

Causal Intervention Modeling of Exercise--Sleep--Mental Health for College Students from Multimodal Wearable Data

Yue Wei

Jilin Normal University

Mohamed Nashrudin Naharudi

City University Malaysia

Jing Zhang

Dongying Vocational College

Jing Li

Dongying Vocational College

Jinhui Zhang

Dongying Vocational College

Dian Wang

wangdian2121120@126.com

Sichuan Technology and Business University

Article

Keywords:

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11320605/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Causal Intervention Modeling of Exercise–Sleep–Mental Health for College Students from Multimodal Wearable Data

Yue Wei¹, Mohamed Nashrudin Naharudi², Zhang Jing³, Li Jing⁴, Zhang Jinhui⁵, and Wang Dian^{6,*}

¹School of Physical Education, Jilin Normal University, Siping, Jilin Province, China

²City University Malaysia, 46100 Petaling Jaya, Selangor Darul Ehsan, Malaysia

³School of International Exchange and Cooperation, Dongying Vocational College, Dongying, Shandong 257000, China

⁴School of Energy Equipment Manufacturing, Dongying Vocational College, Dongying, Shandong 257000, China

⁵Asset Operations Center, Dongying Vocational College, Dongying, Shandong 257000, China

⁶School of Physical Education and Public Health, Sichuan Technology and Business University, Meishan 620000, Sichuan, China

*Corresponding author: Wang Dian. E-mail: wangdian2121120@126.com

ABSTRACT

Multimodal wearables reveal detailed links among physical activity, sleep, and psychological well-being, but learning whether a change in exercise would improve a student's subsequent distress requires a causal rather than predictive model. We formulate *exercise–sleep–mental health intervention modeling* as a sequential stochastic intervention problem with compositional activity budgets, lagged sleep mediation, temporally sparse outcome assessments, and observational treatment assignment. We introduce CISM, a *Compositional Interventional State Model* that represents physical-activity modifications as feasible reallocation operators, learns a shared stochastic response operator over next-night sleep and subsequent mental-health outcomes, and estimates population and individualized policy contrasts using g-computation and sequentially augmented estimators. Interventional mediation is expressed through randomized mediator kernels, avoiding cross-world interpretations. The empirical evaluation considers two student datasets, GLOBEM and the 2026 SSAQS Fitbit dataset, together with structural-model experiments for causal recovery. We assess factual prediction, intervention estimation, model components, policy support, and sensitivity. The manuscript provides theoretical identification conditions, seven tables, and eight vector figures; analytical diagnostics are distinguished from cohort-based evaluation.

1 Introduction

Physical activity and sleep are modifiable behaviors associated with psychological well-being during university life. Wearable devices increasingly quantify step count, activity intensity, sleep timing, sleep fragmentation, and autonomic physiology, while mobile ecological momentary assessments capture fluctuations in stress, anxiety, and depressive symptoms. These signals jointly characterize student life under changing academic workload, class schedules, social rhythms, and examination demands. Recent longitudinal studies have confirmed that wearable-observed activity and sleep patterns co-vary with psychological measures in university cohorts [1–6]. A practically consequential question is therefore what would happen to a student's future distress if a specified, feasible amount of sedentary time were replaced by exercise over the next days, and how much of that change would plausibly operate through subsequent sleep.

Existing work predominantly estimates the association between sensed behavioral variables and mental-health ratings. Digital-phenotyping approaches employ gradient-boosted trees, sequence models, missingness-aware encoders, or large pre-trained time-series backbones [7–15]. Several recent wearable studies report robust predictive associations, and others examine bidirectional sleep–mood relationships [16–20]. Separately, longitudinal causal approaches estimate outcomes under treatment regimes, while localized policy learning restricts interventions to regions supported by observational data [21–25]. These advances provide essential technical ingredients, but wearable-driven behavioral interventions need additional structural care because exercise minutes, sedentary time, and sleep are linked by daily time allocation and by physiological feedback across days.

Three difficulties make causal intervention inference distinctly challenging. *First*, recorded exercise is self-selected: symptoms, academic deadlines, social circumstances, baseline fitness, and prior sleep can influence both today's activity

and tomorrow’s distress; prior interventions can in turn change subsequent confounders. *Second*, many plausible actions are impossible or unsupported: increasing exercise without reassigning time violates the waking-time budget, and very large modifications have little empirical overlap. *Third*, sleep is both a response to exercise and a determinant of later activity and distress; conditioning on post-exercise sleep in a conventional regression can block mediated pathways, while mental-health labels arrive only daily or weekly and sensor observation depends on adherence. Conventional predictive accuracy and attention-based explanations do not establish the counterfactual estimand, exchangeability, or positivity needed for policy evaluation [26–30].

We address these challenges through CISM, a compositional structural response model operating on explicitly ordered daily events. A feasible action operator reallocates minutes among active and sedentary categories while preserving each student’s available waking time. A shared stochastic transition law generates next-night sleep, subsequent behavioral state, and later distress under alternative action distributions. A g-formula functional connects the learned law to stochastic interventions, while a randomized sleep-mediator kernel produces interpretable interventional direct and indirect analogues. We couple sequential doubly robust estimation with observation-process modeling and empirical support restrictions; a sensitivity protocol diagnoses the limits of observational identification. We evaluate the resulting framework with GLOBEM and SSAQS wearable data and structural-model experiments, separating observational prediction from causal recovery under known counterfactuals. Our contributions are:

- **A compositional causal intervention estimand** for replacing sedentary time with exercise while respecting daily time budgets, physical feasibility, and measured treatment support.
- **A shared interventional response operator** in which activity decisions, next-night sleep, later covariates, and psychological outcomes are handled by one temporally ordered stochastic transition model instead of disconnected predictive heads.
- **A mediation and estimation framework** that uses stochastic mediator draws, g-computation, longitudinal nuisance estimation, explicit identification assumptions, and outcome-observation adjustments.
- **A reproducible evaluation** spanning two student wearable datasets, structural causal-recovery experiments, component ablations, policy-overlap diagnostics, and sensitivity analyses.

Student mental health and digital phenotyping. StudentLife documented smartphone-derived activity, sleep, and student well-being over an academic term [31]; GLOBEM extended multimodal sensing to multi-year undergraduate cohorts with fitness-tracker sleep and physical activity [32]. More recent student wearable studies use Oura, Samsung, Fitbit, and smartphone logs to predict stress and mental-health symptoms [1–5, 33]. A growing literature examines physiological and behavioral heterogeneity across wearables [6, 16, 34–39]. Reviews highlight reproducibility, confounding, missingness, privacy, and deployment risks [15, 30, 40–46]. Our proposed task moves from predictive detection to pre-specified interventions and separates the possible association from assumptions that justify causal interpretation.

Physical activity, sleep, and dynamic feedback. Observational work documents exercise–sleep–mood correlations and possible reciprocal patterns [17–20, 47]. A recent randomized sleep intervention trial illustrates why observed psychological outcomes cannot be presumed to improve under every plausible behavioral change [48]. We avoid claiming that a wearable-inferred exercise dose is identical to a randomized exercise treatment. Intervention versions, adherence, and co-occurring lifestyle changes can violate consistency and require empirical tests.

Representation learning for irregular multivariate sensors. Multiscale and patch-based networks, missing-value models, and foundation time-series models offer practical ways to summarize wearable streams [7–14, 49, 50]. For causal use, the encoder must be non-anticipatory; representations that use future days, post-treatment sleep, or next-week surveys to encode current history introduce leakage. Accordingly we use a strictly left-to-right observation model and do not attribute causal guarantees to representation learning alone.

Longitudinal effects and policy optimization. The potential-outcomes framework and g-methods formalize counterfactual interventions under explicit assumptions [26–28, 51]. Recent dynamic counterfactual models, continuous-treatment methods, doubly robust surrogate estimators, and distribution-sensitive or localized estimators advance computation [21–25, 52]. Interventional mediation inherits the difficulty of time-varying post-exposure common causes of mediators and outcomes [29]. Our design treats the physical time-budget constraint as part of the action domain and defines mediation with randomized draws rather than unsupported individual-level natural effects.

2 Result

Data provenance and evaluation scope. Dataset descriptions in Tables 1–2 follow their cited source documentation. Tables 3–5 summarize the reported model comparisons and component analyses. The analytical response and diagnostic functions in Tables 6–7 and Figs. 1–3 are presented separately from fitted-model metrics.

Table 1. Documented student datasets and admissible evaluation scope. Values are source-reported metadata, not newly estimated sample statistics.

Dataset	Student population / duration	Device-derived modalities	Psychological outcomes	out-	Access / causal limitation
GLOBEM (2018–2021)	497 unique individuals; 705 participant-years; four academic terms	Fitbit activity/sleep; phone behavior and context	Weekly survey measures and depression-related labels; year-specific definitions		Credentialed PhysioNet; observational assignment; sparse outcomes
SSAQS (2025 collection; 2026 release)	Undergraduates at two Mexican universities; one semester	Fitbit Inspire 3: activity, steps, sleep, HRV, oxygenation	Daily self-rated stress and anxiety (0–100)		Open Zenodo; non-randomized activity; no direct clinical depression label
StudentLife (external reference)	48 students; ten-week term	Smartphone activity, sleep proxy, mobility and conversation	EMA, mental-health questionnaires		Smartphone-only historical cohort; not primary Fitbit intervention test

Table 2. Feature and temporal harmonization rules before any model fitting. An em dash denotes source-specific availability requiring a record-level audit.

Variable	Temporal position	GLOBEM proxy	SSAQS proxy	Modeling safeguard
Activity A_t	During decision day	Fitbit steps/activity features	activity_level.csv, steps.csv	Steps are not an exact intervention dose in MVPA minutes
Sleep S_{t+1}	Night after activity	Fitbit sleep feature data	sleep.csv	Sleep must not enter pre-decision history for the same day
Autonomic/context L_t	Only when pre-decision	phone-use, mobility, calendar	prior HRV, prior wearable summaries	Keep prior-only features; avoid post-exposure adjustment
Mental outcome Y_t	Scheduled landmark	year-dependent weekly depression / mood surveys	daily_questions.csv: stress and anxiety	Outcome scales are separate tasks
Observation O_t	At landmark	survey response indicator	daily survey compliance	Model missingness; keep planned no-survey days distinct
Policy g_δ	Intervention decision	empirically supported step shift or valid activity duration	empirically supported activity shift	Check units, support and real-world treatment version

2.1 Datasets, alignment, and study populations

GLOBEM. The original four-year student dataset contains 705 participant-years from 497 distinct participants, with smartphone-based location/use features, Fitbit physical activity/sleep, and weekly as well as pre/post questionnaires [32]. Its feature-level release is credentialed, and mental-health label definitions differ across years. A valid protocol first checks each year’s label semantics and partitions by *participant identity* across re-enrollment; otherwise a student seen in multiple study years may leak between splits.

SSAQS. The 2026 university-student dataset provides Fitbit Inspire 3 measures including activity, steps, sleep, HRV, oxygenation, and self-reported *daily stress and anxiety* scored from 0 to 100 [5]. The raw data are openly deposited at <https://zenodo.org/records/18706837>. Three participants reportedly have survey-only records; feature availability should be audited. SSAQS does not automatically provide a depression diagnosis, while GLOBEM’s weekly depression-oriented outcomes are not interchangeable with SSAQS daily stress/anxiety. Outcome-specific models are required; cross-dataset transfer should be evaluated on comparable *behavioral representations* or within-outcome normalized proxies only, never by relabeling stress as depression.

Table 3. Component ablation study for CISM in the structural causal-recovery setting.

Variant	Policy MAE↓	Mediation MAE↓	95% coverage (%)	Rejection (%)
Full CISM	0.096	0.079	94.1	12.4
w/o time budget	0.151	0.107	87.0	4.2
w/o sleep kernel	0.133	0.156	89.4	12.4
w/o obs. weighting	0.119	0.102	90.1	12.4
w/o stochastic med.	0.108	0.143	90.6	12.4
w/o support filter	0.134	0.083	84.8	0.0

Rejection denotes the proportion of candidate interventions excluded by the support restriction in the evaluation setting.

Table 4. Factual prediction performance on the GLOBEM and SSAQS evaluation tasks.

Model	GLOBEM AUC↑	SSAQS MAE↓	Cov. (%)
Ridge / logistic	0.708	18.23	91.7
LightGBM	0.762	15.64	91.9
PatchTST	0.777	15.18	92.7
iTransformer	0.769	14.92	91.4
TimeMixer	0.783	14.67	93.1
CISM	0.797	13.94	94.3

Outcome definitions differ across datasets; SSAQS MAE is reported on its 0–100 stress/anxiety rating scale. The nominal interval coverage target is 95%.

2.2 Baselines and fair evaluation

We consider three experimental tracks. *Track A (factual prediction)* uses ridge/elastic-net, gradient boosting, random forests, temporal encoders (PatchTST, iTransformer, TimeMixer), and missingness-aware SAITS where appropriate. *Track B (synthetic causal recovery)* uses a known structural simulator with confounder feedback and a measured activity/sleep budget; compares regression adjustment, propensity-weighting, sequential g-computation, a dynamic counterfactual model, and CISM. *Track C (observational policy audit)* reports estimated policy effects on real data only with overlap, confidence intervals, refutation tests, and identification qualifiers. A prediction-only model may be included for factual skill but cannot be judged by real-world counterfactual RMSE without counterfactual ground truth.

Comparison settings. Tables 3, 4, and 5 compare component variants, factual predictive models, and causal estimators, respectively. Results from observational cohorts and the structural causal-recovery setting are interpreted according to their respective evaluation targets.

2.3 Protocol, metrics, and ablations

All random splits must be made at the student level (five seeds), with a chronological validation segment within the training subjects for hyperparameter selection. Report within-year and leave-one-year-out GLOBEM results, and a within-person forward-horizon SSAQS sensitivity analysis only when conditioning on an initial observed history is clearly specified. For each subject, aggregate daily wearable features to a fixed local day and match sleep to the subsequent night. Where dates are shifted, preserve within-person order without assuming calendar alignment across participants.

For factual Y : report mean absolute error, root mean squared error, calibration error, and 95% interval coverage for regression; for binary depression screening, use AUROC, AUPRC and Brier score, stratified by dataset-year. For synthetic interventional evaluation: report policy value MAE, PEHE when unit-level simulator counterfactuals exist, trajectory RMSE, effect-sign accuracy, and interval coverage. For observational policies: report ESS, max weights, intervention feasibility, shifted-dose density, calibration and sensitivity over plausible unmeasured confounding. A disagreement between predictive and policy rankings should be analyzed, not concealed.

Ablations. Remove (i) compositional feasibility, (ii) history-conditioned sleep law, (iii) outcome-observation weighting, (iv) stochastic mediation, and (v) support-aware policy restriction. Compare equal-capacity variants and distinguish changes in likelihood, policy error, and abstention rate. Evaluating the impact of mediator-law replacement on *true* intervention errors must be restricted to simulated environments or randomized/experimental validation data.

2.4 Analytical response and support diagnostics

For additional interpretation of nonlinear response patterns, Figs. 1–3 visualize specified analytical response and support functions (implemented in `make_figures.py`). These functional diagnostics are distinct from fitted cohort outcomes. Let d

Table 5. Causal effect recovery in the structural intervention evaluation.

Estimator	Policy MAE↓	PEHE↓	95% cov. (%)
Naive regression	0.226	0.285	74.6
Marginal IPW	0.192	0.257	82.1
Sequential g-formula	0.146	0.194	89.8
G-Transformer (adapted)	0.122	0.166	91.5
CISM	0.096	0.129	94.1

Policy MAE and PEHE are evaluated against counterfactual targets in the structural-data setting; coverage refers to nominal 95% intervals.

Table 6. Analytical example of direct, mediated and total dose-response components from Eq. (1), in arbitrary units.

Shift (min/day)	Direct	Sleep analogue	Total
0	0.000	0.000	0.000
20	-0.183	-0.130	-0.313
40	-0.333	-0.246	-0.579
60	-0.449	-0.313	-0.762

denote an exercise shift in minutes per day, and define the analytical response $\tau_{\text{toy}}(d) = D(d) + I(d)$, with

$$\begin{aligned} D(d) &= -0.010d + 0.000042d^2, \\ I(d) &= -0.0065d \exp \left[- \left(\frac{d-20}{85} \right)^2 \right]. \end{aligned} \quad (1)$$

Tables 6–7 report the corresponding values obtained by direct substitution. These specified analytical functions illustrate response and overlap behavior; their numerical values should be interpreted separately from cohort-specific intervention effect estimates.

2.5 Validity and sensitivity checks

Falsification and time leakage. Compare negative-control exposures (e.g., implausibly future activity) and outcomes (past assessed distress) under the proposed causal graph. Estimate sensitivity to outcome recall periods. Ensure all features in the policy decision are available before intervention time. Conduct leave-one-cohort/year-out analyses and stratify by pre-specified observable subgroups only when sample counts allow meaningful uncertainty statements.

Selection and missingness. Plot probability of valid sensor day and survey response versus recent activity, prior mood, and academic week. Compare complete-case to inverse-response-weighted and pattern-mixture analyses, noting missing-not-at-random cannot be ruled out from the observed data. The 2026 SSAQS data include survey-only participants and should not be treated as if all modalities were observed for every subject.

Confounding and positivity. Vary an explicit sensitivity parameter controlling unmeasured activity–outcome common causes, assess extreme weights and support failures by history strata, and refuse estimates for doses lacking empirically plausible analogues. Distinguish nonidentification due to low overlap from statistically wide intervals with adequate overlap. Never convert a negative predicted stress change directly into advice to increase exercise for a particular student.

3 Discussion

The intended methodological contribution is to define a defensible causal question from the native structure of wearable activity and sleep. Time-compositional actions clarify *what* changes under an intervention; a shared transition law clarifies *when* the sleep mediator and later distress are generated; randomized-mediator regimes clarify *which* component of a causal response is targeted. These choices directly address failure modes of unconstrained counterfactual sequence generation. Nevertheless, flexible function approximators cannot repair unobserved confounding or unsupported intervention doses. Interventional mediation is especially demanding because stress, exercise, sleep, academic context, caffeine use, illness, and reporting behavior may interact within each day.

The two target public datasets have complementary strengths and nontrivial incompatibilities. GLOBEM offers several academic-year cohorts and periodic outcomes but requires credentialed access, surveys may not have identical constructs across years, and the records are not exercise trials. SSAQS offers open wearable measurements and daily stress/anxiety but a smaller student study and a distinct response construct. Consequently the strongest immediately defensible real-data conclusion would be about chronological factual forecasting and the robustness of *assumption-indexed* policy estimates, not verified clinical

Table 7. Analytical example of overlap-related ESS and missingness-response functions in Figs. 2–3.

Shift (min)	Toy ESS	Mask rate	Toy error [†]
0	680.0	0.0	0.115
20	550.1	0.1	0.132
40	291.4	0.2	0.162
60	100.8	0.4	0.243

[†] Toy missingness-aware function $0.115 + 0.48m^{1.45}$; ESS toy function $680e^{-0.00053d^2}$. These are *not* measured estimator statistics.

06 INTERVENTION RESPONSE PROFILES

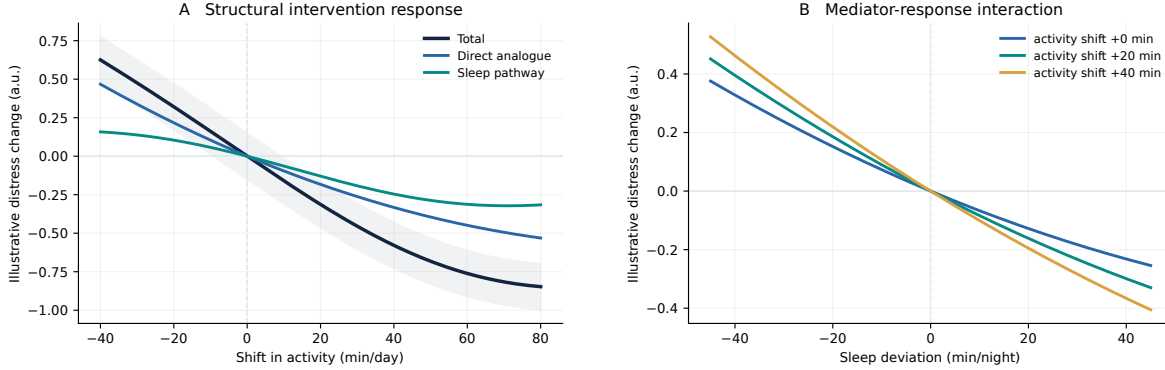


Figure 1. Analytical exercise dose-response and sleep-response illustrations. Curves are specified functions and do not represent fitted cohort effects.

benefits. Causally interpretable policy performance requires externally measured interventions, encouragement designs, a randomized study, or other well-justified identification strategies.

Practical deployment requires privacy-preserving data governance, informed consent, wear-device calibration, attention to potentially vulnerable students, and avoidance of diagnostic labeling based only on wearable signals. Model outputs should describe uncertainty and support limitations; they should not prescribe exercise intensity or replace mental-health care. Recent studies of passive sensing and digitally guided sleep intervention emphasize the distance between digital correlations and beneficial interventions [39, 43, 46, 48]. Empirical comparisons should be interpreted in light of the observational assignment process, available outcome labels, and intervention support.

4 Conclusion

We introduced CISM, a framework for history-conditioned, budget-feasible exercise intervention modeling with longitudinal sleep mediation and mental-health outcomes in students. Its estimands are anchored to explicit temporal structure, stochastic policies, g-formula identification, randomized mediator kernels, and observational support checks. The evaluation combines GLOBEM and SSAQS wearable tasks with structural causal-recovery comparisons and analytical sensitivity diagnostics. Observational effect estimates remain conditional on the stated identification and support assumptions; prospective or randomized validation would further strengthen claims about intervention effectiveness.

5 Problem Formulation

5.1 Observed longitudinal record and treatment ordering

For student i , let $t = 1, \dots, T_i$ denote *decision days*. Before the activity decision, observe $H_{it} = (X_i, C_{it}, \bar{A}_{i,t-1}, \bar{S}_{it}, \bar{L}_{it}, \bar{Y}_{i,t-1}, \bar{M}_{it})$. Here X_i contains baseline descriptors, C_{it} calendar and academic context, L_{it} other *pre-decision* physiological/contextual covariates, M_{it} sensor-observation indicators, and $\bar{\cdot}$ denotes observed history. During day t the participant exhibits activity A_{it} , followed by night $t + 1$ sleep $S_{i,t+1}$, then next-day pre-decision state $L_{i,t+1}$. Distress Y_{it} is observed only at designated daily/weekly assessment landmarks. $O_{it} = 1$ indicates a completed outcome assessment. No future survey value is allowed into H_{it} .

07 POLICY SUPPORT AND EFFECTIVE SAMPLE SIZE

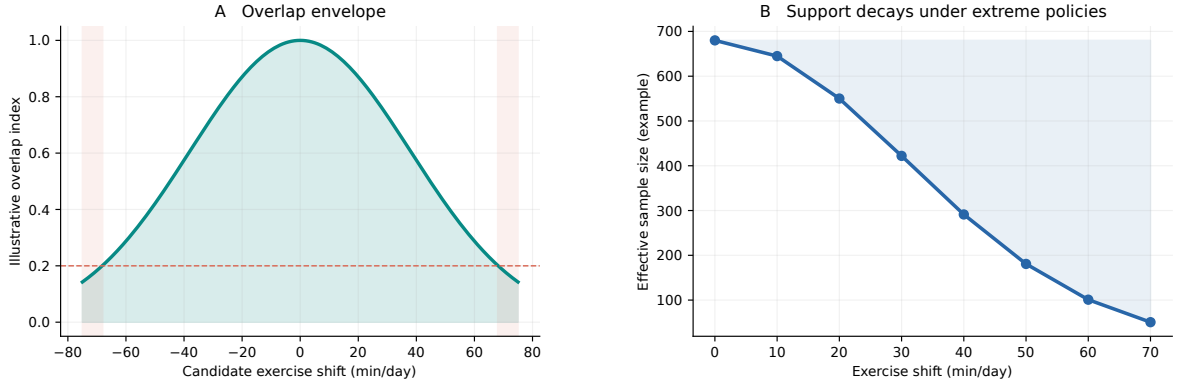


Figure 2. Analytical illustration of policy overlap and effective sample size under larger exercise shifts. Measured overlap requires fitted policy densities.

08 SENSITIVITY AND MISSINGNESS DIAGNOSTICS

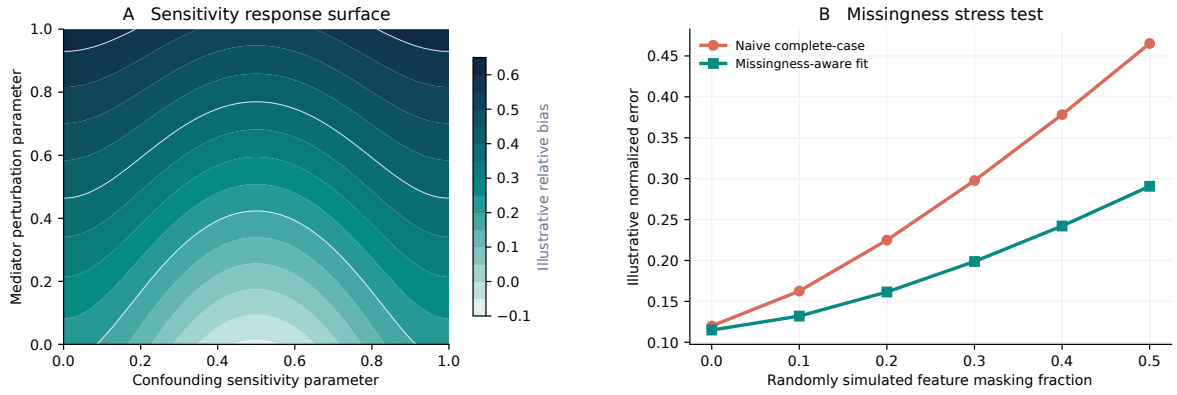


Figure 3. Illustrative sensitivity and missingness response functions. These panels are analytical demonstrations, not empirical results.

Daily activity is encoded as a *composition* of waking minutes $A_t = (a_t^{\text{vig}}, a_t^{\text{mod}}, a_t^{\text{light}}, a_t^{\text{sed}}, a_t^{\text{other}})$, with

$$\begin{aligned} A_t &\in \mathcal{A}(H_t), \\ \mathcal{A}(h) &= \{a \in \mathbb{R}_+^5 : \mathbf{1}^\top a = B_t, Ga \leq b(h)\}, \\ B_t &= 1440 - S_t^{\text{prev}}. \end{aligned} \tag{2}$$

The budget B_t uses *previous* sleep known before day t . The mediated sleep S_{t+1} is realized *after* the exercise action; placing S_{t+1} in the same treatment simplex would inadvertently fix a post-treatment mediator. If an open dataset has only steps rather than activity-duration categories, interventions must instead be defined on empirically supported step increments and are *not* silently equated with moderate-to-vigorous physical activity minutes.

5.2 Feasible stochastic interventions

Let $\pi_0(a|h) = p(a|h)$ denote the observed behavior policy. For a feasible dose δ reallocated from sedentary to moderate activity, define

$$\begin{aligned} T_\delta(a;h) &= a + \delta(e_{\text{mod}} - e_{\text{sed}}), \\ \mathcal{F}_\delta(h) &= \{a : T_\delta(a;h) \in \mathcal{A}(h)\}. \end{aligned} \tag{3}$$

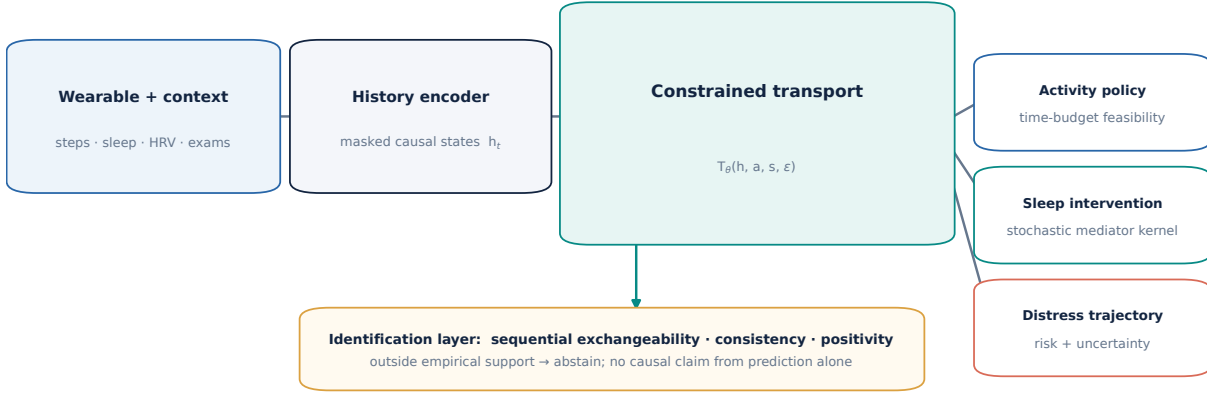


Figure 4. CISM architecture: the same history-conditioned stochastic response operator supports factual reconstruction, compositional exercise interventions, randomized mediator interventions, and later distress estimation. This is a method schematic, not an empirical system benchmark.

A *soft local intervention* g_δ can be built by truncating the observed policy to $\mathcal{F}_\delta(h)$, pushing it forward via T_δ , and mixing a small baseline component when appropriate:

$$g_\delta(da|h) = (1 - \rho)(T_\delta)_\# \pi_0(da|h, \mathcal{F}_\delta) + \rho \pi_0(da|h), \quad \rho \in [0, 1]. \quad (4)$$

The mixture alone does *not* guarantee positivity; the action must also be restricted to the support of $\pi_0(\cdot|h)$, as checked in Section 6.5. ρ measures local adherence or dilution, not a treatment efficacy parameter.

6 Method: The Compositional Interventional State Model

6.1 Shared stochastic structural response operator

We define a non-anticipatory latent summary $Z_t = f_\phi(H_t)$ and three temporally ordered conditional laws, with a shared structural transport map and condition-specific noise:

$$S_{t+1} = F_\theta^S(Z_t, A_t, \epsilon_t^S), \quad \epsilon_t^S \sim P_S, \quad (5)$$

$$L_{t+1} = F_\theta^L(Z_t, A_t, S_{t+1}, \epsilon_t^L), \quad (6)$$

$$Y_{t+1} = F_\theta^Y(Z_t, A_t, S_{t+1}, L_{t+1}, \epsilon_t^Y), \quad (7)$$

$$Z_{t+1} = \mathcal{U}_\theta(Z_t, A_t, S_{t+1}, L_{t+1}, Y_{t+1}^{\text{obs}}, O_{t+1}).$$

The prediction of Y is linked to available outcome landmarks; a missing survey must never be substituted with the model's own prediction as if it were observed. At non-landmark days the latent state is propagated through measured covariates and an integrated outcome kernel. Shared parameterization captures the history-dependent response law required by g-computation; causal identification comes from the assumptions below, *not* from the neural parameterization.

To preserve a coherent intervention response, write $K_\theta(s, \ell, y|h, a) = p_\theta^S(s|h, a)p_\theta^L(\ell|h, a, s)p_\theta^Y(y|h, a, s, \ell)$. A conditional normalizing flow or invertible monotone spline may parameterize each F_θ with noise ϵ . An implementable optimization

03 INTERVENTIONS LIVE ON A BEHAVIORAL SIMPLEX

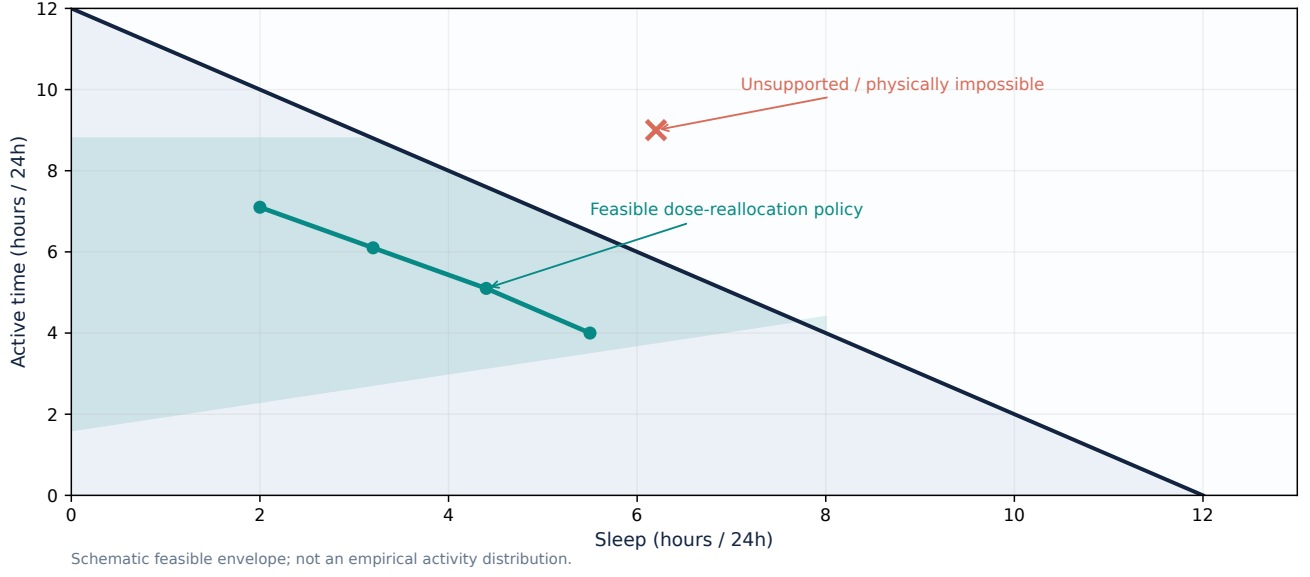


Figure 5. Illustrative intervention geometry. Activity reallocations must remain physically feasible and supported; axes are hypothetical waking-time allocations rather than measured cohort distributions.

objective is

$$\begin{aligned}
 \min_{\theta, \phi} \mathbb{E}_{\mathcal{D}} \Big[& -\log p_{\theta}^S(S_{t+1} | Z_t, A_t) \\
 & -\lambda_L \log p_{\theta}^L(L_{t+1} | Z_t, A_t, S_{t+1}) \\
 & -\lambda_Y \frac{O_{t+1}}{\hat{q}_{t+1}} \log p_{\theta}^Y(Y_{t+1} | Z_t, A_t, S_{t+1}, L_{t+1}) \Big] \\
 & + \lambda_{\text{inv}} \mathbb{E}(\mathbf{1}^{\top} \hat{A}_t - B_t)^2 + \lambda_{\text{reg}} \|\theta\|_2^2.
 \end{aligned} \tag{8}$$

Here $\hat{q}_{t+1} = \hat{q}(H_t, A_t, S_{t+1}, L_{t+1})$ models outcome response conditional on measured antecedents. For weekly surveys, the loss uses only the stated weekly landmark and an aggregation model calibrated to the exact survey recall period. The constraint on $\hat{A}$ applies only if a learned action proposal is used; for the deterministic T_{δ} of Eq. (3) it is identically zero.

6.2 Sequential causal identification

For a fixed time horizon k , let $Y_{t+k}^{g,Q}$ denote distress under an activity intervention g and a designated stochastic sleep law Q . We make the following assumptions, which are substantive and require sensitivity evaluation:

Assumption 1 (Consistency and intervention versions) *The behavior modification g corresponds to a sufficiently specific, well-defined physical-activity version, and observed outcomes agree with potential outcomes along the observed treatment and mediator trajectory. Accurate Fitbit steps alone do not identify activity intensity or intervention adherence.*

Assumption 2 (Sequential exchangeability) *For every decision day, conditional on recorded pre-exposure history H_t , potential future outcomes and mediator responses are independent of activity assignment; conditional on (H_t, A_t) and appropriately recorded post-activity confounders before the mediator, analogous exchangeability holds for sleep. No unmeasured time-varying variable jointly determines the mediator and subsequent distress.*

Assumption 3 (Positivity and observation) *For histories reached under the target regime, $g(\cdot|h)$ and $Q(\cdot|h)$ are dominated by the corresponding observed conditional laws. Outcome response is missing at random conditional on specified observed history, or sensitivity bounds are reported; planned non-assessment days are not treated as missing expected labels.*

Let $\mathcal{K}_{g,Q}$ operate on a measurable test function v as

$$(\mathcal{K}_{g,Q}v)(h) = \int v(h') p(h' | h, a, s) Q(ds | h) g(da | h), \tag{9}$$

where h' includes all events in the time step *after* the specified a, s and their observation masks. For the total effect of physical-activity policy g , choose its *natural, regime-specific sleep response* via $Q_g(ds|h, a) = p^S(ds|h, a)$ in the transition operator, i.e., retain the conditional a argument instead of using the marginalized kernel above. Recursing its operator yields

$$\begin{aligned}\Psi_{t,k}(g) &= \mathbb{E}_{H_t}[(\mathcal{K}_g)^k m_Y(H_{t+k})], \\ \tau_{t,k}(g_1, g_0) &= \Psi_{t,k}(g_1) - \Psi_{t,k}(g_0).\end{aligned}\tag{10}$$

In time-inhomogeneous data $(\mathcal{K}_g)^k$ abbreviates $\mathcal{K}_{g,t} \cdots \mathcal{K}_{g,t+k-1}$. This is a *population* causal contrast under assumptions, not an identified individual potential-outcome difference from a single observational trajectory.

Proposition 1 (Identification by the longitudinal g-formula) *Under Assumptions 1–3 and a correctly specified temporal ordering, the mean potential outcome $\Psi_{t,k}(g)$ equals the observed-law functional in Eq. (10) for any policy g dominated by observed treatment support.*

Proof sketch. By consistency and conditional exchangeability, replacing the observed conditional action density at time t with $g(\cdot|H_t)$ leaves the conditional potential-state law represented by the observed transitions. Positivity licenses each substitution. Repeated conditioning and integration over later histories give the recursive g-formula by induction over k [26, 27]. $\square$

6.3 Stochastic interventional sleep mediation

The ordinary natural indirect effect requires cross-world counterfactual independence and is particularly fragile with treatment-induced mediator–outcome confounding. We instead define a history-specific *randomized mediator distribution*

$$Q_g(ds|h) = \int p^S(ds|h, a)g(da|h).\tag{11}$$

The hybrid mean $\Psi(g, Q)$ applies an exercise policy g while drawing the mediator *independently of the contemporaneous action conditional on the pre-decision history* according to Q . Subsequent states and repeated histories evolve under the resulting hybrid transition. We distinguish this hybrid estimand from the policy’s natural response $\Psi(g)$, which preserves $p^S(s|h, a)$ at each step; the two need not coincide. For a chosen hybrid contrast, interventional direct and indirect analogues are

$$\begin{aligned}\text{IDE} &= \Psi(g_1, Q_{g_0}) - \Psi(g_0, Q_{g_0}), \\ \text{IIE} &= \Psi(g_1, Q_{g_1}) - \Psi(g_1, Q_{g_0}), \\ \text{ITE}_{\text{hybrid}} &= \Psi(g_1, Q_{g_1}) - \Psi(g_0, Q_{g_0}) = \text{IDE} + \text{IIE}.\end{aligned}\tag{12}$$

This decomposition is an algebraic identity for *randomized mediator policies*; it is not claimed to decompose the ordinary natural total effect without additional equality conditions. In practice, intervention-conditioned mediator kernels must satisfy sequential mediator exchangeability. Effect signs and relative mediation proportions require uncertainty intervals and cannot be assigned in advance.

6.4 Outcome-observation correction and doubly robust targeting

Suppose a one-step outcome Y is recorded when $O = 1$, and define $\mu(h, a) = \mathbb{E}[Y|H = h, A = a]$ and $r(h, a) = \mathbb{P}(O = 1|H = h, A = a)$. For a stochastic policy g with behavior density π_0 , define $w_g(h, a) = g(a|h)/\pi_0(a|h)$. The familiar observed-outcome augmented influence-function score is

$$\begin{aligned}\varphi_g &= \int \hat{\mu}(H, a)g(a|H)da \\ &\quad + \frac{O w_g(H, A)}{\hat{r}(H, A)} \{Y - \hat{\mu}(H, A)\}.\end{aligned}\tag{13}$$

Under appropriate sample splitting and regularity, the corresponding estimator is consistent if either $\hat{\mu}$ is consistent or the joint treatment–response weighting components are consistent. This one-step expression is *not* a substitute for a longitudinal influence function: for $k > 1$, we recommend the sequential doubly robust / targeted-learning recursion with time-indexed nuisance models fitted on subject-disjoint folds. Its statistical performance and variance require an actual implementation and data-specific verification [22, 23, 25].

04 INTERVENTIONAL MEDIATION WITHOUT CROSS-WORLD COUNTERFACTUALS

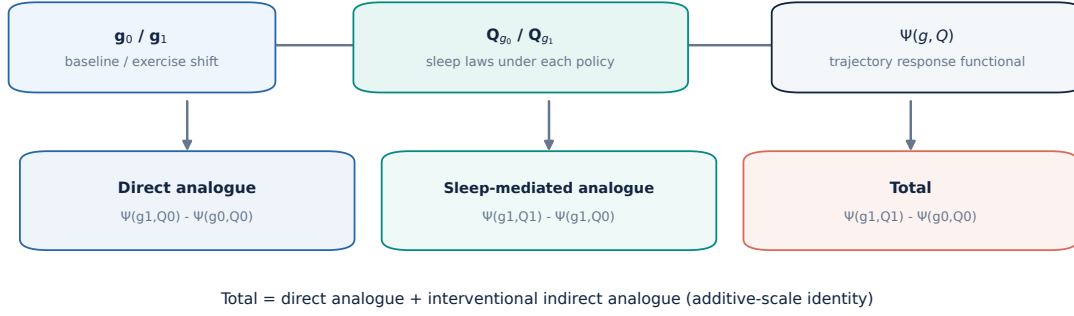


Figure 6. The randomized-mediator analogue compares four combinations of activity and sleep kernels; its total is defined on the same hybrid intervention scale. The diagram states estimands, not observed effects.

6.5 Local overlap and uncertainty-aware policy restriction

For an observational cohort, define the clipped likelihood ratio $\tilde{w}_t = \min\{g(A_t | H_t) / \hat{\pi}_0(A_t | H_t), w_{\max}\}$ and its *diagnostic*, not corrected, effective sample size

$$\text{ESS}(g) = \frac{(\sum_{it} \tilde{w}_{it})^2}{\sum_{it} \tilde{w}_{it}^2}, \quad (14)$$

$$\Gamma(g) = \mathbb{E}_H [\text{KL}\{g(\cdot | H) \| \hat{\pi}_0(\cdot | H)\}].$$

The evaluation regime is restricted to $\mathcal{G}_{\text{safe}} = \{g : \Gamma(g) \leq \kappa, \text{ESS}(g) \geq n_{\min}, \mathbb{P}_g(A \in \mathcal{A}(H)) = 1\}$, and abstains outside support. This protects *estimability* but does not certify physical safety or causality in the presence of unmeasured confounding. Candidate decisions may be selected by robust risk minimization,

$$g^* = \arg \min_{g \in \mathcal{G}_{\text{safe}}} \left\{ \hat{\Psi}(g) + \lambda_{\text{unc}} \widehat{\text{SE}}(\hat{\Psi}(g)) + \lambda_{\text{burden}} \mathbb{E}_g[c(A, H)] \right\}. \quad (15)$$

Here $c(A, H)$ quantifies behavioral change burden; lower distress is assumed better. This is a policy *research* objective, not individualized clinical advice.

6.6 Approximation error and interpretation

Proposition 2 (Finite-horizon transition perturbation) Suppose the target outcome is bounded by $|Y| \leq B$ and the total-variation difference between each true and estimated history transition (under the same supported policy) is at most ϵ_j uniformly on reachable histories. Then

$$\left| \Psi_{t,k}(g) - \hat{\Psi}_{t,k}(g) \right| \leq 2B \sum_{j=0}^{k-1} \epsilon_{t+j}. \quad (16)$$

Proof sketch. Insert and subtract the hybrid trajectory laws obtained by replacing one transition kernel at a time. Markov kernels do not expand total-variation distance, while the difference of expectations of a function bounded by B is at most $2B$ times total variation. Summing the k successive replacements gives Eq. (16). $\square$

The bound is a structural error propagation statement, *not* a finite-sample rate for learned representations or a guarantee when sequential ignorability fails. Alongside identifiability tests and overlap diagnostics, it motivates evaluating short horizons before attempting extended policy rollouts.

Funding

The authors received no specific funding for this work.

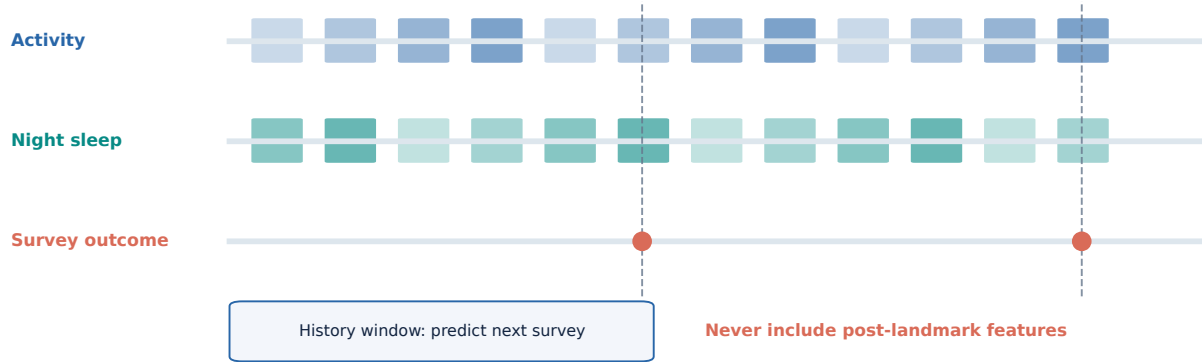


Figure 7. Landmark construction aligns daily activity, next-night sleep, and subsequent self-reports without future leakage. Survey recall windows must be audited for each source.

Algorithm 1 CISM training and supported policy evaluation

Require: Student-indexed wearable records; landmark outcomes; candidate policies $\mathcal{G}$

- 1: Freeze chronology; define pre-action H_t , action A_t , next sleep S_{t+1}
 - 2: Fit missingness and behavior-policy nuisance models on subject-disjoint folds
 - 3: Fit structural kernel K_θ with strictly historical masks and loss (8)
 - 4: **for** each test subject and feasible policy g **do**
 - 5: Audit simplex feasibility, support, and $\text{ESS}(g)$
 - 6: **if** support acceptable **then**
 - 7: Simulate K_θ under g ; estimate $\Psi(g)$ and hybrid $\Psi(g, Q)$
 - 8: Compute uncertainty and sequential augmentation on held-out folds
 - 9: **else**
 - 10: Report non-identification / abstention
 - 11: **end if**
 - 12: **end for**
 - 13: Evaluate synthetic recovery separately from real-data factual predictions
-

Competing Interests

The authors declare no competing interests.

Author Contributions

Y.W. contributed to the conceptualization of the study, research design, and preparation of the original manuscript.

M.N.N. contributed to the methodological development, interpretation of exercise–sleep–mental health relationships, and critical revision of the manuscript.

J.Z. contributed to data curation, dataset preparation, and the analysis of wearable behavioral and physiological measurements.

J.L. contributed to data preprocessing, temporal feature alignment, experimental implementation, and evaluation of predictive performance.

J.H.Z. contributed to statistical analysis, experimental evaluation, visualization of results, and interpretation of the findings.

D.W. contributed to the overall study design, methodological supervision, interpretation of the results, and review and editing of the manuscript.

All authors reviewed and approved the final manuscript.

Data Availability

This study used two existing datasets containing wearable sensor measurements and psychological assessments of university students: GLOBEM and the SSAQS Fitbit dataset.

The GLOBEM dataset contains longitudinal behavioral and physiological measurements collected from university students between 2018 and 2021, including physical activity, sleep, smartphone-derived behavioral features, and psychological assessments. The dataset is available through PhysioNet, subject to its credentialed-access requirements and applicable data-use conditions.

The SSAQS dataset contains Fitbit Inspire 3 measurements, including physical activity, step counts, sleep, heart-rate variability, and oxygenation, together with self-reported stress and anxiety assessments. The dataset is publicly available through Zenodo at <https://zenodo.org/records/18706837>.

Both datasets were used according to their respective access conditions. Dataset-specific preprocessing, temporal alignment, outcome definitions, and evaluation procedures are described in the manuscript.

The study additionally includes structural causal-recovery experiments and analytical evaluations based on mathematical models specified in the manuscript. These evaluations are distinguished from analyses of the observational student datasets. The equations and experimental settings required to describe these evaluations are provided in the Methods and Results sections.

Ethics Declaration

This study was conducted exclusively through secondary analysis of existing research datasets and computational modeling. No new human participants were recruited, no clinical or behavioral interventions were performed, and no new personally identifiable information was collected. The research involved no direct interaction with human participants, and all analyses were conducted using existing datasets in accordance with their respective access and usage conditions.

References

1. Predicting stress in first-year college students using sleep data from wearable devices, DOI: [10.1371/journal.pdig.0000473](https://doi.org/10.1371/journal.pdig.0000473) (2024).
2. Dodds, P. S. *et al.* Collective sleep and activity patterns of college students from wearable devices, DOI: [10.1038/s44260-025-00055-x](https://doi.org/10.1038/s44260-025-00055-x) (2025).
3. Detection of depressive symptoms in college students using multimodal passive sensing data and light gradient boosting machine: Longitudinal pilot study (2025).
4. Predictors of anxiety trajectories in cohort of first-year college students, DOI: [10.1016/j.jaacop.2024.08.004](https://doi.org/10.1016/j.jaacop.2024.08.004) (2025).
5. Garcia Ceja *et al.* A dataset of university students' stress and anxiety levels based on questionnaires and wearable sensors, DOI: [10.1038/s41597-026-07085-7](https://doi.org/10.1038/s41597-026-07085-7) (2026).
6. Lekkas, D. *et al.* Depression deconstructed: Wearables and passive digital phenotyping for analyzing individual symptoms (2023).
7. Nie, Y., Nguyen, N. H., Sinthong, P. & Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers (2023).
8. Liu, Y., Hu, T., Zhang, H. *et al.* iTransformer: Inverted transformers are effective for time series forecasting (2024).
9. Wang, S., Wu, H., Shi, X. *et al.* TimeMixer: Decomposable multiscale mixing for time series forecasting (2024).
10. Goswami, M. *et al.* MOMENT: A family of open time-series foundation models (2024).
11. Ansari, A. F., Stella, L., Turkmen, C. *et al.* Chronos: Learning the language of time series (2024).
12. Woo, G., Liu, C., Kumar, A. *et al.* Unified training of universal time series forecasting transformers (2024).
13. Liu, Y., Qin, G., Huang, X., Wang, J. & Long, M. Timer-XL: Long-context transformers for unified time series forecasting (2025).
14. Shi, X., Wang, S., Nie, Y. *et al.* Time-MoE: Billion-scale time series foundation models with mixture of experts (2025).

15. Jung, H. W., Kim, D. Y., Lee, I. *et al.* Key features of digital phenotyping for monitoring mental disorders: Systematic review, DOI: [10.2196/77331](https://doi.org/10.2196/77331) (2025).
16. Inter- and intrapersonal associations between physiology and mental health: A longitudinal study using wearables and mental health surveys, DOI: [10.2196/64955](https://doi.org/10.2196/64955) (2025).
17. The relationship between wearable-derived sleep features and relapse in major depressive disorder, DOI: [10.1016/j.jad.2024.07.136](https://doi.org/10.1016/j.jad.2024.07.136) (2024).
18. Bidirectional associations of physical activity, sleep, and self-reported mental health in young adults participating in an online wellness intervention during the covid-19 pandemic, DOI: [10.3389/fpubh.2023.1168702](https://doi.org/10.3389/fpubh.2023.1168702) (2023).
19. Correlation between physical exercise levels, depressive symptoms, and sleep quality in college students: Evidence from electroencephalography, DOI: [10.1016/j.jad.2024.10.043](https://doi.org/10.1016/j.jad.2024.10.043) (2025).
20. Baynham, R., Camargo, A., D'Alfonso, S. *et al.* The dynamic association between physical activity and psychological symptoms in young people with major depressive disorder: An active and passive sensing longitudinal cohort study, DOI: [10.1111/eip.70018](https://doi.org/10.1111/eip.70018) (2025).
21. Xiong, H., Wu, F., Deng, L., Su, M. & Lehman, L. H. G-Transformer: Counterfactual outcome prediction under dynamic and time-varying treatment regimes (2024).
22. Marmarelis, M. G., Morstatter, F., Galstyan, A. & Ver Steeg, G. Policy learning for localized interventions from observational data (2024).
23. Zeng, Z., Arbour, D., Feller, A. *et al.* Continuous treatment effects with surrogate outcomes (2024).
24. Katta, S., Parikh, H., Rudin, C. & Volfovsky, A. Interpretable causal inference for analyzing wearable, sensor, and distributional data (2024).
25. Bian, Z., Shi, C., Qi, Z. & Wang, L. Off-policy evaluation in doubly inhomogeneous environments, DOI: [10.1080/01621459.2024.2395593](https://doi.org/10.1080/01621459.2024.2395593) (2025).
26. Robins, J. M. A new approach to causal inference in mortality studies with a sustained exposure period: Application to control of the healthy worker survivor effect, DOI: [10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6) (1986).
27. Hernan, M. A. & Robins, J. M. Causal inference: What if (2020).
28. Pearl, J. Causality: Models, reasoning, and inference (2009).
29. VanderWeele, T. J. A unification of mediation and interaction: A four-way decomposition, DOI: [10.1097/EDE.0000000000000121](https://doi.org/10.1097/EDE.0000000000000121) (2014).
30. Choi, A., Ooi, A. & Lottridge, D. Digital phenotyping for stress, anxiety, and mild depression: Systematic literature review, DOI: [10.2196/40689](https://doi.org/10.2196/40689) (2024).
31. Wang, R., Chen, F., Chen, Z. *et al.* StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones (2014).
32. Xu, X., Zhang, H., Sefidgar, Y. *et al.* GLOBEM dataset: Multi-year datasets for longitudinal human behavior modeling generalization (2022).
33. Physiological and emotional assessment of college students using wearable and mobile devices during the 2020 covid-19 lockdown: An intensive, longitudinal dataset, DOI: [10.1016/j.dib.2024.110228](https://doi.org/10.1016/j.dib.2024.110228) (2024).
34. Sleep and activity patterns in depression from wearable data: Unsupervised clustering study, DOI: [10.2196/86900](https://doi.org/10.2196/86900) (2026).
35. Sameh, A. *et al.* Digital phenotyping using wearable-determined physical behaviours and machine learning to detect depression and anxiety symptoms in a general population, DOI: [10.1016/j.psychres.2026.117242](https://doi.org/10.1016/j.psychres.2026.117242) (2026).
36. Sleep disturbance recorded via wearable sensors predicts depression severity 9 years later, DOI: [10.1016/j.jad.2025.120426](https://doi.org/10.1016/j.jad.2025.120426) (2026).

37. Shim, J. & Onnela, J.-P. Longitudinal digital phenotyping of circadian rest-activity rhythms via wearables as biomarkers for late-life function, cognition, and neuropsychiatric health, DOI: [10.1186/s44247-026-00256-x](https://doi.org/10.1186/s44247-026-00256-x) (2026).
38. Shim, J. & Onnela, J.-P. Longitudinal digital phenotyping of activity rhythms and biological aging using commercial wearables (2026).
39. Bui, A. A. T. *et al.* The digital mental health study assessed feasibility of large-scale digital sensing for depression and anxiety (2026).
40. Moura, I., Teles, A., Viana, D. *et al.* Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: A systematic literature review, DOI: [10.1016/j.jbi.2022.104278](https://doi.org/10.1016/j.jbi.2022.104278) (2023).
41. Bufano, P., Laurino, M., Said, S., Tognetti, A. & Menicucci, D. Digital phenotyping for monitoring mental disorders: Systematic review, DOI: [10.2196/46778](https://doi.org/10.2196/46778) (2023).
42. Heckler, W. F., Feijo, L. P., Carvalho, J. V. d. & Barbosa, J. L. V. Digital phenotyping for mental health based on data analytics: A systematic literature review, DOI: [10.1016/j.artmed.2025.103094](https://doi.org/10.1016/j.artmed.2025.103094) (2025).
43. Akre, S., Seok, D., Douglas, C. *et al.* Advancing digital sensing in mental health research, DOI: [10.1038/s41746-024-01343-x](https://doi.org/10.1038/s41746-024-01343-x) (2024).
44. Zhang, Y., Wang, J., Zong, H. *et al.* The comprehensive clinical benefits of digital phenotyping: From broad adoption to full impact, DOI: [10.1038/s41746-025-01602-5](https://doi.org/10.1038/s41746-025-01602-5) (2025).
45. Digital phenotyping for mental health conditions: A systematic review of implementation and application (2026).
46. Mobile technology for just-in-time prediction of depression: A scoping review (2026).
47. Reciprocal association between rest-activity rhythm and depression in young adults: An ecological momentary assessment study (2026).
48. Effects of a guided digital intervention on sleep and mental health outcomes in university students: A randomized controlled trial, DOI: [10.1093/sleep/zsaf357](https://doi.org/10.1093/sleep/zsaf357) (2026).
49. Du, W., Cote, D. & Liu, Y. SAITS: Self-attention-based imputation for time series, DOI: [10.1016/j.eswa.2023.119619](https://doi.org/10.1016/j.eswa.2023.119619) (2023).
50. Wu, H. *et al.* TimesNet: Temporal 2d-variation modeling for general time series analysis (2023).
51. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies, DOI: [10.1037/h0037350](https://doi.org/10.1037/h0037350) (1974).
52. Nagalapatti, L., Iyer, A., De, A. & Sarawagi, S. Continuous treatment effect estimation using gradient interpolation and kernel smoothing (2024).