CrossShift: Quantifying Interpersonal Differences in Mobile Sensing for Mental Health

PANYU ZHANG, KAIST, Republic of Korea
MINSEO PARK, KAIST, Republic of Korea
TOMIRIS ISMATZODA, KAIST, Republic of Korea
AZIZBEK MUSTAFAKULOV, HumbleBeeAI, Republic of Korea
UZAIR AHMED, KAIST, Republic of Korea
OTABEK NAJIMOV, HumbleBeeAI, Republic of Korea
JUMABEK ALIKHANOV, Gachon University, Republic of Korea
SURJYA GHOSH, BITS Pilani, India
UICHIN LEE*, KAIST, Republic of Korea

Mobile sensing systems for mental health leverage ubiquitous behavioral data to monitor states like stress and anxiety, enabling just-in-time interventions and context-aware care. However, the real-world scalability of these systems is severely limited by poor cross-user generalization—a challenge often attributed to interpersonal differences but rarely quantified systematically. We present CrossShift, a structured empirical framework for decomposing and measuring user-level covariate, label, conditional, and concept shifts in mobile mental-health sensing. CrossShift unifies shift testing, shift quantification, model evaluation, and shift-to-generalization attribution within a common analytical framework tailored to this domain. We evaluate our framework across eight diverse datasets, comprising three newly collected and five public mobile sensing repositories. Our analysis reveals that phone usage, sleep and social-interaction features are the most consistent sources of covariate shift across users. Crucially, we find that concept shift—the variation in the mapping from behavior to mental state—is the primary driver of performance degradation. Our findings suggest that scalable mobile mental health sensing requires moving beyond one-size-fits-all models toward shift-aware system design. We further discuss how CrossShift can support out-of-distribution user detection and selective personalization in real-world deployments.

CrossShift codebase: https://github.com/Kaist-ICLab/CrossShift

CCS Concepts: • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; *Empirical studies in HCI*; • **Computing methodologies** → *Machine learning*; • **Applied computing** → *Health informatics*.

Additional Key Words and Phrases: Affective Computing, Health-Wellbeing, Personalization, Diversity, Mobile Sensing

*Corresponding author.

Authors' Contact Information: Panyu Zhang, KAIST, Daejeon, Republic of Korea, panyu@kse.kaist.ac.kr; Minseo Park, KAIST, Daejeon, Republic of Korea, tim0726@kaist.ac.kr; Tomiris Ismatzoda, KAIST, Daejeon, Republic of Korea, tomiris@kaist.ac.kr; Azizbek Mustafakulov, HumbleBeeAI, Incheon, Republic of Korea, azizbek@humblebee.ai; Uzair Ahmed, KAIST, Daejeon, Republic of Korea, uziahmd@kaist.ac.kr; Otabek Najimov, HumbleBeeAI, Incheon, Republic of Korea, otabek@humblebee.ai; Jumabek Alikhanov, Gachon University, Seongnam, Gyeonggi-do, Republic of Korea, jumabek@humblebee.ai; Surjya Ghosh, BITS Pilani, Zuarinagar, Goa, India, surjyag@goa.bits-pilani.ac.in; Uichin Lee (corresponding author), KAIST, Daejeon, Republic of Korea, uclee@kaist.ac.kr.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2474-9567/2026/9-ART189
https://doi.org/10.1145/3831954

ACM Reference Format:
Panyu Zhang, Minseo Park, Tomiris Ismatzoda, Azizbek Mustafakulov, Uzair Ahmed, Otabek Najimov, Jumabek Alikhanov, Surjya Ghosh, and Uichin Lee. 2026. CrossShift: Quantifying Interpersonal Differences in Mobile Sensing for Mental Health. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 189 (September 2026), 40 pages. https://doi.org/10.1145/3831954

1 INTRODUCTION

Recent advances in mobile technology have improved our ability to monitor and understand human behavior and well-being. Contemporary smartphones and wearable devices integrate various sensors that capture social interactions, phone usage patterns, physical activity, and mobility data. When these rich data streams are combined with self-reported measures, such as stress levels, they provide a continuous, multimodal record that researchers can leverage to model and predict stress [75].

In human-centered computing and mobile computing, mobile sensing for mental health (and mobile human sensing, more generally) is now routinely embedded in personal informatics tools, intervention pipelines, and emotion-support agents [31, 35, 43, 44, 66]. Their impact depends on the generalizability of the model, which governs when interventions are triggered, how feedback is posed, and whether users trust and act on the recommendations. However, prior models have struggled to generalize to unseen users in real-world, in-the-wild settings [55, 87, 95]. Xu et al. attributed this restricted generalizability to substantial differences between individuals [3], that is, to data shifts across users. Prior work has mainly conceptualized users or cohorts as domains (e.g., cultural, demographic, or protocol differences) [9, 36, 53, 55, 59, 62], and thus framed the challenge as cross-dataset or cross-population shifts [87]. However, even within a single dataset, user-independent generalization remains difficult: under leave-one-subject-out evaluations, performance remains limited [38, 55, 87, 95].

In general, data shift is commonly decomposed into four forms: *covariate shift*, *label shift*, *conditional shift*, and *concept shift* [4, 51]. So far, the majority of prior studies on mobile sensing have only focused on covariate and label shift [38, 56, 87]. In this work, we systematically investigate the impact of four types of shifts to deepen our understanding of cross-user shifts. As illustrated in Figure 2, given that X denotes features while Y denotes labels, the four shifts are defined as changes in: (1) the marginal feature distribution: P(X) (i.e., covariate shift), (2) the label distribution: P(Y) (i.e., label shift) (3) the concept mapping or predictive distribution: $P(Y \mid X)$ (i.e., concept shift) (4) the conditional feature distribution: $P(X \mid Y)$ (i.e., conditional shift). These shifts often co-occur across individuals in real-world sensing datasets, posing challenges for robust generalization. Identifying which components have changed is crucial for robust cross-user generalization and for selecting appropriate mitigation or adaptation strategies.

Moreover, existing studies in mobile sensing, particularly those on mental health, often make assumptions about distribution shifts and their impact on model performance based on findings from computer vision (CV) and natural language processing (NLP) [56, 85, 87]. However, the implications and conclusions drawn from these domains may not generalize well to mobile sensing contexts due to fundamental differences in data characteristics and prediction targets. In computer vision settings, for example, fixed input features in X-ray images may consistently indicate the same diagnosis of lung cancers across individuals. In contrast, mental health prediction based on behavioral features derived from mobile sensor data is highly individualized and context-dependent, making such assumptions less reliable, as illustrated in Figure 2. To the best of our knowledge, no prior work has systematically investigated the impact of distribution shifts on model performance within the domain of mobile sensing for mental health.

Based on the identified research gaps, we pose the following research questions:

- **RQ1** What interpersonal differences are present in mobile sensing datasets as revealed by statistical tests?
- **RQ2** Which interpersonal differences contribute most to failures of cross-user model generalization?

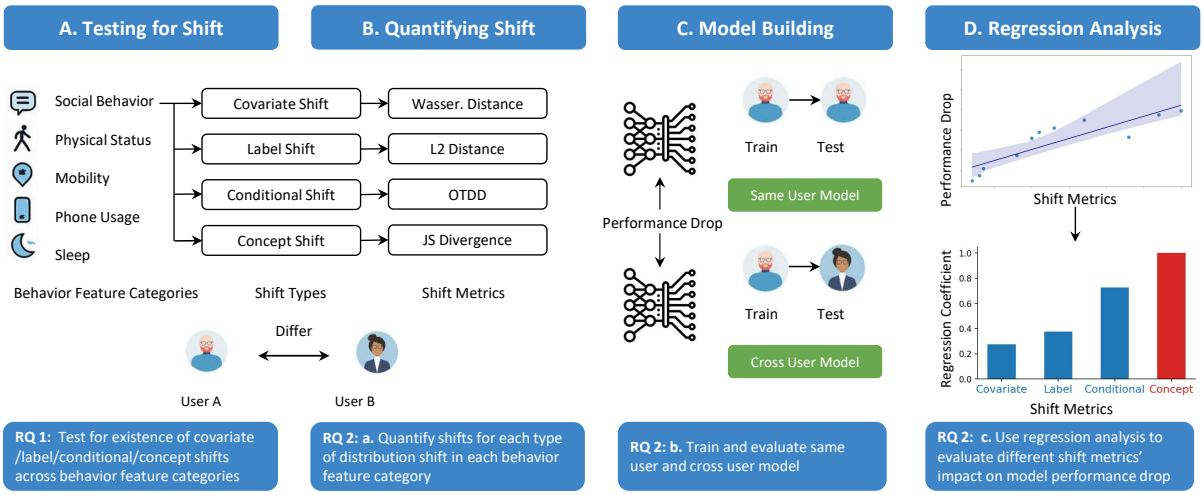


Fig. 1. Overview of the CrossShift analytical framework. CrossShift connects four components: (A) statistical testing for user-level covariate, label, conditional, and concept shifts; (B) pairwise shift quantification across behavior feature categories; (C) within-user and cross-user model evaluation; and (D) regression-based attribution linking shift magnitudes to transferability loss.

To address these questions, this paper makes three primary contributions through experiments on three newly collected datasets and five publicly available datasets, involving 696 valid users. To support both controlled and heterogeneous evaluation, we analyze a family of three comparable in-the-wild cohorts collected by our institution together with five public datasets that vary more substantially in cohort composition, sensing modalities, and collection protocols.

- We operationalize CrossShift as an end-to-end empirical analysis pipeline that connects four components: statistical shift testing, metric-based shift quantification, user-level model evaluation, and regression-based attribution of transferability loss. This organization enables covariate, label, conditional, and concept shifts to be compared under a common framework for mobile mental-health sensing.
- We identify ***Phone Usage***, ***Sleep*** and ***Social Interaction*** as the behavior feature categories exhibiting the greatest between-user diversity.
- We show that ***Concept Shift***, representing differences in the mapping from behavior features to the mental health labels, is the primary factor, limiting cross-user generalizability of current models.

CrossShift provides a structured way to characterize multiple forms of user-level distribution shift across behavioral categories and to relate them to cross-user generalization failures. The released codebase serves as an implementation of this analytical framework. The resulting analyses can inform future work on out-of-distribution (OOD) user detection and selective personalization or adaptation.

2 BACKGROUND AND RELATED WORK

2.1 Behavior Features in Mobile Human Sensing

The following dimensions are commonly included in behavior features for mobile human sensing.

- **User Demographics**: A foundational layer for profiling, this category includes basic information such as age and gender, as well as psychological traits, including the Big Five personality dimensions. Baseline

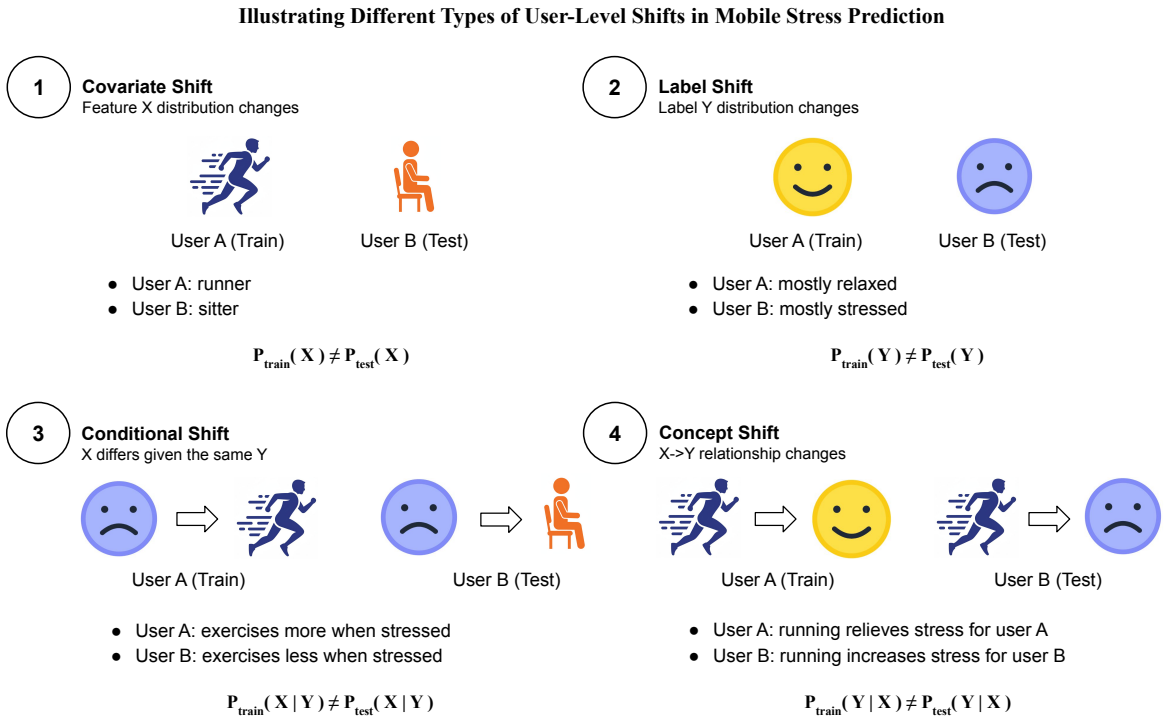


Fig. 2. Types of distribution shifts across people in mobile stress prediction. Using physical activity (X) to predict stress (Y) as an example, the figure illustrates four common user-level shift types: (1) *Covariate shift*, where the feature distribution changes across users ($P_{\text{train}}(X) \neq P_{\text{test}}(X)$); (2) *Label shift*, where the label distribution changes ($P_{\text{train}}(Y) \neq P_{\text{test}}(Y)$); (3) *Conditional shift*, where the distribution of features differs given the same label ($P_{\text{train}}(X \mid Y) \neq P_{\text{test}}(X \mid Y)$); and (4) *Concept shift*, where the relationship between features and labels changes across users ($P_{\text{train}}(Y \mid X) \neq P_{\text{test}}(Y \mid X)$).

mental health status is often captured using validated scales like the GHQ-12 and PSS-10, as well as measures of self-efficacy, optimism, hope, and resiliency [46, 56].

- **Phone Usage Behavior**: This category analyzes a user's digital footprint. It includes AppUsageEvent data, which tracks the count and duration of use for each application category, and ScreenEvent data for screen-on duration and frequency. KeyEvent data, measuring typing distance and time, is also utilized [20, 44, 56, 77, 87].
- **Social Behavior**: Social interaction patterns are quantified through communication logs. This involves analyzing CallEvent data for the duration and count of calls and MessageEvent data for the count of messages sent and received [1].
- **Physical Status**: This category focuses on physical activity and exertion. It includes the count and duration of different activity types (calculated from PhysicalActivityTransition), total step count, and distance traveled. Estimated calories consumed are also sometimes included [46, 56, 67, 87].
- **Mobility**: GPS and location data are used to understand a user's movement patterns. Key features include the different places of visit and the time spent at various types of clustered locations (e.g., home, work) [46, 56, 57, 87].

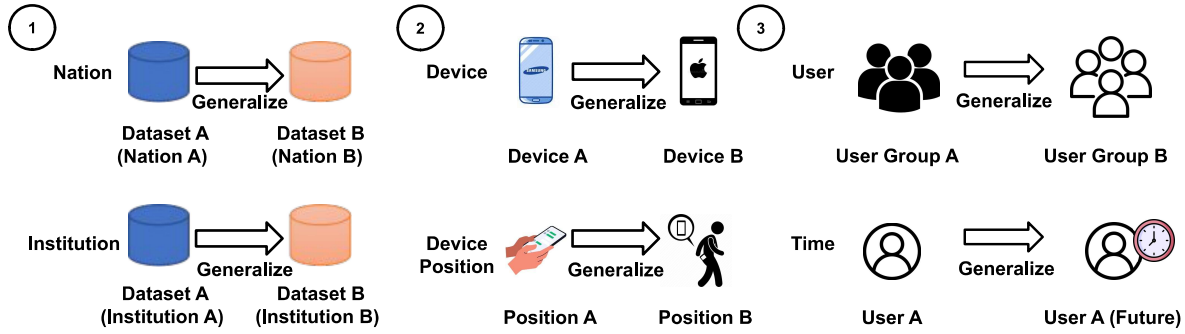


Fig. 3. Types of generalizability: 1) Cross-dataset generalizability: robustness across datasets that differ by country, culture, institution, cohort, or data-collection protocol 2) Device-related generalizability (within same dataset): robustness across device types (e.g., phone/wearable models) and device positions/placements. 3) User-related generalizability (within the same dataset): robustness across users (cross-subject) and across time within the same user (temporal drift).

- **Sleep**: Sleep patterns are a critical component of well-being. Profiles often incorporate daily metrics such as the previous night's sleep duration and sleep onset time, which can be inferred from phone activity [1, 46, 87].

Note that our analysis focused primarily on behavioral variables and excluded demographic factors, as the latter remain constant for each individual.

2.2 In-the-Wild Mobile Sensing for Mental Health

In-the-wild sensing aims to continuously infer users' states in naturalistic conditions, rather than in lab-controlled experiments. However, building reliable in-the-wild models remains challenging due to uncontrolled contexts, varied sensor placements, and strong inter-individual variability [68]. One important aspect of this generalization challenge is the evaluation setting. While many studies adopt a user-dependent setup by training and testing on data from the same individuals, performance often drops substantially in user-independent setups, where models are evaluated on unseen users [30]. Sano et al. analyzed surveys, personality traits, and phone/wearable signals from 18 participants and evaluated models with repeated 10-fold cross-validation [75]. In a participant-level setup (features aggregated per person), they reported 87.5% accuracy for the post-survey baseline (and comparably for sensor-feature models), 81.25% with personality, evening-survey, and call features, and over 75% with screen-on, mobility, call, or activity features—all under a user-dependent evaluation where part of unseen user data is used for model building. Ciman et al. modeled stress from human–smartphone interaction (no self-reports as inputs) [17]. With a 10-fold CV per user, within-subject F-measure is around 0.79–0.85; with a generalized model evaluated by leave-one-subject-out (LOSO), F-measure drops to around 0.70–0.80, making the user-dependent vs. user-independent gap explicit. Recently, Yu et al. reported a macro F1 score of 63.0% using 5-fold group cross-validation (in a user-independent setting) [93]. Toshnazarov et al. achieved 65.8% F1 score; they used physiological signals under LOSO cross-validation (in a user-independent setting) and a pretrained model selected from their best in-lab configuration [80]. Zhang et al. [95] observed 58% AUC-ROC in LOSO. Overall, cross-user evaluation performance, even within a single dataset, remains limited.

2.3 Generalizability Studies in Mobile Computing

As illustrated in Figure 3, generalizability in mobile computing is typically examined along three axes.

- **Cross-dataset generalizability**: datasets often differ by country, culture, institution, or collection protocol, motivating evaluations of cross-dataset generalization and domain adaptation or generalization techniques [9, 13, 36, 53, 55, 56, 59, 69, 81, 92].
- **Cross-device and device-position robustness within a dataset**: devices differ in sensing characteristics and wear/placement locations vary across measurement contexts, motivating robustness techniques to handle sensor and placement variability [14, 52, 96].
- **User-related generalizability**: models trained on data from some participants may not generalize to others (cross-user variability), and even for the same individual, behavior and context evolve over time (temporal drift), leading to performance degradation. These challenges motivate personalization, longitudinal adaptation, and drift-aware modeling [25, 26, 37, 38, 86, 89, 92].

Despite progress, most work proposes remedies without first testing for the presence of distribution shifts and characterizing the specific shifts responsible for generalization failures. Notable exceptions include Nanchen et al. [62], who analyzed covariate shift across datasets; and Xu et al. [87] and Meegahapola et al. [55], who—though not focused on shift analysis—tested for covariate and label shifts.

To our knowledge, there is a lack of systematic investigation of the *relative impact* of different shift types on generalization performance in mobile sensing for mental health. This gap is consequential because many domain generalization and adaptation methods address only a subset of shifts; without prior characterization, mismatches between the method and the underlying shift can limit effectiveness.

2.4 Shift Analysis in Machine Learning Studies

Work in ML tests for distribution shifts and quantifies their impact on performance [16, 24].

For **testing**, Kellis et al. [41] recommended the following: (i) plot outcome distributions for label shift P(Y), (ii) compare feature means, (iii) apply kernel two-sample tests (e.g., Maximum Mean Discrepancy (MMD) [28]), and (iv) train domain-discrimination classifiers for P(X), P($X \mid Y$), and P($Y \mid X$). In mobile human sensing, Nanchen et al. used Permutational Multivariate Analysis of Variance (PERMANOVA) and Permutational Multivariate Analysis of Dispersion (PERMDISP) for multivariate covariate shift [62]; Meegahapola et al. applied t-tests [55]; and Xu et al. trained domain/user classifiers [87].

For **quantification**, Gardner et al. proposed optimal transport dataset distance (OTDD) for covariate shift, Euclidean distance ℓ_2 for label shift, and Fréchet dataset distance (FDD) for conditional/concept shift [24]; Cheng et al. examined correlation, mutual information, and feature-importance surrogates [16]; Polo et al. suggested Jensen-Shannon (JS) divergence for all four shift types [54]. In sensing, Meegahapola et al. used Cohen's d on feature means [55, 56]; Kang et al. measured concept drift via the performance gap between generalized and personalized models [38].

For **impact**, Gardner et al. related OOD performance to in-distribution accuracy and shift magnitudes via correlation or regression [24]. Cheng et al. analyzed the drop fraction versus the magnitude of distribution shift (Δshift) and observed larger degradation when removed or perturbed features are more label-associated [16]. In follow-up work, Cheng et al. reported that TabPFN, a tabular foundation model, shows limited robustness to concept shift; moreover, XGBoost remains the best-performing model under distribution shift in tabular settings [15].

In summary, contemporary studies on shift analysis (i) characterize/quantify *what* changed; (ii) model *how* those changes predict performance loss under deployment conditions. Applying shift analysis to mobile sensing would help characterize user heterogeneity and improve cross-user generalization, but the relative contributions of each shift type still remain unknown.

3 CROSSSHIFT FRAMEWORK

CrossShift is an analytical framework for characterizing interpersonal distribution shifts and relating them to cross-user generalization failures. As illustrated in Figure 1, the framework connects four components: statistical shift testing, shift quantification, user-level model evaluation, and shift-to-generalization attribution. In this section, we describe these components at the framework level. Study-specific datasets, feature construction, preprocessing, and implementation details are presented later in Section 4.

CrossShift is intended as a structured empirical framework for user-level shift analysis in mobile mental-health sensing. While its components draw on established tools from prior work in machine learning and sensing, their integration, domain-specific operationalization, and multi-dataset empirical evaluation are central contributions of this paper.

3.1 Overview and Notation

In CrossShift, each user is treated as a domain. Let X denote the behavioral feature representation and Y denote the prediction target. The framework considers four types of user-level distribution shift: covariate shift ($P(X)$), label shift ($P(Y)$), conditional shift ($P(X \mid Y)$), and concept shift ($P(Y \mid X)$). For the shift-attribution component, we consider a source user a and a target user o, and denote by $d_{a,o}$ the pairwise shift magnitude between them under a given shift definition. We then relate this distance to a transferability loss $\Delta_{a \to o}$, which measures the degradation from within-user to cross-user performance.

3.2 Shift Testing

For the shift-testing component, CrossShift treats each user as a group/domain and asks whether four types of distribution shift are present across users: covariate shift, label shift, conditional shift, and concept shift. Different statistical tests are used depending on the probabilistic object of interest. Covariate and conditional shifts are evaluated using multivariate permutation-based procedures, whereas label and concept shifts are evaluated using association-based tests. The goal of this module is to determine which forms of interpersonal difference are statistically detectable before moving to shift quantification and performance attribution.

3.2.1 Statistical Test. Guided by Nanchen et al. [62], we employ permutation-based multivariate procedures for *covariate* and *conditional* shift, and Pearson's χ^2 tests for *label* and *concept* shift. Unless noted otherwise, analyses are conducted for each predefined feature set on the corresponding multivariate feature representation (Figure 4).

- **Covariate shift** (changes in $P(X)$). For each behavior feature category, we apply PERMANOVA [7] (Permutational Multivariate Analysis of Variance) on an appropriate dissimilarity (Euclidean distance for standardized features), reporting R^2 and permutation p-values.
  In PERMANOVA [7], the distance-based effect size for a grouping factor is often reported as $R^2 = \frac{SS_B}{SS_T}$, where SS_B is the between-groups sum of squares and SS_T is the total sum of squares computed from the chosen inter-sample dissimilarity matrix. This $R^2 \in [0, 1]$ summarizes the proportion of total distance variation attributable to differences in group centroids under the specified distance measure, in our case Euclidean distance. In our context, since a single user's multiple data is treated as a group, a larger R^2 implies greater between-user differences within the dataset. We also run PERMDISP (Permutational Analysis of Multivariate Dispersions) to probe differences in dispersion. PERMANOVA tests whether group centroids in multivariate space differ significantly under a chosen dissimilarity measure, whereas PERMDISP complements this by testing whether observed differences arise from group dispersion (variance) rather than centroid location. Departing from Nanchen et al. [62], we treat dispersion differences as substantive inter-individual variability rather than a nuisance and report them explicitly.

- **Conditional shift** (changes in $P(X \mid Y)$). We stratify the data by the binary label $Y \in \{0, 1\}$ and repeat the PERMANOVA and PERMDISP procedure within each stratum for each behavior feature category.
- **Label shift** (changes in P(Y)). We test for changes in the marginal label distribution across users using a Pearson χ^2 test of independence on contingency tables of Y by user (reporting Cohen's w as an effect size).
- **Concept shift** (changes in $P(Y \mid X)$). For each behavior feature category, we discretize the multivariate features into two interpretable levels via K=2 clustering; the cluster with the higher overall mean is labeled "high" and the other "low." We chose $K = 2$ primarily for interpretability, because this analysis is intended to compare a coarse low/high behavioral state across users rather than optimize unsupervised clustering. Alternative values of K may capture finer-grained concept structure, but would reduce interpretability and comparability across datasets. We then assess the association between Y and the resulting low/high feature level using a Pearson χ^2 test. Differences in this association across user groups provide evidence of changes in $P(Y \mid X)$ (concept shift).

This protocol provides a coherent and interpretable foundation for detecting distinct shift mechanisms. Study-specific feature construction, preprocessing choices, and normalization regimes used to instantiate this testing module are described later in Section 4.

3.3 Shift Quantification Metrics

After testing whether user-level shifts are statistically detectable, CrossShift quantifies the magnitude of pairwise user differences for each shift type. Rather than collapsing user heterogeneity into a single aggregate score, this component computes separate distance measures for covariate, label, conditional, and concept shifts, enabling later attribution analyses to examine which types of shift are most associated with transferability loss.

Given per-user feature representations and labels, CrossShift uses metrics aligned with the probabilistic definition of each shift type (Table 1). These metrics adapt prior shift-analysis tools to the high-dimensional and heterogeneous structure of mobile sensing data.

Covariate shift quantification ($P(X)$). We quantify covariate shift using the Wasserstein distance between users' marginal feature distributions $P(X)$. We use Wasserstein distance rather than OTDD because OTDD

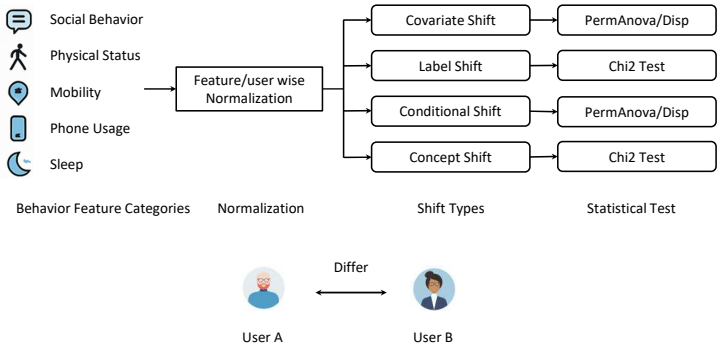


Fig. 4. Overview of RQ1 user-level distribution shift analysis. Behavioral signals from different behavior feature categories are mapped to distinct types of distribution shift (i.e., covariate, label, conditional, and concept), and each shift is evaluated using an appropriate statistical test to assess differences between users. Unless otherwise specified, all analyses use *feature-wise normalization* as the default preprocessing step to ensure that distance-based statistical tests are not dominated by scale differences. We additionally conduct experiments with *user-wise normalization* to examine its impact on shift quantification.

Table 1. Shift quantification framework. We adopt standard distance-based measures for covariate and label shifts, quantified using Wasserstein distance on $P(X)$ and ℓ_2 distance on $P(Y)$. For conditional shift, we extend prior OTDD-based formulations by introducing a feature-importance–normalized conditional term to improve robustness under high-dimensional mobile sensor features. For concept shift, we operationalize differences in $P(Y \mid X)$ using a model-based formulation tailored to mobile sensing: we train user-specific models M_s and M_t on source and target users, respectively, and compute Jensen–Shannon divergence between their predicted label distributions $P_{M_s}(Y \mid X)$ and $P_{M_t}(Y \mid X)$ evaluated on the same pooled sensor inputs from both users. This design provides a stable, distributional proxy for concept shift while avoiding direct estimation of $P(Y \mid X)$, which is unreliable in sparse, high-dimensional sensing data.

Behavior Feature Categories	**Shift Type**	**Shift Metric**
Social Interaction, Sleep, Physical Status, Mobility, Phone Usage	Covariate	Wasserstein distance on $P(X)$ [63]
	Label	ℓ_2 distance on $P(Y)$ [24]
	Conditional	Feature-importance–normalized OTDD conditional term [6, 24]
	Concept	Jensen–Shannon divergence on $P_{M_s}(Y \mid X)$ vs. $P_{M_t}(Y \mid X)$ [54]

jointly captures discrepancies in both $P(X)$ and $P(X \mid Y)$ [24], whereas our goal here is to isolate marginal feature-distribution differences.

Label shift quantification ($P(Y)$). We quantify label shift using the ℓ_2 distance between users' marginal label distributions $P(Y)$. This directly captures differences in label prevalence and is stable for low-dimensional label distributions [24]; unlike KL-based measures, it does not become ill-behaved when one label probability is close to zero.

Conditional shift quantification ($P(X \mid Y)$). We quantify conditional shift using the conditional component of OTDD, which measures discrepancies between label-conditioned feature distributions. To make this measure more robust for high-dimensional mobile sensing features, we use a feature-importance-normalized variant of the OTDD conditional term, reducing the influence of marginal feature-scale effects.

Concept shift quantification ($P(Y \mid X)$). We quantify concept shift using a predictive-function-based formulation, since directly estimating $P(Y \mid X)$ is unstable in sparse, high-dimensional sensing data [50, 72]. We train a source model M_s on pooled source-user data and a target model M_t on held-out target-user data, evaluate both on the same combined feature set $X_s \cup X_t$, and compute the Jensen–Shannon divergence between their predicted label distributions [54]. Unlike embedding-similarity formulations such as FDD [24], this predictive-distribution view targets user-level changes in $P(Y \mid X)$ and separates concept shift from conditional shift, which is important in mobile mental-health sensing where these two shifts can otherwise appear empirically entangled.

This decomposition supports more precise attribution of cross-user performance degradation to specific shift types. Study-specific preprocessing pipelines and feature representations are described in Section 4.

3.4 Model Evaluation and Shift-to-Generalization Attribution

After quantifying pairwise user-level shifts, CrossShift relates these shifts to cross-user model degradation. As illustrated in Figure 5, each user $u \in \mathcal{U}$ is treated as a domain. For a source user a, we train a personalized model f_a using only a's training data and record its within-user performance as $\mathrm{AUROC}_{a,\mathrm{self}}$. We then apply the same model to each target user $o \neq a$ and record the cross-user performance $\mathrm{AUROC}_{a,o}$. The transferability loss is defined as

$$\Delta_{a \to o} = \mathrm{AUROC}_{a,\mathrm{self}} - \mathrm{AUROC}_{a,o},$$

which measures the degradation from within-user to cross-user evaluation. To attribute this degradation to specific shifts, CrossShift fits a linear model of the form $\Delta_{a \to o} = \beta d_{a,o} + \varepsilon$, where $d_{a,o}$ is the pairwise distance for a given shift metric. The coefficient β estimates how strongly that shift metric is associated with transferability loss. Study-specific model choices, prediction targets, and preprocessing details are described in Section 4.

4 METHODOLOGY

This section describes how the CrossShift framework is instantiated in our study, including the datasets, feature construction, preprocessing, and experimental procedures used for RQ1 and RQ2.

4.1 Datasets

We analyze eight in-the-wild mobile sensing datasets in total. These datasets play two complementary roles in the paper. First, D#1–D#3 provide a relatively controlled family of comparable cohorts collected by our institution under closely related study designs, allowing us to examine whether the main shift patterns persist even when major cohort-level confounds are reduced. Second, five public datasets provide broader external heterogeneity in cohort composition, sensing availability, and collection protocols, allowing us to assess whether the same observations extend beyond our own collection context. D#1–D#3 have also recently been released as open datasets[1], expanding the set of publicly available in-the-wild cohorts for future research in mobile mental-health sensing.

Dataset construction vs. analytic eligibility. Across datasets, we distinguish between *dataset-level quality control* and *analysis-specific eligibility filtering*. Dataset-level quality control determines which participants and data streams are retained after collection and preprocessing. Analysis-specific eligibility filtering is then applied later depending on the research question. In particular, RQ2 introduces additional per-user eligibility requirements for personalized modeling and transferability analysis. We report these stages separately to make the construction of each analytic cohort explicit.

Collected datasets (D#1-D#3). We analyze three in-the-wild mobile sensing cohorts collected annually from 2020 to 2022 at our institution. Each cohort followed a broadly comparable four-week study design with university students, using participants' own Android smartphones together with wearable sensors and dense Experience Sampling Method (ESM) based in-situ labeling. Across D#1, D#2, and D#3, the core design was intentionally kept similar with overlapping smartphone sensing streams, Fitbit-based wearable sensing, and repeated emotion self-reports while later waves progressively refined the protocol and sensing coverage. This makes D#1-D#3 useful as a controlled family of related cohorts for testing whether shift patterns replicate across comparable real-world collections rather than arising only from highly heterogeneous public datasets.

D#1, D#2, and D#3 include 102, 112, and 114 enrolled participants, respectively. After dataset-level quality control, 92, 99, and 106 participants remain, with 10,259, 21,042, and 21,838 ESM responses retained after QC, respectively. Dataset-level quality control removed participants with insufficient data coverage or ESM compliance and excluded streams with inadequate collection coverage for reliable analysis. All three datasets were collected under IRB-approved procedures; participants provided written informed consent, participation was voluntary, and participants received approximately USD 100 compensation. Fitbit Inspire HR was used in all three datasets. For more details of study design, protocol differences, and retained counts, please refer to Appendix A.1 and Table 2.

Public datasets. To assess whether our findings generalize beyond our original cohorts, we additionally analyze five public in-the-wild datasets: K-EmoPhone [39], GLOBEM [88], CrossCheck [10], StudentLife [82], and the

[1] https://doi.org/10.7910/DVN/QGRVU8

Table 2. Overview of institution-collected in-the-wild datasets (D#1–D#3) and public benchmarks used in this study. For RQ1, we perform statistical analyses using hourly aggregated features. Predictive modeling (RQ2) follows the preprocessing pipeline of [95]. "#Users" denotes the number of valid users before applying the dataset-specific AUROC threshold in Table 5.

Dataset	#Users	Sampling Freq.	Target	Sensing Modalities	Preprocessing
D#1	92	15 ESM/day	Stress	**Behav:** app usage, notifications, calls/SMS, activity; **Physio:** Fitbit HR; **Context:** mobility, light, BT/Wi-Fi, Fitbit steps/distance/calories	RQ1: Hourly; RQ2: [95]
D#2	99	12 ESM/day	Stress	**Behav:** app usage, notifications, calls/SMS, activity, keystroke dynamics; **Physio:** Fitbit HR; **Context:** mobility, light, BT/Wi-Fi, Fitbit steps/distance/calories	RQ1: Hourly; RQ2: [95]
D#3	106	10 ESM/day	Stress	**Behav:** app usage, notifications, calls/SMS, activity, keystroke dynamics; **Physio:** Fitbit HR; **Context:** mobility, light, BT/Wi-Fi, Fitbit steps/distance/calories	RQ1: Hourly; RQ2: [95]
K-EmoPhone	47	~10 ESM/day	Stress	**Physio:** HR, skin temp, EDA, IBI; **Behav:** accel, steps, phone usage; **Context:** location, ambient light	RQ1, RQ2: [95]
StudentLife	49	3–13 EMAs/day	Stress	**Behav:** activity, conversation, sleep; **Context:** location/mobility	RQ1, RQ2: [83]
CrossCheck	62	Every 2–3 days	Stress	**Behav:** activity, speech, app usage, sleep; **Context:** location, ambient env.	RQ1, RQ2: [2]
GLOBEM	80	Weekly	Depression	**Behav:** usage, calls, BT, activity, sleep; **Context:** location/mobility	RQ1, RQ2: [87]
College Exp.	161	Weekly (EMA)	Stress	**Behav:** activity, usage, comms, sleep; **Context:** mobility	RQ1, RQ2: [45]

College Experience Study [64]. For these datasets, we use the released data and follow the preprocessing and feature extraction procedures adopted in the corresponding original studies [2, 45, 83, 87, 95], rather than re-deriving raw collection-stage quality-control rules from scratch.

For GLOBEM, we use a merged multi-year version rather than individual yearly subsets. This choice is specific to RQ2: although each yearly subset contains a sufficient number of users, the number of labeled observations per user is often too limited for the personalized modeling required in our transferability analysis. We therefore merge the sub-datasets first and then apply the same RQ2 eligibility logic used elsewhere. As summarized in Table 5, 80 users remain after basic per-user label availability and label-balance checks (e.g., more than one label class per user), and only 5 users remain after the additional self-AUROC criterion used for RQ2. Stress is the primary target label throughout the paper whenever sufficiently reliable per-user analysis is possible. We retain GLOBEM because it substantially broadens the external evaluation setting; however, under our personalized RQ2 setup, the per-user stress labels in GLOBEM provide insufficient reliable support for stable analysis. We therefore use depression as a dataset-specific exception for GLOBEM, while stress remains the target for all other datasets.

Detailed information for all eight datasets is provided in Table 2. As clarified below, these datasets are instantiated differently in RQ1 and RQ2 because the two components of CrossShift require different feature representations and analysis pipelines.

Unless otherwise noted, RQ1 operates on the dataset-level preprocessed cohorts, whereas RQ2 applies additional user-level eligibility filters required for personalized modeling and transferability analysis; these retained counts are summarized in Table 5.

4.2 RQ1 Instantiation in This Study

We now describe how the shift-testing component of CrossShift is instantiated in this study, including the feature representation and preprocessing choices used for RQ1. More specifically, we have the following three questions.

- Which behavior feature categories (i.e., phone usage, social interaction, mobility, physical status, and sleep) show the biggest between-user differences under covariate, conditional, and concept shift?
- Which types of distribution shifts are statistically significant across users?
- To what extent can user-wise normalization mitigate covariate and conditional shifts, and does it have any effect on concept shift?

Experiments explore which behavior feature category exhibits the largest magnitude of change across different types of shifts and which shifts are statistically significant. In addition, we added another experiment to examine the impact of user-wise normalization. User-wise normalization can mitigate cross-user *covariate shift* by aligning feature statistics [49, 78]. However, such preprocessing generally assumes a stable distribution of $P(Y \mid X)$ and thus cannot fully correct *conditional shift*, where $P(X \mid Y)$ itself changes across users [60, 71]. Likewise, normalization does not handle *label shift* (changes in $P(Y)$), which typically requires prior-shift estimation and reweighting [50], nor *concept shift*, which calls for explicit model adaptation to evolving targets [22, 84]. Accordingly, we aim to investigate whether user-wise normalization can address covariate shift, as well as conditional and conceptual shift, in mobile sensing for mental health.

The analysis process of RQ1 is explained in Figure 4. We will explore 20 shift instances (five behavior feature categories × four shift types).

4.2.1 Normalization Regimes. RQ1 compares two preprocessing regimes for the feature-based shift analyses. The first is *feature-wise normalization*, where each feature is standardized using global statistics computed across all users within the training set. This regime preserves between-user differences while preventing features with larger numeric ranges from dominating distance-based tests [8, 62].

The second is *user-wise normalization*, where each feature is standardized separately within each user's training set using that user's own mean and standard deviation, to be specific, per-user z-score normalization applied feature-by-feature. This regime is evaluated separately rather than sequentially after feature-wise normalization. Its purpose is to test whether the observed between-user differences are primarily driven by user-specific location and scale effects.

4.2.2 Feature Representation and Preprocessing. The normalization regimes used in RQ1 are described separately above. Here, we focus only on feature construction, aggregation, and grouping for the shift analyses.

During the data processing stage for D#1, D#2, and D#3 datasets, we adopt an hourly aggregation window over the full data-collection period. The analysis is restricted to a set of explicitly interpretable features, as detailed in Table 3. We exclude user demographic variables from the final feature set, as this study focuses on behavioral data collected from mobile and wearable devices. Features are organized into five behavior feature categories: *phone usage*, *social interaction*, *mobility*, *physical status*, and *sleep*. These categories have been shown in prior work to correlate with or help predict self-reported labels in mobile sensing (Table 3).

Table 3. Behavior feature categories & feature design for RQ1.

Data Sources	Interpretable Features	References	
Phone Usage Behavior	AppUsageEvent; ScreenEvent; KeyEvent; DeviceEvent	App-usage event counts and durations; screen-on duration and count; typing distance and time per app category	[20, 44, 56, 77, 87]
Social Interaction	CallEvent; MessageEvent	Call duration and count; message count	[1]
Physical Status	PhysicalActivityTransition; step/calorie/distance counters	Counts and durations of physical-activity types; step count; distance traveled; calories expended	[46, 56, 67, 87]
Mobility	Location (clustered/labeled via Google Maps API)	Places visited; dwell time at location clusters	[46, 56, 57, 87]
Sleep	Daily sleep survey; estimates from ScreenEvent	Prior-night sleep duration and onset	[1, 46, 87]

Before conducting shift analyses, we apply *feature-wise normalization* across all users, scaling each feature based on global statistics.

For each behavior feature category, we examine four types of distribution shift. Widely used literature shows user-wise normalization helps to prevent covariate shift [29, 42, 61, 91, 94]. To assess the impact of user-specific alignment, we apply *user-wise standard normalization* (z-scoring features separately for each user) and report results under two preprocessing regimes for covariate, conditional, and concept shift: (i) feature-wise normalization only, and (ii) user-wise normalization.

4.3 RQ2 Instantiation in This Study

We now describe how the shift-attribution component of CrossShift is instantiated in this study. The prediction targets used for each dataset are summarized in Table 4, and the overall study-specific RQ2 pipeline is illustrated in Figure 5. We then describe the corresponding preprocessing, model construction, transferability evaluation, and statistical analysis procedures in detail. After testing for the existence of different types of shifts in RQ1, we aim to test the impact of different types of shifts. More specifically, we will address these two questions in RQ2.

- Among different types of shifts, which behavior feature category is contributing to the model performance drop?
- Which type of shift is contributing most to the model performance drop?

Before evaluating the effects of the 20 shift instances (five behavior feature categories × four shift types), we need to formalize and instantiate the metric for each shift type. Then we can pair the computed shift metrics with model performance to attribute cross-user degradation to either (i) specific behavior feature categories or (ii) specific shift types.

4.3.1 Feature Representation and Preprocessing. For RQ2, we employ a different feature design for D#1, D#2, D#3 compared with RQ1 to ensure a strong in-domain baseline for assessing the impact of distribution shifts on model performance. Without such a baseline, cross-user generalization results become difficult to interpret, as poor performance may stem from inherent model limitations rather than cross-user shifts. Our preprocessing and feature extraction procedures follow the standardized pipeline of Zhang et al. [95]. In brief, the pipeline uses extracted features spanning multiple sensing domains, including social interaction, activity, context, phone usage, physiological signals, and sleep-related behavior, followed by model-based top-K feature selection and personalized XGBoost evaluation. Temporal splitting and normalization are performed using training-set statistics only.

Table 4. Summary of datasets, preprocessing pipelines used for predictive modeling (RQ2), and corresponding prediction targets.

Datasets	RQ2 Preprocessing	Prediction Target
D#1, D#2, D#3 (New)	[95]	Stress
K-EmoPhone	[95]	Stress
StudentLife	[83]	Stress
CrossCheck	[2]	Stress
GLOBEM	[87]	Depression
College Experience	[45]	Stress

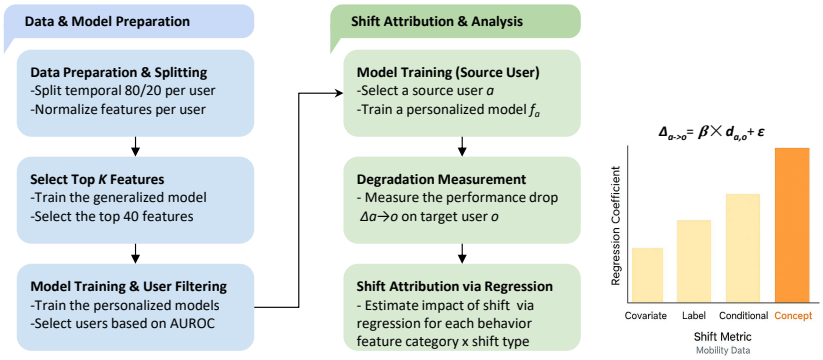


Fig. 5. RQ2 pipeline for quantifying the effects of interpersonal differences on cross-user model generalization.

4.3.2 Prediction Targets and Experimental Pipeline. We describe the temporal splitting, user eligibility, feature selection, and modeling procedures used in the RQ2 pipeline.

Temporal splitting and normalization. We split each user's data into temporally ordered training and testing subsets using an 80/20 ratio, with the first 80% of samples used for training and the remaining 20% reserved for testing. All normalization is performed using training-set statistics only; for each user, we estimate the normalization parameters from that user's training data and apply them unchanged to the corresponding test data, thereby preventing information leakage.

User eligibility. RQ2 imposes additional user-level eligibility criteria beyond dataset-level preprocessing. For each dataset, users must first satisfy basic per-user label availability and label-balance criteria, including requiring more than one label class per user. We then first split each user's observations chronologically into 80% train and 20% test subsets and exclude users whose test split contains fewer than ten samples. Figure 14 in the appendix shows the pooled distribution of per-user test-set sample counts after the temporal 80/20 split; the 10-sample cutoff removes only a sparse tail of users and was used to avoid unstable AUROC estimates. Among the eligible users, we then retain only those whose personalized within-user model achieves a dataset-specific self-AUROC threshold, so that transferability loss is evaluated relative to a non-random in-distribution baseline. Dataset-specific thresholds and retained user counts are summarized in Table 5.

Table 5. User-retention summary. For collected datasets, we report raw users and users remaining after dataset-level quality control (QC). For public datasets, we used the released data and followed the preprocessing procedures of the original studies; therefore, raw collection-stage QC counts are not reported. We then apply temporal eligibility checks and a dataset-specific self-AUROC threshold to define the final analytic cohort.

Dataset	Raw	Post-QC	Eligible	Thr.	Final
D#1	102	92	92	0.65	33
D#2	112	99	99	0.65	43
D#3	114	106	106	0.65	46
K-EmoPhone	–	–	47	0.50	34
CrossCheck	–	–	62	0.65	21
StudentLife	–	–	49	0.50	14
GLOBEM	–	–	80	0.65	5
College Exp.	–	–	161	0.60	135

Feature selection and user filtering. We first perform global top-K feature selection on the normalized training data using model-based feature importance from XGBoost (default K=40). Here, *global top-K* denotes feature selection based on a single generalized model trained across all users' training set, rather than personalized models. We choose XGBoost because it performs strongly on tabular features and has demonstrated competitive results in our domain [15, 95]. For RQ2, we use a fixed XGBoost protocol across users and datasets, without dataset-specific hyperparameter tuning or an additional validation split. This choice preserves comparability for shift-attribution analysis, since our goal is to relate shift types to transferability loss rather than optimize predictive performance separately for each dataset. Full details of the RQ2 preprocessing steps, XGBoost hyperparameter settings, and training procedure are provided in Appendix A.2.1. The selected features are grouped into five behavior feature categories—*phone usage*, *physical status*, *sleep*, *mobility*, and *social interaction*. Using this fixed feature subset, we then train a personalized XGBoost model for each eligible user and evaluate its in-distribution performance on that user's test set. The final RQ2 cohort consists of users whose self-AUROC exceeds the dataset-specific threshold reported in Table 5. This filtering is used to ensure that cross-user degradation is analyzed only when the source user exhibits a meaningful within-user predictive signal.

4.3.3 Transferability Evaluation and Statistical Analysis. We now describe how the generic transferability analysis in Section 3.4 is instantiated in this study.

Transferability evaluation and regression. Each retained source-target pair contributes one observation to the corresponding regression, and all pairs are treated equally; we do not apply residual weighting by sample size. To reduce instability from sparse users, we instead control data sufficiency upstream through the temporal eligibility and user-retention criteria described in Section 4.3.2.

Statistical testing. To assess whether shift distances systematically predict transferability loss, we test whether the per-source regression coefficients differ from zero using one-sided Wilcoxon signed-rank tests and sign tests (null hypothesis: median $\beta = 0$). We apply Benjamini-Hochberg false discovery rate correction across all category-shift combinations [11].

Additional methodological details. To streamline the main manuscript, we defer detailed rationales (e.g., user filtering, non-negative coefficients, and regression vs. correlation) to the Methodological Transparency & Reproducibility Appendix (META; Appendix A.2.1).

5 RESULTS

5.1 RQ1 What Interpersonal Differences are Present in Mobile Sensing Datasets as Revealed by Statistical Tests?

Major findings. Our RQ1 analysis yields three key findings. First, covariate and conditional shifts are widespread across datasets, with Phone Usage, Sleep, and Social Interaction showing the most consistent between-user differences (Section 5.1.1). Second, label shift is detectable across users but does not by itself explain the observed behavior–label heterogeneity (Section 5.1.2). Third, user-wise normalization substantially reduces covariate and conditional centroid differences, but concept shift remains comparatively persistent, indicating that feature alignment alone cannot remove user-specific behavior–label mappings (Section 5.1.3).

5.1.1 Which Behavior Feature Categories Show the Strongest between-User Differences under Covariate, Conditional, and Concept Shift? In this subsection, we analyze between-user differences using *feature-wise normalization* only, which applies a global normalization across all users to characterize intrinsic behavioral heterogeneity prior to any user-specific alignment. Results for all shift types under feature-wise normalization across all datasets are shown in Figure 6. Label shift is reported separately in Table 6.

Covariate shift ($P(X)$). Under feature-wise normalization (Figure 6, panels 1–2), we observe strong and highly consistent between-user differences across all datasets. Aggregating PERMANOVA effect sizes (R^2) across the eight datasets, *Phone Usage* showed the largest mean effect size ($\bar{R}^2 = 0.350$), followed by *Social Interaction* (0.332) and *Sleep* (0.308). *Physical Status* and *Mobility* were slightly lower but comparable (0.288 and 0.287). All category–dataset PERMANOVA tests were significant ($p \leq 0.01$), and the ordering should be interpreted as a cross-dataset tendency rather than strict per-dataset dominance. Thus, covariate shift is most pronounced in routine- and communication-related behavioral domains.

PERMDISP effects were smaller and less stable than PERMANOVA effects. Averaging over non-NA cells, *Social Interaction* showed the largest mean dispersion effect ($\bar{R}^2 = 0.242$), followed by *Phone Usage* (0.189) and *Mobility* (0.180), while *Sleep* and *Physical Status* were lower (0.163 and 0.153). PERMDISP produced six NA entries (*Social Interaction*: 4; *Mobility*: 1; *Sleep*: 1), because some dataset–category combinations did not contain sufficient non-constant within-user variation after normalization to estimate dispersion reliably. We therefore report these cases as NA rather than imputing or forcing unstable estimates. Overall, covariate shift is dominated by differences in feature centroids rather than differences in within-user variability.

Conditional shift ($P(X \mid Y)$). For conditional shifts given $Y = 0$ and $Y = 1$ (Figure 6, panels 3–6), PERMANOVA effects remained substantial after stratifying by stress label and largely mirrored the covariate-shift pattern. Averaging across datasets and both strata, *Phone Usage* again showed the largest mean effect size ($\bar{R}^2 = 0.349$), followed by *Social Interaction* (0.315) and *Sleep* (0.302), with *Physical Status* and *Mobility* close behind (0.299 each). Nearly all category–dataset tests were significant ($p \leq 0.01$), with one exception for *Social Interaction* in D#1 under the high-stress stratum.

PERMDISP effects for conditional shift were again smaller and less systematic. Averaging over non-NA cells across both strata, the largest dispersion effects were observed for *Social Interaction* ($\bar{R}^2 = 0.254$) and *Phone Usage* (0.207), followed by *Mobility* (0.178), *Physical Status* (0.163), and *Sleep* (0.161). PERMDISP yielded 15 NA entries across the two strata, especially for *Social Interaction*.

Concept shift ($P(Y \mid X)$). Category-specific patterns differed qualitatively for concept shift (Figure 6, panels 7–8). Across both concept-shift panels, mean Cohen's W^2 effect sizes were largest for *Social Interaction* ($\bar{W}^2 = 0.455$), followed by *Sleep* (0.405), *Physical Status* (0.397), *Phone Usage* (0.393), and *Mobility* (0.385). Most defined category–dataset tests were statistically significant (73/77 at $p < 0.05$; 72/77 at $p < 0.01$), and effect-size distributions substantially overlapped across categories.

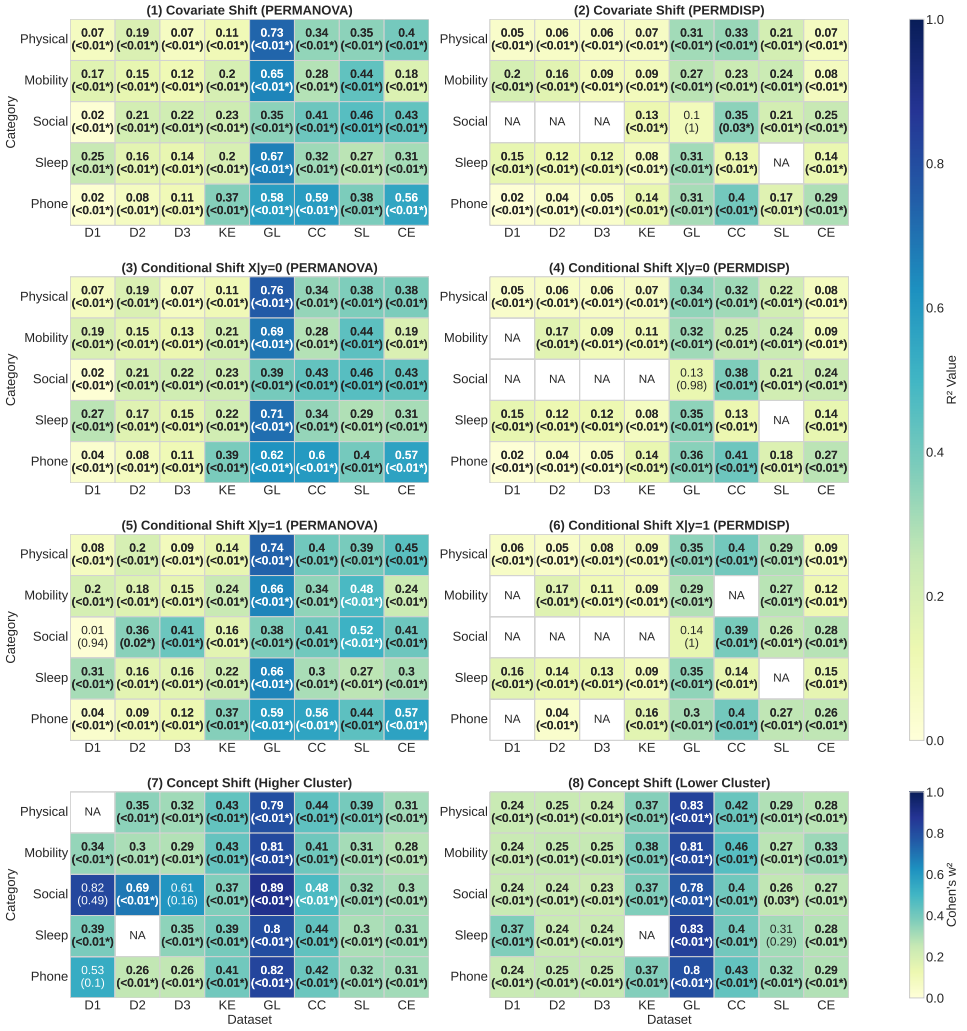


Fig. 6. **RQ1 (Feature-wise normalization).** Summary of between-user differences across shift types under feature-wise normalization (i.e., global normalization across users). Columns correspond to datasets: **D1–D3** (three newly collected in-the-wild datasets from the same institution), **KE** (K-EmoPhone), **GL** (GLOBEM), **CC** (CrossCheck), **SL** (StudentLife), and **CE** (College Experience). Panels report effect sizes from PERMANOVA (centroid differences) and PERMDISP (dispersion differences): (1) Covariate shift (PERMANOVA), (2) Covariate shift (PERMDISP), (3) Conditional shift $P(X \mid Y{=}0)$ (PERMANOVA), (4) Conditional shift $P(X \mid Y{=}0)$ (PERMDISP), (5) Conditional shift $P(X \mid Y{=}1)$ (PERMANOVA), (6) Conditional shift $P(X \mid Y{=}1)$ (PERMDISP), (7) Concept shift (higher cluster), and (8) Concept shift (lower cluster). For covariate and conditional shifts, each cell reports the PERMANOVA or PERMDISP effect size (R^2) with its associated p-value. For concept shift, each cell reports Cohen's W^2 (effect size for label–feature association) with its associated p-value.

Unlike covariate and conditional shifts, concept shift was not concentrated in a small subset of behavioral domains. Instead, it appeared broadly across behavior feature categories, indicating pervasive differences in $P(Y|X)$.

Table 6. **RQ1 (Label shift).** Between-user differences in label distributions $P(Y)$, reported separately because normalization choices for X do not apply to Y. Each bar reports Cohen's W^2 (effect size for between-user differences in label prevalence) together with its associated p-value.

Dataset	D#1	D#2	D#3	KEmoP.	GLOBEM	CrossCh.	StudentLi.	CollegeEx.
ω^2	0.24	0.27	0.31	0.36	0.77	0.40	0.24	0.27
p-value	$< .01$	$< .01$	$< .01$	$< .01$	$< .01$	$< .01$	$< .01$	$< .01$

Label shift ($P(Y)$). Label shift results are reported separately in Table 6. Label prevalence differed significantly across users in all datasets (all $p < .001$), with a mean effect size of $\bar{W}^2 = 0.357$ and values ranging from 0.243 to 0.765. However, label-shift effects were smaller on average than concept-shift effects (0.357 vs. 0.407) and were measured only at the dataset level, suggesting that variation in $P(Y)$ alone does not explain the dominant cross-user differences.

Conditional shift often mirrored covariate shift in category ordering and magnitude, but the distinction remains useful: conditional shift asks whether user differences persist after holding the label fixed, whereas covariate shift does not.

Implications. Covariate and conditional shifts are strongest in *Phone Usage*, *Sleep*, and *Social Interaction* and are driven mainly by centroid differences, whereas concept shift is broader across categories and therefore requires model-level or user-level adaptation rather than feature-level preprocessing alone.

5.1.2 Which Types of Distribution Shifts are Statistically Significant Across Users? Across datasets, all four shift types show statistically detectable between-user heterogeneity under feature-wise normalization (Figure 6, Table 6). This confirms that cross-user variability in mobile mental-health sensing is not limited to covariate or label shift: conditional and concept shifts are also pervasive, with concept shift showing the broadest and most persistent effects. We discuss the methodological and deployment implications in Sections 6.1 and 6.2.

5.1.3 Can We Handle Covariate, Conditional, and Concept Shifts Using User-Wise Normalization? As described in Section 4, we repeat all RQ1 analyses after applying *user-wise standard normalization*, where features are z-scored independently for each user prior to statistical testing. This experiment evaluates whether per-user alignment is sufficient to mitigate different forms of between-user distribution shift.

Covariate shift ($P(X)$). Under user-wise normalization (Figure 7, panels 1–2), covariate-shift PERMANOVA effects were nearly eliminated. Mean effect sizes were close to zero for *Physical Status* ($\bar{R}^2 = 0.0004$), *Mobility* (0.0005), *Social Interaction* (0.0005), and *Phone Usage* (0.0004). The only significant category–dataset test was *Sleep* in *College Experience* ($R^2 = 0.302$, $p < 0.01$), raising the mean for *Sleep* to $\bar{R}^2 = 0.0377$. Thus, user-wise normalization removes nearly all between-user centroid differences in $P(X)$, indicating that most covariate shift under feature-wise normalization comes from user-specific location and scale effects.

PERMDISP effects were weaker and less consistent, indicating that user-wise normalization removes most feature-centroid differences but not all within-user variability.

Conditional shift ($P(X \mid Y)$). Conditional-shift PERMANOVA effects were also greatly reduced after user-wise normalization (Figure 7, panels 3–6). Mean effect sizes were $\bar{R}^2 = 0.0169$ for Y=0 and 0.0503 for Y=1, with 3/40 and 12/40 significant tests, respectively. This shows that label-conditioned centroid differences are largely attenuated, although modest residual effects remain in the high-stress stratum.

This residual asymmetry also clarifies why conditional shift is not fully redundant with covariate shift. Conditional shift broadly mirrors covariate shift after normalization, but the stronger residual effects for Y=1

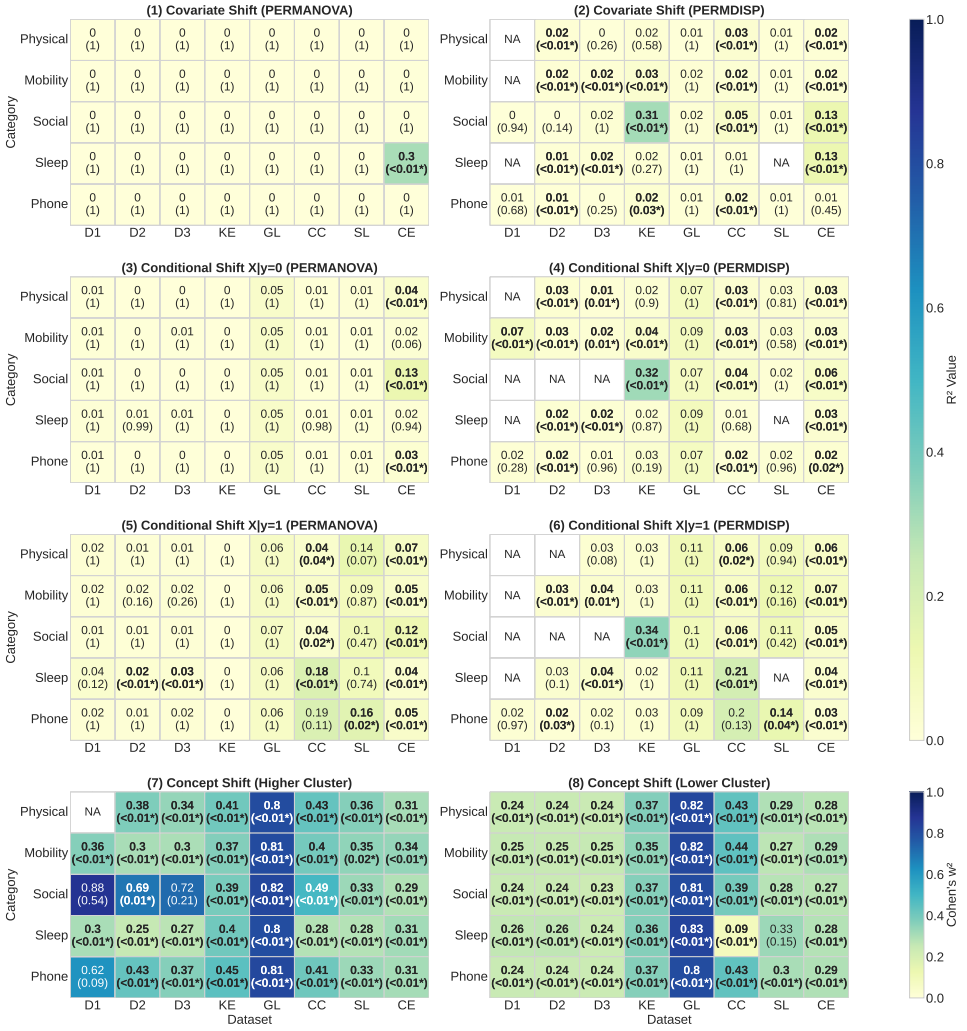


Fig. 7. **RQ1 (User-wise normalization).** Summary of between-user differences across shift types after user-wise normalization (per-user alignment prior to testing). Columns correspond to datasets: **D1–D3** (three newly collected in-the-wild datasets from the same institution), **KE** (K-EmoPhone), **GL** (GLOBEM), **CC** (CrossCheck), **SL** (StudentLife), and **CE** (College Experience). Panels follow the same layout as Figure 6: (1–2) covariate shift, (3–4) conditional shift $P(X \mid Y{=}0)$, (5–6) conditional shift $P(X \mid Y{=}1)$, and (7–8) concept shift (higher/lower cluster). For covariate and conditional shifts, each cell reports the PERMANOVA or PERMDISP effect size (R^2) with its associated p-value. For concept shift, each cell reports Cohen's W^2 with its associated p-value.

suggest that some between-user heterogeneity is label-state-dependent rather than purely marginal. PERMDISP effects remained small and dataset-dependent, indicating that most conditional shift observed under feature-wise normalization arises from marginal differences in $P(X)$ rather than robust dispersion differences in $P(X \mid Y)$.

Concept shift ($P(Y \mid X)$). In contrast, concept shift remained strong under user-wise normalization (Figure 7, panels 7–8). Across both concept-shift panels, the overall mean Cohen's W^2 was 0.401, close to the feature-wise value (0.407). Aggregating across datasets and clusters, *Social Interaction* showed the largest mean effect size ($\bar{W}^2 = 0.464$), followed by *Phone Usage* (0.415), *Physical Status* (0.396), *Mobility* (0.384), and *Sleep* (0.347). Most defined tests remained significant (75/79 at $p < 0.05$; 73/79 at $p < 0.01$).

This persistence indicates that user-wise feature alignment does not mitigate concept shift. The remaining differences reflect user-specific mappings from behavioral features to mental-health labels, i.e., differences in $P(Y \mid X)$ that are largely invariant to feature centering or scaling.

Implications. User-wise normalization is effective for reducing covariate and conditional centroid differences, but it has limited impact on dispersion heterogeneity and does not mitigate concept shift. This pattern supports the interpretation that feature-level alignment can address user-specific scale/location effects, whereas persistent differences in the $X \rightarrow Y$ relationship require model-level or user-level adaptation rather than normalization alone [49, 60, 71, 78].

5.2 RQ2 Which Specific Interpersonal Differences Contribute Most to Failures of Cross-User Model Generalization?

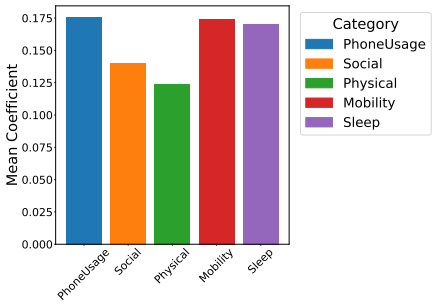
Details of the statistical tests used to assess the regression coefficients for all datasets are reported in Appendix Table 10. All results in this section are robust to variations in the feature dimensionality and user-filtering thresholds; a systematic robustness analysis is reported in Section 5.3.1. In addition, detailed regression coefficients per behavior feature category/shift type/dataset can be found in Figure 15 in Appendix.

Major findings. Our RQ2 analysis yields two key findings. First, behavior feature categories differ in their contribution to cross-user degradation; Phone Usage, Sleep, and Social Interaction often contribute strongly, but the dominant category varies across datasets rather than following a single universal ordering (Section 5.2.1). Second, concept shift shows the strongest and most consistent association with transferability loss across datasets, indicating that user-specific behavior–label mappings are the main driver of cross-user generalization failure (Section 5.2.2).

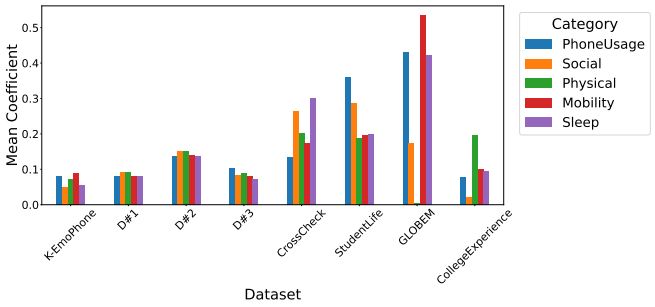
5.2.1 Which Behavior Feature Categories Contribute Most to Performance Degradation? Figure 8 summarizes how different behavior feature categories contribute to performance degradation under distribution shift. When aggregated across all shift types and datasets using the mean regression coefficient (Figure 8a), *Phone Usage* showed the largest overall effect ($\bar{\beta} = 0.176$), followed closely by *Mobility* ($\bar{\beta} = 0.174$) and *Sleep* ($\bar{\beta} = 0.170$). *Social Interaction* was somewhat lower ($\bar{\beta} = 0.140$), whereas *Physical Status* showed the smallest overall mean coefficient ($\bar{\beta} = 0.124$). These results indicate that cross-user heterogeneity in smartphone interaction patterns is, on average, the strongest contributor to generalization failure, although the differences among the top three categories are modest.

The dataset-wise breakdown in Figure 8b reveals substantial heterogeneity underlying this global trend. Although *Phone Usage* remains the largest contributor in some datasets (e.g., *StudentLife*, $\bar{\beta} = 0.359$, and D#3, $\bar{\beta} = 0.102$), other categories dominate in specific contexts—for example, *Sleep* in *CrossCheck* ($\bar{\beta} = 0.302$), *Mobility* in *GLOBEM* ($\bar{\beta} = 0.535$), and *Physical Status* in *College Experience* ($\bar{\beta} = 0.195$). This variability indicates that the most destabilizing behavioral domain is not universal, but depends on dataset-specific sensing protocols, cohort characteristics, and behavioral coverage.

Overall, these results suggest that performance degradation under distribution shift is often driven by a small subset of behavioral domains, but the identity of the dominant domain varies substantially across datasets.

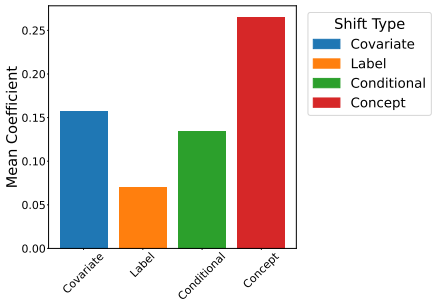


(a) behavior feature category–level summary aggregated across shift types and datasets

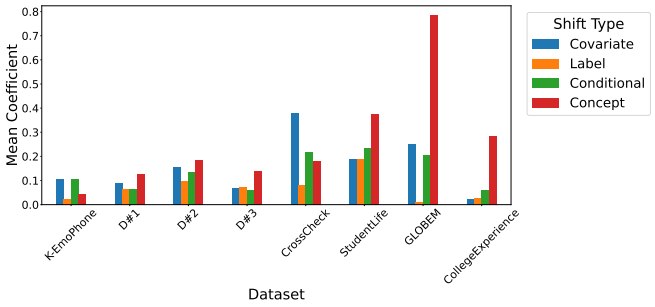


(b) behavior feature category–level summary aggregated across shift types

Fig. 8. Category-level summaries of regression coefficients quantifying the contribution of different behavior feature categories to detected distribution shifts. (a) Mean regression coefficients aggregated across all shift types and datasets, highlighting the overall importance of each behavior feature category. (b) Mean regression coefficients aggregated across shift types for each dataset, illustrating how the relative influence of behavior feature categories varies across datasets. As a descriptive pooled fit summary, R^2 values are uniformly small (mean 0.007–0.009 per category), indicating that the mean coefficients are statistically reliable but the source-target linear fit explains only a small share of variance. See Appendix Table 10 for the corresponding regression coefficients and p-values.



(a) Regression coefficients aggregated across behavior feature categories and datasets



(b) Dataset-wise regression coefficients aggregated across behavior feature categories

Fig. 9. Summary of regression coefficient magnitudes for different distribution shift types. (a) Mean regression coefficients aggregated across all behavior feature categories and datasets, comparing the relative strength of covariate, label, conditional, and concept shifts. (b) Dataset-wise mean regression coefficients aggregated across behavior feature categories, illustrating how the prominence of each shift type varies across datasets. See Appendix Table 10 for the corresponding regression coefficients and p-values. As a descriptive pooled fit summary, R^2 rises from Covariate ($\approx$ 0.001) to Concept ($\approx$ 0.010), consistent with the coefficient ordering.

Implications. Rather than treating all behavioral features uniformly, deployable mobile mental-health systems should adopt category-aware monitoring and adaptation strategies, with particular attention to phone-usage, mobility, and physical-activity features whose impact varies most strongly across datasets.

5.2.2 Which Type of Shift Contributes Most to Performance Degradation? Figure 9 compares the relative impact of different shift types on cross-user performance degradation. When aggregated across behavior feature categories

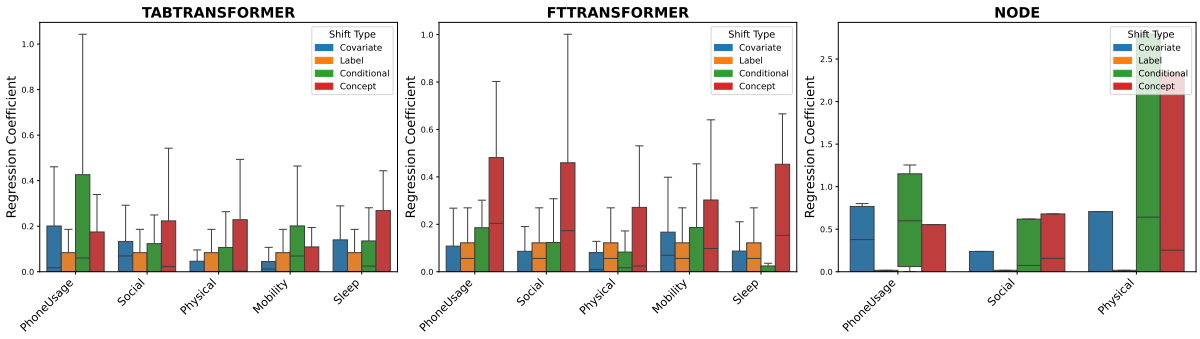


Fig. 10. RQ2 results on D#3 across different model architectures: TabTransformer, FTTransformer, and NODE.

and datasets (Figure 9a), *concept shift* showed the largest overall mean regression coefficient ($\bar{\beta} = 0.265$), exceeding *covariate shift* ($\bar{\beta} = 0.158$), *conditional shift* ($\bar{\beta} = 0.134$), and *label shift* ($\bar{\beta} = 0.070$). This indicates that mismatches in the learned input–output relationship $P(Y \mid X)$ are, on average, the strongest drivers of cross-user generalization failure.

The dataset-wise results in Figure 9b show that this dominance is global rather than universal across datasets. Concept shift is especially pronounced in *GLOBEM* ($\bar{\beta} = 0.785$), where it is markedly larger than all other shift types and substantially elevates the overall mean in Figure 9a. Concept shift also remains the largest mean component in *StudentLife* ($\bar{\beta} = 0.375$), *College Experience* ($\bar{\beta} = 0.286$), D#2 ($\bar{\beta} = 0.185$), D#3 ($\bar{\beta} = 0.138$), and D#1 ($\bar{\beta} = 0.126$). In contrast, *CrossCheck* and *K-EmoPhone* show weaker separation, with *covariate shift* exceeding concept shift on average ($\bar{\beta} = 0.380$ vs. 0.180 in *CrossCheck*; $\bar{\beta} = 0.107$ vs. 0.044 in *K-EmoPhone*).

These results indicate that while covariate and conditional shifts can meaningfully exacerbate performance degradation in specific datasets, concept shift shows the strongest overall association with transferability loss across datasets, even though its dominance is not uniform in every individual dataset.

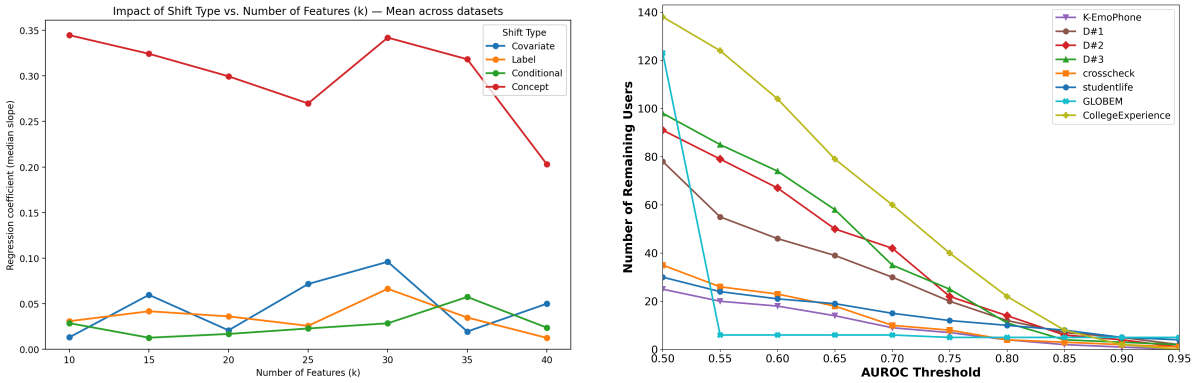
Implications. Because concept shift is both the strongest and the most dataset-sensitive driver of cross-user performance degradation, particularly in large-scale datasets such as *GLOBEM*, deployable systems should prioritize early detection of concept shift and trigger user-specific adaptation or retraining when such shifts are detected.

5.3 Additional Analyses: Robustness, Scope Extension, and Concept-Shift Handling

We next report four follow-up analyses around the main RQ2 result. We first test the robustness of the conclusion that concept shift is the dominant predictor of transferability loss, then examine whether this pattern persists for a more objective label, and finally explore two concept-shift-oriented directions for handling cross-user failures: stronger generalization models and OOD-guided personalization.

5.3.1 Robustness of the Main RQ2 Finding. We first examine whether the main RQ2 conclusion that concept shift is the dominant predictor of transferability loss remains stable under changes in model architecture, feature dimensionality, and user-filtering thresholds.

Impact of model types. Beyond XGBoost, which is well-suited to sparse tabular data with limited labels, we evaluated robustness to model choice by training three widely-used deep tabular architectures on D#3: TabTransformer, FT-Transformer, and NODE [27, 33, 70]. Using the same self-AUROC threshold of 0.65 across architectures, we find that the coefficient patterns vary by model (Figure 10). Concept shift remains substantial



(a) Feature number K (b) User-filtering threshold

Fig. 11. Robustness analyses for the RQ2 shift-attribution results. (a) Mean regression coefficients across different numbers of selected features K. (b) Number of retained users under different in-distribution AUROC filtering thresholds. Together, these analyses show that the main RQ2 conclusion is not driven by a single feature-set size or filtering threshold.

across architectures, although conditional shift is also large in several settings. Overall, the exact ordering is model-dependent, but behavior–label mismatch remains a major source of cross-user performance degradation.

Impact of the feature number K. To assess the sensitivity of our findings to the choice of feature dimensionality, we repeated the full RQ2 pipeline using multiple values of $K \in \{10, 15, 20, 25, 30, 35, 40\}$. Across all datasets and tested values of K, concept shift consistently remained the leading shift type, with the largest regression coefficients and strongest association with transferability loss (Figure 11a). The relative ordering among shift types was stable, indicating that our main conclusions are not driven by a particular feature-set size.

Impact of the user-filtering threshold. As outlined in Figure 5, we apply user screening based on in-distribution AUROC. The threshold balances construct validity—retaining users whose personalized models capture a meaningful $X \rightarrow Y$ relationship—against statistical precision in downstream regressions. We target at least 30 retained users per dataset to support stable coefficient estimates and adequate statistical power [12]. As shown in Table 5 and Figure 11b, this target is achieved for all datasets except *StudentLife*, *K-EmoPhone*, and *GLOBEM*. For these datasets, imposing an AUROC threshold of ≥ 0.50 reduces the retained cohort below 30 users. Importantly, the qualitative pattern remains unchanged: concept shift continues to dominate covariate, label, and conditional shifts in predicting transferability loss, except for the CrossCheck dataset.

5.3.2 Objective-Label Case Study with Step-Count Prediction. Because self-reported mental-health labels may contain subjective reporting bias [19, 23], we additionally test whether the RQ2 pattern persists for an objective, sensor-derived target. On D#3, we predict the next 5-minute Fitbit step-count level from preceding behavioral features, avoiding contemporaneous accelerometer signals that would make the task trivial. We binarize each user's target into HIGH/LOW using that user's median step count. Predicting short-term physical activity can provide an objective behavioral signal for identifying imminent low physical activity, which may inform future just-in-time intervention opportunities. Although this objective-label task achieves substantially higher LOSO performance than stress prediction (AUROC ≈ 0.70), concept shift remains the dominant contributor to cross-user degradation (Figure 12). This suggests that user-specific $P(Y \mid X)$ variability is not only an artifact of subjective stress self-reports.

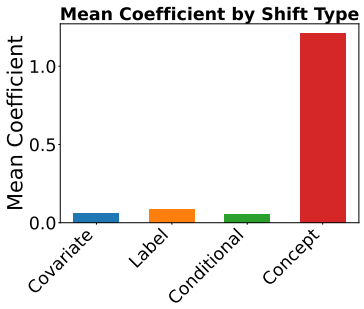


Fig. 12. RQ2 results for step count prediction.

Table 7. Test AUROC comparison on D#3 (LOSO) across baseline, domain generalization, and foundation model.

Category	Model	Test AUROC
Baseline	XGBoost	0.5693 ± 0.0886
Domain Generalization	DRO	0.5805 ± 0.0902
	Adv. Label Group DRO	0.5817 ± 0.0882
	VREx	0.5819 ± 0.0895
	Group DRO	0.5839 ± 0.0880
	IRM	0.5646 ± 0.0749
	ERM	0.5822 ± 0.0867
	HHISS	**0.5934 ± 0.0918**
Tabular Foundation Model	TabPFN	0.5698 ± 0.0843

5.3.3 Toward Handling Concept Shift: Stronger Generalization Models on D#3. We conduct an additional experiment on *D#3* only, using a strict leave-one-subject-out (LOSO) evaluation protocol. We compare several domain generalization baselines adopted from the prior work [24], recent domain generalization models for mobile sensing [86], and the foundation model TabPFN, using the same input features and evaluation metric (test AUROC). The results are summarized in Table 7. Overall, these results suggest that while domain generalization objectives and foundation-model priors can improve cross-user performance relative to our XGBoost baseline, they do not fully eliminate the generalization gap. This is consistent with our broader conclusion that handling user-level *concept shift* requires explicit mechanisms beyond feature-level alignment, including reliable shift detection and user-triggered adaptation in deployed mobile mental-health systems.

5.3.4 Toward Handling Concept Shift: OOD-Guided Personalization. To make the deployment implications of concept shift more concrete, we conducted an illustrative OOD-guided personalization experiment on the College Experience dataset, which provides a suitable longitudinal setting for simulating early adaptation followed by later-user evaluation. For each target user, we used a temporal 50/50 split: the first half was treated as early target-user data available for adaptation, and the second half was held out for evaluation. We first trained a source-only model using all non-target users, then computed a pre-adaptation concept-shift score for the target user using Jensen-Shannon (JS) divergence on the early target-user slice.

We compare three adaptation policies. *Partial personalization* trains a model using both the source users' data and the early target-user data. *Target-user-only personalization* trains a new model from scratch using only

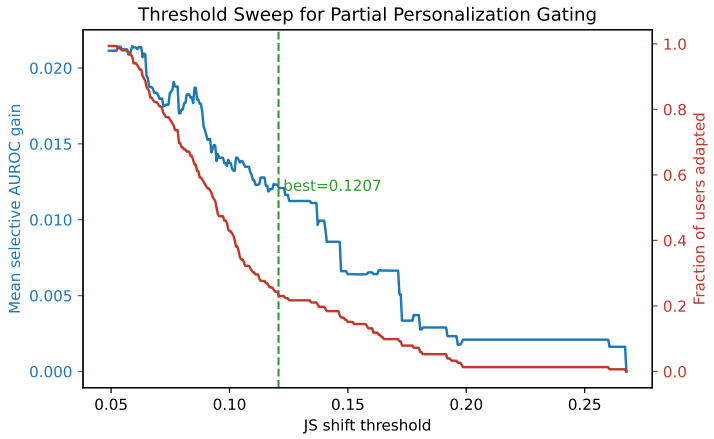


Fig. 13. Illustrative threshold sweep for OOD-guided partial personalization on the College Experience dataset. The best offline operating point occurs at JS = 0.1207, adapting 35/152 users (23.0%) and yielding an estimated selective AUROC gain of +0.0121 over the source-only baseline. This figure is intended as an offline policy illustration rather than a learned deployment controller.

the early target-user data. *Fine-tuning* starts from the source model and further updates it using only the early target-user data [30]. Thus, all three policies use the same early target-user slice, but differ in how aggressively they rely on it.

Across 152 evaluated users, partial personalization was the only policy that improved average performance, increasing AUROC from 0.5778 to 0.5988 (+0.0209). By contrast, target-user-only personalization and fine-tuning reduced average AUROC to 0.5379 (-0.0399) and 0.5569 (-0.0209), respectively. This pattern is expected because the early target-user slice is often small and label-sparse. Training only on this slice can discard useful population-level structure and overfit to limited target-user labels, while fine-tuning can also over-adapt the source model to noisy or imbalanced early observations. Partial personalization is more conservative: it preserves the broader source-user signal while still allowing the model to adjust toward the target user.

Table 8. Comparison of downstream adaptation policies in the illustrative OOD-guided personalization case study on the College Experience dataset. The source-only baseline AUROC is 0.5778.

Policy	Always-adapt AUROC	Δ AUROC vs. source-only	OOD-gated selective gain	Interpretation
Partial personalization	0.5988	+0.0209	+0.0086	Retains source signal while adapting to target user
Target-user-only personalization	0.5379	-0.0399	-0.0004	Data-starved; overfits early target labels
Fine-tuning	0.5569	-0.0209	-0.0027	Over-adapts to sparse early target data

When partial personalization was applied only to users flagged as OOD, it still yielded a positive selective gain over the source-only baseline (AUROC: +0.0086). A post-hoc threshold sweep identified an illustrative best offline operating point at JS shift = 0.1207, adapting 35 of 152 users (23.0%) and yielding an estimated selective AUROC gain of +0.0121 (57.9% of the always-adapt improvement) over the source-only baseline (Figure 13). Together, these results show that CrossShift-style diagnostics can help decide when a validated conservative adaptation policy should be considered, without themselves prescribing the adaptation policy.

6 DISCUSSION

6.1 Lessons Learned: Concept Shift as the Dominant Source of Failure

As illustrated in Figures 6 and 9, mobile sensing for mental health exhibits pronounced *concept shift*, i.e., user-level mapping variation in $P(Y \mid X)$. In RQ1, concept shift is statistically significant across most behavior feature categories; in RQ2, regression analysis identifies concept shift as the dominant contributor to the performance gap between in-distribution (ID; train/test on the same user) and out-of-distribution (OOD; train on one user, test on another). Moreover, this phenomenon is not unique to mobile sensing: even in EEG-based mental-workload prediction, subject-to-subject differences in $P(Y \mid X)$ have been documented and shown to impede cross-subject generalization [5].

This phenomenon is consequential: our results suggest that failure to generalize across users is driven less by differences in label prevalence alone and more by mismatches in the user-specific mapping from behavior to mental-health labels, in other words, concept shift. Achieving adequate generalization under concept shift would require impractically large and diverse cohorts that span the full range of user-level mappings. Moreover, simply adding more data or contextual variation may not resolve this problem; when the mapping between behavior and labels shifts substantially between the training and test sets, additional data can even worsen prediction performance [65].

Part of the concept shift we observe likely reflects genuine differences in behavior: the same behavioral pattern can carry different meaning for different users, since wearable and behavioral signals are strongly context-sensitive [34, 74]. However, some of this variation is plausibly attributable to how labels are collected rather than to behavior itself. Mental-health labels rely on self-report, which is subject to recall bias, mood-dependent reporting, social desirability, and inconsistent use of rating scales [21, 23, 40]; in-the-wild ESM protocols introduce additional noise from missed prompts, delayed responses, and ambiguity in the reporting window [47, 58, 73]. Since the current data do not allow these two sources to be fully separated, we treat the concept shift measured here as reflecting a combination of genuine behavior–label heterogeneity and labeling noise.

The four shift types should therefore be interpreted as complementary rather than mutually exclusive: covariate and conditional shifts characterize differences in behavioral distributions, label shift captures differences in outcome prevalence, and concept shift captures whether similar behavioral patterns carry different meanings across users. Our results suggest that covariate and conditional shifts often co-occur, whereas concept shift remains functionally distinct because it persists after feature alignment and most directly explains cross-user performance degradation.

6.2 How to Mitigate Concept Shift?

Prior work on distribution shifts in mobile mental-health sensing has primarily relied on domain adaptation and domain generalization methods [48, 76], which often assume that the predictive relationship $P(Y \mid X)$ is stable across domains. Our RQ1 and RQ2 results challenge this assumption: concept shift is prevalent across users and is the dominant driver of cross-user generalization failure. This suggests that feature alignment or domain-invariant learning alone is insufficient when the behavior–label mapping itself changes.

Our results therefore refine, rather than merely repeat, the common personalization argument [18, 30, 79, 95]. Rather than treating personalization as a general remedy for cross-user variability, our diagnostics reinforce the case for it by identifying why shared models tend to fail across users: differences in the learned behavior-to-label mapping persist after standard preprocessing and are the aspect of cross-user heterogeneity most associated with performance loss. The design question this raises is therefore not who needs personalization, but how diagnostics like CrossShift can help explain the specific source of a shared model's failure for a given user or subgroup.

A complementary system-oriented strategy is *out-of-distribution (OOD) user detection* with event-triggered adaptation [22, 90]. Instead of retraining on a fixed schedule, as in Time2Stop [66], a deployed system could monitor

incoming data for meaningful distributional deviations and trigger adaptation only when needed, especially when concept shift undermines the learned decision boundary. Stronger generalization models, including tabular foundation models [32], may improve robustness, but our results in Section 5.3.3 suggest that they do not eliminate user-level concept shift. We therefore view CrossShift-style diagnostics as a way to decide when validated adaptation policies should be considered; Section 5.3.4 illustrates this idea in an offline OOD-guided personalization case study.

6.3 Deployment Implications of CrossShift

CrossShift is primarily an analytical framework for retrospective and offline shift analysis. Its outputs may also inform future model monitoring and adaptation decisions in deployable mobile sensing systems, but this paper does not present CrossShift itself as a complete deployment system. By quantifying the magnitude and functional impact of different shift types across behavior feature categories, CrossShift helps clarify which forms of user-level shift are most strongly associated with transferability loss in a given setting.

A practical deployment workflow can be framed as follows. First, CrossShift is applied offline to the training cohort to identify which shift types are most strongly associated with transferability loss. Then, during deployment, incoming user data are monitored for the corresponding shift signals. If covariate or conditional shift increases while the concept-shift proxy remains low, the system may continue monitoring or apply lightweight recalibration or feature alignment. If persistent concept shift is detected, the system may trigger selective user-level adaptation or personalization. If the same shift pattern is observed repeatedly across many users, the cohort model may be updated offline. We emphasize that this is an illustrative workflow rather than a tested closed-loop deployment system.

Our illustrative policy experiment in Section 5.3.4 also highlights an important limitation: a useful shift diagnostic does not by itself determine the correct adaptation policy. In the College Experience case study, the same OOD signal supported a positive selective gain when paired with partial personalization, but not when paired with target-only personalization or fine-tuning. Thus, CrossShift is best viewed as a decision-support layer for policy selection rather than a standalone solution. It can help indicate when a user is atypical and which type of shift is most likely responsible, but effective deployment still requires downstream policies whose usefulness must be validated empirically.

6.4 Cross-Dataset Generalizability

Results from RQ1 and RQ2 reveal substantial cross-dataset variation in both the magnitude and structure of detected distribution shifts. Figure 6 shows that while similar categories of shift recur across datasets, their effect sizes and statistical stability differ markedly, indicating that shift patterns are not directly transferable across cohorts.

Covariate and conditional shifts. For covariate shift $P(X)$ and conditional shift $P(X \mid Y)$, PERMANOVA effect sizes exhibit broadly consistent qualitative patterns across datasets, with nontrivial between-user centroid differences observed in most behavioral categories. However, the *magnitude* of these effects varies substantially. The three collected datasets (D#1–D#3) generally show smaller effect sizes than the public datasets, whereas *GLOBEM* consistently exhibits the largest and most stable centroid differences across categories.

Dispersion effects measured by PERMDISP are more heterogeneous. Across both covariate and conditional shifts, PERMDISP results include more NA entries and less consistent patterns, reflecting the sensitivity of dispersion-based tests to sample size, temporal coverage, and within-user variability.

Concept shift. Cross-dataset differences are most pronounced for concept shift $P(Y \mid X)$. As shown in Figure 6 (panels 7–8), all datasets exhibit statistically significant Cohen's W^2 values across most behavioral categories,

indicating pervasive between-user variation in feature–label relationships. However, *GLOBEM* and *CrossCheck* consistently yield larger effect sizes than the student datasets, with *GLOBEM* showing the strongest concept-shift effects overall.

Why are GLOBEM and CrossCheck larger? A plausible explanation differs across the two datasets. CrossCheck includes a clinical psychiatric population, which plausibly increases between-user behavioral heterogeneity relative to the university-student cohorts. For GLOBEM, the larger effect sizes should be interpreted in light of the fact that we analyze a merged multi-cohort dataset rather than a single homogeneous cohort. This construction increases cross-user diversity and can amplify detectable shift magnitudes. This caution is especially important for RQ2, where only five GLOBEM users remain after the additional eligibility and self-AUROC filtering in Table 5, limiting the stability of dataset-specific interpretations. These larger effect sizes should therefore not be interpreted as directly comparable measures of pure behavioral heterogeneity, since cohort composition, psychiatric/clinical status, dataset construction, sensing modalities, labeling practices, and analytic filtering differ substantially across datasets.

Implications. Overall, these findings demonstrate that shift effect sizes are strongly *dataset-dependent* and should not be interpreted as directly comparable across cohorts. Differences in sensing modalities, labeling practices, cohort composition, and dataset construction can all substantially influence the magnitude and stability of detected shifts. Consequently, shift analysis and mitigation strategies should be evaluated within each dataset's population and data-collection context, motivating dataset-aware and user-adaptive modeling approaches rather than one-size-fits-all solutions.

6.5 Exploratory Demographic and Psychometric Follow-Up

Demographic and psychometric variables are static within a user and therefore play no role in the core CrossShift evaluations, which quantify changes in behavioral feature and label distributions. As a post hoc check, we examined whether these static attributes are associated with the shift magnitudes and slopes already computed in RQ1 and RQ2, without rerunning either pipeline. For RQ1, we related age, gender, and psychometric traits to each user's existing covariate and conditional shift scores; for RQ2, we treated the same variables as moderators of the existing shift–performance slopes. The resulting associations are modest and inconsistent rather than indicative of a clear signal: in RQ1, female users tend toward lower shift magnitudes in phone usage and sleep in several datasets, whereas extraversion and conscientiousness show limited consistency across datasets. In RQ2, female gender and higher conscientiousness are associated with lower covariate-shift slope sensitivity, while higher PHQ-9 scores are associated with greater sensitivity. These results should be interpreted as exploratory and non-causal; rather than revising the shift metrics themselves, they suggest that demographic and trait variables may be useful as grouping signals for selective personalization in future work.

6.6 Limitations & Future Work

Limitations. RQ2 analyzes each shift distance independently using separate regressions, following prior work [24]. This supports interpretable shift-specific attribution, but it cannot model interactions among shifts that often co-occur. Our exploratory joint linear models suffered from strong collinearity and unstable coefficients, suggesting that future work should use more robust multivariate approaches such as regularized, hierarchical, or non-linear models.

RQ2 also uses a fixed XGBoost protocol without dataset-specific tuning, and concept shift is estimated through predictive models that approximate $P(Y \mid X)$. This improves cross-dataset comparability, but absolute performance, coefficient magnitudes, and concept-shift estimates may still depend on model quality and training stability. Our robustness checks across alternative model families partially mitigate this concern, as the main conclusion that concept shift dominates remains unchanged.

For *GLOBEM*, we use a merged multi-cohort dataset to obtain enough users with identifiable personalized models for RQ2. This aggregation may increase detectable heterogeneity, so the *GLOBEM* results should be interpreted as evidence from a heterogeneous multi-cohort setting rather than from a single homogeneous cohort.

Finally, our analyses rely on harmonized binary or binarized prediction targets across datasets with different original label scales and reporting frequencies. This improves comparability, but it may obscure finer-grained affective variation that would be visible under ordinal or continuous labels.

Future work. This study focuses on *between-user* shifts. Future work should extend CrossShift to *within-user* temporal drift, especially in long-term sensing or intervention settings where behavior, mental states, and $P(Y \mid X)$ may change over time. The proposed shift metrics could also be reused as user-pair distances for cohort discovery and group-personalized modeling, such as cluster-based or multi-task adaptation.

7 CONCLUSION

Mobile sensing for mental health exhibits pronounced inter-individual variability, and current models and systems scale poorly across users. We find that *Phone Usage*, *Social Interaction*, and *Sleep* exhibit the most pronounced between-user differences across datasets. More critically, *concept shift*—the user-dependent mapping $P(Y \mid X)$ from behavioral sensor features to mental-wellbeing labels—emerges as the dominant driver of cross-user performance degradation. The primary contribution of this paper is therefore a structured empirical framework and corresponding cross-dataset evidence for decomposing user-level covariate, label, conditional, and concept shifts in mobile mental-health sensing. We hope this analytical framing and its released implementation support future work on shift-aware evaluation, monitoring, and adaptation in mobile mental-health sensing and related mobile human sensing domains.

Acknowledgments

The corresponding author of this work is Uichin Lee. This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2025-02305801 and RS-2026-25552484).

References

[1] Lala Acharya, Lan Jin, and William Collins. 2018. College life is stressful today – Emerging stressors and depressive symptoms in college students. *Journal of American College Health* 66, 7 (2018), 655–664. arXiv:https://doi.org/10.1080/07448481.2018.1451869 doi:10.1080/07448481.2018.1451869 PMID: 29565759.

[2] Daniel A Adler, Fei Wang, David C Mohr, and Tanzeem Choudhury. 2022. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE* 17, 4 (2022), e0266516. doi:10.1371/journal.pone.0266516

[3] Daniel A. Adler, Yuewen Yang, Thalia Viranda, Xuhai Xu, David C. Mohr, Anna R. Van Meter, Julia C. Tartaglia, Nicholas C. Jacobson, Fei Wang, Deborah Estrin, and Tanzeem Choudhury. 2024. Beyond Detection: Towards Actionable Sensing Research in Clinical Mental Healthcare. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 160 (Nov. 2024), 33 pages. doi:10.1145/3699755

[4] Mohamed Akrout, Amal Feriani, Faouzi Bellili, Amine Mezghani, and Ekram Hossain. 2023. Domain Generalization in Machine Learning Models for Wireless Communications: Concepts, State-of-the-Art, and Open Issues. *Commun. Surveys Tuts.* 25, 4 (Oct. 2023), 3014–3037. doi:10.1109/COMST.2023.3326399

[5] Isabela Albuquerque, João Monteiro, Olivier Rosanne, and Tiago H. Falk. 2022. Estimating distribution shifts for predicting cross-subject generalization in electroencephalography-based mental workload assessment. *Frontiers in Artificial Intelligence* 5, Article 992732 (Oct. 2022). doi:10.3389/frai.2022.992732

[6] David Alvarez-Melis and Nicolò Fusi. 2020. Geometric dataset distances via optimal transport. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1799, 12 pages.

[7] Marti J. Anderson. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26, 1 (2001), 32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x

[8] Marti J. Anderson and Daniel C. I. Walsh. 2013. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs* 83, 4 (2013), 557–574. doi:10.1890/12-2010.1

[9] Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Miriam Bidoglia, Sally Stares, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, José Luis Zarza, Alethia Hume, Luca Cernuzzi, Ivano Bison, Marcelo Dario Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Fausto Giunchiglia, and Daniel Gatica-Perez. 2023. Complex Daily Activities, Country-Level Diversity, and Smartphone Sensing: A Study in Denmark, Italy, Mongolia, Paraguay, and UK. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 506, 23 pages. doi:10.1145/3544548.3581190

[10] Dror Ben-Zeev, Rui Wang, Sanjukta Abdullah, Rachel Brian, Eleanor A Scherer, Laura A Mistler, and Andrew T Campbell. 2017. CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric Rehabilitation Journal* 40, 3 (2017), 266.

[11] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.

[12] D. D. Boos and J. M. Hughes-Oliver. 2000. How Large Does n Have to be for Z and t Intervals? *The American Statistician* 54, 2 (2000), 121–128. doi:10.1080/00031305.2000.10474532

[13] Jordan Li Cahoon and Luis Antonio Garcia. 2023. Continuous Stress Monitoring for Healthcare Workers: Evaluating Generalizability Across Real-World Datasets. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Houston, TX, USA) *(BCB '23)*. Association for Computing Machinery, New York, NY, USA, Article 31, 5 pages. doi:10.1145/3584371.3612974

[14] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A Systematic Study of Unsupervised Domain Adaptation for Robust Human-Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 39 (March 2020), 30 pages. doi:10.1145/3380985

[15] Zi-Jian Cheng, Zi-Yi Jia, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. 2025. Realistic Evaluation of TabPFN v2 in Open Environments. arXiv:2505.16226 [cs.LG] https://arxiv.org/abs/2505.16226

[16] Zi-Jian Cheng, Zi-Yi Jia, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. 2025. TabFSBench: Tabular Benchmark for Feature Shifts in Open Environments. arXiv:2501.18935 [cs.LG] https://arxiv.org/abs/2501.18935

[17] Matteo Ciman and Katarzyna Wac. 2018. Individuals' Stress Assessment Using Human-Smartphone Interaction Analysis. *IEEE Transactions on Affective Computing* 9, 1 (2018), 51–65. doi:10.1109/TAFFC.2016.2592504

[18] Ting Dang, Dimitris Spathis, Abhirup Ghosh, and Cecilia Mascolo. 2023. Human-centred artificial intelligence for mobile health sensing: challenges and opportunities. *Royal Society Open Science* 10, 11 (Nov. 2023), 230806. doi:10.1098/rsos.230806

[19] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M. Mattingly, Gregory D. Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 374, 16 pages. doi:10.1145/3491102.3502037

[20] Raihana Ferdous, Venet Osmani, and Oscar Mayora. 2015. Smartphone app usage as a predictor of perceived stress levels at workplace. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. 225–228. doi:10.4108/icst.pervasivehealth.2015.260192

[21] Adrian Furnham and Monika Henderson. 1982. The Good, the Bad and the Mad: Response Bias in Self-report Measures. *Personality and Individual Differences* 3, 3 (1982), 311–320.

[22] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.* 46, 4, Article 44 (March 2014), 37 pages. doi:10.1145/2523813

[23] Nan Gao, Soundariya Ananthan, Chun Yu, Yuntao Wang, and Flora D. Salim. 2023. Critiquing Self-report Practices for Human Mental and Wellbeing Computing at Ubicomp. arXiv:2311.15496 [cs.HC] https://arxiv.org/abs/2311.15496

[24] Josh Gardner, Zoran Popovic, and Ludwig Schmidt. 2023. Benchmarking Distribution Shift in Tabular Data with TableShift. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 53385–53432. https://proceedings.neurips.cc/paper_files/paper/2023/file/a76a757ed479a1e6a5f8134bea492f83-Paper-Datasets_and_Benchmarks.pdf

[25] Taesik Gong, Yewon Kim, Adiba Orzikulova, Yunxin Liu, Sung Ju Hwang, Jinwoo Shin, and Sung-Ju Lee. 2023. DAPPER: Label-Free Performance Estimation after Personalization for Heterogeneous Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 55 (June 2023), 27 pages. doi:10.1145/3596256

[26] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. 2019. MetaSense: few-shot adaptation to untrained conditions in deep mobile sensing. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems* (New York, New York) *(SenSys '19)*. Association for Computing Machinery, New York, NY, USA, 110–123. doi:10.1145/3356250.3360020

[27] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34.

Curran Associates, Inc., 18932–18943. https://proceedings.neurips.cc/paper_files/paper/2021/file/9d86d83f925f2149e9edb0ac3b49229c-Paper.pdf

[28] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research* 13, 25 (2012), 723–773.

[29] Hirotaka Hachiya, Masashi Sugiyama, and Naonori Ueda. 2012. Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. *Neurocomputing* 80 (2012), 93–101.

[30] Yunjo Han, Panyu Zhang, Minseo Park, and Uichin Lee. 2024. Systematic Evaluation of Personalized Deep Learning Models for Affect Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 206 (Nov. 2024), 35 pages. doi:10.1145/3699724

[31] Karishma Hegde and Hemadri Jayalath. 2025. Emotions in the Loop: A Survey of Affective Computing for Emotional Support. arXiv:2505.01542 [cs.HC] https://arxiv.org/abs/2505.01542

[32] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2022. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. arXiv:2207.01848 [cs.LG]

[33] Xin Huang, Aakanksha Khetan, Marian Cvitkovic, and Zohar Karnin. 2020. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. *arXiv preprint arXiv:2012.06678* (2020).

[34] Kaixin Ji. 2025. *Measuring Information Behaviors and Cognitive Bias Using Wearables.* Ph. D. Dissertation. RMIT University, Melbourne, Australia.

[35] Gyuwon Jung and Uichin Lee. 2025. CounterStress: Enhancing Stress Coping Planning through Counterfactual Explanations in Personal Informatics. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 714, 20 pages. doi:10.1145/3706598.3713730

[36] Nathan Kammoun, Lakmal Meegahapola, and Daniel Gatica-Perez. 2023. Understanding the Social Context of Eating with Multimodal Smartphone Sensing: The Role of Country Diversity. In *Proceedings of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23).* Association for Computing Machinery, New York, NY, USA, 604–612. doi:10.1145/3577190.3614129

[37] Hua Kang, Qingyong Hu, and Qian Zhang. 2024. SF-Adapter: Computational-Efficient Source-Free Domain Adaptation for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 164 (Jan. 2024), 23 pages. doi:10.1145/3631428

[38] Pixi Kang, Julian Moosmann, Mengxi Liu, Bo Zhou, Michele Magno, Paul Lukowicz, and Sizhen Bian. 2025. Bridging Generalization and Personalization in Wearable Human Activity Recognition via On-Device Few-Shot Learning. arXiv:2508.15413 [cs.LG] https://arxiv.org/abs/2508.15413

[39] Soowon Kang, Woohyeok Choi, Cheul Young Park, Narae Cha, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Heepyung Kim, Yong Jeong, and Uichin Lee. 2023. K-EmoPhone: A Mobile and Wearable Dataset with In-Situ Emotion, Stress, and Attention Labels. *Scientific Data* 10, 1 (2023), 351. doi:10.1038/s41597-023-02248-2

[40] Kate Kelley, Belinda Clark, Vivienne Brown, and John Sitzia. 2003. Good Practice in the Conduct and Reporting of Survey Research. *International Journal for Quality in Health Care* 15, 3 (2003), 261–266. doi:10.1093/intqhc/mzg031

[41] Manolis Kellis. 2023. Lecture 14: Dataset Shift. https://mlhcmit.github.io/2023/lectures/lecture14_dataset_shift.pdf. Machine Learning for Healthcare course materials, accessed on August 29, 2025.

[42] Azhar Ali Khaked, Nobuyuki Oishi, Daniel Roggen, and Paula Lago. 2025. In Shift and In Variance: Assessing the Robustness of HAR Deep Learning Models against Variability. *Sensors* 25, 2 (2025), 430. doi:10.3390/s25020430

[43] Muhammad Kifayathullah, Rajeev Sekar, Abinaya R, and Vijayaprabakaran K. 2025. Personalized Mental Health Assistance: Integrating Emotion Prediction with GPT-Based Chatbot. In *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS).* 1–6. doi:10.1109/SCEECS64059.2025.10940203

[44] Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, and Hwajung Hong. 2022. Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 279, 20 pages. doi:10.1145/3491102.3517701

[45] Ilina Kirovska and Veljko Pejović. 2025. The Reports of My Capabilities Are Greatly Exaggerated - Small LLMs for Depression Inference from Mobile Sensing Data. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Finland) *(UbiComp Companion '25).* Association for Computing Machinery, New York, NY, USA, 1223–1228. doi:10.1145/3714394.3756240

[46] Nicholas D. Lane, Ye Xu, Hong Lu, Shaohan Hu, Tanzeem Choudhury, Andrew T. Campbell, and Feng Zhao. 2011. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11).* Association for Computing Machinery, New York, NY, USA, 355–364. doi:10.1145/2030112.2030160

[47] Ha Le, Veronika Potter, Akshat Choube, Rithika Lakshminarayanan, Varun Mishra, and Stephen Intille. 2025. A Context-assisted, Semi-automated Activity Recall Interface Allowing Uncertainty. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 4, Article 186 (2025), 34 pages. doi:10.1145/3770710

[48] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[49] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. 2016. Revisiting Batch Normalization For Practical Domain Adaptation. *arXiv preprint arXiv:1603.04779* (2016). https://arxiv.org/abs/1603.04779

[50] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018. Detecting and Correcting for Label Shift with Black Box Predictors. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 3122–3130. https://proceedings.mlr.press/v80/lipton18a.html

[51] Xiaofeng Liu, Zhenhua Guo, Site Li, Fangxu Xing, Jane You, C.-C. Jay Kuo, Georges El Fakhri, and Jonghye Woo. 2021. Adversarial Unsupervised Domain Adaptation with Conditional and Label Shift: Infer, Align and Iterate. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 10347–10356. doi:10.1109/ICCV48922.2021.01020

[52] Wang Lu, Jindong Wang, Yiqiang Chen, Sinno Jialin Pan, Chunyu Hu, and Xin Qin. 2022. Semantic-Discriminative Mixup for Generalizable Sensor-based Cross-domain Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 65 (July 2022), 19 pages. doi:10.1145/3534589

[53] Aurel Ruben Mäder, Lakmal Meegahapola, and Daniel Gatica-Perez. 2024. Learning About Social Context From Smartphone Data: Generalization Across Countries and Daily Life Moments. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 537, 18 pages. doi:10.1145/3613904.3642444

[54] Felipe Maia Polo, Rafael Izbicki, Evanildo Gomes Lacerda, Juan Pablo Ibieta-Jimenez, and Renato Vicente. 2023. A unified framework for dataset shift diagnostics. *Information Sciences* 649 (2023), 119612. doi:10.1016/j.ins.2023.119612

[55] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 176 (Jan. 2023), 32 pages. doi:10.1145/3569483

[56] Lakmal Meegahapola, Hamza Hassoune, and Daniel Gatica-Perez. 2024. M3BAT: Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 46 (May 2024), 30 pages. doi:10.1145/3659591

[57] Varun Mishra, Tian Hao, Si Sun, Kimberly N. Walter, Marion J. Ball, Ching-Hua Chen, and Xinxin Zhu. 2018. Investigating the Role of Context in Perceived Stress Detection in the Wild. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 1708–1716. doi:10.1145/3267305.3267537

[58] Varun Mishra, Byron Lowens, Sarah Lord, Kelly Caine, and David Kotz. 2017. Investigating Contextual Cues as Indicators for EMA Delivery. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. Association for Computing Machinery, 935–940. doi:10.1145/3123024.3124571

[59] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. 2020. Evaluating the Reproducibility of Physiological Stress Detection Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 147 (Dec. 2020), 29 pages. doi:10.1145/3432220

[60] José G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. 2012. A Unifying View on Dataset Shift in Classification. *Pattern Recognition* 45, 1 (2012), 521–530. doi:10.1016/j.patcog.2011.06.019

[61] Borum Nam, Joo Young Kim, In Young Kim, and Baek Hwan Cho. 2022. Selective Prediction With Long Short-term Memory Using Unit-Wise Batch Standardization for Time Series Health Data Sets: Algorithm Development and Validation. *JMIR Medical Informatics* 10, 4 (2022), e30587.

[62] Alexandre Nanchen, Lakmal Meegahapola, William Droz, and Daniel Gatica-Perez. 2023. Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 217–228. doi:10.1145/3600211.3604688

[63] Philip Naumann, Jacob Kauffmann, and Grégoire Montavon. 2025. Wasserstein Distances Made Explainable: Insights into Dataset Shifts and Transport Phenomena. arXiv:2505.06123 [cs.LG] https://arxiv.org/abs/2505.06123

[64] Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F. Huckins, Courtney Rogers, Meghan L. Meyer, and Andrew T. Campbell. 2024. Capturing the College Experience: A Four-Year Mobile Sensing Study of Mental Health, Resilience and Behavior of College Students during the Pandemic. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 38 (March 2024), 37 pages. doi:10.1145/3643501

[65] Alex Nguyen, David J. Schwab, and Vudtiwat Ngampruetikorn. 2026. Generalization vs Specialization under Concept Shift. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=kKPP6q5vBs

[66] Adiba Orzikulova, Han Xiao, Zhipeng Li, Yukang Yan, Yuntao Wang, Yuanchun Shi, Marzyeh Ghassemi, Sung-Ju Lee, Anind K Dey, and Xuhai Xu. 2024. Time2Stop: Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 250, 20 pages. doi:10.1145/3613904.3642747

[67] Lakyntiew Pariat, Angelyne Rynjah, Joplin, and M G Kharjana. 2014. Stress Levels of College Students: Interrelationship between Stressors and Coping Strategies. *IOSR Journal of Humanities and Social Science* 19 (2014), 40–45. https://api.semanticscholar.org/CorpusID:27178242

[68] Rosalind W Picard. 2016. Automating the recognition of stress and emotion: From lab to real-world impact. *IEEE MultiMedia* 23, 3 (2016), 3–7.

[69] Arvind Pillai, Subigya Kumar Nepal, Weichen Wang, Matthew Nemesure, Michael Heinz, George Price, Damien Lekkas, Amanda C. Collins, Tess Griffin, Benjamin Buck, Sarah Masud Preum, Trevor Cohen, Nicholas C. Jacobson, Dror Ben-Zeev, and Andrew Campbell. 2024. Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 174 (Jan. 2024), 38 pages. doi:10.1145/3631452

[70] Sergei Popov, Stanislav Morozov, and Artem Babenko. 2020. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1eiu2VtwH

[71] Joaquín Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (Eds.). 2009. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA. https://direct.mit.edu/books/edited-volume/3841/Dataset-Shift-in-Machine-Learning

[72] Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. 2019. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In *Advances in Neural Information Processing Systems*, Vol. 32. 13962–13973. https://proceedings.neurips.cc/paper/2019/file/846c260d715e5b854ffad5f70a516c88-Paper.pdf

[73] Mashfiqui Rabbi, Katherine Li, H. Yanna Yan, Kelly Hall, Predrag Klasnja, and Susan Murphy. 2019. ReVibe: A Context-assisted Evening Recall Approach to Improve Self-report Adherence. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4, Article 149 (Dec. 2019), 27 pages. doi:10.1145/3369806

[74] René Riedl, Fred D. Davis, and Alan R. Hevner. 2014. Towards a NeuroIS Research Methodology: Intensifying the Discussion on Methods, Tools, and Measurement. *Journal of the Association for Information Systems* 15, 10 (2014), I–XXXV. doi:10.17705/1jais.00377

[75] Akane Sano and Rosalind W. Picard. 2013. Stress Recognition Using Wearable Sensors and Mobile Phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 671–676. doi:10.1109/ACII.2013.117

[76] Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime Carbonell, and Kun Zhang. 2021. Domain Adaptation with Invariant Representation Learning: What Transformations to Learn?. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 24791–24803. https://proceedings.neurips.cc/paper_files/paper/2021/file/cfc5d9422f0c8f8ad796711102dbe32b-Paper.pdf

[77] Thomas Stütz, Thomas Kowar, Michael Kager, Martin Tiefengrabner, Markus Stuppner, Jens Blechert, Frank H. Wilhelm, and Simon Ginzinger. 2015. Smartphone Based Stress Prediction. In *User Modeling, Adaptation and Personalization*, Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless (Eds.). Springer International Publishing, Cham, 240–251.

[78] Baochen Sun and Kate Saenko. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. *arXiv preprint arXiv:1607.01719* (2016). doi:10.48550/arXiv.1607.01719

[79] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2020. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* 11, 2 (2020), 200–213. doi:10.1109/TAFFC.2017.2784832

[80] Kobiljon Toshnazarov, Uichin Lee, Byung Hyung Kim, Varun Mishra, Lismer Andres Caceres Najarro, and Youngtae Noh. 2024. SOSW: Stress Sensing With Off-the-Shelf Smartwatches in the Wild. *IEEE Internet of Things Journal* 11, 12 (2024), 21527–21545. doi:10.1109/JIOT.2024.3375299

[81] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. 2023. Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices. *Journal of Biomedical Informatics* 148 (2023), 104556. doi:10.1016/j.jbi.2023.104556

[82] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14.

[83] Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2017. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 110 (Sept. 2017), 24 pages. doi:10.1145/3130976

[84] Gerhard Widmer and Miroslav Kubat. 1996. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning* 23, 1 (1996), 69–101. doi:10.1007/BF00116900

[85] Lauren J. Wong and Alan J. Michaels. 2022. Transfer Learning for Radio Frequency Machine Learning: A Taxonomy and Survey. *Sensors* 22, 4 (2022), 1416. doi:10.3390/s22041416

[86] Yi Xiao, Harshit Sharma, Sawinder Kaur, Dessa Bergen-Cico, and Asif Salekin. 2025. Human Heterogeneity Invariant Stress Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 142 (Sept. 2025), 42 pages. doi:10.1145/3749465

[87] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 190 (Jan. 2023), 34 pages. doi:10.1145/3569485

[88] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E. Morris, Eve Riskin, Jennifer Mankoff, and Anind K. Dey. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. In *36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://arxiv.org/abs/2211.02733

[89] Meng Xue, Yinan Zhu, Wentao Xie, Zhixian Wang, Yanjiao Chen, Kui Jiang, and Qian Zhang. 2025. MobHAR: Source-free Knowledge Transfer for Human Activity Recognition on Mobile Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 22 (March 2025), 24 pages. doi:10.1145/3712620

[90] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. 2022. OpenOOD: benchmarking generalized out-of-distribution detection. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2362, 14 pages.

[91] Xiaoyu Ye, Kai Sakurai, Nitika Nair, and Kevin I-Kai Wang. 2024. Machine Learning Techniques for Sensor-Based Human Activity Recognition with Data Heterogeneity—A Review. *Sensors* 24, 24 (2024), 7975. doi:10.3390/s24247975

[92] Hyungjun Yoon, Jaehyun Kwak, Biniyam Aschalew Tolera, Gaole Dai, Mo Li, Taesik Gong, Kimin Lee, and Sung-Ju Lee. 2025. SelfReplay: Adapting Self-Supervised Sensory Models via Adaptive Meta-Task Replay. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems* (UC Irvine Student Center., Irvine, CA, USA) *(SenSys '25)*. Association for Computing Machinery, New York, NY, USA, 226–239. doi:10.1145/3715014.3722066

[93] Han Yu and Akane Sano. 2023. Semi-Supervised Learning for Wearable-based Momentary Stress Detection in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 80 (June 2023), 23 pages. doi:10.1145/3596246

[94] Tahmina Zebin, Matthew Sperrin, Niels Peek, and Alexander J Casson. 2018. Human activity recognition from inertial sensor time-series using batch normalized deep LSTM recurrent networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1–4.

[95] Panyu Zhang, Gyuwon Jung, Jumabek Alikhanov, Uzair Ahmed, and Uichin Lee. 2024. A Reproducible Stress Prediction Pipeline with Mobile Sensor Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 143 (Sept. 2024), 35 pages. doi:10.1145/3678578

[96] Zhijun Zhou, Yingtian Zhang, Xiaojing Yu, Panlong Yang, Xiang-Yang Li, Jing Zhao, and Hao Zhou. 2020. XHAR: Deep Domain Adaptation for Human Activity Recognition with Smart Devices. In *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. 1–9. doi:10.1109/SECON48991.2020.9158431

A APPENDIX

A.1 Institution-Collected Cohorts: D#1–D#3

D#1, D#2, and D#3 are three longitudinal in-the-wild mobile sensing cohorts collected across 2020–2022 from university students over approximately four-week periods. Each cohort combines in-situ ESM-based affective labels with smartphone and wearable sensing collected in naturalistic daily-life settings.

A.1.1 Recruitment and Study Context. Participants were recruited from the same university context and completed the study using their own Android smartphones. Before data collection began, participants attended an offline briefing session and provided written informed consent under an IRB-approved protocol. The data were collected in everyday settings rather than laboratory conditions.

A.1.2 Collection Conditions. Across the three annual waves, the overall collection design remained broadly similar: all cohorts used dense ESM prompting, continuous smartphone sensing, and Fitbit-based wrist sensing. ESM prompt frequency was intentionally reduced from 15 to 12 to 10 prompts per day across D#1–D#3, respectively. Keystroke sensing was introduced from D#2 onward.

A.1.3 Dataset Construction. The final analyzed samples for D#1, D#2, and D#3 were 92, 99, and 106 participants, respectively. These cohorts were derived through dataset-level preprocessing and screening procedures intended

to ensure sufficient usable sensing and ESM data quality for downstream analysis. We summarize the resulting cohort sizes and data volumes in Table 2.

Dataset-level quality control for D#1–D#3. For the three newly collected datasets, participants were excluded during dataset construction if their wearable coverage, smartphone coverage, or ESM compliance fell below the predefined quality thresholds. In addition, sensor streams with inadequate collection coverage were excluded from modeling. These dataset-level QC procedures were applied before any RQ-specific analysis.

A.1.4 Study Design (D#1, D#2, D#3). We next summarize the study objectives, participant procedures, ESM schedules, sensing instrumentation, compliance support, and data-collection outcomes for the three institution-collected cohorts.

D#1 (4-week cohort; Android 7.0+).

- **Objectives.** This dataset incorporates a reasonable logging period of four weeks and a larger cohort size, while maintaining frequent in-situ affect sampling for richer temporal resolution.
- **Participants & ethics.** Campus recruitment; in-person briefing (1 hour); IRB-approved informed consent. Targeted daily active logging ≥ 12 h.
- **ESM schedule.** 15 random prompts/day during 10:00–22:00 (12 h window). Average spacing ≈ 50 min; minimum gap 30 min. Each prompt expired after 10 min (non-response removed to reduce recall bias).
- **Label scope.** 7-point (−3 . . . + 3) scales for Valence, Arousal, Stress, Attention, Task Disturbance, Mental Load; free-text/interval Emotion Duration; post-answer Emotion Change item.
- **Instrumentation.** Personal Android smartphone sensing plus two wearables: Fitbit Inspire HR (continuous). Smartphone app collected mobility, device, usage, and network context; wearables captured heart rate and activity.
- **Compliance support.** Remote progress checks every 3 days; targeted reminders if uploads or ESM response rates lag.
- **Outcome.** N = 102 (32 female); mean age 22.9 (SD = 3.4; 18–32); total ESM = 10,773 (mean/person 105.6, SD = 55.3). Participants received ∼ $100.

D#2 (4-week cohort; Android 8.0+; richer labels and inputs).

- **Objectives.** Replicate and scale D#1 with (i) additional behavioral signals (smartphone keystroke dynamics) and (ii) refined ESM capturing *changed* affect while answering.
- **Participants & ethics.** Same onboarding/consent flow; daily target ≥ 12 h active logging.
- **ESM schedule.** 12 random prompts/day (10:00–22:00); average spacing ≈ 60 min; minimum gap 30 min; 10 min expiry.
- **Label scope.** As in D#1 plus two "now, while answering" items: Changed Valence and Changed Arousal (both −3 . . . + 3).
- **Instrumentation.** Smartphone sensing (as in D#1) *plus* KeyEvent logging (keyboard type, key-category transitions, inter-key latency, grid distance). Fitbit throughout; Server uploads batched hourly.
- **Compliance support.** Monitoring every 2 days; additional nudges for participants trending below a 50% ESM response rate.
- **Outcome.** N = 112 (27 female); mean age 24.1 (SD = 3.6; 18–33); total ESM = 22,142 (mean/person 197.7, SD = 61.1). Compensation ∼ $100.

D#3 (4-week cohort; Android 8.0+; no chest strap; affect lexicon).

- **Objectives.** Maintain scale while improving wearability (no chest strap) and broaden self-report via per-emotion intensities alongside dimensional affect.
- **Participants & ethics.** Same onboarding/consent; daily target ≥ 12 h active logging.

- **ESM schedule.** 10 random prompts/day (10:00–22:00); average spacing ≈ 70 min; minimum gap 50 min; 10 min expiry.
- **Label scope.** Eight affect-word intensities (0 . . . 6: Happy, Relaxed, Cheerful, Content, Sad, Anxious, Depressed, Angry), plus Valence (−3 . . . + 3), Arousal (−3 . . . + 3), Stress (0 . . . 6), Task Disturbance (0 . . . 6).
- **Instrumentation.** Smartphone sensing (as in D#2, including KeyEvent). Fitbit Inspire HR. Hourly uploads.
- **Compliance support.** Monitoring every 2 days with targeted reminders for low ESM compliance.
- **Outcome.** $N = 115$ (42 female); mean age 23.5 ($SD = 3.4$; 18–34); total ESM = 23,250 (mean/person 202.2, $SD = 38.8$). Compensation ∼ $100.

A.1.5 Data Collection Details (D#1, D#2, D#3). We next describe the shared data-collection conventions, dataset-specific sensing and ESM payloads, and operational monitoring procedures used across D#1, D#2, and D#3.

Common conventions across datasets.

- **Time and units.** All timestamps are Unix epoch milliseconds in UTC. Numeric units are standardized: battery temperature in deci-degree Celsius, light in lux, heart rate in BPM, speed in m/s, distance in meters, and minute summaries for steps and calories on Fitbit.
- **Phone OS and app.** Participants used personal Android devices (D#1: 7.0+; D#2/D#3: 8.0+). The study app logged sensors continuously, recorded ESM fields, and scheduled random prompts within the daily 12-hour window.
- **Privacy transforms.** GPS longitudes received a per-participant fixed random displacement (preserves relative geometry); call/SMS identifiers were one-way hashed (MD5); Wi-Fi/Bluetooth MAC addresses replaced by UUIDs.

D#1 sensor and ESM payloads.

- **ESM fields.** `ResponseTime`, `TriggerTime`, `ReactionTime`; Valence, Arousal, Stress, Emotion Duration (0–60 or NA), Attention, Task Disturbance, Mental Load, Emotion Change (all −3 . . . + 3 unless duration).
- **Smartphone (system).** `BatteryEvent` (status, level%, temperature, plug type); `InstalledApp` (name, package, system flags, install/update times).
- **Smartphone (interaction).** `AppUsageEvent` (Android usage event types); `NotificationEvent` (visibility, category); `CallEvent` (type, hashed contact, duration, presentation, dataUsage); `MessageEvent` (class SMS/MMS, box INBOX/SENT, hashed contact); `DeviceEvent` (screen, power, ringer, Bluetooth, phone-state, etc.).
- **Smartphone (context).** `Location` (accuracy, altitude, latitude, longitude, speed); `ActivityEvent`/ `Activity Transition` (type, confidence, enter/exit); `Light` (5 Hz lux); `WiFiScan` (frequency MHz, RSSI dBm, BSSID UUID); `BluetoothScan` (RSSI dBm, addr UUID); `DataTraffic` (rx/tx bytes, mobile rx/tx, window times).
- **Wearables.** `Fitbit-HeartRate` 1 Hz BPM; `Fitbit-StepCount`/ `Distance`/ `Calorie` minute summaries (steps/min; meters/min; cal/min with activity level 0–3 and METs). Data uploaded in background batches.

D#2 additions and refinements.

- **ESM fields.** `IntendedTime` (scheduled time), `TriggerTime` (actual enable), `ReactionTime`, `ResponseTime`; Valence, Arousal, Stress, Emotion Duration, Attention, Task Disturbance, Mental Load, `ChangedValence`, `ChangedArousal` (all −3 . . . + 3 except duration).
- **Keystroke dynamics.** `KeyEvent`: app package; keyboardType (CHUNJIIN/QWERTY_KOR/OTHERS); prevKeyType and currKeyType (BACKSPACE/KOR/ENG/NUMBER/SPECIAL); inter-key `timeTaken` (ms); key-grid `distance`.
- **Wear schedule.** Fitbit throughout.
- **Upload cadence.** Hourly server synchronization for all streams and ESM.

- **Other streams.** Same smartphone system/interaction/context payloads and sampling as in D#1.

D#3 label redesign and device simplification.

- **ESM fields.** `IntendedTriggerTime`, `ActualTriggerTime`, `ReactionTime`, `ResponseTime`; eight affect-word intensities (0 . . . 6: Happy, Relaxed, Cheerful, Content, Sad, Anxious, Depressed, Angry); Valence (−3 . . . + 3); Arousal (−3 . . . + 3); Stress (0 . . . 6); Task Disturbance (0 . . . 6).
- **Wearables.** Fitbit Inspire HR only; `Fitbit-HeartRate` 1 Hz BPM; minute-level `StepCount`/`Distance`/`Calorie`.
- **Smartphone streams.** Same as D#2 (including `KeyEvent`). Hourly uploads maintained.

Operational monitoring and quality gates (all datasets).

- **Participant monitoring.** Server-side dashboards checked every 3 days (D#1) or every 2 days (D#2/D#3) for ESM compliance, stream coverage, and upload lags; reminders sent as needed.
- **Minimum expectations.** Daily active logging ≥ 12 h; prompt expirations enforced at 10 min; minimum inter-prompt gap enforced (30 min in D#1/D#2; 50 min in D#3).
- **Post-collection.** Devices returned; app uninstalled; post surveys completed (K-BFI-15, PSS-10; plus PHQ-9 and GHQ-12 where applicable).

A.2 Methodological Transparency & Reproducibility Appendix (META)

A.2.1 RQ2: Additional Methodological Details. We provide additional methodological rationale for the RQ2 user-filtering and feature-reduction procedures, treatment of negative regression coefficients, use of regression-based attribution, and fixed XGBoost protocol.

Rationale for user filtering and feature reduction. Cross-user generalization in mobile mental-health sensing is highly unstable when a user's in-distribution signal is weak. When the feature–label association for a user is noisy, the learned decision boundary may align spuriously with those of distant users, yielding unpredictable or misleading out-of-distribution performance. Such users break the expected monotonic relationship between interpersonal distance $d_{a,o}$ and transferability loss $\Delta_{a\to o}$ and severely destabilize regression-based attribution.

We therefore exclude users whose in-distribution AUROC falls below a conservative threshold, retaining only those whose personalized models exhibit a meaningful and learnable signal. In addition, we reduce the feature space to the top-K most informative features prior to distance computation. In high-dimensional mobile sensing data, irrelevant features are weighted equally by most distance metrics and can dominate distances, obscuring the contribution of behaviorally meaningful dimensions. Concentrating on a compact, informative subset mitigates this distortion and yields distances that more faithfully reflect functional similarity between users. Thresholds are chosen to balance stability and statistical power, targeting a final cohort of approximately thirty retained users to support reliable inference [12].

Rationale for non-negative regression coefficients. Although counterintuitive, we occasionally observe negative slopes in preliminary analyses, indicating larger estimated shifts coinciding with smaller performance drops. Similar anti-correlations have been reported in recent tabular shift benchmarks [16, 24]. In our setting, such effects primarily arise from residual noise, heterogeneous feature relevance across users, and unstable in-distribution baselines. We therefore interpret negative slopes as indicating no reliable association and truncate them to zero prior to aggregation. This choice stabilizes inference while preserving the intended interpretation of β as the sensitivity of transferability loss to increasing interpersonal distance.

Why regression coefficients rather than correlation. A natural alternative to regression is to summarize the association between shift distance $d_{a,o}$ and transferability loss $\Delta_{a\to o}$ using a correlation coefficient (e.g., Pearson or Spearman). However, correlation quantifies only the *strength* of association, not its *magnitude*. Our goal in RQ2 is not merely to test whether larger shifts are associated with worse performance, but to quantify *how*

Table 9. Fixed XGBoost configuration used throughout RQ2. The same configuration is used for feature selection, personalized within-user evaluation, cross-user transfer evaluation, and model-based shift estimation.

Parameter	Value
n_estimators	500
max_depth	3
learning_rate	0.1
subsample	1.0
colsample_bytree	1.0
reg_alpha	0.0
reg_lambda	1.0
min_child_weight	1
gamma	0
objective	binary:logistic
eval_metric	auc
tree_method	hist
early_stopping_rounds	None
enable_categorical	False
random_state	42
n_jobs	-1

much performance degrades per unit increase in a specific type of shift. The regression slope β has a direct and interpretable meaning as a sensitivity measure: it estimates the expected increase in transferability loss for a unit increase in interpersonal distance under a given shift type and feature category. This enables direct comparison of effect sizes across shift types and behavioral domains, which correlation coefficients do not support.

In addition, correlation is scale-invariant and conflates effect magnitude with variance structure, whereas our regression formulation preserves the scale of $\Delta_{a \to o}$ (AUROC drop) and yields coefficients that are comparable across datasets, categories, and shift metrics.

RQ2 XGBoost protocol. For RQ2, we use a single fixed XGBoost protocol across all datasets. This protocol is applied consistently to K-EmoPhone, D#1, D#2, D#3, CrossCheck, StudentLife, GLOBEM, College Experience, and the step-count case study. We do not perform dataset-specific hyperparameter tuning, cross-validation, ensembling, or threshold optimization. This design prioritizes comparability for shift-attribution analysis rather than maximizing the predictive AUROC of each individual dataset. The same configuration is used for global top-K feature selection, personalized within-user evaluation, cross-user transfer evaluation, and model-based shift estimation.

A.3 Figures & Tables

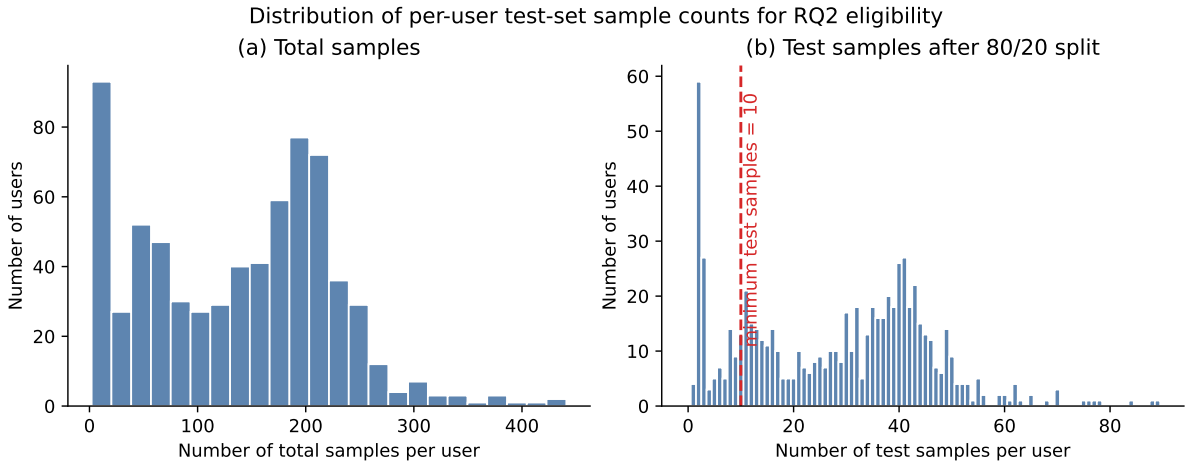


Fig. 14. Per-user sample-count distributions pooled across all datasets included in the RQ2 temporal-split eligibility analysis. (a) Total samples per user before temporal splitting. (b) Test-set sample counts per user after the temporal 80/20 split. The dashed vertical line in (b) marks the minimum test-set size requirement of 10 samples. This cutoff removes only a sparse tail of users with very limited evaluation data and was used to avoid unstable AUROC estimates. User-retention counts after subsequent filtering steps are summarized separately in Table 5.

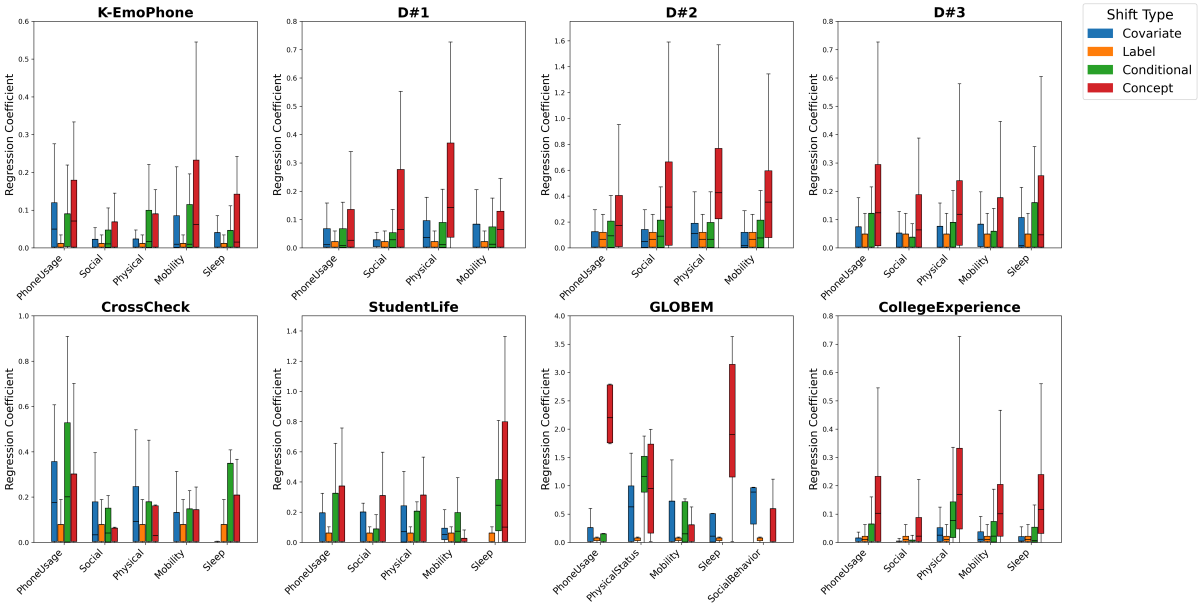


Fig. 15. Results for RQ2: Regression coefficient box plot per behavior feature category & shift type.

Table 10. Consolidated coefficient tests across eight datasets (full). Entries marked with * indicate statistical significance after Benjamini–Hochberg false discovery rate correction (adjusted $p < .05$).

		StudentLife		K-EmoPhone		CrossCheck		D#3		D#2		D#1		GLOBEM		CollegeExp.	
Cat.	**Shift**	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P
Mobility	Concp.	0.325	.001*	0.041	.015*	0.186	.050	0.154	.000*	0.185	.000*	0.126	.000*	0.740	.034*	0.281	.000*
	Cond.	0.134	.002*	0.108	.000*	0.210	.060	0.045	.000*	0.146	.000*	0.071	.000*	0.902	.034*	0.062	.000*
	Covar.	0.140	.022*	0.188	.000*	0.212	.024*	0.047	.000*	0.133	.000*	0.063	.000*	0.485	.054	0.031	.000*
	Label	0.187	.001*	0.021	.013*	0.082	.046*	0.073	.000*	0.098	.000*	0.064	.000*	0.012	.054	0.025	.000*
Phone	Concp.	0.523	.001*	0.059	.005*	0.195	.044*	0.141	.000*	0.185	.000*	0.126	.000*	1.710	.034*	0.287	.000*
	Cond.	0.367	.003*	0.102	.001*	0.049	.085	0.074	.000*	0.060	.000*	0.055	.000*	0.000	.090	0.000	.000*
	Covar.	0.359	.001*	0.144	.000*	0.216	.136	0.120	.000*	0.210	.000*	0.077	.000*	0.000	.090	0.000	.000*
	Label	0.187	.001*	0.021	.013*	0.082	.046*	0.073	.000*	0.098	.000*	0.064	.000*	0.012	.054	0.025	.000*
Phys.	Concp.	0.317	.002*	0.040	.017*	0.187	.043*	0.134	.000*	0.185	.000*	0.127	.000*	0.000	.090	0.474	.000*
	Cond.	0.193	.006*	0.092	.000*	0.215	.101	0.069	.000*	0.202	.003*	0.065	.000*	0.000	.159	0.214	.000*
	Covar.	0.056	.023*	0.132	.001*	0.318	.066	0.085	.000*	0.117	.000*	0.111	.000*	0.000	.090	0.068	.000*
	Label	0.187	.001*	0.021	.013*	0.082	.046*	0.073	.000*	0.098	.000*	0.064	.000*	0.012	.054	0.025	.000*
Sleep	Concp.	0.372	.001*	0.048	.007*	0.160	.045*	0.122	.000*	0.185	.000*	0.126	.000*	1.474	.034*	0.325	.000*
	Cond.	0.155	.007*	0.124	.001*	0.500	.103	0.055	.002*	0.060	.000*	0.055	.000*	0.117	.054	0.018	.000*
	Covar.	0.085	.000*	0.022	.008*	0.464	.017*	0.036	.000*	0.210	.000*	0.077	.000*	0.086	.054	0.010	.000*
	Label	0.187	.001*	0.021	.013*	0.082	.046*	0.073	.000*	0.098	.000*	0.064	.000*	0.012	.054	0.025	.000*
Social	Concp.	0.338	.002*	0.034	.003*	0.174	.047*	0.138	.000*	0.185	.000*	0.127	.000*	0.000	.090	0.061	.000*
	Cond.	0.323	.000*	0.093	.001*	0.116	.096	0.057	.000*	0.202	.003*	0.065	.000*	0.000	.159	0.000	.000*
	Covar.	0.300	.000*	0.049	.003*	0.689	.028*	0.060	.000*	0.117	.000*	0.111	.000*	0.687	.034*	0.000	.000*
	Label	0.187	.001*	0.021	.013*	0.082	.046*	0.073	.000*	0.098	.000*	0.064	.000*	0.012	.054	0.025	.000*