Extended Abstract: Reinforcement Learning for Mental Health Interventions Using Unlabeled Smartphone Data

Motivation From 2013 to 2021, the share of students meeting criteria for at least one mental health condition rose by nearly 50% [Lipson et al., 2022]. Effective mental health interventions require individual-level contextual awareness, yet generic recommendations such as "sleep more" or "exercise more" ignore a student's current state and recent trajectory. Expert labeling for traditional supervised or RL-based mental health systems is expensive and hard to scale. We ask whether offline reinforcement learning can personalize mental health interventions using passively collected smartphone data, with no expert-annotated intervention labels.

Methods We model the problem as a Markov Decision Process (MDP) over the StudentLife dataset [Wang et al., 2014], which tracks 48 college students across mood, sleep, activity, and social behavior over ten weeks via Android smartphones. Because the dataset provides no prescribed interventions, we construct a discrete action space by inferring proxy actions from daily behavioral shifts: the largest normalized deviation across sleep, activity, and social dimensions determines which of 7 actions (`none`, $\pm$sleep, $\pm$activity, $\pm$social) we assign to each transition. We define three reward variants (sparse, dense, and observed-only) that combine short-term and long-term mood improvement, normalized behavioral signals, and observability masks. State vectors encode 3-day behavioral histories, mood, and observability flags ($s_t \in \mathbb{R}^{17}$). We train and evaluate DQN, Double DQN, CQL, BCQ, AWAC, and IQL against behavior cloning, majority action, rule-based, random, action-frequency, and contextual bandit baselines.

Implementation We aggregate StudentLife EMA responses and sensing logs into daily summaries, z-scoring sleep, activity, and social features per student. We build transitions with 3-day lagged state history and split episodes at gaps greater than 2 days. The final dataset contains 3,045 transitions across 165 episodes, split 80/10/10 chronologically. We implement DQN, Double DQN, BCQ, CQL, AWAC, and IQL along with other baselines. We evaluate all policies using Per-Decision Importance Sampling (PDIS), Weighted PDIS, Doubly Robust (DR) estimation, and within-support Δmood.

Results Double DQN achieves the highest estimated value (PDIS $+0.020$, DR $+0.560$), but only 20% of its recommendations overlap with the logging policy. BCQ offers the best safety-performance tradeoff: positive DR ($+0.21$), $\sim$76% behavioral support, and stable training. CQL's DR of -6.01 reflects overconservative Q-value underestimation rather than policy failure, with $\sim$78% behavioral support. All RL methods achieve positive within-support Δmood. A Double DQN reward ablation shows that sparse and observed-only signals (2.5% nonzero) yield near-identical policies with near-zero PDIS, whereas densifying the reward with sleep, activity, and social proxies is required for positive off-policy value (PDIS $+0.020$) and higher logging-policy overlap (20% vs. 13%).

Discussion Our results show that offline RL can surface promising behavior-aware recommendation policies from unlabeled smartphone data. That said, the low effective sample size (ESS $\approx$2–2.5 episodes) in importance-weighted estimators limits confidence in OPE results. We treat actions as inferred proxies rather than true interventions, mood observations cover only $\approx$7% of days, and the dataset includes only 48 students. We present these findings as hypothesis-generating rather than clinically actionable.

Conclusion We build an end-to-end offline RL pipeline for personalizing mental health interventions from unlabeled smartphone data. By converting passive behavioral trajectories into structured MDP transitions and comparing multiple conservative offline RL algorithms, we find that Double DQN and BCQ offer complementary tradeoffs between estimated policy value and distributional safety. Future work should explore other RL models like: TD3+BC, alternative methods of reward modeling, prospective validation through small-scale intervention studies, and potential exploration of other mental health datasets.

Reinforcement Learning for Mental Health Interventions Using Unlabeled Smartphone Data

Alfred Yu
Department of Computer Science
Stanford University
alfredyu@stanford.edu

Elisabeth Holm
Department of Computer Science
Stanford University
eholm@stanford.edu

Juan Pablo Pacheco
Department of Computer Science
Stanford University
pacheco7@stanford.edu

Abstract

Mental health challenges among college students are increasingly common, yet effective interventions depend on an individual's current context and behavior. Generic recommendations such as "sleep more" or "exercise more" may not be equally beneficial for every person at every moment. We investigate whether offline reinforcement learning (RL) can personalize mental health interventions using actively and passively collected smartphone data, with no expert-annotated intervention labels. Using the StudentLife dataset (48 college students tracked over ten weeks), we build a Markov Decision Process by inferring proxy actions from daily behavioral shifts and constructing dense reward signals from mood, sleep, activity, and social interaction. We train and compare DQN, Double DQN, Conservative Q-Learning (CQL), Batch-Constrained Q-Learning (BCQ), Advantage Weighted Actor-Critic (AWAC), and Implicit Q-Learning (IQL) against six baselines, evaluating all policies under Per-Decision Importance Sampling (PDIS), Doubly Robust (DR) estimation, and within-support mood improvement. Double DQN achieves the highest estimated value (PDIS $+0.020$, DR $+0.560$) while BCQ provides the best safety-performance tradeoff with $\sim$76% behavioral support and positive DR ($+0.21$). A reward variant ablation reveals that extreme reward sparsity ($\approx$2.5% nonzero signals) causes the behavior cloning loss to dominate BCQ training, preventing Q-networks from differentiating reward signals. Our results demonstrate the feasibility of offline RL for mental health personalization without expert-labeled data, but highlight the central challenges: sparse rewards, limited sample sizes, and distribution shift under offline evaluation.

1 Introduction

Mental health challenges among college students are increasingly common. From 2013 to 2021, there was nearly a 50% increase in students meeting criteria for at least one mental health condition Lipson et al. (2022). Yet effective interventions are highly specific to an individual's current context and behavior. Generic recommendations such as "try to get good sleep" or "exercise throughout the week" may not be as impactful as concrete, specific, individualized interventions (e.g. "Prioritize increasing sleep today").

Given a student's recent behavioral trajectory and current state, our goal is to learn a policy that recommends the action most likely to improve future well-being. Personalized, data-driven interventions

could enable scalable, accessible mental health support, helping people receive the right support at the right time.

The main obstacle connecting RL algorithms to mental health interventions is data. Training RL-based mental health systems typically requires either randomized controlled trials with prescribed interventions Arévalo Avalos et al. (2025) or expert-labeled datasets, but expert labeling of mental health interventions is costly given the monetary and time cost of hiring these experts. Passive smartphone datasets, by contrast, contain rich behavioral trajectories but no prescribed actions.

Instead of having experts label interventions, we propose studying students' data, detecting shifts, and using those shifts as proxies for interventions. For example, consider a student who started socializing more: we label this as an intervention (if implemented in an app, this would pop up as "Chat with a friend"). We then study what happened after to learn the correlation between such an "intervention" and future outcomes. Our contribution is to convert unlabeled smartphone behavior into an offline RL problem by inferring proxy actions from behavioral shifts and evaluating whether learned policies can improve future mood under support constraints. Specifically, we contribute:

1. An end-to-end offline RL pipeline that constructs MDP transitions (states, proxy actions, reward signals, and episode structure) from the unlabeled StudentLife dataset.
2. Implementation and comparison of DQN, Double DQN, CQL, BCQ, AWAC, and IQL against six baselines, evaluated with three offline policy evaluation estimators.
3. Analysis of results and reward sparsity effects, showing that fewer than 2.5% nonzero reward signals cause the BCQ behavior cloning loss to dominate Q-learning and wash out reward variant differences. We furthermore discuss our results and suggest plausible next steps for exploration.

2 Related Work

StudentLife Dataset. Wang et al. created the StudentLife dataset, containing mental well-being records of 48 students at Dartmouth College across a ten-week period using Android phones Wang et al. (2014). They tracked variables such as stress, sleep, activity, mood, and sociability. This dataset is public and serves as our primary data source. Its mix of passive behavioral signals and periodic self-reports suits offline RL well, though EMA sparsity ($\approx$7% of days with mood observed) poses significant reward modeling challenges.

RL for Mental Health Interventions. Arévalo Avalos et al. demonstrated the promise of RL for mental health interventions Arévalo Avalos et al. (2025). They studied 1,121 adults between December 2021 and July 2022, using random assignment to personalized messaging, static messaging, or messages only monitoring their mood. The researchers found that across study groups, participants demonstrated a 25% reduction in depression symptoms (PHQ-8) and a 24% reduction in anxiety symptoms (GAD-7) post-intervention. Their work establishes the clinical potential of personalized RL-based messaging, but depends on randomized experimental data with prescribed interventions, which passive sensing datasets like StudentLife do not provide.

Our Contribution. While Wang provides the trajectories, they show no intervention or learning. Avalos demonstrate clinical RL benefits but only in a randomized trial with explicit, expert-designed messages (with expert labeling not being feasible for every given case). Our contribution is to bridge the gap, utilizing unlabeled smartphone data as an offline RL problem by inferring proxy actions via day-to-day shifts, removing dependence on otherwise expensive expert annotation. To our knowledge this is the first end-to-end offline RL pipeline built on StudentLife that constructs MDP transitions, reward variants, and episode structure purely from observational sensing data.

3 Methods

3.1 MDP Formulation and Dataset Construction

We model the problem as a finite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$ over daily behavioral summaries. Since we never directly observe the true mental state of a student (only reported mood),

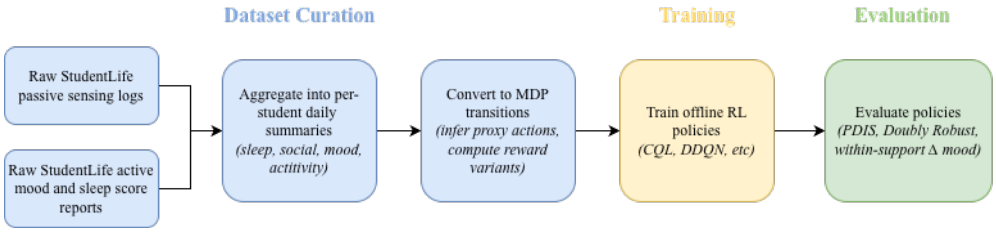

Figure 1: Pipeline overview

the problem is closer to a Partially Observable MDP (POMDP). However, we approximate it as a standard MDP by including a short history of recent observations in the state, capturing temporal patterns without explicitly modeling hidden states.

We aggregate StudentLife EMA responses and sensing logs (accelerometer, WiFi, and Bluetooth proximity) into daily summaries, normalizing sleep, physical activity, and social interactions per student. Gaps greater than 2 days split trajectories into new episodes. We build transitions (s_t, a_t, r_t, s_{t+1}) with 3-day lagged state history. The final dataset contains 3,045 transitions across 165 episodes, split chronologically 80/10/10 into train, validation, and test.

3.1.1 State Space.

The state $s_t \in \mathbb{R}^{17}$ encodes mood, z-scored sleep/activity/social (3 features), 3-day lags of each feature (12 features), and `mood_observed` (1 feature):

- Current-day features: mood score, z-scored sleep duration, z-scored physical activity, z-scored social interaction frequency (4 features)
- Three-day lags of each behavioral feature listed above (3x4 = 12 features)
- Binary indicator: `mood_observed` (1 feature)

We z-score all behavioral features per student to enable cross-student policy learning while respecting individual baselines.

3.1.2 Action Space.

Since StudentLife provides no prescribed interventions, we define a discrete action space $\mathcal{A}$ with 7 actions: `none`, ±sleep, ±activity, ±social. We infer these from the largest normalized daily shift, assigning `none` when the shift falls below 0.6:

$$a_t = \begin{cases} \texttt{none} & \text{if } \max_d |\Delta z_d| < 0.6 \\ \texttt{increase\_}d^* & \text{if } \Delta z_{d^*} > 0.6 \\ \texttt{decrease\_}d^* & \text{if } \Delta z_{d^*} < -0.6 \end{cases}$$

where $d^* = \arg\max_d |\Delta z_d|$ across sleep, activity, and social dimensions. We treat significant behavioral shifts as proxies for intentional behavioral interventions.

3.1.3 Reward Design.

Mood was only directly reported ≈7% of days, so reward sparsity was our central challenge, which we realized after the milestone. We define three reward variants to assess sensitivity to reward formulation:

$$r_{\text{sparse}} = \Delta\text{mood}_{\text{short}} + 0.25\,\Delta\text{mood}_{\text{long}} \tag{1}$$

$$r_{\text{dense}} = 0.60\,\Delta\text{mood}_{\text{short}} + 0.15\,\Delta\text{sleep} + 0.15\,\Delta\text{social} + 0.10\,\Delta\text{activity} \tag{2}$$

$$r_{\text{obs.only}} = r_{\text{sparse}} \text{ only when mood observed at both steps} \tag{3}$$

The dense reward (r_{dense}) incorporates behavioral signals on all days, reducing zero-reward transitions to ≈4.8% (versus ≈97.5% for sparse).

Table 1: Dataset and reward diagnostics.

Total rows	3151
Observed mood (count, fraction)	213 (6.8%)
Reward (dense)	
Valid rows	3151 (100.0%)
Nonzero rows	3001 (95.2%)
Mean (std)	0.0014 (0.2670)
Min / 25% / Median / 75% / Max	-2.589 / -0.119 / 0.000 / 0.116 / 2.989
Reward (sparse)	
Valid rows	188 (6.0%)
Nonzero rows	79 (2.5%)
Mean (std)	0.0735 (0.8515)
Min / 25% / Median / 75% / Max	-4.167 / 0.000 / 0.000 / 0.000 / 4.139
Reward (observed only)	
Valid rows	85 (2.7%)
Nonzero rows	80 (2.5%)
Mean (std)	0.1362 (1.2913)
Min / 25% / Median / 75% / Max	-4.167 / -0.575 / 0.042 / 1.042 / 4.139

Action	Proportion	Mean reward (dense)
none	34.08%	-0.0015
decrease_activity	13.33%	-0.1973
increase_activity	13.20%	0.1801
decrease_social	11.68%	-0.2392
increase_social	11.08%	0.2486
increase_sleep	8.76%	0.2855
decrease_sleep	7.87%	-0.2568

3.2 Iterative Development and Hypothesis Revision

Our original hypothesis was that offline RL on smartphone sensing data from StudentLife could learn intervention policies that improve mental health outcomes and outperform simpler baselines. Unfortunately, our initial runs revealed that we had consistent PDIS estimates of zero, which was mathematically correct but effectively meaningless, since only 2.5% of timesteps carried non-zero rewards. We noted no reliable improvements and incredibly weak signals. Across all policies, only 7 matched steps had real mood data, too few for meaningful inference.

These results revealed that reward sparsity and missing mood labels dominated the evaluation signal. We revised our hypothesis accordingly: due to sparse rewards and limited coverage in StudentLife, offline RL may not reliably learn effective intervention policies; instead, we aim to evaluate whether RL provides any advantage over simpler baselines and under what conditions it becomes viable. This motivated our shift to the dense reward formulation, the addition of other RL variants, and the expanded baseline suite including contextual bandits.

3.3 Algorithms

DQN and Double DQN. We leverage a neural network fitted-Q with experience replay and a target network. Double DQN decouples action selection from value estimation to reduce overoptimism:

$$y_i = r_i + \gamma\, Q_{\phi'}(s'_i, \arg\max_{a'} Q_\phi(s'_i, a')), \quad \mathcal{L}(\phi) = \frac{1}{N}\sum_i \|Q_\phi(s_i, a_i) - y_i\|^2$$

We tune learning rate, batch size, and target-update interval on the validation set.

BCQ (Batch-Constrained Q-Learning). BCQ trains a BC network alongside the Q-network to stay close to logged behavior. Only actions satisfying

$$\frac{\hat{P}(a|s)}{\max_{a'} \hat{P}(a'|s)} \geq \tau = 0.3$$

are eligible; the rest are masked to $-\infty$, preventing OOD overestimation. Unlike CQL's global penalty, BCQ conditions on what π_β would plausibly do.

CQL (Conservative Q-Learning). CQL adds a regularizer that penalizes Q-values on OOD state-action pairs:

$$\mathcal{L}_{\text{CQL}}(\phi) = \mathcal{L}_{\text{TD}}(\phi) + \alpha \left(\mathbb{E}_{s,a\sim\hat{\mu}}[Q_\phi(s,a)] - \mathbb{E}_{s,a\sim\mathcal{D}}[Q_\phi(s,a)] \right).$$

We train for 10,000 steps via d3rlpy with a 2-layer, 256-unit MLP.

IQL (Implicit Q-Learning). We implement a discrete IQL variant in PyTorch (d3rlpy supports only continuous actions). The goal is to improve over π_β without ever querying Q on OOD actions. A SARSA-style update keeps the critic in-support,

$$\hat{Q}^{\pi_\beta} \leftarrow \arg\min_Q \; \mathbb{E}_{(s,a,s',a')\sim\mathcal{D}} \Big[\big(Q(s,a) - (r + \gamma\, Q(s',a')) \big)^2 \Big],$$

but this only evaluates π_β. To improve over it, IQL fits a value network V_ψ to an upper expectile of Q over logged actions via asymmetric regression,

$$\ell_2^\lambda(x) = \begin{cases} (1-\lambda)\, x^2 & x < 0 \\ \lambda\, x^2 & \text{otherwise,} \end{cases}$$

so $\lambda > 0.5$ pulls the fit toward the best in-support actions while still averaging only over logged data. The critic then bootstraps from this value:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}} \big[\ell_2^\lambda \big(Q_\phi(s,a) - V_\psi(s) \big) \big], \quad \mathcal{L}_Q(\phi) = \mathbb{E} \big[\big(r + \gamma\, V_\psi(s') - Q_\phi(s,a) \big)^2 \big].$$

We set $\lambda = 0.9$ and advantage temperature 1.0 by validation grid search (5,000 steps), recovering the policy as an advantage-weighted softmax over $Q_\phi - V_\psi$.

AWAC (Advantage-Weighted Actor-Critic). AWAC shares IQL's goal but extracts an *explicit* actor π_θ. Its critic bootstraps from the next-state value under the current policy rather than a $\max$,

$$\hat{Q}^{\pi_\theta} \leftarrow \arg\min_\phi \; \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}} \left[\Big(Q_\phi(s,a) - \big(r + \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}[Q_\phi(s',a')] \big) \Big)^2 \right],$$

evaluating π_θ while still training only on dataset actions. The actor is updated by advantage-weighted regression,

$$\theta \leftarrow \arg\max_\theta \; \mathbb{E}_{(s,a)\sim\mathcal{D}} \Big[\log \pi_\theta(a \mid s)\, \exp\Big(\tfrac{1}{\lambda}\, \hat{A}^{\pi_\theta}(s,a) \Big) \Big], \quad \hat{A}^{\pi_\theta}(s,a) = Q_\phi(s,a) - V_\psi(s),$$

where λ controls how far the actor departs from π_β ($\lambda \to 0$ is greedy, large λ approaches cloning). We warm-start the critic with the IQL expectile objective (1,000 steps), use its V_ψ as baseline, clip $\exp(\frac{\hat{A}}{\lambda})$ at 100, and select $\lambda = 1.0$ with critic expectile 0.7 over 5,000 steps. Like BCQ, AWAC stays in-support, but via soft weighting rather than a hard mask.

3.4 Offline Policy Evaluation

Because deploying an untested policy on students poses ethical and safety risks, we evaluate all learned policies entirely offline using three complementary estimators. Let π_b denote the logging policy, π the learned policy, and r_t the reward at step t.

Per-Decision Importance Sampling (PDIS). PDIS upweights logged timesteps where the learned policy agrees with the observed action:

$$\hat{V}^{\text{PDIS}}(\pi) = \frac{1}{|D|} \sum_{H\in D} \sum_{t=0}^{T} \left(\prod_{k=0}^{t} \frac{\pi(a_k|s_k)}{\pi_b(a_k|s_k)} \right) r_t.$$

We also report *Weighted* PDIS (WPDIS), which normalizes by the sum of importance weights to reduce variance. Cumulative IS ratios are clipped at $w_{\max} = 5$ and deterministic policies are smoothed with $\epsilon = 0.05$ to avoid zero-probability denominators.

Doubly Robust (DR) Estimation. DR combines a direct Q-network model with an importance-weighted correction, remaining consistent when either the model or the importance weights is correctly specified and achieving lower variance than pure IS. We report DR only for methods that expose a trained Q-network.

Within-support Δmood. On days where the student's logged action matches the policy's recommendation, we compute the mood change to the next observed report. This metric sidesteps IS variance entirely but is informative only when the matched-step count n is large. Because mood is observed on only $\approx$7% of days, n is frequently small ($n \leq 14$ across all methods), so we treat this as a secondary diagnostic rather than a primary ranking criterion.

Behavior Policy. Our logging policy π_b is a logistic-regression behavior-cloning model trained on standardized states (test action match: 89.2%), used in place of empirical action frequencies to avoid near-zero denominator instability on rare actions.

4 Experimental Setup

Baselines. We compare all RL methods against six non-sequential baselines, all trained on the same chronlogical level 80/10/10 split:

- **Random uniform**: selects uniformly over the 7 actions.
- **Majority action**: always predicts the modal action, "none" ($\approx$34% of data).
- **Action frequency**: samples proportional to marginal action counts in the training set.
- **Behavior cloning**: multinomial logistic regression on standardized states (`StandardScaler + LogisticRegression`, `max_iter`=1000, seed 42) predicting the logged action $\pi(a \mid s)$.
- **Rule-based**: corrects the behavioral dimension farthest from the student's personal baseline ($|z| > 0.6$), recommending the corresponding increase/decrease action (else "none").
- **Contextual bandit**: ridge regression ($\alpha = 1.0$) on $[s,\ \mathrm{onehot}(a),\ s \odot \mathrm{onehot}(a)]$ estimating immediate reward $\hat{q}(s, a)$, choosing $\arg\max_a \hat{q}(s, a)$ with no future planning; one-step IPS weights clipped at 10.

Shared setup. All sequential RL methods use 17-dimensional state features (current mood, z-scored sleep/activity/social behavior with 3-day lags, and a mood-observed flag), 7 discrete actions, and discount $\gamma = 0.99$. Missing states and rewards are imputed with 0. We use a chronological train/validation/test split: models are selected on validation and reported on the held-out test split. Offline policy evaluation uses a logistic-regression behavior-cloning logging policy (89.2% test action match) with cumulative importance ratios clipped at $w_{\max} = 5$; deterministic learned policies are softened with $\epsilon = 0.05$.

Training configuration. Unless noted, networks are 256×256 ReLU MLPs trained with Adam. For each method, we have the following setups:

- **CQL** (d3rlpy): lr 6.25×10^{-5}, batch 32, penalty $\alpha = 1.0$, target update every 8000 steps, 10,000 gradient steps. No grid search (early results trailed DQN/Double DQN).
- **BCQ** (PyTorch): separate Q- and BC-networks, lr 1×10^{-4}, batch 64, hard target update every 100 steps, 15,000 steps, $\tau = 0.3$. No tuning.
- **IQL / AWAC** (PyTorch): lr 1×10^{-4}, batch 64, target update every 100 steps, 5,000 steps. Validation grid search over IQL expectile $\{0.7, 0.9\}$ and temperature $\{1.0, 3.0\}$, and AWAC $\lambda \in \{0.5, 1.0\}$.
- **DQN / Double DQN** (d3rlpy): Adam, 256×256 ReLU encoder, 5,000 steps; grid over lr $\{1 \times 10^{-4}, 6.25 \times 10^{-5}\}$, batch $\{32, 64\}$, target update $\{1000, 8000\}$, hidden $\{[64, 64], [256, 256]\}$ on the dense reward (selected by validation PDIS). Best: DQN lr 6.25×10^{-5}/update 8000, Double DQN lr 1×10^{-4}/update 1000, both batch 32.

5 Results

5.1 Quantitative Evaluation

Table 2 summarizes offline evaluation of all policies on the test split under the dense reward formulation. We report PDIS, Weighted PDIS, Doubly Robust estimates where Q-values are available,

Table 2: General performance across all algorithms, with 95% CIs.

Policy	PDIS	WPDIS	DR	Δmood	n	BehSup
CQL	-0.019 ± 0.056	-0.004 ± 0.023	-6.014 ± 2.136	$+0.12 \pm 0.51$	13	0.778
BCQ	-0.005 ± 0.040	-0.005 ± 0.026	$+0.208 \pm 0.538$	$+0.12 \pm 0.51$	13	0.758
Behavior Cloning	-0.003 ± 0.015	-0.003 ± 0.018	N/A	$+0.14 \pm 0.48$	14	0.811
IQL	$+0.006 \pm 0.024$	-0.011 ± 0.073	$+1.167 \pm 0.368$	$+1.83$	1	0.157
AWAC	$+0.008 \pm 0.048$	$+0.004 \pm 0.024$	-0.195 ± 0.304	$+0.04 \pm 0.54$	12	0.765
Majority action	$+0.001 \pm 0.010$	$+0.017 \pm 0.028$	N/A	$+0.80 \pm 0.65$	5	0.399
Random uniform	$+0.006 \pm 0.025$	-0.007 ± 0.076	N/A	$+0.58 \pm 0.82$	2	0.126
Action frequency	$+0.004 \pm 0.004$	$+0.024 \pm 0.015$	N/A	N/A	0	0.212
Rule-based	$+0.011 \pm 0.020$	$+0.030 \pm 0.077$	N/A	-0.83	1	0.055
Contextual bandit	$+0.017 \pm 0.025$	$+0.026 \pm 0.075$	$+0.121^{\dagger}$	N/A	0	0.180
DQN	$+0.017 \pm 0.031$	$+0.034 \pm 0.072$	$+0.279 \pm 0.224$	N/A	0	N/A
Double DQN	$\mathbf{+0.020 \pm 0.030}$	$\mathbf{+0.044 \pm 0.063}$	$\mathbf{+0.560 \pm 0.253}$	$\mathbf{+1.00}$	1	N/A

within-support mood improvement (Δmood), number of matched observations (n), and mean behavior support $\pi_b(a|s)$.

No policy achieves a sugnificant improvement over behavior under offline evaluation. PDIS and WPDIS estimates are near zero for most methods, and confidence intervals generally straddle zero. The strongest estimated values come from lower-support policies: Double DQN has the highest PDIS (+0.020) and DR (+0.560), but only 20% of its recommendations overlap with logged behavior. In contrast, behavior cloning, CQL, BCQ, and AWAC remain much closer to the behavior policy, with support around 0.76–0.81, but their PDIS estimates are near zero. This suggests a core tradeoff: policies that depart from logged behavior may achieve higher estimated value, but their OPE estimates are less trustworthy because they rely on weaker support.

Within-support Δ mood should be treated only as a secondary diagnostic. Although most methods have positive raw mood, matched counts are extremely small: behavior cloning has the largest reliable match count at only n=14, while IQL and Double DQN rely on n=1. Thus, Δ mood does not provide strong evidence of policy improvement. Overall, the main quantitative result is diagnostic rather than confirmatory: the dataset is sufficient to produce differentiated policies, but not sufficient to establish reliable mental-health improvement.

5.2 Reward Variant Ablation

Because mood is observed on only $\sim 7\%$ of days, the training reward is a major design choice. We ablate three reward variants using our best policy by validation PDIS: Double DQN. We train three separate models with identical hyperparameters (best validation-PDIS config on `reward_dense`: lr 10^{-4}, batch 32, target update 1000, 256×256 MLP, 5000 steps) and vary only the reward signal: **sparse** (mood change only), **dense** (mood plus sleep/activity/social proxies; default), and **observed-only** (sparse, masked when mood is unobserved on t or $t-1$). All models are evaluated on the test split with PDIS, WPDIS, and within-support Δmood.

Table 3: Double DQN reward ablation (test split).

Reward	**PDIS**	**WPDIS**	**Match**	**Δmood**	n
Sparse	-0.0001	$+0.0073$	13.1%	$+1.00$	1
Dense	$\mathbf{+0.0198}$	$\mathbf{+0.0435}$	**20.0%**	$+1.00$	1
Observed-only	-0.0001	$+0.0073$	13.1%	$+1.00$	1

The dense reward produces a clearly positive PDIS and higher logging-policy overlap, while sparse and observed-only rewards collapse to near-zero PDIS. Sparse and observed-only are identical (PDIS ≈ 0, match 13.1%), so explicit mood masking does not change the learned policy.

Overall, the results suggest that densifying the reward with behavioral proxies improves Double DQN learning in this low-mood-observation regime, whereas mood-only sparse signals provide too little TD supervision to learn a policy that generalizes under off-policy evaluation.

5.3 Subgroup Policy Analysis

We evaluate the best saved Double DQN model (`reward_dense`) on 14 state-conditioned subgroups derived from the transition data: mood observed/missing; low/high mood; low/normal sleep (sleep_z ≤ -0.5); low/high activity and social ($z \leq \pm 0.5$); weekday/weekend; and early/late term. For each subgroup we report action match with logged behavior, targeted recommendation rates, and offline metrics (PDIS, WPDIS) computed on filtered pseudo-episodes; subgroups with $N < 20$ rows are excluded from the table. These are diagnostic checks of where the policy behaves differently.

Table 4: Double DQN test subgroups with $N \geq 20$.

Subgroup	N	Match	Sleep↑	Soc.↑	PDIS	WPDIS
Mood observed	26	7.7%	0.0%	0.0%	−0.0005	**+0.156**
Mood missing	279	**21.1%**	33.3%	13.6%	+0.022	+0.048
Low sleep	117	19.7%	**35.9%**	9.4%	−0.003	−0.008
Normal sleep	188	20.2%	27.1%	14.4%	+0.033	+0.062
Low social	125	16.8%	27.2%	4.0%	+0.017	+0.047
High social	99	18.2%	24.2%	**26.3%**	+0.040	**+0.206**
High activity	103	16.5%	35.0%	8.7%	**+0.065**	+0.092
Early term	130	23.1%	26.2%	13.1%	+0.007	+0.030

Action match stays below 24% in every reliable subgroup, confirming that Double DQN generally departs from the passive logging policy rather than imitating it.

The clearest state-conditional pattern is for sleep: on low-sleep days the policy recommends `increase_sleep` at 35.9% despite zero such actions in the logged data for that subgroup, suggesting the Q-network uses sleep z-scores to steer toward a sensible intervention direction even when students rarely increased sleep on their own. For engagement the pattern reverses: low-social days receive fewer social-increase recommendations (4.0%) than logged behavior (13.6%), so the policy does not compensate for social deficits.

When mood is directly observed, overlap with logged actions drops to 7.7% and recommendations concentrate on `decrease_social`, indicating the policy behaves qualitatively differently on the small fraction of rows with mood labels, which are states where true outcome feedback is available.

Subgroup PDIS/WPDIS are highest on high-activity and high-social rows, but effective sample size remains low, so these estimates reflect support-limited variance rather than robust subgroup rankings. Per-student analysis shows the same heterogeneity, as only one test student meets the 20-row threshold, and individual action match ranges from 0% to 67%. Overall, the policy shows partial contextual structure for sleep but not social engagement.

5.4 Qualitative Analysis

Action Distribution. The action distribution heatmap (Figure 2) shows that CQL and BCQ select broadly similar actions, with both favoring `none`. Given that BCQ achieves substantially higher DR, the performance gap is unlikely to be explained by coarse differences in action coverage alone. This suggests the main difference lies in value estimation / off-policy evaluation quality, rather than in visibly different action-selection patterns at the aggregate level. However, because the heatmap collapses over states, this does not rule out important state-conditional policy differences, especially in the relatively rare non-none timesteps.

Behavior Support vs. Mood Improvement. Figure 3 plots mean behavior support $\pi_b(a|s)$ against mood improvement for each policy. Notably, support and PDIS are strongly negatively correlated. The RL methods and behavior cloning cluster in the upper right, showing high support alongside positive mood improvement. Low-support policies like rule-based and random produce highly variable mood estimates because their matched observation counts are small ($n \leq 2$), making these estimates unreliable.

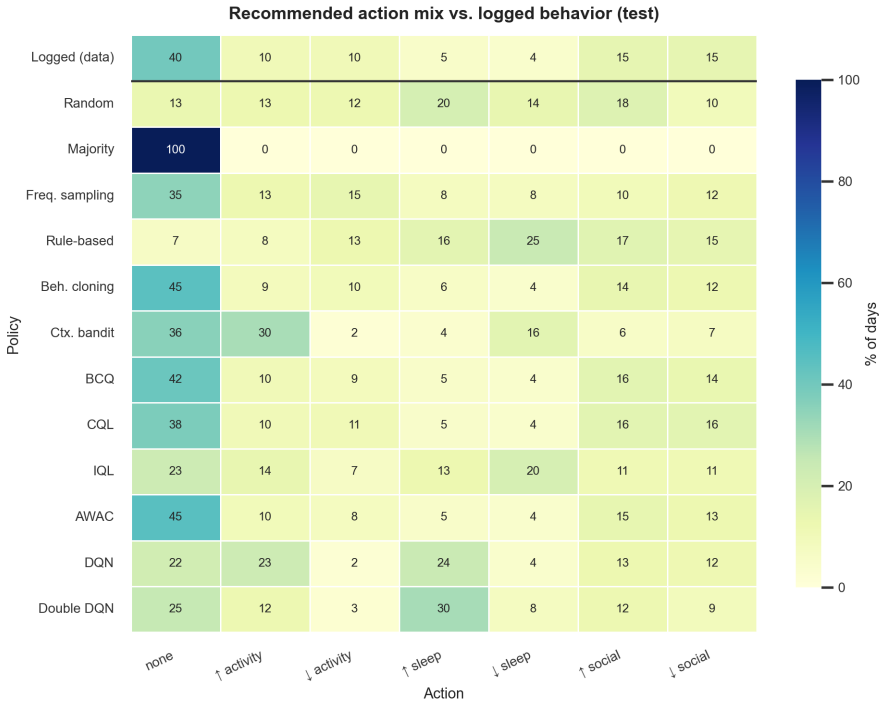

Figure 2: Action distribution heatmap across all policies. Each cell shows the percentage of timesteps a policy selected that action.

6 Discussion

6.1 Safety–Performance Tradeoffs

Our results show a tradeoff between the estimated policy value and distributional safety. Double DQN achieves the highest PDIS ($+0.020$) and DR ($+0.560$), but only 20% of its recommendations overlap with logged behavior. BCQ offers a more deployable alternative: positive DR ($+0.208$) with high behavior support (0.758). CQL stays even closer to the logging policy (0.778 support) yet yields strongly negative DR (-6.014), consistent with overconservative Q underestimation rather than policy failure. For mental health applications of RL, it appears that high OPE scores alone are insufficient if recommendations depart far from observed student behavior.

6.2 Reward Design

Reward specification largely determines what offline RL can learn. With mood observed on only $\sim 7\%$ of days, sparse and observed-only rewards leave fewer than 2.5% of transitions with nonzero signal, causing BCQ's behavior-cloning loss to dominate and producing nearly identical policies across reward variants. With more dense rewards that use behavioral proxies, we can reduce zero-reward transitions to $\sim 4.8\%$ and we see the clearest learning signal. With the dense reward, Double DQN achieves positive PDIS/WPDIS, while sparse variants collapse to near zero. While dense rewards improve learnability, we'd like to acknowledge that they do not establish clinical validity and further testing is necessary before real-life validation studies.

6.3 Evaluation Reliability

Aggregate gains should be interpreted cautiously because PDIS and DR confidence intervals mostly straddle zero for well-supported policies, and importance-weighted estimators suffer from very low effective sample size ($\sim$2–2.5 episodes). Within-support Δmood is positive for most methods but rests on few matched observations ($n \leq 14$; $n = 1$ for several top-scoring policies). Behavior cloning matches or exceeds conservative RL methods on Δmood with the highest support (0.811), suggesting

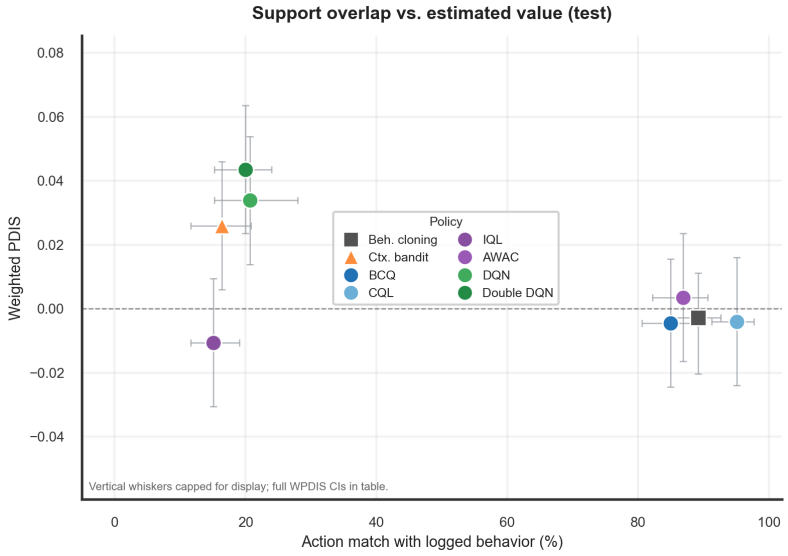

Figure 3: Action match with logged behavior vs PDIS. The idea is that methods on the higher left are less like logged behavior, but demonstrate higher reward, actions on the right (which include baselines like Behavioral Cloning and CQL) are predictably more like the actual actions executed by the dataset.

logged proxy actions are learnable from state alone and that sequential planning adds limited benefit under current coverage.

6.4 Contextual Structure and Limitations

Subgroup analysis shows partial contextual structure, as Double DQN recommends `increase_sleep` on low-sleep days (35.9%) despite no logged sleep increases in that subgroup, but does not compensate for low social engagement. Per-student heterogeneity is large (action match 0%–67%). Broader limitations include proxy actions that conflate intentional change with natural variation, a POMDP approximated as an MDP via short history, only 48 students and $\sim$3,045 transitions, and dense behavioral proxies that may not reflect true well-being. We treat these findings as hypothesis-generating.

7 Conclusion

We built an end-to-end offline RL pipeline that converts unlabeled StudentLife smartphone trajectories into MDP transitions with inferred proxy actions and three reward formulations. We compared DQN, Double DQN, CQL, BCQ, AWAC, and IQL against six baselines using PDIS, weighted PDIS, doubly robust estimation, and within-support mood improvement.

Our results demonstrate potential for offline RL to learn state-conditioned policies without expert-labeled mental-health interventions, but gains depend on reward design and distributional support. Double DQN achieves the highest estimated value at the cost of low logging-policy overlap; BCQ appears to provide the best safety–performance balance. Extreme mood sparsity causes conservative methods to collapse toward behavior cloning, while densified behavioral rewards enable clearer policy differentiation.

Our evaluations show the central barriers to this research problem (sparse outcomes, low effective sample size under importance weighting, proxy actions, and student-level heterogeneity) that must be addressed before deployment. Offline RL is a promising framework for mental health personalization from passive sensing, provided safety, reward specification, and evaluation reliability are treated as first-class concerns alongside estimated policy value.

7.1 Future Work

Next steps include collecting datasets with denser mood labels and actual intervention assignments, learning reward models that separate behavioral proxies from true well-being outcomes, using recurrent or personalized models to handle student-level heterogeneity. Without denser outcomes and stronger support, simply adding stronger offline RL algorithms is unlikely to solve the central evaluation problem.

8 Team Contributions

- **Alfred Yu:** Handled model and baseline training, and did implementation of DQN, Double DQN, baselines, IQL, AWAC, CQL. Coordinated integration of the full pipeline. Ran hyperparameter search. Included confidence error bars. Included subgroup analysis on evaluations. Implemented baselines (rule-based, contextual bandit, majority action, random, action frequency). Generated/updated visuals.
- **Elisabeth Holm:** Led the MDP approximation design with temporal state history, RL dataset construction from StudentLife EMA and sensing data, episode segmentation logic, and evaluation splits. Designed three reward function variants (sparse, dense, observed-only) to evaluate reward specification sensitivity and ran reward ablation.
- **Juan Pablo Pacheco:** He led the offline policy evaluation infrastructure. He wrote core functions (`compute_discounted_return`, `compute_pdis_estimate`, `compute_direct_method`, `compute_mood_improvement`). He also ran BCQ. He led the writeup for key parts such as offline policy evaluation and related work.

Changes from Proposal and Milestone. Our final division of labor remained largely consistent with the roles outlined in the proposal, with Alfred focusing on model development and training, Elisabeth focusing on dataset construction and MDP design, and Juan Pablo focusing on evaluation. However, the project revealed that building a reliable RL dataset and evaluation framework required substantially more effort than anticipated. Much of our work shifted toward addressing challenges unique to real-world mental health data, including sparse and delayed rewards, trajectory construction, episode segmentation, offline policy evaluation, and robust experimental design. As a result, the dataset and evaluation components grew in scope relative to our original expectations, while the overall role structure remained stable. This evolution reflects our learnings during the research process, as we found that rigorous data preparation and evaluation methodology were just as important as the RL algorithms themselves for producing meaningful and trustworthy results.

AI Tools Disclosure. Cursor and Claude (Anthropic) were used throughout this project for activities such as graph generation, text editing and phrasing, and coding. Claude generated boilerplate infrastructure in the model and evaluation pipeline (data loading, policy wrapper classes, model wrapper code, the evaluation driver, and figure generation code).

References

Marvyn R. Arévalo Avalos, Karina Rosales, Chris Karr, Caroline A. Figueroa, Tiffany Luo, Suchitra Sudarshan, Vivian Yip, and Adrian Aguilera. 2025. Personalizing a mental health texting intervention using reinforcement learning. *npj Mental Health Research* 4, 1 (2025), 64. doi:10.1038/s44184-025-00173-3

Sarah Ketchen Lipson, Sasha Zhou, Sara Abelson, Justin Heinze, Matthew Jirsa, Jasmine Morigney, Akilah Patterson, Meghna Singh, and Daniel Eisenberg. 2022. Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national healthy minds study, 2013–2021. *Journal of affective disorders* 306 (2022), 138–147.

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 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 3–14. doi:10.1145/2632048.2632054

A Hyperparameter Details

Table 5: Hyperparameters for all trained models. DQN/Double DQN and IQL/AWAC values are the configurations selected by validation-PDIS grid search on the dense reward; BCQ and CQL were not tuned. All networks use Adam.

Model	Hyperparameter	Value
DQN	Architecture	MLP (256, 256), ReLU
(d3rlpy)	Learning rate	6.25×10^{-5}
	Batch size	32
	Target update interval	8,000 steps
	Training steps	5,000
	γ	0.99
Double DQN	Architecture	MLP (256, 256), ReLU
(d3rlpy)	Learning rate	1×10^{-4}
	Batch size	32
	Target update interval	1,000 steps
	Training steps	5,000
	γ	0.99
DQN/Double DQN grid:		lr $\{1\times10^{-4}, 6.25\times10^{-5}\}$, batch $\{32, 64\}$, target update $\{1000, 8000\}$, hidden $\{64\times64, 256\times256\}$
BCQ	Q-/BC-network	MLP (256, 256), ReLU
(PyTorch)	Learning rate	1×10^{-4}
	Batch size	64
	Target update interval	100 steps
	Training steps	15,000
	Action threshold τ	0.3
	γ	0.99
	Tuning	none (single config)
CQL	Architecture	MLP (256, 256), ReLU
(d3rlpy,	Learning rate	6.25×10^{-5} (default)
defaults)	Batch size	32
	Target update interval	8,000 steps
	Conservative weight α	1.0
	Training steps	10,000
	γ	0.99
	Tuning	none (library defaults)
IQL	Architecture	MLP (256, 256), ReLU
(PyTorch)	Critic learning rate	1×10^{-4}
	Batch size	64
	Target update interval	100 steps
	Training steps	5,000
	Expectile τ	0.9
	Temperature	1.0
	γ	0.99
AWAC	Architecture	MLP (256, 256), ReLU
(PyTorch)	Critic / actor learning rate	1×10^{-4} / 1×10^{-4}
	Batch size	64
	Target update interval	100 steps
	Training steps	5,000 (1,000 critic warm-up + 4,000 actor)
	Advantage temperature λ	1.0
	Expectile τ (critic)	0.7
	Max IS weight	100
	γ	0.99
IQL/AWAC grid:		IQL expectile $\{0.7, 0.9\}\times$ temp $\{1.0, 3.0\}$; AWAC $\lambda \in \{0.5, 1.0\}$
OPE	IS weight clip $w_{\max}$	5.0
	Soft ϵ (deterministic policies)	0.05
	γ	0.99
	Behavior policy	Logistic-regression BC (89.2% match)