automated major depressive disorder identification. *Med. Image Anal.* **84**, 102707.
53. Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, 1989 (University of Chicago Legal Forum), pp. 139–167.
54. Ibragimov, B., and Vakhrushev, A. (2024). Uplift Modelling via Gradient Boosting. In *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (ACM)*, pp. 1177–1187. <https://doi.org/10.1145/3637528.3672019>.
55. Schröder, M., Frauen, D., and Feuerriegel, S. (2024). Causal fairness under unobserved confounding: A neural sensitivity framework. *International Conference on Learning Representations 2024*, 41986–42022.
56. Kim, K., and Zubizarreta, J.R. (2022). Toward fair and robust optimal treatment regimes. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.
57. Cinelli, C., and Hazlett, C. (2020). Making sense of sensitivity: extending omitted variable bias. *J. Roy. Stat. Soc. B* **82**, 39–67. <https://doi.org/10.1111/rssb.12348>.
58. Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning vol. 70 of ICML 2017 (PMLR)*, pp. 214–223.
59. Richter, M., Emden, D., Leenings, R., Winter, N.R., Mikolajczyk, R., Masag, J., Zwicky, E., Borgers, T., Redlich, R., Koutsouleris, N., et al. (2025). Generalizability of clinical prediction models in mental health. *Mol. Psychiatry* **30**, 3632–3639. <https://doi.org/10.1038/s41380-025-02950-0>.
60. Guerreiro, J., Garriga, R., Lozano Bagén, T., Šnajder, H., Karnik, N.S., and Matic, A. (2024). Transatlantic transferability and replicability of machine-learning algorithms to predict mental health crises. *npj Digit. Med.* **7**, 227. <https://doi.org/10.1038/s41746-024-00830-4>.
61. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Proceedings of the Advances in Neural Information Processing Systems* **32**.
62. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.
63. Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on*

- Knowledge Discovery & Data Mining, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, eds. (ACM), pp. 2623–2631.
64. Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701. <https://doi.org/10.1037/h0037350>.
 65. Devriendt, F., Moldovan, D., and Verbeke, W. (2018). A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data* 6, 13–41. <https://doi.org/10.1089/big.2017.0104>.
 66. Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton Mifflin).
 67. Maslach, C., and Jackson, S.E. (1981). The measurement of experienced burnout. *J. Organ. Behav.* 2, 99–113.
 68. Lazarus, R.S., and Folkman, S. (1984). *Stress, Appraisal, and Coping* (Springer).
 69. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. <https://doi.org/10.1145/2090236.2090255>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Identification of Major Psychiatric Disorders Dataset	Park et al. ⁴²	https://osf.io/8bsvr/
Medical Students Mental Health Dataset	Carrard et al. ⁴⁰	https://zenodo.org/records/5702895
Mental Health Depression During Quarantine Life Dataset	Amin et al. ⁴¹	https://data.mendeley.com/datasets/pxjmjyfdh2/1
Software and algorithms		
Python 3.10	Python Software Foundation	https://www.python.org
PyTorch 1.13	Paszke et al. ⁶¹	https://pytorch.org
scikit-learn	Pedregosa et al. ⁶²	https://scikit-learn.org
Optuna	Akiba et al. ⁶³	https://optuna.org
Code for this study	This paper	https://github.com/Nayem9274/domain-adaptive-uplift-mh

METHOD DETAILS

Notation and formal definitions

Potential outcomes and individual treatment effect

Following the Neyman–Rubin potential-outcomes framework,⁶⁴ let $T \in \{0, 1\}$ denote a binary treatment indicator, $Y(t)$ the potential outcome under treatment value t , and $\mathbf{x} \in \mathbb{R}^d$ a vector of pre-treatment covariates. The *Conditional Average Treatment Effect* (CATE) is

$$\text{CATE}(\mathbf{x}) = \mathbb{E}[Y(1) - Y(0) | \mathbf{X} = \mathbf{x}]. \quad (\text{Equation 1})$$

Because only one potential outcome is observed per subject, CATE cannot be computed directly from observed data alone.

Uplift

In uplift modeling,⁶⁵ the model estimates a plug-in approximation of CATE for binary outcomes:

$$\hat{u}(\mathbf{x}) = \hat{P}(Y = 1 | \mathbf{X} = \mathbf{x}, T = 1) - \hat{P}(Y = 1 | \mathbf{X} = \mathbf{x}, T = 0). \quad (\text{Equation 2})$$

A positive value indicates that the treatment is predicted to increase the outcome probability for individual $\mathbf{x}$; a negative value indicates the reverse. Individuals are ranked by $\hat{u}$ to prioritize those predicted to show the greatest differential vulnerability under the exposure proxy—that is, those most likely to deteriorate under treatment relative to control—who are therefore the primary targets for monitoring or low-intensity support.

Area under the uplift curve (AUUC)

The *uplift curve* plots the observed mean treatment effect as progressively larger fractions of the test set are targeted in descending order of $\hat{u}$. At targeting fraction f , the net uplift is

$$\text{NetUplift}(f) = \mathbb{E}[Y | T = 1, \hat{u} \geq q_f] - \mathbb{E}[Y | T = 0, \hat{u} \geq q_f], \quad (\text{Equation 3})$$

where q_f is the f -th upper quantile of $\hat{u}$. The AUUC is the area between this curve and the random-baseline curve (targeting without uplift information), computed via the trapezoidal rule over ten equal-fraction bins. A *positive* AUUC indicates that the model's ranking outperforms random selection; a *negative* AUUC indicates an inverted ranking—the model assigns high predicted uplift to individuals who actually benefit less.

Leave-one-domain-out (LODO) evaluation

In LODO evaluation, all samples from one study population are held out as the test set; the model is trained on the remaining populations and the held-out population's unlabeled features are optionally used for unsupervised domain alignment. This protocol approximates the deployment scenario in which a model trained on available populations is applied to a new population with no labeled outcome data.¹²

Domain-adversarial training

Domain-Adversarial Neural Networks (DANN⁴³) learn a feature representation that is simultaneously predictive of the outcome and invariant with respect to study population membership. A gradient-reversal layer (GRL) multiplies the gradient by $-\lambda_D$ during

backpropagation, causing the encoder to maximize the domain discriminator's loss while the discriminator itself is optimized normally. The scalar $\lambda_D \geq 0$ controls alignment strength; $\lambda_D = 0$ recovers the non-adapted baseline (NoDA).

Datasets and problem setup

We use three heterogeneous study populations that differ in population, context, and measurement: medical students (academic stress), a general-population sample collected during COVID-19 quarantine, and a clinical psychiatric sample spanning healthy controls and major disorders. The medical-student sample comprises cross-sectional survey data collected in Switzerland with demographics and psychometrics (e.g., burnout, anxiety, empathy).⁴⁰ The quarantine sample contains population-level survey responses collected in Bangladesh during the COVID-19 lockdown period, with stress-related indicators (e.g., frustration, habit change, mood swings).⁴¹ The psychiatric sample was collected in South Korea and includes diagnostic labels, cognitive scores, and EEG-derived features.⁴² A descriptive cross-dataset characterization of these three populations—including demographic heterogeneity analyses and cross-context comparisons—is provided in our prior published work.¹³ Table 1 summarizes the key characteristics of each study population; Table 2 maps the available features across datasets.

Data harmonization proceeded as follows. Column names were lowercased, numeric fields were imputed with the median and categorical fields with the mode, and remaining non-numeric columns were one-hot encoded. A single `ColumnTransformer` fit on the concatenated frame produced a common design matrix $X \in \mathbb{R}^{n \times d}$ across domains. Domain IDs $D \in \{0, 1, 2\}$ index the medical, quarantine, and psychiatric study populations respectively. For the quarantine dataset, the growing-stress item was originally recorded on a three-level scale (yes/maybe/no); both *yes* and *maybe* responses were mapped to $T = 1$ on the grounds that self-reported uncertainty about increasing stress still reflects a degree of perceived stress elevation, while *no* responses were mapped to $T = 0$. For fairness diagnostics we retain binary *gender* and an *education* bin (e.g., ≥ 12 years). Five medical students who identified as non-binary (sex = 3) were excluded prior to analysis to ensure valid binary gender encoding for the fairness diagnostic, yielding $n = 881$ retained participants in that sample and a total analytical sample of $n = 2,624$; the published descriptive analysis of these datasets¹³ reports $n = 886$ and $n = 2,629$ respectively, as it precedes this exclusion step.

Limitations of harmonization

Merging three surveys into a shared feature matrix entails several trade-offs that merit explicit acknowledgment. First, *missing-by-design* columns (Table 2) are imputed with the training-set median, which obscures domain-specific information: for example, the absence of trait-anxiety scores (STAI-T) in the quarantine and psychiatric datasets means the model cannot draw on this clinically important predictor when either of those populations is the test target. Second, one-hot encoding of categorical variables such as occupation in the quarantine dataset inflates feature dimensionality and may introduce spurious correlations that are absent in smaller cohorts. Third, pooled standardization using statistics computed across all three datasets may distort within-population distributions when cohorts differ substantially in means or variances—as is likely between self-report Likert-scale items and the EEG-derived continuous features in the psychiatric dataset. Future work should explore population-specific normalization strategies or modality-specific sub-encoders to mitigate these effects.

Construct validity and comparability

The three study populations differ not only in their feature distributions but in the *constructs* that the treatment and outcome proxies represent—a recognized challenge in multi-site machine-learning research.⁶⁶

Measurement heterogeneity

In the medical-student sample, the treatment proxy (burnout exhaustion at or above the 75th percentile) is a validated psychometric score derived from the Maslach Burnout Inventory,⁶⁷ offering relatively high measurement reliability. In the quarantine sample, the treatment proxy (growing stress) is a single dichotomous self-report item, which carries greater measurement noise and may conflate stress appraisal with stress exposure⁶⁸—a distinction that matters for causal identification. In the psychiatric sample, the treatment proxy is a clinical diagnosis—a qualitatively different type of variable that reflects a clinician-assigned category rather than a continuous self-report. These differences in proxy type and reliability limit the degree to which treatment effects are directly comparable across study populations.

Mitigation strategies

To partially address construct heterogeneity, we adopt three practices. First, all treatment and outcome variables are binarized to a common $\{0, 1\}$ scale, reducing measurement-scale differences. Second, the semi-synthetic benchmark (Uplift target and semi-synthetic ground truth section) injects a known treatment-effect signal on top of real features, allowing model behavior to be evaluated independently of proxy quality. Third, cross-dataset AUUC comparisons are treated as *ordinal rankings* of differential vulnerability rather than as absolute causal effect estimates.

Interpretation caveat

A positive AUUC in this study means that the model ranks individuals by *proxy-defined* differential vulnerability better than chance—it does not mean the model identifies individuals who would benefit from a real-world clinical intervention. Causal identification requires the assumptions of ignorability (no unmeasured confounding), overlap, and SUTVA (Stable Unit Treatment Value Assumption), none of which can be verified in these observational datasets. In the psychiatric dataset a further structural issue applies. The treatment proxy ($T = 1$: anxiety-spectrum diagnosis) and the outcome proxy ($Y = 1$: mood disorder or MDD) are both derived from the same diagnostic-label columns (`main.disorder` and `specific.disorder`), which are also retained as predictor features after

one-hot encoding. In the processed dataset, the two proxy categories are mutually exclusive under the present proxy construction: $\Pr(Y = 1|T = 1) = 0.000$ versus $\Pr(Y = 1|T = 0) = 0.344$. The model therefore has access to features that strongly encode the proxy relationship between T and Y , creating a structural shortcut that conflates proxy-label leakage with causal signal. We flag this as a proxy-design limitation specific to the psychiatric fold and return to it in Systematic λ_D analysis and diagnostics section and Limitations and Future Work subsection.

Uplift target and semi-synthetic ground truth

To evaluate individual treatment effects (ITE) under controlled conditions, we construct a semi-synthetic outcome. Each subject is assigned a Bernoulli treatment $T \in \{0, 1\}$, and the observed binary outcome $Y \in \{0, 1\}$ is sampled from a response surface that embeds a known ITE signal. This preserves the empirical feature distribution while providing a reference conditional average treatment effect

$$u^*(x) = \mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0]. \quad (\text{Equation 4})$$

Critically, this benchmark should be interpreted as a *model behavior check*: it tests whether the architecture can recover a known injected signal without systematic distortion, not whether the model achieves real-world causal identification. Ground-truth ITEs are unavailable in the observational datasets, and the assumptions required for causal identification (ignorability, overlap, SUTVA) cannot be verified in those datasets; we discuss this gap explicitly in Limitations and Future Work subsection. We report ranking-based uplift metrics on the observed outcomes (AUUC) and CATE error (RMSE) on the semi-synthetic benchmark.

Model: DANN_Uplift (domain-adversarial uplift network)

Our model comprises three components: a shared feature extractor $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^m$, two treatment-specific heads $h_1, h_0 : \mathbb{R}^m \rightarrow (0, 1)$ that predict potential outcomes, and a domain discriminator $g_\phi : \mathbb{R}^m \rightarrow \Delta^{K-1}$ connected via a Gradient Reversal Layer (GRL).

Encoder and potential-outcome heads

Unless stated otherwise, f_θ is a two-layer MLP ($128 \rightarrow 64$, ReLU) with dropout 0.30 after the first hidden layer. The potential-outcome heads are single linear layers with sigmoid activations:

$$\hat{y}^{(1)} = h_1(f_\theta(x)), \quad (\text{Equation 5})$$

$$\hat{y}^{(0)} = h_0(f_\theta(x)), \quad (\text{Equation 6})$$

$$\hat{u}(x) = \hat{y}^{(1)} - \hat{y}^{(0)}. \quad (\text{Equation 7})$$

For an observed tuple (x, t, y) the supervised loss is computed only on the exposed head:

$$\mathcal{L}_{\text{uplift}} = \text{BCE} \left(\underbrace{t \cdot \hat{y}^{(1)} + (1 - t) \cdot \hat{y}^{(0)}}_{y_{\text{obs}}}, y \right). \quad (\text{Equation 8})$$

Adversarial domain alignment and fairness regularizer

The discriminator receives features passed through a GRL and predicts the domain label d . Let $\mathcal{L}_{\text{dom}}$ denote the cross-entropy domain loss; the overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{uplift}} + \lambda_D \mathcal{L}_{\text{dom}}^{(\text{rev})} + \lambda_F \mathcal{L}_{\text{fair}}, \quad (\text{Equation 9})$$

where $\mathcal{L}_{\text{dom}}^{(\text{rev})}$ indicates the domain loss backpropagated through the GRL (i.e., maximizing domain confusion). The fairness term penalizes disparities in estimated uplift between protected subgroups. For uplift vector $u = \hat{y}^{(1)} - \hat{y}^{(0)}$ and binary masks $G \in \{0, 1\}$ (gender) and $E \in \{0, 1\}$ (education),

$$\mathcal{L}_{\text{fair}} = (\bar{u}_{G=0} - \bar{u}_{G=1})^2 + (\bar{u}_{E=0} - \bar{u}_{E=1})^2, \quad (\text{Equation 10})$$

evaluated only when both subgroup cells are present in the minibatch or, for stability, aggregated at epoch level.

Scope and limitations of the fairness metric

The fairness penalty in Equation 10 captures *group-mean uplift parity*—a first-order check that the average predicted benefit does not systematically differ between demographic groups. This is a necessary but not sufficient condition for equitable treatment allocation. The metric does not capture: (i) *intersectional* disparities (e.g., women with low education versus men with high education); (ii) distributional inequities within groups, such as differences in the variance or tail behavior of uplift scores rather than only the mean; or (iii) downstream allocation fairness, which depends on how ranked uplift scores are actually used by decision-makers.⁶⁹ We treat the current metric as a preliminary audit and note that more comprehensive fairness evaluation—including intersectional analysis and allocation-level metrics—is an important direction for future work; we return to this in Limitations and Future Work subsection.

Baselines and variants

All variants use the same encoder architecture (128–64, dropout 0.30). The compared methods are.

- **NoDA**: no domain alignment ($\lambda_D = 0$), standard two-head uplift network.
- **DANN**⁴³: domain-adversarial alignment with tuned fixed or scheduled $\lambda_D(p)$.
- **CDAN**⁴⁴: conditional adversarial alignment where the discriminator conditions on predictions by concatenating $[f(x), \hat{y}^{(0)}, \hat{y}^{(1)}]$.
- **CORAL**⁴⁵: non-adversarial covariance alignment that penalizes covariance mismatch between domain embeddings.
- **MMD**⁴⁶: kernel mean embedding alignment (MMD) between source and target embeddings.

All methods use Adam with weight decay in $[10^{-5}, 10^{-4}]$, gradient clipping at 1.0, and early stopping on validation AUUC. As shown in Figure 1, all methods use a common shared encoder backbone; what differs is the alignment mechanism, allowing a controlled comparison under identical conditions.

Evaluation protocol

We evaluate two transfer settings.

1. **Single-source transfer**: train on a single labeled source domain, adapt using unlabelled target batches, and evaluate on held-out target labels.
2. **Leave-one-domain-out (LODO)**: train on two domains and test on the third, rotating across folds.

The primary metric is AUUC (area under the uplift curve) on the target; we also plot uplift curves. On the semi-synthetic benchmark we compute CATE RMSE between $\hat{u}(x)$ and $u^*(x)$. Fairness is summarized by absolute uplift gaps $|\bar{u}_{G=0} - \bar{u}_{G=1}|$ and $|\bar{u}_{E=0} - \bar{u}_{E=1}|$ computed on treated subsets. Methods are compared using paired statistical tests on fold-wise AUUC (nonparametric Wilcoxon signed-rank tests with bootstrap confidence intervals and Cliff's δ where applicable).

Experimental details

Encoder selection

We evaluated several backbones (32–16, 64–32, 128–64) with dropout choices {0.0, 0.30}; the 128–64 backbone with dropout 0.30 yielded the most stable positive mean AUUC under LODO and was adopted as default.

Adversarial weight protocol

To ensure reproducibility we followed a systematic protocol for selecting λ_D : (i) dense grid sweeps of λ_D over a plausible interval to visualize AUUC sensitivity per fold; (ii) selection of per-fold optimal λ_D and a shared λ_D that maximizes mean AUUC; (iii) evaluation of scheduled ramps (linear and sigmoid) as principled alternatives to fixed values.

Diagnostics

The psychiatric study population was the most fragile target: large λ_D values often led to negative AUUC (uplift inversion). However, subsequent proxy checks showed that this inversion is not attributable to domain adaptation alone. In the psychiatric dataset, the treatment and outcome proxies are derived from diagnostic-label fields that are also retained as input features, creating a structurally negative proxy relationship that can induce shortcut learning. We therefore interpret the psychiatric fold as an informative stress test of proxy design rather than as a clean measure of adaptation failure.

Hyperparameter search and reporting

We varied the adversarial weight λ_D , the fairness penalty weight λ_F , and the learning rate (lr), among others. Specifically, λ_D was swept across a dense grid (0.1, 0.2, ..., 0.8), and both linear and sigmoid schedules for λ_D were tested. The fairness weight λ_F was also varied via grid search, and we tried several learning rates. The hyperparameter search ranges used across all domain adaptation methods are summarized in Table 5.

QUANTIFICATION AND STATISTICAL ANALYSIS

The primary performance metric is the area under the uplift curve (AUUC), computed via the trapezoidal rule over ten equal targeting fractions on the held-out domain under leave-one-domain-out (LODO) evaluation. A positive AUUC indicates that the model's uplift ranking outperforms random targeting; a negative AUUC indicates an inverted ranking. On the semi-synthetic benchmark, model accuracy is quantified by CATE root-mean-square error (RMSE) between the predicted uplift $\hat{u}(x)$ and the known injected treatment effect $u^*(x)$.

All adaptation methods were evaluated across 5 random seeds $\times$ 3 held-out folds, yielding $n = 15$ paired seed $\times$ fold observations per method comparison. Pairwise method comparisons use the two-sided Wilcoxon signed-rank test on fold-wise AUUC. Effect sizes are reported as Cliff's δ . Bootstrap 95% confidence intervals (1,000 resamples) are reported for the median paired difference. No corrections for multiple comparisons were applied given the exploratory nature of the comparisons; all p -values and confidence intervals are reported as-is. The threshold for statistical significance was set at $\alpha = 0.05$; no comparison met this threshold.

Fairness is summarized by the absolute mean uplift gap between demographic subgroups (gender; education) among treated individuals within each fold. No formal significance testing was applied to fairness gaps given the small subgroup sample sizes; these values are reported as descriptive audit statistics.

Full statistical details—including exact n per fold, AUUC values per seed and fold, confidence interval bounds, Wilcoxon p -values, Cliff's δ , and CATE-RMSE values across runs—are reported in the Results section and Tables.