Hierarchical Modeling of mHealth Intervention Effects with Missing-Data Considerations in the Intern Health Study

Hanzhen (Jenny) Zhu

Supervisors: Luke Francisco and Prof. Ambuj Tewari

April 2025

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Honors Data Science at the University of Michigan

Abstract

We evaluate how daily push-notification interventions (targeting mood, activity, and sleep) affect medical interns' well-being using Intern Health Study data from 2022–2023. We applied hierarchical time-series models (using `Pymer4` and `PyMC`) to estimate both individual and population-level intervention effects under varying data completeness thresholds (30%, 50%, 70%). To address missing data, which are likely related to behavioral and mental health outcomes, we compare complete case analysis (CCA), mean imputation, linear interpolation, and multiple imputation by chained equations (MICE). Using a held-out predictive framework, we find that MICE combined with next-day zeroing-out of interventions yields the most reliable effect estimates, with the lowest out-of-sample prediction error. Under this best-performing setup, a mood notification lifted the next-day mood by about 0.05 SD on average, and the improvement grew by another 0.09 SD for every 1 SD drop in the prior-day mood; in contrast, a sleep notification reduced that night's sleep by approximately 0.06 SD. Our results demonstrate the importance of principled missing-data handling and offer actionable guidance for designing adaptive, personalized mHealth interventions in high-stress populations.

Keywords: hierarchical mixed-effects modeling, Bayesian MCMC, causal inference, intervention effect estimation, missing data, multiple imputation, longitudinal wearable data, mobile health (mHealth), mental health

1 Introduction

Medical interns face intense stress during training, leading to much higher rates of depression than the general population (Tennant, 2001). Addressing this is crucial for their well-being and professional performance. In high-stress environments like medical training, prevention-focused mental health interventions are important, especially when time and resources for traditional mental health counseling are limited. Wearable devices provide a practical solution by passively monitoring behaviors such as sleep and activity and offering support through targeted push notifications. The NIH-funded Intern Health Study (IHS) investigates how mobile health (mHealth) interventions can improve interns' mental health during these stressful periods. Our study aims to investigate how randomized interventions from the micro-randomized trial (MRT) can inform adaptive strategies.

A prior IHS study from NeCamp et al. (2020), which analyzed 2018 IHS data, highlights how the effectiveness of mHealth notifications depends on interns' prior outcomes (i.e., moderation effects). For example, they observed that after receiving notifications, interns with low mood scores in the preceding week tended to experience mood improvements, whereas those with higher prior-week mood scores tended to experience mood declines, highlighting the need for targeted strategies. Furthermore, the study's Discussion section emphasizes that future efforts should focus on designing tailored notifications, which motivates our study.

Building on these findings, our study investigates time-series models to *isolate* intervention effects and thus approximate *causal* relationships from observational data, which is a meaningful step toward designing adaptive interventions. Our study uses IHS data from 2022 and 2023, which share a consistent randomization scheme: each evening, participants had a 50% chance of receiving a notification targeting one of three areas: mood, activity, or sleep. Unlike previous years, these interventions were not tailored to participants' prior outcomes, providing a cleaner framework for evaluating causal effects.

Initially, we used Bayesian structural time-series causal inference methods by applying Google's `CausalImpact` model to estimate individual-level causal effects of various interventions (targeting activity, mood, sleep, and support versus consequences) (Brodersen et al., 2015). Although effective in quantifying trends, `CausalImpact` had challenges in creating robust synthetic baselines due to the high frequency of interventions (once every two days on average).

To deal with this limitation, we transitioned to a hierarchical Bayesian framework and employed two models to cross-verify results. Using `PyMC`, we implemented a mixed-effects model with MCMC sampling (Salvatier et al., 2015). `PyMC` not only isolates each intervention's impact when interventions overlap but also balances individual-level and global intervention effects through flexible multilevel priors. We estimated the slope coefficients of intervention indicators (0, 1) on outcomes using Markov Chain Monte Carlo (MCMC), isolating and quantifying intervention effects, and visualizing their distributions across individuals. To validate our findings, we implemented a comparable multilevel hierarchical model using `Pymer4` (Jolly, 2018), which employs a frequentist restricted maximum likelihood (REML) approach. We verified that both models produced nearly identical moderation effects (see Appendix A for examples), so for brevity we present only the `Pymer4` results. By demonstrating consistency across Bayesian and frequentist frameworks, this dual-model strategy enhances the robustness of our conclusions.

Another significant challenge in longitudinal wearable studies is missing data due to intermittent non-response and participant dropout. Conventional methods—such as complete-case analysis or mean imputation—can reduce statistical power and introduce bias when data are not missing completely at random (MCAR) (Schafer and Graham, 2002; White and Carlin, 2010). In the IHS, missingness may depend not only on interns' mental health status (e.g., low mood reducing engagement) but also on their sleep and activity patterns or sensor measures, thereby violating the MCAR assumption. To address this, we employ multiple imputation by chained equations (MICE) (Azur et al., 2011; van Buuren and Groothuis-Oudshoorn, 2011), which assumes data are missing at random (MAR). MICE iteratively

imputes missing values using regression models conditioned on observed covariates, preserving data structure and relationships. Importantly, it accounts for uncertainty by generating multiple plausible datasets that are later pooled for final estimates. To empirically evaluate imputation accuracy, we designed a predictive experiment comparing MICE with simpler methods, including complete-case analysis (CCA, which drops any day with a missing value), linear interpolation, and mean imputation. This experiment ranks these imputation methods by withholding a set of fully observed days, imputing the remaining data with each method, and then measuring how accurately each one predicts the held-out values. The method with the lowest error is considered the closest proxy to a fully observed dataset (i.e., a 'never-missing' dataset).

Our study has two primary objectives: (1) to isolate and quantify causal effects of mHealth notifications on daily mood, steps, and sleep using hierarchical time-series models, and (2) to determine the optimal strategy for handling missing data in longitudinal wearable studies. By integrating robust causal inference with principled missing data handling, our work offers actionable, data-driven guidance for designing personalized, accessible, and cost-effective mHealth interventions for medical interns and other high-stress populations.

2 Methods

2.1 Data Sources and Preprocessing

The Intern Health Study (IHS) provides, for each participant and study day (2022–2023), (i) sensor measurements, including sleep minutes, step count, mood rating (1-10), and (ii) three binary indicators (0/1) showing whether a sleep-, step-, or mood-focused notification was delivered the previous evening.

The sensor dataset contains missing data for two main reasons: (1) participant dropout, resulting in few or no measurements after a certain day, and (2) intermittent non-response. To address dropout, we applied a 14-day sliding window to assess data completeness. For each participant, we retained data up to the point where fewer than a specified *data completeness threshold*, $X\%$ (where $X = 30\%, 50\%, \text{or } 70\%$), of days within the window had valid measurements; data beyond this cutoff were excluded. Below this threshold, day-level imputations become unreliable, and an intern who reports so rarely is unlikely to notice or respond to the push notifications.

After removing dropout days using the sliding-window filter, we addressed the remaining intermittent gaps with three imputation methods and complete-case analysis (CCA). Specifically, missing values were imputed using linear interpolation, multiple imputation by chained equations (MICE), and mean imputation. CCA excluded all records with missing values, retaining only fully observed data. Further details on imputation workflows are provided in Section 2.3.2.

To prepare the cleaned datasets for `PyMC` and `Pymer4` model fitting, we implemented a systematic preprocessing pipeline. First, sensor data (steps, mood, and sleep) were standardized within each participant to ensure comparability across individuals and metrics. For instance, a standardized step value of 1 on a given day indicates that the participant's steps were 1 standard deviation (SD) above their average. Next, we generated lagged variables (`lag1` and `lag2`) for mood, steps, and sleep to capture trends from the previous two days, thereby capturing temporal dependencies. Lagged variables (`lag1` and `lag2`) for intervention indicators were created to account for potential delayed effects.

Furthermore, for the imputed sensor datasets (using data completeness thresholds of 30%, 50%, or 70%), we compared two approaches to handling intervention data: *next-day zeroing-out* and *no zeroing-out*. In the next-day zeroing-out strategy, an intervention indicator is set to 0 when the outcome it targets is missing on the following day, because we want to avoid attributing any effect to an intervention whose impact cannot be verified. In contrast, the no zeroing-out approach retained the original intervention indicators regardless of missingness.

By treating the no zeroing-out method as a baseline, we could evaluate both the relative significance and the direction of intervention effects using the next-day zeroing-out strategy.

To control for to and quantify the diminishing effectiveness of interventions over time, we introduced a day counter variable and calculated day-intervention interaction terms (e.g., $\texttt{day}_{ij} \times \texttt{intervention\_steps\_lag1}_{ij}$).

Initially, the sensor dataset comprised 1,404 interns. Under the *next-day zeroing-out* strategy and preprocessing, applying a 70% completeness threshold retained 321 interns, each with 167 days on average. This yielded 3,665 step, 4,142 mood, and 5,073 sleep interventions. Relaxing the threshold to 50% increased the sample to 719 interns (on average 178 days each), with 7,612 step, 8,532 mood, and 10,389 sleep interventions. The 30% threshold retained 951 interns (on average 190 days each), with 9,465 step, 10,650 mood, and 12,932 sleep interventions. Using the *no-zeroing-out* strategy roughly doubles the intervention counts at each threshold. See Table 1 for full sample statistics.

Table 1: Sample statistics after preprocessing under each data-completeness threshold and intervention-handling strategy.

Thres.	# Subjects	Avg. days	Next-day zeroing-out			No zeroing-out		
			Step Int.	Mood Int.	Sleep Int.	Step Int.	Mood Int.	Sleep Int.
70%	321	167	3,665	4,142	5,073	6,418	7,170	9,294
50%	719	178	7,612	8,532	10,389	15,725	17,530	22,097
30%	951	190	9,465	10,650	12,932	22,320	24,950	31,243

2.2 Modeling Approaches: `PyMC` and `Pymer4`

We implemented hierarchical mixed-effects models with `PyMC` and `Pymer4`.

`PyMC` is a probabilistic programming library for Bayesian modeling that allows for the specification of hierarchical models using Markov Chain Monte Carlo (MCMC) algorithms with the No-U-Turn Sampler (NUTS), an extension of Hamiltonian Monte Carlo (HMC). MCMC is used to approximate complex posterior distributions when direct computation is infeasible. They rely on a Markov chain that explores the parameter space by sampling in a way that reflects the posterior distribution. After many iterations, the chain converges to the target posterior, which enables accurate parameter estimation.

In our `PyMC` models, we set priors for most coefficients as $\mathcal{N}(0, 1)$ since we normalized the sensor data. Variance parameters used HalfNormal priors to ensure numerical stability by constraining variance estimates to be positive. Subject-level priors were modeled as Normal distributions with mean=0 and standard deviations defined by these HalfNormal priors, which capture individual variability around global effects (modeled with $\mathcal{N}(0, 1)$ priors). This creates a hierarchical relationship between global and subject-level parameters. The NUTS sampler (target acceptance rate 0.95) and posterior predictive checks evaluate uncertainty and generate p-values for parameter estimates.

`Pymer4` is a `Python` library for fitting multilevel regression models, similar to `lme4` in `R`. By building another hierarchical model without relying on Bayesian MCMC, `Pymer4` allowed us to compare results through cross-comparison and model refinement, ensuring result robustness.

Multilevel Modeling Function

In this project, both `PyMC` and `Pymer4` were used to analyze the effects of interventions (steps, mood, sleep) on outcomes using multilevel (hierarchical) models. These models incorporate global (fixed) effects to capture population-level trends and subject-level (random) effects to account for individual variability. For instance, if the outcome is today's steps (i.e., steps lag

0), the model equation is:

$$
\begin{aligned}
\text{steps}_{ij} = \beta_0 + \\
&\textbf{Lagged Effects:} \\
&\beta_1 \cdot \text{mood\_lag1}_{ij} + \beta_2 \cdot \text{mood\_lag2}_{ij} + \beta_3 \cdot \text{sleep\_lag1}_{ij} + \\
&\beta_4 \cdot \text{sleep\_lag2}_{ij} + \beta_5 \cdot \text{steps\_lag1}_{ij} + \beta_6 \cdot \text{steps\_lag2}_{ij} + \\
&\textbf{Interventions:} \\
&\beta_7 \cdot \text{intervention\_mood\_lag1}_{ij} + \beta_8 \cdot \text{intervention\_mood\_lag2}_{ij} + \\
&\beta_9 \cdot \text{intervention\_sleep\_lag1}_{ij} + \beta_{10} \cdot \text{intervention\_sleep\_lag2}_{ij} + \\
&\beta_{11} \cdot \text{intervention\_steps\_lag1}_{ij} + \beta_{12} \cdot \text{intervention\_steps\_lag2}_{ij} + \\
&\textbf{Moderators:} \\
&\beta_{13} \cdot (\text{intervention\_steps\_lag1}_{ij} \cdot \text{steps\_lag1}_{ij}) + \beta_{14} \cdot (\text{intervention\_steps\_lag2}_{ij} \cdot \text{steps\_lag2}_{ij}) + \\
&\beta_{15} \cdot (\text{intervention\_steps\_lag1}_{ij} \cdot \text{mood\_lag1}_{ij}) + \beta_{16} \cdot (\text{intervention\_steps\_lag2}_{ij} \cdot \text{mood\_lag2}_{ij}) + \\
&\beta_{17} \cdot (\text{intervention\_steps\_lag1}_{ij} \cdot \text{sleep\_lag1}_{ij}) + \beta_{18} \cdot (\text{intervention\_steps\_lag2}_{ij} \cdot \text{sleep\_lag2}_{ij}) + \\
&\textbf{Day-Intervention Interaction Terms:} \\
&\beta_{19} \cdot (\text{day}_{ij} \cdot \text{intervention\_steps\_lag1}_{ij}) + \beta_{20} \cdot (\text{day}_{ij} \cdot \text{intervention\_steps\_lag2}_{ij}) + \\
&\beta_{21} \cdot (\text{day}_{ij} \cdot \text{intervention\_sleep\_lag1}_{ij}) + \beta_{22} \cdot (\text{day}_{ij} \cdot \text{intervention\_sleep\_lag2}_{ij}) + \\
&\beta_{23} \cdot (\text{day}_{ij} \cdot \text{intervention\_mood\_lag1}_{ij}) + \beta_{24} \cdot (\text{day}_{ij} \cdot \text{intervention\_mood\_lag2}_{ij}) + \\
&\textbf{Subject-Level Effects:} \\
&u_i + u_{i1} \cdot \text{intervention\_steps\_lag1}_{ij} + u_{i2} \cdot \text{intervention\_steps\_lag2}_{ij} + \\
&u_{i3} \cdot \text{steps\_lag1}_{ij} + u_{i4} \cdot \text{steps\_lag2}_{ij} + \epsilon_{ij}
\end{aligned}
$$

Both `PyMC` and `Pymer4` models output the estimated coefficients for each slope coefficient (global β and subject-level u) with their standard errors, confidence intervals, and p-values. These outputs enable statistical evaluation of predictor effects on the outcome, which in this case is today's step ($steps_{ij}$). Note that all numerical variables, including the outcome ($steps_{ij}$) and numerical predictors (*mood*, *sleep*, and *steps* lags), are normalized before model fitting to ensure comparability.

Here is a break-down explanation of the formula above:

- **Lagged effects** represent the effects of mood, sleep, and steps from the previous two days on today's step count.
- **Interventions** assess the overall effectiveness of three types of notifications (step-, mood-, and sleep-targeting), if sent one or two evenings prior, on improving step levels. These variables are binary indicators (0/1).
- **Moderators** capture how prior-day outcomes interact with interventions targeting steps to influence the steps outcome over the previous two days. Similarly, when mood or sleep is the response, the moderators adjust to moderate interventions targeting mood or sleep.
- **Day-intervention interaction terms** not only control for, but also quantify the diminishing effect of interventions over time, which can occur due to participant fatigue.
- **Subject-Level terms** capture individual variability specific to the response variable. For example, if steps are the response, these terms account for subject-specific baseline step levels (u_i), responsiveness to step-related interventions (u_{i1}, u_{i2}), and lagged effects of steps (u_{i3}, u_{i4}), where $i = 1, 2, \ldots, n$ (n=number of subjects). Similarly, when mood or sleep is the response, the subject-level terms adjust to reflect baseline levels, intervention responsiveness, and lagged effects of those outcomes.

2.3 Imputation Evaluation

2.3.1 Theoretical Framework: Missing Data

Let $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ represent the *full* data that would be observed if no data were missing, where Y_{obs} denotes observed values and Y_{mis} denotes missing values. Let X denote fully observed covariates. For each observation y_{ik} (participant i, variable k), define an indicator d_{ik} such that $d_{ik} = 1$ if y_{ik} is observed and $d_{ik} = 0$ otherwise.

Missing Completely at Random (MCAR). Data are MCAR if the probability of missingness is independent of both observed and unobserved data:

$$\Pr(d_{ik} = 1 \mid Y_{\text{obs}}, Y_{\text{mis}}, X) \;=\; \Pr(d_{ik} = 1).$$

Under MCAR, the complete cases are a random subset of the full data, meaning that CCA yields unbiased estimates.

Missing at Random (MAR). Data are MAR if the probability of missingness depends solely on observed data (Y_{obs}) and fully observed covariates X, but not on unobserved values:

$$\Pr(d_{ik} = 1 \mid Y_{\text{obs}}, Y_{\text{mis}}, X) = \Pr(d_{ik} = 1 \mid Y_{\text{obs}}, X).$$

In this case, any systematic differences between observed and missing data can be explained by the variables we have. Imputation methods using observed data (e.g., multiple imputations like MICE) are theoretically valid under MAR.

Missing Not at Random (MNAR). Data are MNAR if the chance of missingness depends on unobserved values even after conditioning on the observed data:

$$\Pr(d_{ik} = 1 \mid Y_{\text{obs}}, Y_{\text{mis}}, X) = f(Y_{\text{mis}}, Y_{\text{obs}}, X).$$

where f depends on Y_{mis}.

Here, the missingness mechanism is influenced by the missing values themselves. This makes MNAR difficult to detect and model with standard methods.

2.3.2 CCA and 3 Imputation Methods: Mean, Interpolation, and MICE

After applying the completeness threshold (30%, 50%, or 70% sliding window) and before standardizing variables within subjects, we addressed intermittent missing data using four strategies. Each strategy makes different assumptions and has unique strengths and weaknesses:

1. Complete Case Analysis (CCA):

In our study, the *Complete Dataset* refers to data retained through complete-case analysis (CCA), where any day missing mood, steps, or sleep values is excluded. CCA yields unbiased estimates only under the Missing Completely at Random (MCAR) assumption. Otherwise, CCA reduces sample size and statistical power and can introduce bias in either direction—since the true missingness mechanism is unknown. For example, if lower-mood days are more likely to go unreported, dropping them could inflate the estimated mood improvements. We therefore use CCA primarily as a conservative baseline for comparing imputation methods, rather than a primary analytic approach.

2. Mean Imputation:

In mean imputation, missing values are replaced by the individual's observed mean for that variable. This approach preserves the overall average at the individual level but neglects temporal trends and underestimates variability. As a result, regression estimates may be biased. For example, imputing missing step counts with a participant's average value may mask important day-to-day fluctuations that correlate with mood or intervention effects.

3. Linear Interpolation:

Linear interpolation (`interpolate("linear")`) estimates missing data points based on the linear trend between adjacent observed points in each participant's time series. For instance, a sequence such as 7, NaN, NaN, 4 would be imputed as 7, 6, 5, 4. Although interpolation maintains continuity, it can oversmooth data, potentially obscuring true variability and artificially affecting intervention effect estimates.

4. Multiple Imputation by Chained Equations (MICE):

MICE addresses missing data by iteratively imputing plausible values based on relationships observed in the data. MICE assumes that data are Missing at Random (MAR), meaning that the probability of missingness depends only on these specified covariates. This assumption is more plausible in the context of the IHS, in which a participant's likelihood of reporting may be influenced by recent observations. For example, a notably low mood on the previous day might reduce a participant's motivation to report mood or track steps on the following day.

In our MICE implementation, each missing value was predicted using 12 variables: the current value (lag0) and three past values (lag1–lag3) for mood, steps, and sleep. To stabilize variance, steps and sleep were square-root transformed during imputation and later squared to return to the original scale. We evaluated 5, 10, 15, and 20 iterations and found that the imputed distributions converged by the 5th iteration. Therefore, we selected the dataset produced after 5 iterations with one imputation, because its distribution closely resembled the observed data—acknowledging that perfect alignment is unlikely due to the non-MCAR nature of the missingness. Compared to simpler methods, MICE introduces additional uncertainty and supports more robust statistical inference.

It is important to note that, while MAR is a more reasonable assumption than MCAR in our context, it is inherently untestable from the observed data alone. If the data are Missing Not at Random (MNAR), where missingness depends on unobserved factors, MICE may introduce bias. For example, if the likelihood that participants will miss their step count on a given day increases when their step count on the previous day is missing—and that previous day's missing value is associated with even earlier missing step counts—the imputation model may repeatedly rely on an incomplete history. This cascading dependency can lead to systematic over- or under-estimation of the imputed step counts. Therefore, we designed a **predictive experiment** below to evaluate how well each imputation method preserves the data's predictive structure.

2.3.3 Predictive Experiment

First, we split the pre-processed data (70% data completeness threshold and next-day zeroing-out) into a 90% training set and a 10% test set, ensuring that each intern appears in only one split to assess how well each method generalizes to new individuals.

Next, using the training set, we generated four sensor datasets: the non-imputed complete-case data and three versions imputed by mean imputation, linear interpolation, and MICE. After merging intervention indicators and standardizing variables within subjects, we fitted a multivariate linear model with the predictor set described in Section 2.1, omitting subject-level random effects, separately for each outcome (mood, steps, sleep). This produced 4 imputations $\times$ 3 outcomes = 12 fitted models.

Then, we applied the same imputation procedures to the test set and used the coefficients estimated from the training data to predict each outcome for the held-out interns. Prediction error was summarized by the mean-squared error (MSE) calculated on test observations with *non-missing* true values. One MSE was obtained for every outcome–imputation combination, yielding a total of 12 MSE values.

Since all data were normalized within subject to have a variance of 1, a naive model that always predicts the subject mean (0) would yield an MSE of 1. Therefore, an MSE below 1 indicates that the model captures additional predictive information beyond the baseline, whereas an MSE of 1 or higher suggests that the model predicts no better than simply guessing the subject mean. This serves as an important benchmark for assessing imputation quality.

Finally, we compared the MSE values across the imputation methods to identify the method that best preserves the predictive structure of the data. The method with the lowest MSE is considered to most reliably represent the hypothetical never-missing scenario.

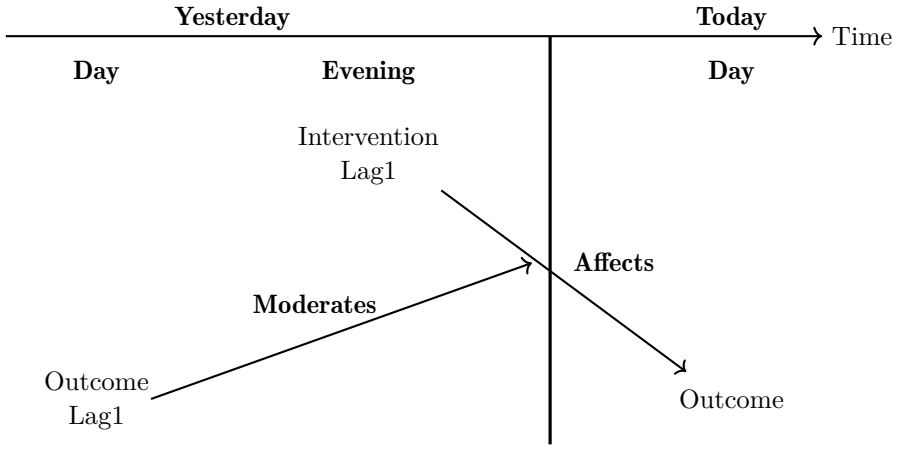
3 Results

3.1 Intervention Effects

To evaluate intervention effects, we conducted a comprehensive analysis examining how methodological choices in data handling influence results. Our investigation includes (i) an overview of how zeroing-out strategies, data completeness thresholds, and imputation methods interact to shape effect estimates and (ii) an in-depth comparison of results under two imputation methods, interpolation and MICE.

In our analysis, we focus on two key components of the intervention effect estimates:

- **Average Intervention Effects:** The y-intercept in our moderation effect plots represents the overall impact of the targeted notifications. It quantifies the baseline change in today's outcome that can be attributed to the intervention sent yesterday evening when yesterday's outcome is at the individual's mean. This isolates the net effect of the intervention.
- **Moderation Effects:** The slope in the moderation effect plots shows how an individual's outcome from the previous day (Lag 1) modulates the impact of the intervention on today's outcome (Lag 0). A negative slope indicates that the benefit of the intervention diminishes as the baseline outcome increases, while a positive slope suggests that higher baseline outcomes enhance the intervention's efficacy. In essence, the slope measures the extent to which yesterday's performance influences the intervention's effect.



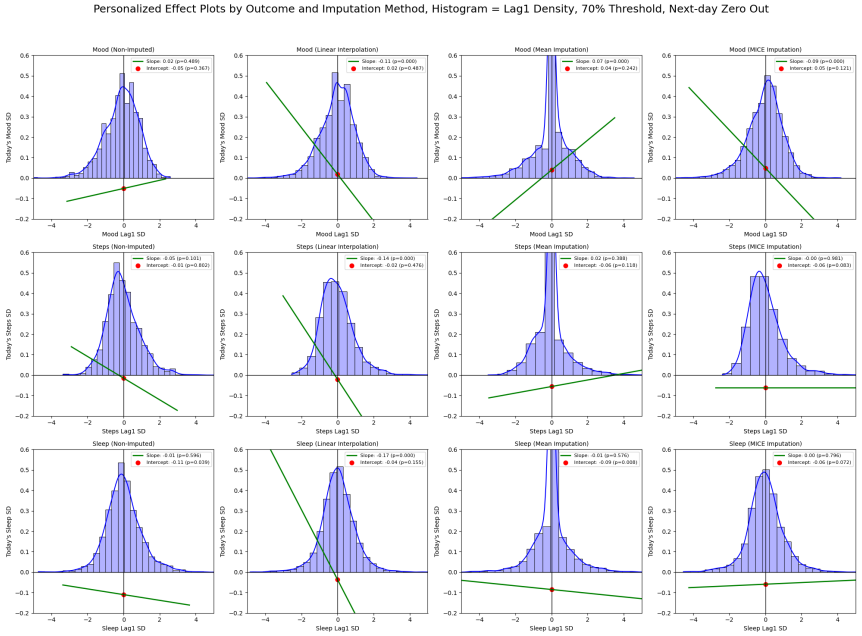
Conceptual timeline: An intervention sent yesterday evening (Intervention Lag1) may affect today's outcome, and yesterday's outcome (Outcome Lag1) may moderate that effect.

3.1.1 Overview of Methodological Impacts

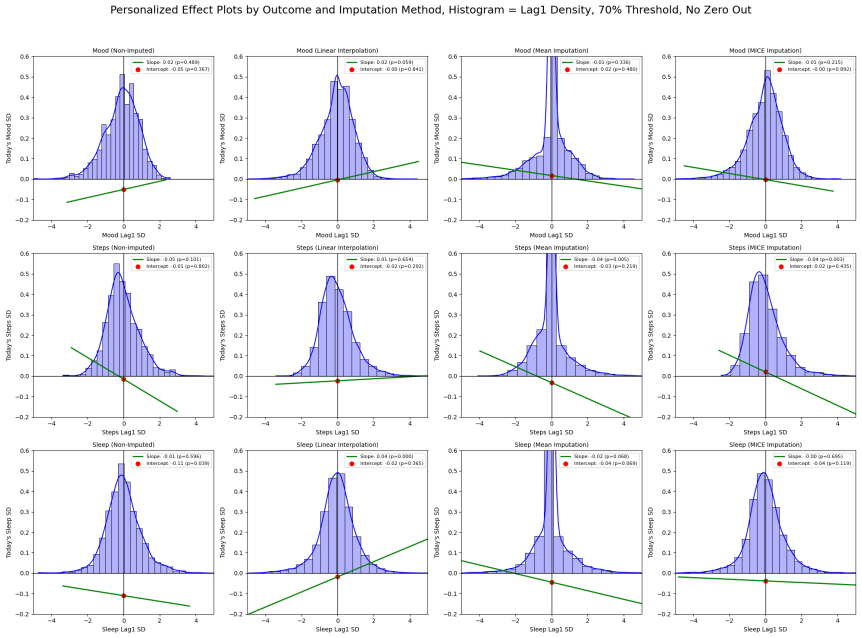
Each subplot in Figure 1 displays the estimated moderation effects (slopes) and average intervention effects (y-intercepts) for steps, mood, and sleep outcomes, across CCA and three imputation methods. Overlaid histograms illustrate the distribution of Lag1 outcome values, standardized within individuals.

To enhance readability, we focus on results at the 70% data completeness threshold, presented in Figure 1, which compares intervention effects under two zeroing-out strategies:

next-day zeroing-out and no zeroing-out. Corresponding results at the 30% threshold are included in Appendix B for reference.



(a) Next-day Zeroing-Out, 70% Threshold



(b) No Zeroing-Out, 70% Threshold

Figure 1: Composite graphs comparing intervention effects under two zeroing-out strategies (next-day vs. no zeroing-out) at the **70% data completeness threshold**.

Importantly, we evaluate how three factors influence the estimated intervention effects:

1. **Zeroing-Out Strategy:** When comparing the next-day zeroing-out approach with the no zeroing-out, our graphs reveal that not zeroing out results in average effects near zero

and flat moderation slopes, showing weaker or no effects. This supports our hypothesis that keeping intervention flags as 1 even when the next-day outcome is missing (i.e., no zeroing-out) overestimates engagement by treating missing outcomes as engaged, which dilutes the estimated effect and pulls the average toward zero. In contrast, where intervention indicators are set to 0 on days when the next-day outcome is missing, produces more pronounced effects. This comparison demonstrates that the intervention indeed has a measurable impact.

2. **Data Completeness Threshold:** When applying a 70% data completeness threshold—which retains fewer data points and involves a lower proportion of imputed values—the net effects (y-intercepts) tend to be more pronounced than those observed with a 30% threshold. This trend is particularly notable in complete-case analysis (CCA) and mean imputation.

3. **Imputation Method:** All composite graphs reveal clear differences among imputation methods. Under next-day zeroing-out, linear interpolation consistently produces strong, negative moderation slopes at both the 70% and 30% thresholds, meaning that as the Lag1 value increases, the beneficial effect of the intervention declines sharply. Although MICE exhibits flatter slopes (for steps and sleep), its y-intercepts are more pronounced, suggesting stronger average intervention effects compared to linear interpolation. Meanwhile, mean imputation yields unstable effects under different data processing conditions.

3.1.2 In-Depth Intervention Effect Analysis: Interpolation V.S. MICE

To isolate the impact of imputation methods, we compared linear interpolation and MICE under a **70% data completeness threshold** and **next-day zeroing-out** strategy.

To evaluate the effects of interventions, we examined the slope coefficients of intervention variables in our hierarchical models. Coefficients were considered significant if their p-value < 0.05.

Outcome: Steps. For interpolation (Table 2), both prior day's steps (0.522, $p < 0.001$) and prior two days' mood (0.010, $p < 0.05$) significantly predict today's step count. Step interventions also show significant moderation by prior steps (lag1: -0.135; lag2: -0.059; both $p < 0.001$), suggesting that interventions are less effective when recent physical activity is already high.

In contrast, under MICE (Table 3), prior mood (lag1: 0.017; lag2: 0.012), and prior steps (lag1: 0.039; lag2: 0.041) all significantly predict today's step count (all $p < 0.05$). Notably, intervention effects on mood (-0.064) and sleep (0.065) are also significant under MICE, while these effects were not observed under interpolation. However, the moderation effects observed under interpolation are absent here.

Figure 2 visually contrasts these findings. These moderation effect plots illustrate how the impact of step interventions on today's step count depends on yesterday's steps relative to an individual's average. The horizontal axis represents the deviation of yesterday's steps from the individual mean, measured in standard deviations (SD). The vertical axis displays the expected change in today's steps (in SDs) resulting from a step intervention delivered the previous evening. An overlaid histogram shows the distribution of lag1 steps (in SDs) from the individual mean. The y-intercept, slope, and x-intercept in these plots provide important insights. To clarify this, consider the interpolation plot (left panel of Figure 2) as an example:

- **Y-intercept (-0.021).** This is the estimated intervention effect when yesterday's steps are exactly at the individual's mean. Here, receiving a step intervention leads to a small negative change of -0.021 SDs in today's steps, but this effect is not statistically significant ($p = 0.476$).

- **Slope (-0.135).** This is the estimate of the moderation effect. For each additional SD above the mean in yesterday's steps, the effect of the intervention decreases by 0.135 SDs. Conversely, for each SD below the mean, the intervention effect increases by 0.135 SDs. Mathematically:

$$\text{Intervention Effect} = \beta_{\text{I}} + \beta_{\text{M}} \times (\text{Yesterday's Steps in SDs})$$

  where $\beta_{\text{I}} = -0.021$ (intercept) and $\beta_{\text{M}} = -0.135$ (slope).

  For example, if an individual had 1 SD below their average steps yesterday ($x = -1$) and received a step intervention, the expected increase in today's steps is:

$$-0.021 + (-0.135) \times (-1) = 0.114 \text{ SDs}.$$

- **X-intercept (-0.156).** This is the point where the intervention effect becomes zero. Hence, if yesterday's steps exceed about -0.156 SDs above the individual's mean, the effect of sending a step intervention becomes negative. In other words, above this threshold, a step notification is unlikely to improve today's step count and may instead reduce its effectiveness.

Table 2: Significant Global Slope Coefficients (Outcome = Steps), Interpolation

Variable	**Estimate**	**2.5% CI**	**97.5% CI**	**P-value**
`steps_lag1`	0.522	0.502	0.541	0.000***
`steps_intervention_steps_moderator_lag2`	-0.059	-0.088	-0.029	0.000***
`steps_intervention_steps_moderator_lag1`	-0.135	-0.164	-0.105	0.000***
`mood_lag2`	0.010	0.001	0.018	0.035*
`day_intervention_interaction_steps_lag2`	-0.000	-0.001	-0.000	0.045*

Table 3: Significant Global Slope Coefficients (Outcome = Steps), MICE

Variable	**Estimate**	**2.5% CI**	**97.5% CI**	**P-value**
`mood_lag1`	0.017	0.008	0.026	0.000***
`steps_lag1`	0.039	0.028	0.050	0.000***
`steps_lag2`	0.041	0.030	0.051	0.000***
`(Intercept)`	0.017	0.006	0.027	0.002**
`mood_lag2`	0.012	0.003	0.021	0.011*
`intervention_sleep_lag1`	-0.065	-0.124	-0.007	0.029*
`intervention_mood_lag1`	-0.064	-0.124	-0.003	0.038*
`steps_intervention_mood_moderator_lag2`	-0.035	-0.070	-0.001	0.046*
`intervention_steps_lag1`	-0.062	-0.132	0.008	0.083

Outcome: Mood. For mood outcomes, the linear interpolation model (Table 4) shows that yesterday's mood (`mood_lag1`, estimate = 0.522) is a strong predictor of today's mood. Significant negative moderation effects are observed for both lag1 (slope = -0.114) and lag2 (slope = -0.051), indicating that the benefit of a mood intervention diminishes as prior mood increases (see left panel of Figure 3).

MICE results (Table 5) reveal a similar significant negative moderation effect for lag1 (slope = -0.093, $p < 0.001$), confirming the consistency of the moderation effects across the two imputation methods. The right panel of Figure 3 illustrates that under MICE, when

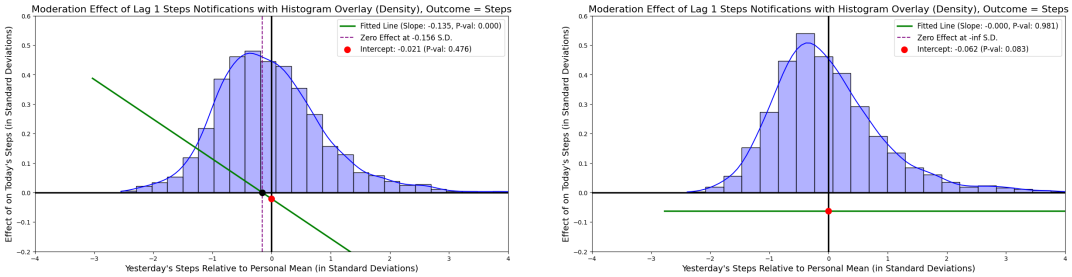


Figure 2: Moderation Effect of Lag1 Step Intervention, Outcome=Step.
Left: interpolation, Right: MICE
(Duplicate of enlarged panels already shown in Figure 1)

yesterday's mood is exactly at the individual's average (0 SD), the net intervention effect is 0.048 SD. As yesterday's mood increases, the effect decreases by 0.093 SD for each additional SD. Notably, the higher y-intercept under MICE suggests that, at baseline, mood interventions have a stronger positive impact compared to the interpolation model.

Table 4: Significant Global Slope Coefficients (Outcome = Mood), Interpolation

Variable	**Estimate**	**2.5% CI**	**97.5% CI**	**P-value**
`mood_lag1`	0.522	0.502	0.541	0.000***
`mood_lag2`	0.038	0.024	0.051	0.000***
`mood_intervention_mood_moderator_lag2`	-0.051	-0.078	-0.025	0.000***
`mood_intervention_mood_moderator_lag1`	-0.114	-0.141	-0.088	0.000***
`sleep_lag1`	0.013	0.004	0.021	0.003**
`day_intervention_interaction_steps_lag1`	-0.000	-0.001	-0.000	0.031*
`day_intervention_interaction_mood_lag1`	-0.000	-0.001	-0.000	0.037*

Table 5: Significant Global Slope Coefficients (Outcome = Mood), MICE

Variable	**Estimate**	**2.5% CI**	**97.5% CI**	**P-value**
`mood_lag1`	0.182	0.169	0.195	0.000***
`mood_lag2`	0.133	0.122	0.145	0.000***
`mood_intervention_mood_moderator_lag1`	-0.093	-0.124	-0.063	0.000***
`steps_lag1`	0.012	0.004	0.021	0.005**
`sleep_lag1`	0.011	0.002	0.019	0.013*
`day_intervention_interaction_mood_lag1`	-0.000	-0.001	-0.000	0.019*
`day_intervention_interaction_steps_lag1`	-0.000	-0.001	-0.000	0.027*
`intervention_steps_lag1`	0.054	-0.009	0.116	0.091.

Outcome: Sleep. Interpolation (Table 6) identifies yesterday's sleep as a dominant predictor (0.494, $p < 0.001$), and mood and step lags are also significant predictors. Significant moderation effects of sleep interventions (lag1: -0.171; lag2: -0.070, both $p < 0.001$) the intervention is more effective when participants had fewer minutes asleep the previous day.

Under MICE (Table 7), yesterday's steps (-0.033, $p < 0.001$), two-day lagged sleep (0.019, $p < 0.001$), and two-day lagged mood (0.016, $p = 0.001$) significantly predict today's sleep. Additionally, intervention effects on steps (-0.108, $p = 0.001$) and mood (-0.091, $p < 0.01$)

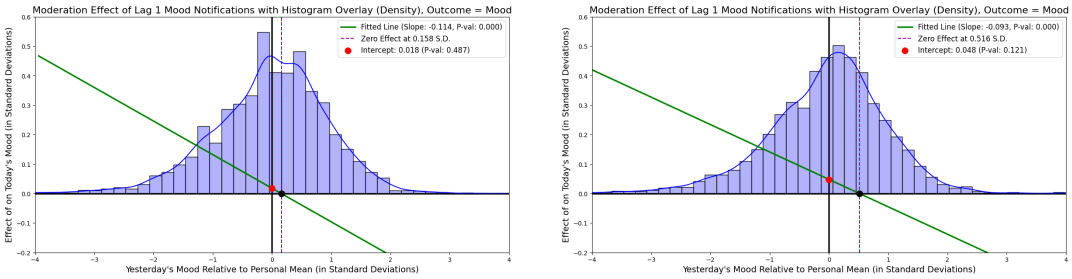


Figure 3: Moderation Effect of Lag1 Mood Intervention, Outcome=Mood.
Left: interpolation, Right: MICE
(Duplicate of enlarged panels already shown in Figure 1)

are significant, suggesting that step and mood interventions reduce sleep duration. However, moderation effects for sleep interventions are not significant under MICE.

The moderation plots (Figure 4) clearly contrast these findings, with significant moderation under interpolation and negligible effects under MICE.

Table 6: Significant Global Slope Coefficients (Outcome = Sleep), Interpolation

Variable	**Estimate**	**2.5% CI**	**97.5% CI**	**P-value**
`mood_lag1`	0.022	0.013	0.031	0.000***
`sleep_lag1`	0.494	0.473	0.516	0.000***
`steps_lag2`	0.026	0.017	0.035	0.000***
`steps_lag1`	-0.057	-0.065	-0.048	0.000***
`sleep_intervention_sleep_moderator_lag2`	-0.070	-0.096	-0.044	0.000***
`sleep_intervention_sleep_moderator_lag1`	-0.171	-0.197	-0.145	0.000***
`sleep_intervention_steps_moderator_lag2`	-0.029	-0.054	-0.004	0.022*
`sleep_intervention_steps_moderator_lag1`	-0.027	-0.052	-0.002	0.036*

Table 7: Significant Global Slope Coefficients (Outcome = Sleep), MICE

Variable	**Estimate**	**2.5% CI**	**97.5% CI**	**P-value**
`sleep_lag2`	0.019	0.010	0.028	0.000***
`steps_lag1`	-0.033	-0.042	-0.024	0.000***
`mood_lag2`	0.016	0.007	0.025	0.001***
`intervention_steps_lag1`	-0.108	-0.173	-0.043	0.001**
`intervention_mood_lag2`	-0.091	-0.152	-0.031	0.003**
`day_intervention_interaction_steps_lag1`	0.001	0.000	0.001	0.003**
`sleep_lag1`	0.016	0.005	0.026	0.004**
`sleep_intervention_mood_moderator_lag2`	-0.033	-0.063	-0.003	0.029*
`mood_lag1`	0.010	0.001	0.019	0.030*
`day_intervention_interaction_mood_lag2`	0.000	0.000	0.001	0.037*
`day_intervention_interaction_sleep_lag2`	0.000	-0.000	0.001	0.068.
`intervention_sleep_lag1`	-0.059	-0.122	0.005	0.072.
`intervention_sleep_lag2`	-0.054	-0.114	0.005	0.075.

Overall Comparison:

Comparing the results of linear interpolation and MICE, we observe three key differences:

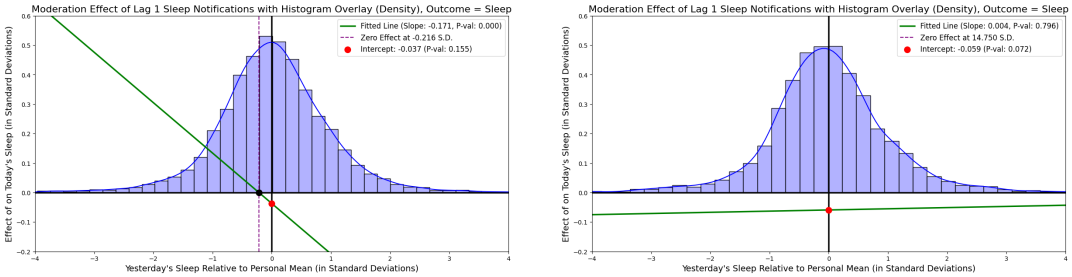


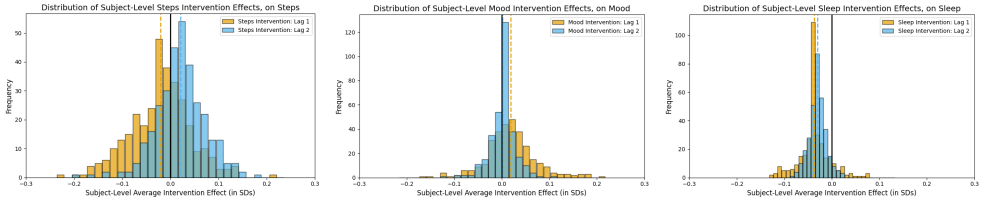
Figure 4: Moderation Effect of Lag1 Sleep Intervention, Outcome=Sleep.
Left: interpolation, Right: MICE
(Duplicate of enlarged panels already shown in Figure 1)

First, **moderation effects.** For steps and sleep outcomes, the linear interpolation model shows significant negative moderation effects, while the MICE model shows little to no moderation for these outcomes. For mood outcomes, however, both imputation methods reveal significant negative moderation effects.

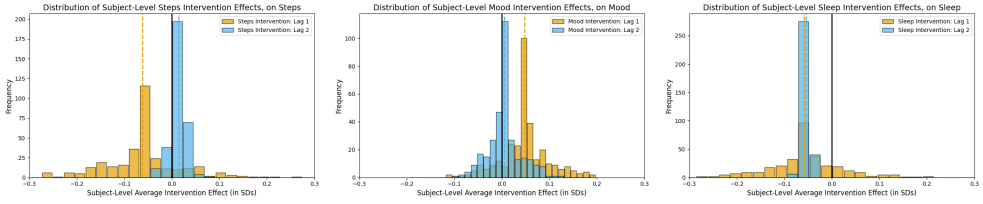
Second, **average intervention effects.** The net effect (y-intercept) of yesterday's intervention (Lag 1 intervention) on today's outcomes (Lag 0 outcome) is more pronounced when using MICE compared to linear interpolation. Specifically, using MICE, mood exhibits a more positive baseline effect, while steps and sleep show more negative baseline effects.

Figure 5 illustrates these differences and also compares the impact of interventions from two days ago (Lag 2) on today's outcomes. Panels (a)–(c) show the distributions of subject-level average intervention effects for steps, mood, and sleep under linear interpolation, and panels (d)–(f) present the corresponding MICE results. In each plot, the yellow histogram represents Lag 1 intervention effects, the blue histogram represents Lag 2 intervention effects, and the dashed vertical lines mark their corresponding mean intervention effect.

This comparison reveals that under MICE (bottom row), the Lag 1 distributions shift further away from zero relative to those under linear interpolation (top row). This shift is especially noticeable for mood (see panels (b) vs. (e)), where the average Lag 1 effect is more positive. Additionally, Lag 1 intervention effects (yesterday, yellow) typically show greater variability compared to Lag 2 (two days ago, blue), indicating that more recent interventions produce more varied immediate outcomes.



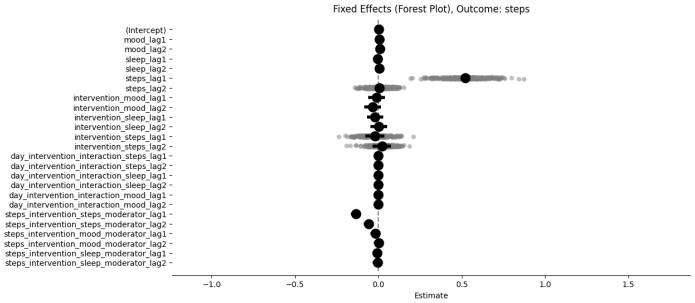
(a) Steps (Interpolation) (b) Mood (Interpolation) (c) Sleep (Interpolation)



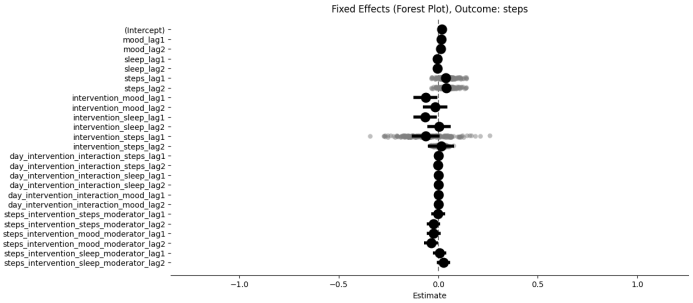
(d) Steps (MICE) (e) Mood (MICE) (f) Sleep (MICE)

Figure 5: Distribution of subject-level intervention effects on target outcome (x in SDs; y = subjects). Top: interpolation; bottom: MICE. All histograms share a common x-axis scale.

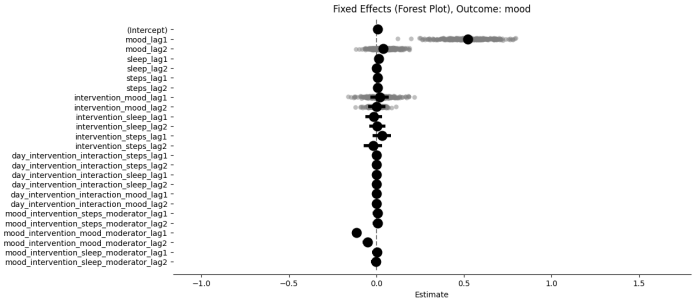
Third, **predictor estimates.** Linear interpolation tends to identify outcome lag predictors as statistically significant with substantially larger effect sizes than other covariates, likely reflecting bias from the interpolation process that amplifies the influence of recent observations. In contrast, MICE produces smaller, more balanced effect estimates and identifies a different set of significant predictors. These differences are visualized in the forest plots in Figure 6.



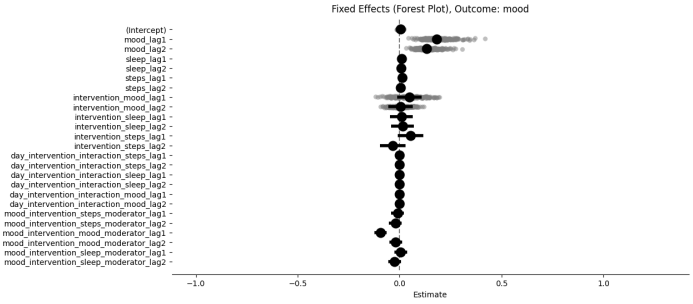
(a) Steps (Interpolation)



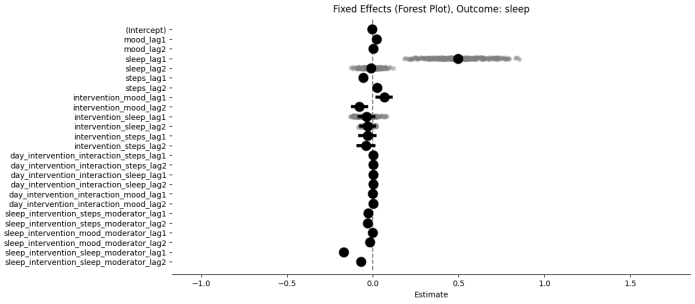
(b) Steps (MICE)



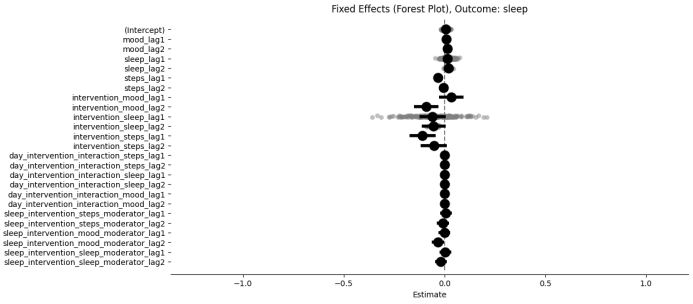
(c) Mood (Interpolation)



(d) Mood (MICE)



(e) Sleep (Interpolation)



(f) Sleep (MICE)

Figure 6: Forest plots of predictor estimates. Left: linear interpolation; right: MICE. The vertical dashed line marks the null effect (zero).

Together, these findings highlight that the choice of imputation method greatly influences effect estimates and their subsequent interpretations. Determining the most reliable imputation strategy is therefore crucial. This motivates us to compare the predictive accuracy of models trained on datasets imputed using different methods, to determine which approach best preserves the underlying data structure.

3.2 Imputation Evaluation Findings

3.2.1 Predictive Accuracy Across Imputation Methods

Table 8 and Figure 7 summarize the mean squared error (MSE) values achieved by four approaches to handling missing data: Complete Data (CCA), Linear Interpolation, Mean Imputation, and MICE Imputation. For each imputation method, a predictive model was trained to forecast three outcomes (steps, mood, and sleep) on the test set. All outcomes were standardized within each subject, yielding a variance of 1 and a naive baseline MSE of 1.

Table 8: Prediction Experiment: MSE Comparison for Different Imputation Methods

Outcome	Complete Data	Linear Interpolation	Mean Imputation	MICE Imputation
Steps	0.960	1.001	1.634	0.814
Mood	0.847	0.840	1.536	0.655
Sleep	0.998	1.085	1.672	0.856

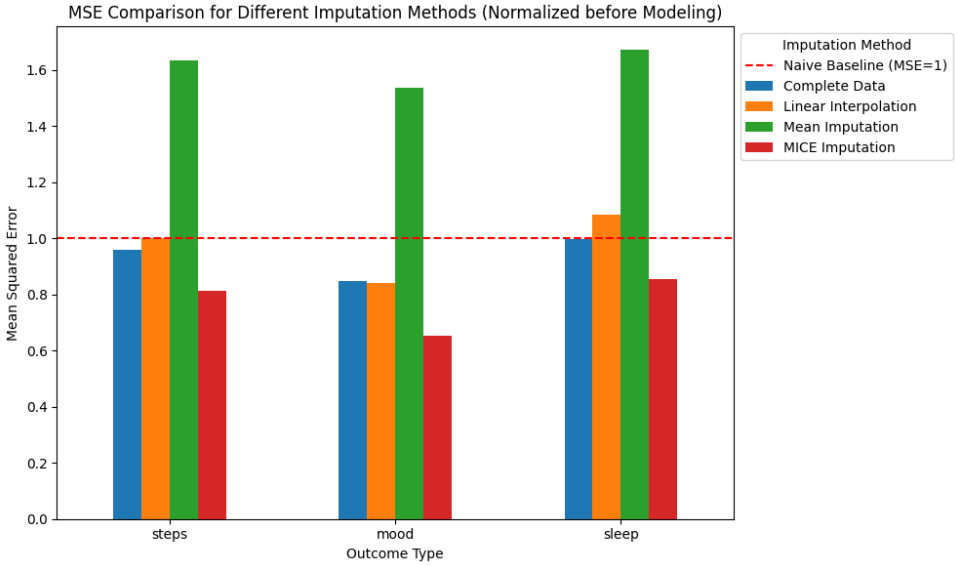


Figure 7: MSE values for each outcome under different imputation methods. The horizontal dashed line at MSE=1 represents the naive baseline.

Interpretation. Because an MSE of 1 corresponds to always predicting the subject mean (i.e., zero after normalization), an MSE below 1 indicates that the model captures additional predictive information beyond this naive baseline. Conversely, an MSE of 1 or above suggests performance comparable to or worse than guessing the mean.

From Table 8, **MICE Imputation** outperforms the naive baseline across all outcomes, yielding MSE values below 1 (0.814 for steps, 0.655 for mood, and 0.856 for sleep). This indicates that the MICE-imputed models consistently capture meaningful data signals, leading to *better* predictive accuracy than the naive model that predicts the mean. Notably, MICE also yields lower MSE than the **Complete Data** (CCA) approach, indicating that the data loss from CCA may impair predictive performance more than the uncertainty introduced by MICE imputation.

Linear Interpolation produces MSE values near 1 for steps and sleep (1.001 and 1.085, respectively), indicating performance that is at or just below the baseline, but it does

outperform the naive baseline for mood (0.840). By contrast, **Mean Imputation** yields MSE values above 1 for all outcomes (1.634 for steps, 1.536 for mood, and 1.672 for sleep), indicating that prediction models based on mean-imputed data perform *worse* than simply predicting the mean.

In summary, **MICE Imputation** provides the most accurate predictions, consistently producing MSE values below 1 and outperforming both the naive baseline and the other imputation strategies. Therefore, for various versions of the result interpretations we discussed above, we recommend adopting the MICE interpretation described in Section 3.1.2.

3.2.2 Distributional Comparisons

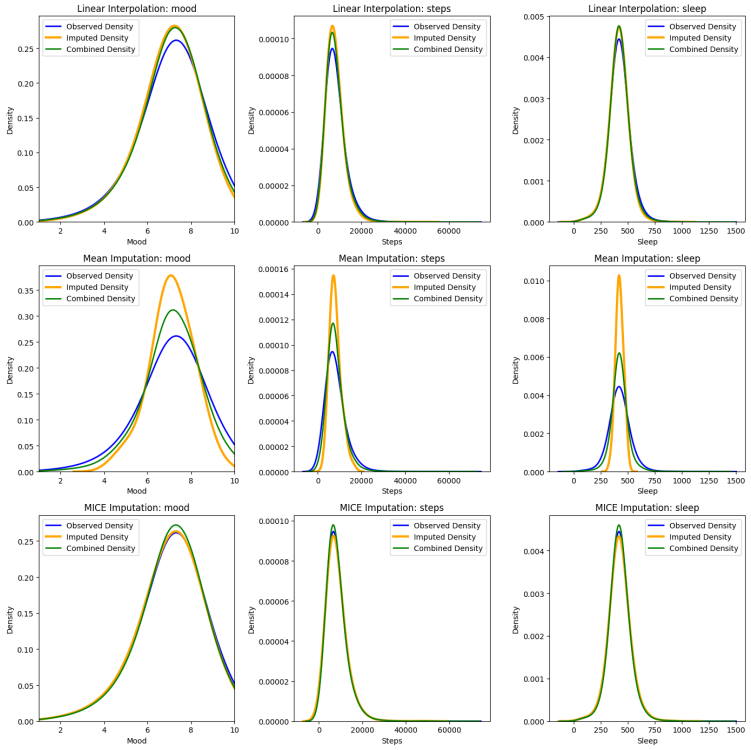


Figure 8: Smoothed density curves of observed, imputed, and combined values by variable and imputation method, before normalizing. data completeness threshold = 70%

Appendix C presents smoothed density curves for the observed data, the imputed data, and their combined distribution, for each imputation method (rows) and outcome variable (columns), prior to within-subject normalization, under 3 different data completeness thresholds (70%, 50%, 30%). Our goal is to assess how well each imputation strategy replicates the distribution of the observed data. Note that perfect overlap is not expected, given that the missing data pattern is likely non-MCAR.

For example, for the sensor dataset retained at the 70% data completeness threshold (Figure 8), the top row (Linear Interpolation) shows imputed densities that are generally higher than the observed distributions, suggesting that this method may oversmooth the data, hence reducing variability (spread). The middle row (Mean Imputation) produces narrow density peaks centered at the means, indicating a larger loss of variability. In contrast, the bottom row (MICE imputation) displays imputed densities that closely match the shape and spread of the observed data for all three outcomes. Thus, MICE appears to be the most effective method for preserving the original distributional properties of the IHS dataset.

4 Discussion

4.1 Key Findings on Intervention Effects

We discuss results based on MICE-imputed data (for its lowest MSE) with a 70% data completeness threshold and next-day zeroing-out strategy to derive practical insights for designing personalized, adaptive mHealth interventions that help manage mental health under stress.

First, average intervention effects (y-intercepts) show that mood notifications improve mood outcomes by about 0.048 SD above individual means. Conversely, sleep notifications reduce sleep by around 0.059 SD on average, suggesting that direct reminders to "get better sleep" may unintentionally increase stress, making falling asleep even harder. Instead of directly pointing out sleep deficits, indirect strategies like encouraging relaxation exercises might be a more effective approach.

Second, moderation effect slopes indicate how intervention effectiveness varies with prior outcomes. Notably, mood notifications are most beneficial when the previous day's mood is below 0.516 SD (x-intecept) above the individual average, improving by 0.093 SD (slope) per additional SD decrease from the individual mean. Above this threshold, notifications may become counterproductive. This emphasizes tailoring notifications to individual circumstances, prioritizing mood support during periods of below-average mood, and avoiding interventions during positive emotional states.

Finally, we found no adaptive moderation effects for targeted step and sleep interventions. However, step outcomes are influenced by previous-day mood and step counts. Thus, integrating emotional context into activity-focused interventions may enhance their effectiveness.

4.2 Imputation's Impact on Results

4.2.1 Methodological Implications

Across all three outcomes, MICE produced the lowest out-of-sample prediction error: about 20% lower than CCA and linear interpolation, and more than 50% lower than mean imputation. As a widely adopted method in longitudinal health research, MICE better preserves temporal patterns and the relationships between interventions and outcomes for the IHS data than simpler imputation approaches.

Many mobile health studies either drop incomplete cases via complete-case analysis or default to simplistic methods like mean imputation, implicitly assuming a missing-data mechanism that may not hold. Our comparison of four methods shows that imputation choice can substantially alter estimated intervention effects and introduce bias. To avoid these pitfalls, we ranked methods by their out-of-sample predictive performance rather than relying on unverifiable missing-data assumptions. We recommend that future longitudinal studies adopt this predictive validation framework in place of arbitrary imputation practices, thus enhancing the reproducibility and credibility of findings.

4.2.2 Comparison with Prior Work

Experimental Design and Intervention Effect Estimates

Compared to the prior 2018 IHS analysis by NeCamp et al. (2020), our study extends their work by introducing several methodological refinements.

First, unlike the 2018 IHS, which only used global coefficients, our hierarchical modeling frameworks incorporated subject-level random effects. This approach captures individual variability more effectively, providing personalized estimates of intervention effects. Tailoring intervention strategies to individual needs is an important step toward the study's goal of developing personalized interventions.

Second, the two studies also differ in their randomization schemes in the MRT design. In the 2018 IHS, participants received the same type of intervention each day for a full week, with equal likelihood of being assigned to mood, activity, sleep, or no notification. Effects were then analyzed using the weekly average of outcomes from the following week. In contrast, the 2022–2023 IHS used a more randomized design where participants had a 50% chance of receiving a notification each day, with equal likelihood for step, mood, or sleep interventions. This design captures more shorter-term effects and interactions between different outcomes and intervention types, offering a more nuanced understanding of their impacts.

Missing Data Handling and Sensitivity Analysis

Our results align with the sensitivity analyses from the 2018 IHS study by NeCamp et al. (2020), which emphasized how missing data handling methods can substantially influence moderation effects. The earlier work compared multiple imputation (MICE) with two complete case analysis (CCA) strategies: (1) *dropout sensitivity*, which excluded data after participants' dropout dates, and (2) *weekly missingness sensitivity*, which removed weeks with more than five days of missing data. Both CCA approaches produced notably smaller moderation effects than MICE. For example, activity moderation effects dropped from a significant -0.039 (p=0.013) under MICE to -0.003 (p=0.874) under *dropout CCA* and 0.004 (p=0.858) under *weekly missingness CCA*, highlighting the impact of methodological choices.

Expanding on this, our study introduces a predictive evaluation framework that assesses imputation methods by their out-of-sample prediction accuracy. This empirical approach sidesteps unverifiable missing-data assumptions (MCAR, MAR, MNAR) and offers a more objective basis for choosing imputation strategies.

Interestingly, while NeCamp et al. (2020) reported significant negative moderation effects across all three outcomes using MICE, our study did not identify such effects for steps and sleep under the same method. This difference likely arises from variations in experimental design and analysis. In the 2018 IHS, participants received the same type of notification every day for a full week, potentially reinforcing the intervention and increasing the likelihood of detecting effects. They also analyzed moderation effects using weekly averages of daily data, which helps identify stable trends but can obscure important daily variations. In contrast, the 2022–2023 IHS used daily randomization, allowing us to capture short-term changes and individual responses more precisely through hierarchical modeling. These design differences likely contributed to the more conservative moderation estimates in our study.

4.3 Limitations and Future Directions

While MICE achieved the best imputation performance in our study, surpassing CCA in predictive accuracy, it has limitations when applied to high-dimensional, nonlinear, and longitudinal data. First, MICE imputes missing data iteratively, one variable at a time, which makes it difficult to capture complex relationships. Second, its reliance on predictive mean matching and logistic regression can limit its ability to capture nonlinear effects (Dong et al., 2021). Third, it assumes that data are missing at random (MAR). If missingness depends on unobserved factors or evolves over time, this assumption is violated and imputations may be biased. In fact, no imputation method can fully correct for missing not at random (MNAR) scenarios, since the underlying mechanism cannot be directly validated. Given the heterogeneous, time-series nature of our wearable data, we are motivated to explore more advanced imputation approaches.

Emerging machine learning approaches show potential to address these gaps. For instance, Generative Adversarial Imputation Nets (GAIN) employs an adversarial framework to learn nonlinear relationships without requiring complete cases (Yoon et al., 2018). Its hint mechanism helps distinguish observed from imputed values, enabling it to handle mixed data types and high dimensionality (Goodfellow et al., 2014). However, GAIN faces challenges in

computational efficiency and training stability, and it lacks built-in mechanisms to handle time-series correlations.

To better handle time-dependent sensor data, the individualized dynamic latent factor model by Zhang et al. (2023) offers a more tailored solution. These models project multi-resolution time series data into a shared latent space with subject-specific trajectories, capturing individual heterogeneity while borrowing strength across variables. Their early results show improvements over traditional methods for irregular wearable data. Thus, we anticipate that applying this model to the IHS dataset will improve the imputation of irregular sensor measurements and more reliable intervention effect estimates.

In future work, we will compare these advanced deep learning and latent factor approaches against MICE using our predictive evaluation framework. This will allow us to examine if these methods better capture nonlinear and temporal patterns in wearable data. We also plan to generate a synthetic IHS dataset and apply our models to both real and synthetic data. By comparing parameter estimates across datasets, we aim to evaluate the fidelity of synthetic data generation. If results are consistent, this would not only validate our imputation methods but also support privacy preservation in future studies.

Our study highlights the importance of adaptive interventions to support medical interns' mental health. Through wearable devices, these solutions have the potential to provide accessible, personalized mental health care by dynamically adjusting interventions to individual needs. Although our study focuses on medical interns, the same principles and methods could benefit anyone lacking timely and accessible mental health support. Ultimately, by integrating wearable devices and data modeling, we hope our work paves the way for a more inclusive, proactive, and efficient mental health care system.

5 Acknowledgements

First of all, I'm grateful to my mentors, Luke Francisco and Professor Ambuj Tewari.

Luke has been an incredible mentor since January 2024, guiding me through two projects—first on predictive modeling in the PROMPT Precision Health Study, and now on causal inference in the Intern Health Study (IHS). He introduced me to statistical research and encouraged me to dive into questions I was curious about. After I expressed interest in causal inference and intervention design, he supported my transition to the IHS project and gave me the freedom to explore new methods. From him, I learned how to break down complex problems and approach research with curiosity and creativity.

I am also thankful to Professor Ambuj Tewari for his steady support and for offering me this valuable research opportunity, which gave me the confidence to pursue my long-term passion for applying data science to mental health research.

My Undergraduate Research Journey

My sresearch journey in Statistics began with PROMPT (PROviding Mental Health Precision Treatment), where we tackled an urgent challenge: how to support psychiatric patients facing months-long waits for clinical care? During these delays, patients' mental health can deteriorate significantly. Using wearable device data, I have built machine learning models—including logistic regression, random forests, SVM, and neural networks—to predict symptoms of depression (PHQ-9), anxiety (GAD-7), and suicidal ideation. These models aimed to detect high-risk patients early, helping clinics prioritize care more effectively.

The prediction models only solved part of the issue. They relied on manual clinical response, which limited speed and scalability. After spending a summer at CVS Health working on time-series causal inference models, I became interested in a deeper question: when is the right moment to intervene? That internship also helped me see a big challenge in traditional healthcare systems: they often can't respond in real time due to logistical constraints, like the coordination needed to reallocate resources. In mental health care, such delays can be fatal.

That's when I started thinking about wearable devices in a new way. Their ability to *detect* mental health crises means they *understand* users' real-time conditions; if they could further *act on* this knowledge to deliver personalized interventions, care could be provided instantly and at no cost. This idea motivated me to shift to this IHS project, which gave me access to real-world intervention data. I'm grateful that my advisors supported this transition.

My Insights

Across both projects, I have identified two major challenges in today's mental health care:

1. **Long appointment wait times:** Caused by a mismatch between demand and supply, these delays increase costs, limit access, and can worsen symptoms and raise suicide risk.

2. **Ineffective mobile interventions if generic**: The same message could have different effects depending on a person's current state.

This strengthened my belief in the promise of an intelligent intervention system—one that adapts to sensor data in real time and provides personalized support. Such a system could directly address both challenges. I strongly believe that integrating mobile health technology with data science has the potential to transform the future of mental health care.

Meanwhile, I observed that even well-timed messages can lose their effectiveness if they feel robotic or impersonal. This sparked my curiosity about integrating warmth and empathy into digital interventions through large language models (LLMs). This semester, I joined a project developing a chatbot that identifies negative attributions in individuals experiencing depression and reframes them with compassion, much like a supportive therapist would.

These experiences helped me discover a path I'm deeply passionate about. I'm thankful to have found it early in my undergraduate journey. I sincerely thank the Department of Statistics at the University of Michigan for organizing the Undergraduate Research Program in Statistics (URPS). Lastly, to my lovely family—thank you for supporting me all the time!

References

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274.

Dong, W., Fong, D. Y. T., Yoon, J.-s., Wan, E. Y. F., Bedford, L. E., Tang, E. H. M., and Lam, C. L. K. (2021). Generative adversarial networks for imputing missing data for big data clinical research. *BMC Medical Research Methodology*, 21(1).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Jolly, E. (2018). Pymer4: Connecting r and python for linear mixed modeling. *Journal of Open Source Software*, 3(31):862.

NeCamp, T., Sen, S., Frank, E., Walton, M. A., Ionides, E. L., Fang, Y., Tewari, A., and Wu, Z. (2020). Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: Micro-randomized trial. *Journal of Medical Internet Research*, 22(3):e15033.

Salvatier, J., Wiecki, T., and Fonnesbeck, C. (2015). Probabilistic programming in python using pymc.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Tennant, C. (2001). Work-related stress and depressive disorders. *Journal of Psychosomatic Research*, 51(5):697–704.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

White, I. R. and Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28):2920–2931.

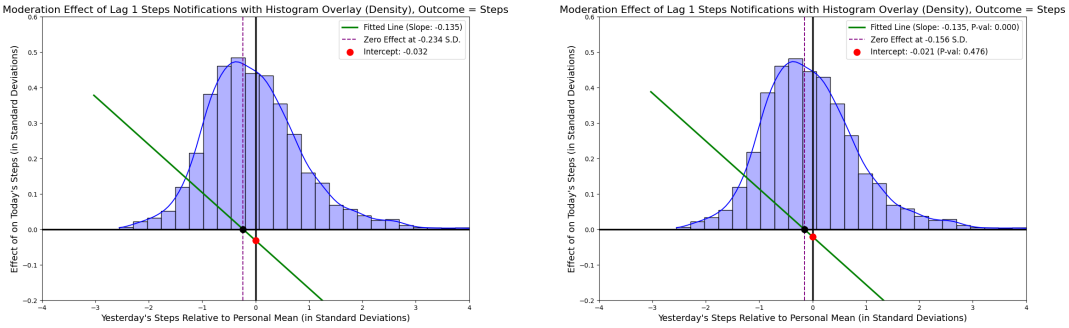
Yoon, J., Jordon, J., and van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5689–5698. PMLR.

Zhang, J., Xue, F., Xu, Q., Lee, J.-A., and Qu, A. (2023). Individualized dynamic latent factor model for multi-resolutional data with application to mobile health.

Appendix

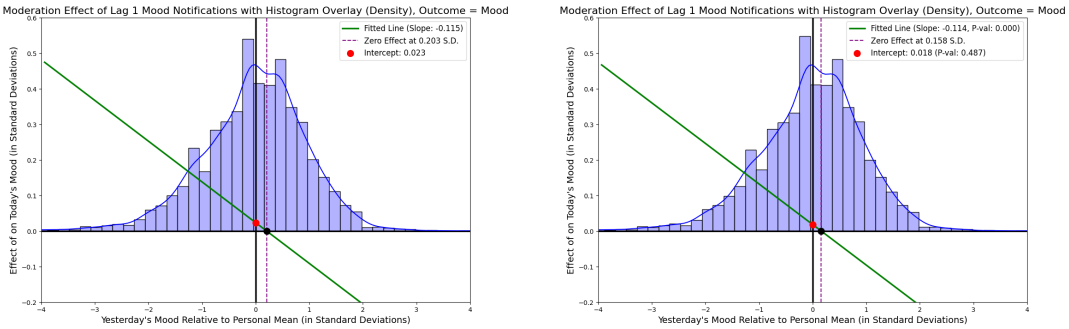
A Moderation Effects Comparison: `PyMC` vs. `Pymer4`

These panels show examples of moderation effect plots at a 70% data-completeness threshold with next-day zeroing-out, imputed using interpolation. `PyMC` (Bayesian MCMC) and `Pymer4` (frequentist REML) yield nearly identical moderation effects estimates (slopes).



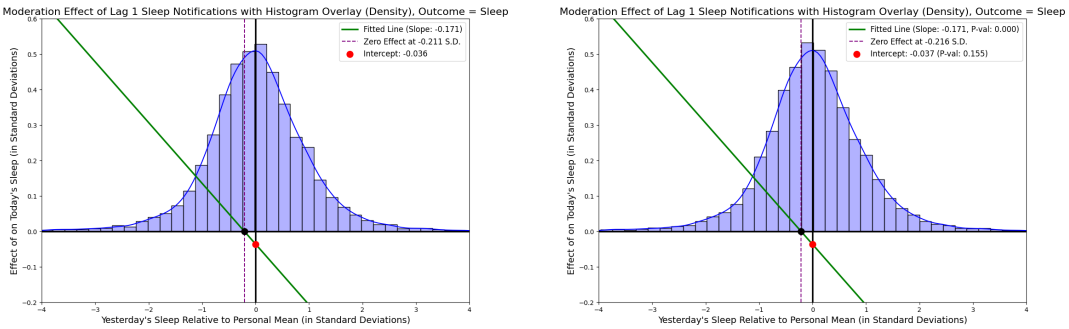
(a) Lag1 Step Intervention on Steps (`PyMC`) (b) Lag1 Step Intervention on Steps (`Pymer4`)

Figure 9: Moderation effects of the Lag1 step intervention across `PyMC` vs. `Pymer4`.



(a) Lag1 Mood Intervention on Mood (`PyMC`) (b) Lag1 Mood Intervention on Mood (`Pymer4`)

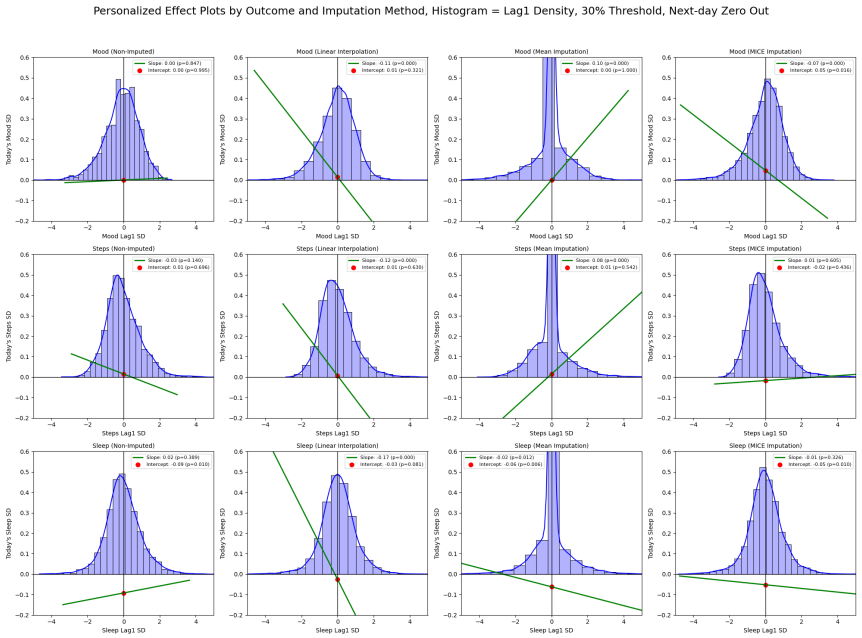
Figure 10: Moderation effects of the Lag-1 mood intervention across `PyMC` vs. `Pymer4`.



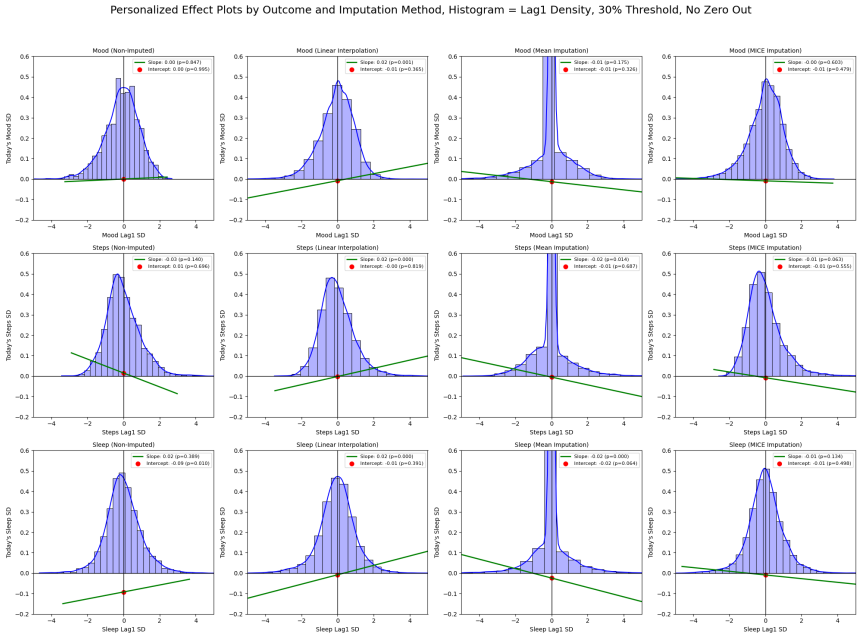
(a) Lag1 Sleep Intervention on Sleep (`PyMC`) (b) Lag1 Sleep Intervention on Sleep (`Pymer4`)

Figure 11: Moderation effects of the Lag-1 sleep intervention across `PyMC` vs. `Pymer4`.

B Supplementary Composite Graphs at 30% Data Completeness Threshold



(a) Next-day Zeroing-Out, 30% Threshold



(b) No Zeroing-Out, 30% Threshold

Figure 12: Composite graphs comparing intervention effects under two zeroing-out strategies at the **30% data completeness threshold**. These graphs complement the 70% results shown in the main text.

C Distribution Comparisons of Different Imputation Methods at Different data completeness thresholds

The three figures below compare the observed, imputed, and combined distributions of mood, steps, and sleep under three imputation methods: linear interpolation (top rows), mean imputation (middle rows), and MICE (bottom rows). Each figure corresponds to a different data completeness threshold: 70%, 50%, and 30%, respectively.

Across all thresholds, linear interpolation produces imputed densities that are slightly elevated compared to the observed distributions, suggesting reduced variability and potential oversmoothing. Mean imputation yields sharply peaked distributions centered at the mean, indicating a substantial loss of variance. In contrast, MICE produces imputed distributions that closely align with the observed data in both shape and spread across all outcomes.

As the data completeness threshold decreases (e.g., 30%), the alignment between observed and imputed distributions weakens. However, MICE maintains better alignment than the other methods, demonstrating a greater ability to recover original distributional characteristics even under substantial missingness.

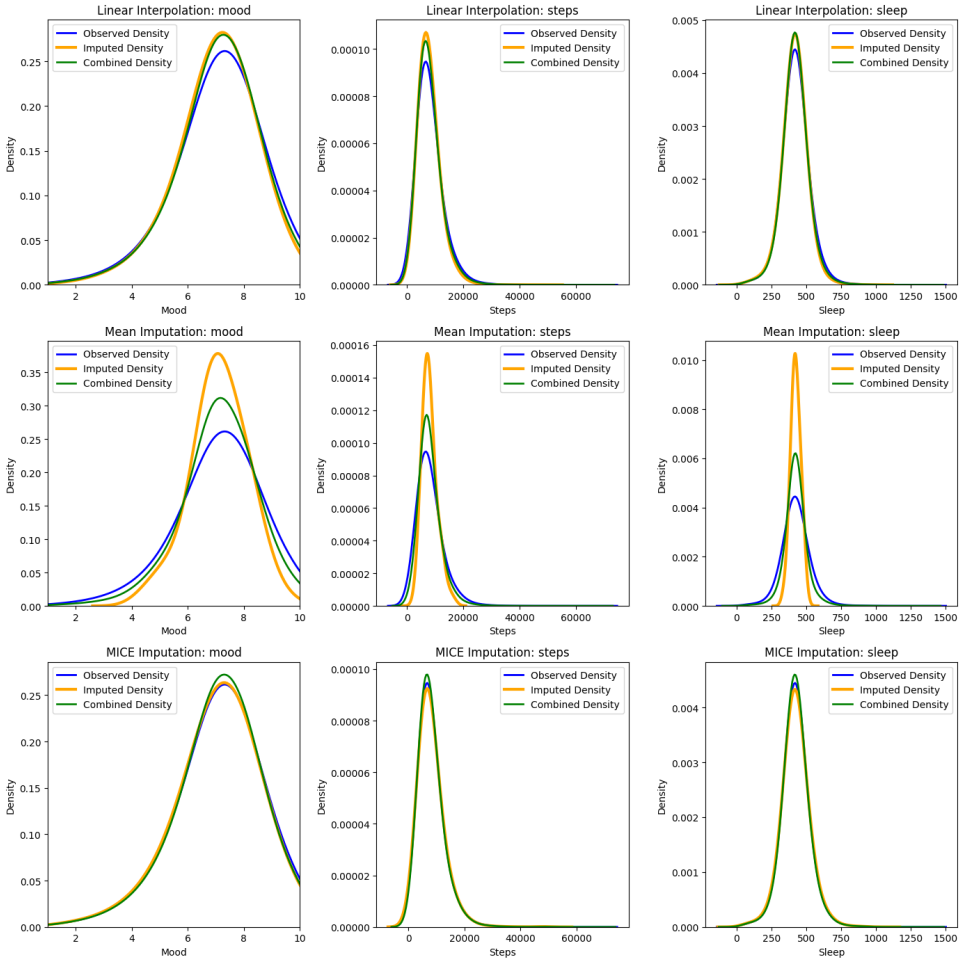


Figure 13: Observed, imputed, and combined data distributions for all outcomes under linear interpolation, mean imputation, and MICE at the **70% data completeness threshold**

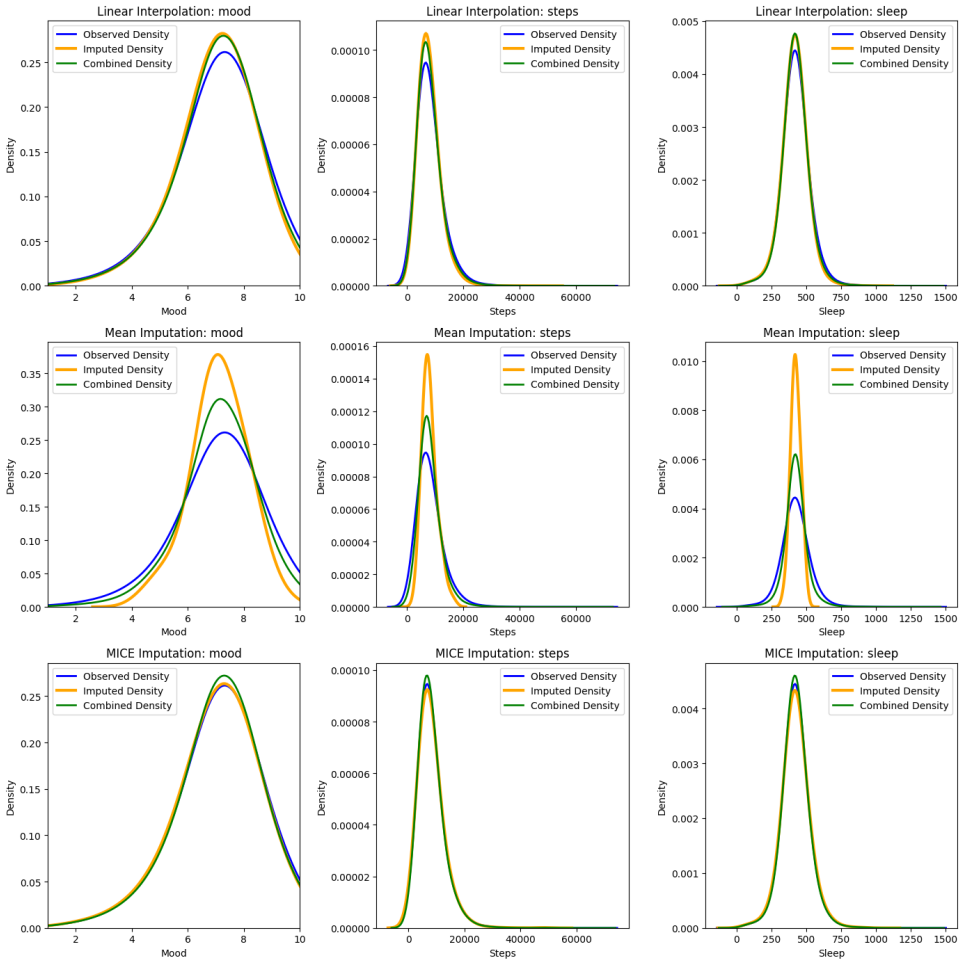


Figure 14: Observed, imputed, and combined data distributions for all outcomes under linear interpolation, mean imputation, and MICE at the **50% data completeness threshold**

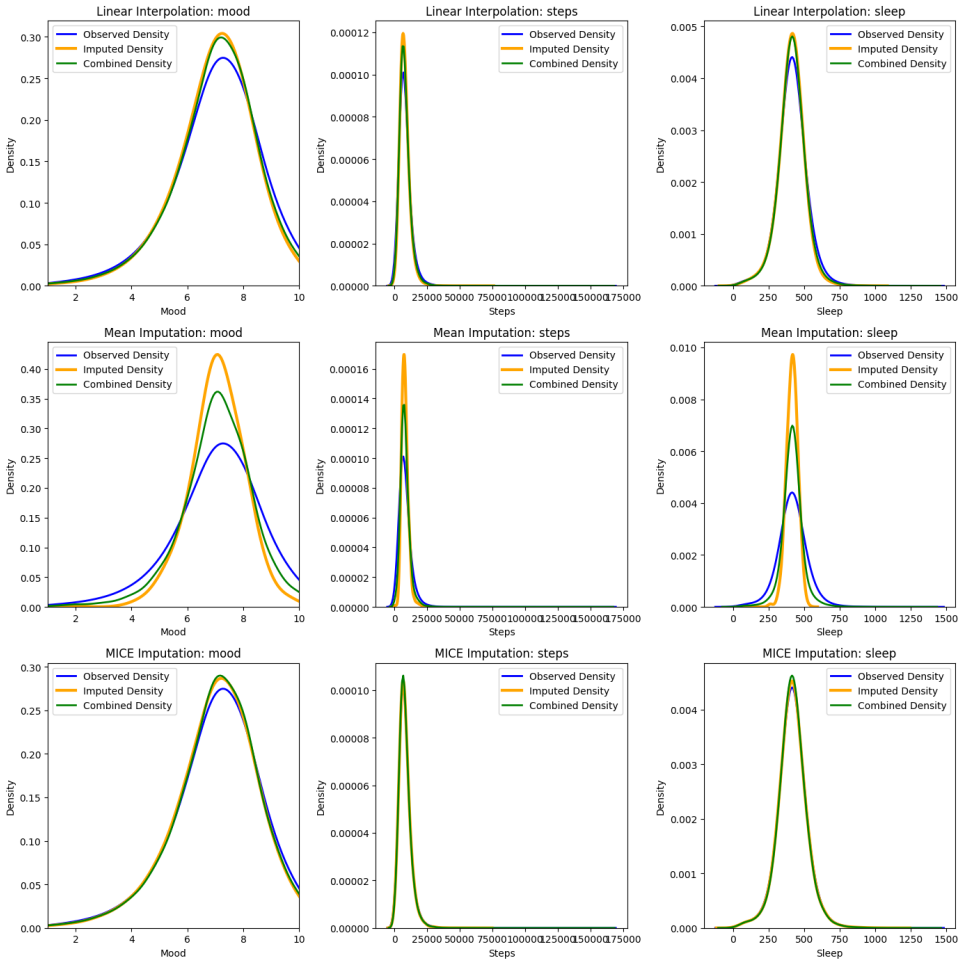


Figure 15: Observed, imputed, and combined data distributions for all outcomes under linear interpolation, mean imputation, and MICE at the **30% data completeness threshold**