"I attest that upon completion of this exam, I have destroyed the data file and I acknowledge that these data should not be used for any other purposes." 6768

Mobile Health Study

Author: 6768

University of California, Irvine

June 25, 2025

Abstract: This study evaluates the effectiveness of motivational messaging on physical activity using data from a two-week randomized controlled trial. Participants were randomly assigned to one of three groups: Control (no messages), Standard (generic daily messages), or Tailored (personalized daily messages). Daily step counts, recorded via wearable devices, served as the primary outcome. To address four scientific questions, we applied a range of statistical methods including linear regression, generalized estimating equations (GEE), predictive modeling with multiple imputation, and linear mixed-effects models.

Our findings indicate that both Standard and Tailored interventions increased average daily step counts during the first week, with the greatest gains observed among participants who were somewhat active at baseline. However, these improvements diminished in the second week, suggesting waning intervention effects. Predictive models demonstrated high accuracy for second-week step counts, though only a few days significantly contributed to predictions. Finally, substantial individual-level heterogeneity in step trajectories was observed. Results also showed the magnitude of this variability differed across intervention arms.

Overall, while motivational messaging—especially for already active individuals—can elicit short-term increases in activity, sustaining these gains may require more adaptive or personalized intervention strategies.

1 Introduction

Physical inactivity remains a major public health concern, contributing to the burden of chronic conditions such as cardiovascular disease, type 2 diabetes, and cancer. Mobile health (mHealth) technologies—including wearables and smartphone applications—offer scalable solutions to promote physical activity by providing continuous behavioral monitoring and real-time feedback. Among these, motivational messaging has shown potential to increase physical activity, with personalized (tailored) messages theorized to outperform generic messages due to their greater relevance and specificity.

To evaluate the effectiveness of motivational messaging, a team of behavioral scientists conducted a two-week randomized controlled trial comparing three intervention groups: (1) a Control group receiving no messages, (2) a Standard group receiving daily generic motivational messages, and (3) a Tailored group receiving personalized daily messages based on each participant's recent behavior or baseline profile. Daily step count, recorded via wearable devices, served as the primary outcome.

This report addresses four core scientific questions posed by the researchers:

1. How do the three intervention groups compare in average daily step count during the first week? How does this comparison vary by baseline activity level?
2. How do the three intervention groups compare in the 14-day trajectories of daily step count?
3. Using each participant's baseline covariates (baseline activity level, age, sex), intervention group, and their daily step counts on Day 1, Day 2, ..., and Day 7, how well can you predict their average daily step count averaged over Days 8–14?

4. To what extent do individuals differ in their daily step count over the course of the 14-day study? Does the magnitude of this individual-level variability differ between the three arms?

The following sections outline the methods used to analyze the data, the results obtained, and the implications for personalized mHealth interventions in promoting physical activity.

2 Methods

All the statistical analysis provided were conducted using R version 5.1 with packages ggplot2, dplyr, lme4, geepack, naniar, mice, and lmerTest. Overleaf was used for editing purposes.

2.1 Data Description

The `stepcount.csv` dataset contains longitudinal physical activity data from 300 participants enrolled in a two-week randomized trial. Each row represents a unique participant-day observation, with up to 14 rows per participant. The dataset includes the following variables:

- `id`: Unique participant identifier.
- `Group`: Randomized intervention arm (Control, Standard, Tailored), indicating whether participants received no messages, generic messages, or personalized messages based on recent behavior or baseline profile.
- `Act`: Baseline activity level (not active, somewhat active, very active).
- `Age`: Age in years (continuous).
- `Female`: Binary sex indicator (1 = female, 0 = male).
- `day`: Study day (1 to 14).
- `dow`: Day of the week (e.g., Mon, Tue), included to assess weekday/weekend patterns.
- `stepcount`: Daily step count measured by a wearable device; the primary outcome.

2.2 Statistical Methods

All models were evaluated for Normality and homoscedasticity with residual–versus–fitted and QQ plots (See Appendix) and were deemed acceptable.

2.2.1 Preliminary Data Exploration

Table 1 summarizes baseline characteristics of participants, stratified by intervention arms. It presents continuous variables as means (standard deviations) and categorical variables as counts (percentages). This table allows us to examine if the groups are comparable at baseline. The spaghetti plot in 1 is used to visualize mean trajectories of step counts over time, allowing us to observe patterns and variability across groups and obtain an overall sense of data.

Baseline covariates—age, sex, and self-reported activity level—were extracted for each participant based on their Day 1 record. These variables were used to assess potential confounding and effect modification. Sex was coded as a binary factor (0 = male, 1 = female) and activity level was treated as an ordered factor with three levels: not active, somewhat active, and very active.

We created a summary table using the `CreateTableOne()` function from the `tableone` R package to compare baseline characteristics and Week 1 step count averages across intervention groups. This table stratifies by treatment arm and presents both categorical and continuous covariates, allowing for inspection of group balance.

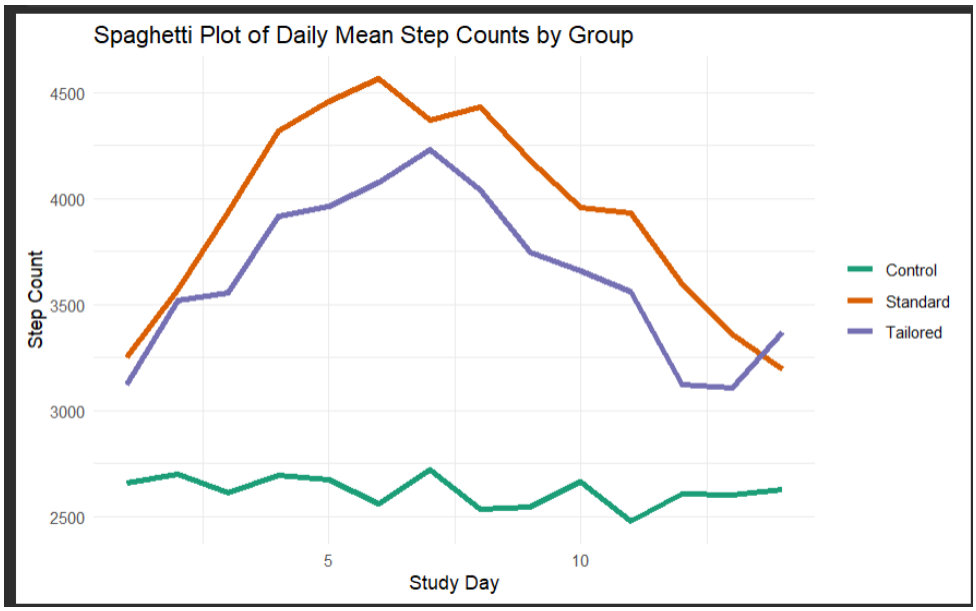


Figure 1: Spaghetti plot of daily step counts over time. Each line represents a group's trajectory

Table 1: Baseline Characteristics Stratified by Intervention Group

	Control	**Standard**	**Tailored**
n	100	95	99
Age (mean (SD))	28.39 (5.72)	27.47 (5.66)	27.84 (5.55)
Female (%)			
Male	60 (60.0)	58 (61.1)	59 (59.6)
Female	40 (40.0)	37 (38.9)	40 (40.4)
Mean Step Count (mean (SD))	2566.75 (1628.42)	3769.35 (2087.98)	3376.28 (2276.15)

2.2.2 Missing Data

Most participants had recordings on each day of the trial, however a majority recorded at least one zero step count day. These 0 step count days likely reflect non-wear or non-compliance rather than true inactivity. This constitutes the reason why we chose to treat these observations as missing rather than 0. There were also six participants recorded with missing gender. Since these participants were only 2% of the total study, their observations were removed to maintain simplicity.

To assess the nature of missingness in the data, we conducted a test for Missing Completely at Random (MCAR) using the `mcar_test()` function from the `naniar` package in R. The resulting p-value was extremely small (1.9×10^{-27}), providing strong evidence against the MCAR assumption. This suggests that the missing data are more likely to be Missing at Random (MAR) or Not Missing at Random (NMAR).

Given this result—and the fact that linear mixed-effects models (LMMs) remain valid under the MAR assumption—we proceeded with the analysis under the assumption that the data are MAR. For most scientific questions, the impact of missing data was mitigated through the use of robust modeling techniques. For Scientific Question 3, where missingness was more substantial, we employed multiple imputation to avoid excessive data loss and preserve predictive accuracy. Removing all missing observations would have substantially reduced the sample size and statistical power, so imputation was a necessary step to maintain the integrity of the analysis.

2.2.3 SQ1

To evaluate differences in physical activity across the three intervention groups during the first week of the study, we focused on the average daily step count for Days 1 through 7. First, we calculated each participant's mean step count during this period. To avoid artificially deflating averages due to potential non-wear days, we treated step counts equal to zero as missing and excluded them from the calculation of individual means. Two subjects had no non-zero recordings and were removed from this analysis completely.

To assess whether the effects of the intervention differed by baseline activity level, we fit a linear regression model using the `lm()` function in R, with the average Week 1 step count as the outcome. The model included main effects for treatment group and baseline activity level, as well as their interaction, along with age and sex as covariates:

$$\text{mean\_steps}_i = \beta_0 + \boldsymbol{\beta}_G^\top \cdot \text{Group}_i + \boldsymbol{\beta}_A^\top \cdot \text{Act}_i + \boldsymbol{\beta}_{GA}^\top \cdot (\text{Group}_i \times \text{Act}_i) + \beta_8 \cdot \text{Age}_i + \beta_9 \cdot \text{Female}_i + \epsilon_i \quad (1)$$

- Group_i is a vector of indicators for Standard and Tailored.
- $f(\text{day}_t)$ includes both day_t and the spline term $(\text{day}_t - 7)_+$.
- Act_i is a vector of indicators for "somewhat active" and "very active."
- The $\boldsymbol{\beta}_X$ terms are coefficient vectors for grouped effects.
- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

This model allows for formal hypothesis testing of group differences and effect modification by baseline activity level. Standard errors and p-values were used to evaluate statistical significance. The interaction term tests whether baseline activity modifies group differences in average week-1 steps.

A two sample t-test was performed to examine whether the mean steps during the first week of the control group differs from that of the intervention groups. The test effectively examines whether the means of two independent groups differ significantly from each other. To quantify the added value of the Group × Act interaction, we compared the full model with a reduced additive model (`Group+Act+Age+Female`) using a likelihood-ratio F-test.

2.2.4 SQ2

To evaluate differences in daily step count trajectories across the three intervention groups over the 14-day study period, we fit a longitudinal regression model using Generalized Estimating Equations (GEE). A linear spline was included into the model to capture the nonlinear trend observed among non-control subjects (Figure 1), where steps count rise sharply during the first week, then drop sharply in the second.. A knot was placed at day 7 to reflect this pattern. Daily step count was modeled as a function of intervention group, age, time, baseline activity level, and sex, including all two- and three-way interactions between group, time, and baseline activity. The outcome model is specified in equation 2.

GEE was used for this analysis because the data have repeated measurements of step counts over time for each subject, and GEE can account for within-subject correlation. One advantage of GEE is its robustness: as long as the mean structure is correctly specified, estimates of both parameters and their standard errors remain valid even if the working correlation structure is misspecified.

Within an individual, we assume measurements that are closer to one another in time are more correlated. As such errors were assumed to follow an AR1 structure across time points within subjects. The model was fit using the `geeglm()` function from the **geepack** package in `R`. A Wald-type F-test using the `glht()` function from the **multcomp** package tested the joint null hypothesis that the intervention effects on step count slopes were equal across the two post-randomization time periods against the alternative that at least one of these coefficients differ.

$$H_0 : \beta_9 = \beta_{10} = \beta_{11} = \beta_{12}$$

Using the same notation from 1:

$$\begin{aligned}\text{steps}_{it} = \beta_0 &+ \boldsymbol{\beta}_G^\top \cdot \text{Group}_i + \boldsymbol{\beta}_T^\top \cdot f(\text{day}_t) + \boldsymbol{\beta}_A^\top \cdot \text{Act}_i + \beta_7 \cdot \text{Female}_i + \beta_8 \cdot \text{Age}_i \\ &+ \boldsymbol{\beta}_{GT}^\top \cdot (\text{Group}_i \times f(\text{day}_t)) + \boldsymbol{\beta}_{GA}^\top \cdot (\text{Group}_i \times \text{Act}_i) \\ &+ \boldsymbol{\beta}_{TA}^\top \cdot (\text{Act}_i \times f(\text{day}_t)) + \boldsymbol{\beta}_{GTA}^\top \cdot (\text{Group}_i \times \text{Act}_i \times f(\text{day}_t)) + \epsilon_{it}\end{aligned} \tag{2}$$

Where $(\text{day}_t - 7)_+ = max(\text{day}_t - 7, 0)$.

2.2.5 SQ3

To evaluate how well baseline activity patterns predict subsequent behavior, we constructed models to predict each participant's average daily step count during the second week of the study (Days 8–14) using baseline covariates and daily step counts from the first week (Days 1–7). The analysis accounted for missingness due to non-wear periods by treating zero step count values as missing and applying multiple imputation.

We first reshaped the data to wide format, with separate columns for each of the 14 days, then performed multiple imputation using the `mice` package in `R` with predictive mean matching (PMM) to generate 40 imputed datasets. Convergence diagnostics showed stable means and variances across imputations. For downstream analysis, I combined the results from all imputations using Rubin's rules to create more reliable estimates by properly accounting for the uncertainty from missing data From the completed dataset, we calculated two sets of predictors:

- Baseline covariates: treatment group assignment, baseline activity level, age, and sex.
- Daily step counts from Days 1 through 7 (`wk1_step_1` to `wk1_step_7`).

The outcome variable was each participant's mean daily step count over Days 8–14.

We trained a linear regression models using a 70/30 train-validation split. The equation is displayed using the same notation as model 1.

$$\text{week2\_mean}_i = \beta_0 + \boldsymbol{\beta}_G^\top \cdot \text{Group}_i + \boldsymbol{\beta}_A^\top \cdot \text{Act}_i + \beta_5 \cdot \text{Age}_i + \beta_6 \cdot \text{Female}_i + \sum_{d=1}^{7} \gamma_d \cdot \text{wk1\_step}_{id} + \epsilon_i \tag{3}$$

Model fitting was conducted using the `lm()` function in `R`. Predictive performance was evaluated on the held-out validation set using root mean squared error (RMSE) and R^2. For the linear model, diagnostics were performed similarly to SQ1.

2.2.6 SQ4

To evaluate the extent of individual-level variability in physical activity and whether this variability differs across intervention groups, we fit linear mixed-effects models to participants' daily step counts using the `lmer()` function from the **lme4** package in `R`. The model specification allowed for random effects at the participant level and included fixed effects for intervention group, time, baseline activity level, and sex.

To model temporal trends, we included a linear spline at Day 7, decomposing the time effect into two parts. This allowed us to estimate separate slopes for Days 1–7 and Days 8–14. The full model incorporated a random intercept and participant-specific random slopes for both weeks:

$$\begin{aligned}\text{steps}_{it} = \beta_0 &+ \boldsymbol{\beta}_G^\top \cdot \text{Group}_i + \boldsymbol{\beta}_T^\top \cdot f(\text{day}_t) + \boldsymbol{\beta}_A^\top \cdot \text{Act}_i + \beta_7 \cdot \text{Female}_i + \beta_8 \cdot \text{Age}_i \\ &+ \boldsymbol{\beta}_{GT}^\top \cdot (\text{Group}_i \times f(\text{day}_t)) + \boldsymbol{\beta}_{GA}^\top \cdot (\text{Group}_i \times \text{Act}_i) \\ &+ \boldsymbol{\beta}_{TA}^\top \cdot (\text{Act}_i \times f(\text{day}_t)) + \boldsymbol{\beta}_{GTA}^\top \cdot (\text{Group}_i \times \text{Act}_i \times f(\text{day}_t)) \\ &+ b_{0i} + \boldsymbol{b}_1^\top \cdot f(\text{day}_t + \epsilon_{it}\end{aligned} \tag{4}$$

where b_{0i} is the subject-specific random effects on the intercept,$\boldsymbol{b}_1^\top$ denotes the random slope vector for week 1 and week 2, and ϵ_{it} denotes residual error.

To assess whether individual-level variability in daily step counts differed by intervention group, we fit separate linear mixed-effects models for each of the three groups (Control, Standard, and Tailored). For each model, we included fixed effects for study day (using linear spline terms for Day 1 and Day 2), baseline activity level, and sex. We also included random intercepts and random slopes for Day 1 and Day 2 at the participant level to account for subject-specific deviations in baseline activity and time trends. After fitting the models separately by group, we extracted the standard deviations of the random effects to compare the magnitude of variability in baseline step counts and trajectories across participants within each group. This allowed us to evaluate whether the tailored or standard interventions induced greater heterogeneity in individual responses, as measured by the variance components of the random intercepts and slopes.

3 Results

3.1 SQ1

Participants in the Standard and Tailored groups walked significantly ($p < 0.05$ for both β_1 and β_2) more than those in the Control group, with estimated increases of 486.5 (95% CI: 84.8, 888.1) and 514.3 (95% CI: 128.9, 899.7) steps per day, respectively. The likelihood ratio test results in Table 3 showed that the interaction between intervention arm and baseline activity was significant ($p < 0.05$). The effect of the intervention varied by baseline activity level. For instance, somewhat active participants in the Standard group had an additional increase of 1,723.3 steps/day (95% CI: 1,067.4, 2,379.2) beyond the main effects. Similar patterns were observed for the Tailored group.

Among participants classified as somewhat active at baseline, both the Standard and Tailored interventions produced significantly greater increases in average daily steps during the first week compared to the Control group. Specifically, participants in the Standard group who were somewhat active walked an additional 1,723 steps per day (95% CI: 1,067, 2,379) beyond the main effects, while those in the Tailored group showed an additional increase of 1,306 steps per day (95% CI: 656, 1,957). These findings suggest that individuals who were already somewhat active were particularly responsive to both types of motivational messages, with the Standard group showing the largest interaction effect in this subgroup.

For participants classified as very active at baseline, only those in the Tailored group exhibited a statistically significant additional benefit. Very active individuals in the Tailored group walked an estimated 1,096 additional steps per day (95% CI: 311, 1,880) beyond the main effects. In contrast, the interaction effect for very active individuals in the Standard group was smaller and not statistically significant (estimated increase: 442 steps/day; 95% CI: –298, 1,183). Table 2 shows the coefficients calculated for the linear regression.

Table 2: Linear Regression Coefficients for Model 1 with 95% Confidence Intervals

Term	**Estimate**	**Pr(>|t|)**	**2.5% CI**	**97.5% CI**
(Intercept)	4058.13	< 2e–16	3385.89	4730.37
GroupStandard	486.48	0.01778	84.85	888.11
GroupTailored	514.30	0.00909	128.93	899.66
Actsomewhat_active	1593.68	4.15e–11	1137.00	2050.36
Actvery_active	3557.46	< 2e–16	3022.44	4092.49
Age	-80.14	5.11e–13	-100.96	-59.33
Female	-568.62	4.65e–06	-808.24	-329.01
GroupStandard:Actsomewhat_active	1723.33	4.41e–07	1067.45	2379.21
GroupTailored:Actsomewhat_active	1306.15	9.77e–05	655.76	1956.55
GroupStandard:Actvery_active	442.46	0.24062	-298.20	1183.12
GroupTailored:Actvery_active	1095.52	0.00635	311.23	1879.81

Table 3: Likelihood Ratio Test Comparing Additive vs. Interaction Model

Model	Res.Df	RSS	Df	Sum Sq	F	Pr(>F)
Additive (H_0)	285	312241656				
Interaction (H_1)	281	277940023	4	34301633	8.67	1.30×10^{-6}

3.2 SQ2

Participants in both the Standard and Tailored intervention groups exhibited significantly increasing step count trends during the first week compared to Control:

- **Standard group:** $\beta_9 = 138.80$, $p < 0.001$
- **Tailored group:** $\beta_{10} = 117.07$, $p < 0.001$

However, these early gains were not sustained into the second week. Both intervention groups experienced significant declines in step counts during Days 8–14:

- **Standard group:** $\beta_{11} = -334.49$, $p < 0.001$
- **Tailored group:** $\beta_{12} = -243.00$, $p < 0.001$

A Wald test comparing these group-specific trends confirmed that the time trajectories differed significantly across groups ($F(4, 3893) = 38.0$, $p < 0.001$). Altogether during the first week, these results suggest that the rate of physical activity increase differs between intervention groups. The results from the second week show that rate of return to baseline physical activity differs by intervention group. A complete display of all model coefficients is shown in table 5.

3.3 SQ3

The Linear Regression Model (3) showed strong performance in predicting second-week activity, with an adjusted $R^2 = 0.886$ and a root mean squared error (RMSE) of approximately 792 steps on the validation set.

Several covariates were significantly associated with Week 2 mean step count. Compared to the Control group:

- Participants in the **Standard group** had lower predicted step counts ($\beta = -502$, $p = 0.001$),
- Participants in the **Tailored group** also had lower predicted step counts ($\beta = -659$, $p < 0.001$).

Participants who were *somewhat active* at baseline had lower Week 2 averages than those classified as *not active* ($\beta = -429$, $p = 0.024$). Neither Age ($p = 0.992$) nor sex ($p = 0.992$) were significant predictors. The interaction between the Tailored intervention group and the somewhat active category approached statistical significance ($p = 0.051$), suggesting a potential differential benefit of tailored messages for this subgroup. All other interaction terms, including those involving the Standard group and both somewhat active and very active categories, were not statistically significant ($p > 0.25$), indicating no clear evidence that the intervention effects differed by baseline activity for those subgroups.

Daily step counts from the first week were weak predictors of Week 2 outcomes. Only, step counts on Day 3 ($\beta = 0.222$, $p = 0.017$), Day 6 ($\beta = 0.194$, $p = 0.035$), and Day 7 ($\beta = 0.506$, $p < 0.001$) were significantly associated with higher Week 2 means. This outcome suggests that step counts on days later in the week are stronger predictors of a participant's week 2 performance.

A scatterplot (Figure 2) of observed versus predicted Week 2 step counts showed close alignment along the identity line, indicating good predictive calibration. These findings demonstrate that a simple linear model using baseline covariates and early step data can effectively forecast near-term physical activity. Coefficients for each covariate can be seen in Table 6

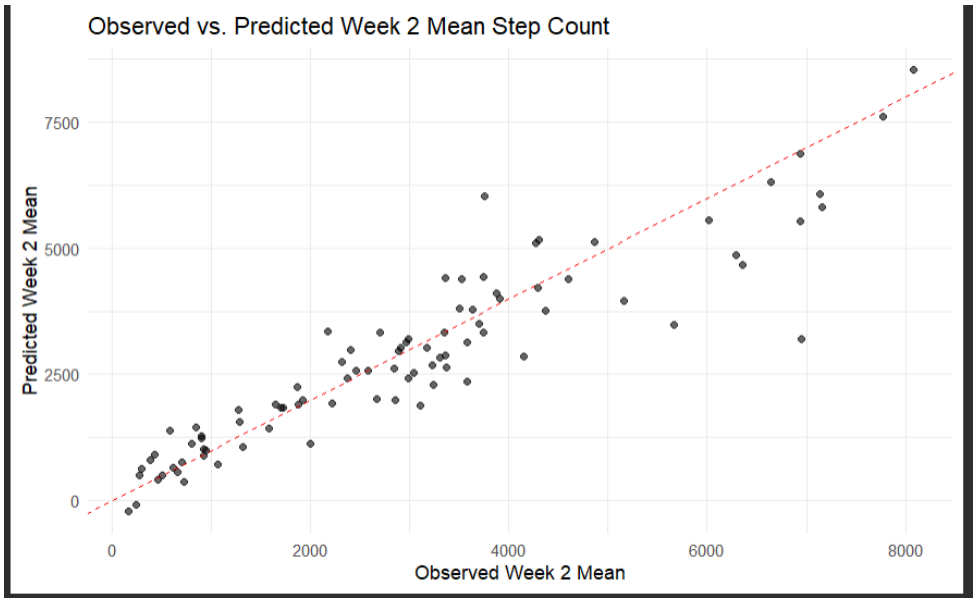


Figure 2: Model 3 Observed vs Predicted Week 2 Mean Step Count

3.4 SQ4

We first compared a random intercept model (Model A) to a full model 4 including random slopes for both week 1 and week 2 (Model B). The likelihood ratio test indicated strong evidence in favor of the more complex model ($\chi^2 = 897.5$, $p < 0.001$), suggesting substantial within-subject heterogeneity in time trends. An intermediate model including only a random slope for week 1 was also significantly outperformed by the full model ($\chi^2 = 69.8$, $p < 0.001$), indicating that variability in second-week trajectories (week 2) also differed across individuals.

For participants in the Control group, the estimated standard deviation of the random intercept was 982.4, indicating substantial heterogeneity in baseline step counts. Assuming normality, approximately 95% of individuals in this group are expected to have baseline step counts within the range $\beta_0 \pm 1.96 \times 982.4$. Similarly, the standard deviation of the random slope for the first week was estimated at 104.8, implying that 95% of individuals are expected to have daily changes in step count during this period between $\beta_1 \pm 1.96 \times 104.8$.

After Day 7, the time trend changes, and the variability in post-week 1 slopes reflects the combined standard deviations of both spline components. Under the same assumptions, 95% of individuals are expected to have daily changes in step count during the second week within the range $(\beta_1 + \beta_2) \pm 1.96 \times (104.8 + 129.6)$, where 129.6 is the estimated standard deviation of the Day 2 slope. These intervals illustrate the considerable subject-level variation in both baseline levels and activity trajectories over time.

The Tailored group exhibited greater individual-level variability in baseline step counts and changes over time, as reflected in the higher standard deviations of the random effects. This suggests that participants in the Tailored intervention responded more heterogeneously to the messaging, possibly due to the personalized nature of the intervention. In contrast, the Control group had the least variability, consistent with the absence of any messaging. The Standard group was intermediate between the two. Table 4 has the variability for each group displayed. Coefficients for the complete model (Model B) are show in Table 7.

Table 4: Standard deviations of random effects by intervention group from Model 4

Effect	**Control**	**Standard**	**Tailored**
Random Intercept (id)	915.9	940.6	1123.0
Random Slope (Day 1)	59.2	88.1	160.0
Random Slope (Day 2)	97.1	135.6	175.0
Residual	573.5	730.4	704.0

4 Discussion

4.1 Conclusion

This study aimed to evaluate the impact of motivational messaging—both standard and tailored—on physical activity over a two-week period. Using a randomized controlled design and daily step count data from wearable devices, we found that both interventions significantly increased average daily steps during the first week compared to control, with the largest gains observed among participants who were somewhat active at baseline. Tailored messages showed additional benefit for very active individuals, suggesting some subgroup-specific responsiveness.

However, we did not observe sustained improvements into the second week. Both intervention groups exhibited significant declines in step counts after Day 7, indicating that while messaging can prompt short-term increases, its effects may diminish over time—possibly due to habituation or message fatigue. Predictive modeling of Week 2 behavior revealed that only a subset of early step counts (notably Days 3, 6, and 7) were meaningfully associated with later activity, and that baseline covariates explained much of the variance. This points to a limited role for early engagement metrics in forecasting sustained behavioral change.

Lastly, while there was substantial individual heterogeneity in activity trajectories, linear mixed-effects models did not show systematic differences in variability across intervention arms. This suggests that messaging strategies affected mean behavior but not the degree of individual responsiveness.

In conclusion, motivational messaging—particularly for those already somewhat or very active—can effectively boost physical activity in the short term, but maintaining engagement likely requires more adaptive, personalized strategies. Future work should explore dynamic interventions that adjust to individual patterns and feedback to support sustained behavior change.

4.2 Limitations

First, the QQ and residual plots are displayed in the Appendix. Of note is the QQ Plot of Model 2. The rising right end of the graph could be an indicator of non-normality. Secondly, the choice of linear splines and the specific knot placement at day 7 was somewhat arbitrary; alternative modeling approaches, such as nonparametric models and other model selection techniques, might provide a better fit for the observed trends. Although the data was treated as Missing at Random (MAR) and used multiple imputation and the MCAR test result ($p < 1e\text{-}27$) strongly rejected MCAR, no formal test can distinguish MAR from Not Missing at Random (NMAR), so assuming MAR introduces uncertainty. If data were NMAR (e.g., people wore devices less on low-activity days), the estimates may be biased. It was assumed zero step count equals non-wear or non-compliance, but this wasn't verified. Some true zero-activity days may have been wrongly excluded, or low-but-valid activity days may have been misclassified.

5 Appendix

Table 5: GEE Coefficients for Model 2 with 95% Confidence Intervals

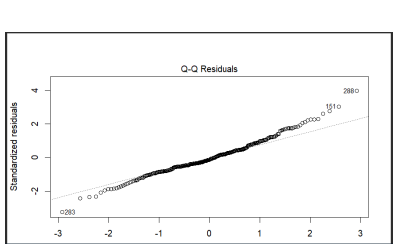
Term	**Estimate**	**Pr(>|W|)**	**2.5% CI**	**97.5% CI**
(Intercept)	3951.95	< 2e–16	3245.9	4658.0
GroupStandard	-66.19	0.688	-389.1	256.8
GroupTailored	61.09	0.707	-257.0	379.2
week1	-2.74	0.823	-26.7	21.2
week2	20.03	0.246	-13.8	53.8
Actsomewhat_active	1560.35	5.3e–12	1116.9	2003.8
Actvery_active	3707.74	< 2e–16	3116.2	4299.3
Female	-630.04	6.1e–07	-877.5	-382.5
Age	-75.61	2.9e–11	-97.9	-53.3
GroupStandard:week1	138.80	1.6e–07	86.9	190.7
GroupTailored:week1	117.07	1.5e–05	64.0	170.1
GroupStandard:week2	-334.49	< 2e–16	-399.6	-269.4
GroupTailored:week2	-243.00	2.7e–12	-311.1	-174.9
GroupStandard:Actsomewhat_active	839.85	0.020	130.4	1549.4
GroupTailored:Actsomewhat_active	591.75	0.108	-129.2	1312.7
GroupStandard:Actvery_active	349.99	0.417	-494.5	1194.5
GroupTailored:Actvery_active	1103.75	0.064	-65.6	2273.1
week1:Actsomewhat_active	8.07	0.775	-47.2	63.3
week1:Actvery_active	-33.87	0.515	-135.8	68.0
week2:Actsomewhat_active	-47.36	0.351	-146.9	52.2
week2:Actvery_active	-43.08	0.530	-177.6	91.4
GroupStandard:week1:Actsomewhat_active	212.07	8.2e–05	106.5	317.6
GroupTailored:week1:Actsomewhat_active	165.07	0.010	39.4	290.8
GroupStandard:week1:Actvery_active	39.56	0.613	-113.8	192.9
GroupTailored:week1:Actvery_active	5.60	0.958	-203.6	214.8
GroupStandard:week2:Actsomewhat_active	-385.51	5.0e–06	-551.1	-220.0
GroupTailored:week2:Actsomewhat_active	-386.50	3.2e–05	-568.6	-204.4
GroupStandard:week2:Actvery_active	2.73	0.982	-237.2	242.7
GroupTailored:week2:Actvery_active	-85.01	0.531	-350.8	180.8

Table 6: Linear Regression Coefficients for Model 3 with 95% Confidence Intervals

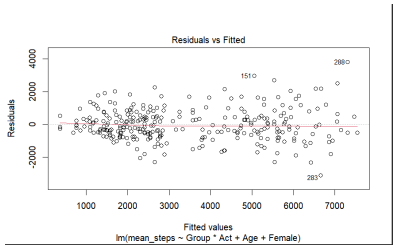
Term	**Estimate**	**Pr(>|t|)**	**2.5% CI**	**97.5% CI**
(Intercept)	111.29	0.78009	-674.00	896.49
GroupStandard	-530.06	0.00529	-901.00	-159.53
GroupTailored	-482.34	0.00898	-843.00	-121.98
Actsomewhat_active	-175.68	0.46493	-649.00	297.60
Actvery_active	-40.49	0.90415	-703.00	621.93
Age	-0.11	0.99227	-21.90	21.66
Female	-1.16	0.99188	-226.00	223.90
wk1_step_1	-0.09	0.24945	-0.23	0.06
wk1_step_2	-0.03	0.69279	-0.21	0.14
wk1_step_3	0.25	0.00736	0.07	0.44
wk1_step_4	-0.05	0.62724	-0.24	0.14
wk1_step_5	0.16	0.11027	-0.04	0.36
wk1_step_6	0.18	0.06253	-0.01	0.37
wk1_step_7	0.52	1.46×10^{-13}	0.39	0.65
GroupStandard:Actsomewhat_active	-371.87	0.25172	-1010.00	266.19
GroupTailored:Actsomewhat_active	-636.30	0.05057	-1270.00	1.61
GroupStandard:Actvery_active	351.60	0.31376	-335.00	1038.30
GroupTailored:Actvery_active	-187.71	0.60044	-893.00	518.07

Table 7: Fixed Effects Estimates with 95% Confidence Intervals for Model 4

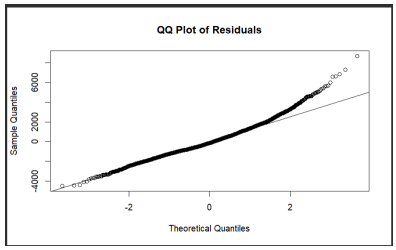
Effect	**Estimate**	**Pr(>|t|)**	**2.5%**	**97.5%**
(Intercept)	4034.832	< 2e-16	3379.31	4690.36
GroupStandard	-57.705	0.786	-476.69	361.28
GroupTailored	55.317	0.786	-343.63	454.26
week1	-1.944	0.926	-42.77	38.88
week2	17.547	0.567	-42.46	77.55
Actsomewhat_active	1592.557	1.20e-10	1119.16	2065.95
Actvery_active	3720.262	< 2e-16	3164.74	4275.78
Female	-496.349	3.16e-05	-726.46	-266.24
Age	-80.703	5.18e-14	-100.74	-60.67
GroupStandard:week1	140.142	8.22e-06	78.97	201.31
GroupTailored:week1	122.109	4.82e-05	63.63	180.59
GroupStandard:week2	-345.876	3.58e-13	-436.01	-255.74
GroupTailored:week2	-277.399	1.04e-09	-364.41	-190.39
GroupStandard:Actsomewhat_active	800.414	0.021	121.25	1479.58
GroupTailored:Actsomewhat_active	531.394	0.122	-141.72	1204.51
GroupStandard:Actvery_active	291.447	0.458	-479.29	1062.18
GroupTailored:Actvery_active	1118.905	0.007	307.72	1930.09
week1:Actsomewhat_active	6.221	0.859	-62.74	75.18
week1:Actvery_active	-35.965	0.384	-117.09	45.16
week2:Actsomewhat_active	-46.284	0.369	-147.10	54.53
week2:Actvery_active	-39.996	0.509	-158.89	78.90
GroupStandard:week1:Actsomewhat_active	213.801	2.69e-05	114.56	313.04
GroupTailored:week1:Actsomewhat_active	169.050	0.001	71.03	267.07
GroupStandard:week1:Actvery_active	42.030	0.466	-71.84	155.90
GroupTailored:week1:Actvery_active	1.245	0.984	-117.47	119.96
GroupStandard:week2:Actsomewhat_active	-373.775	7.74e-07	-520.24	-227.31
GroupTailored:week2:Actsomewhat_active	-385.235	2.36e-07	-529.32	-241.15
GroupStandard:week2:Actvery_active	11.395	0.893	-154.12	176.91
GroupTailored:week2:Actvery_active	-47.829	0.593	-223.37	127.71



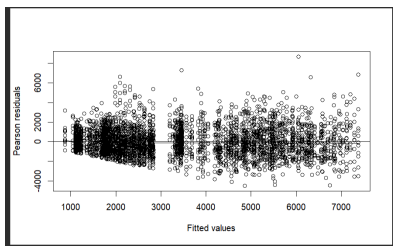
(a) Residuals vs. Fitted Model 1



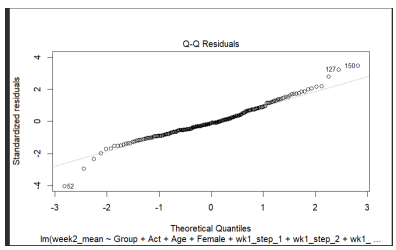
(b) QQ Plot Model 1



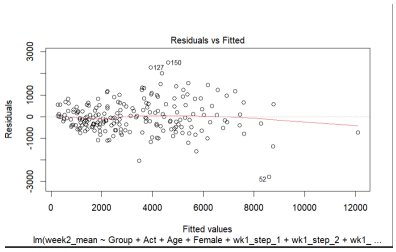
(c) Residuals vs. Fitted Model 2



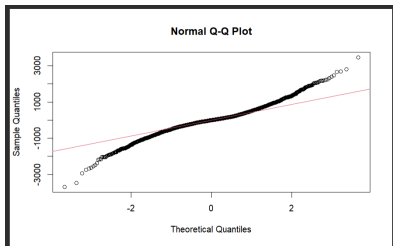
(d) QQ Plot Model 2



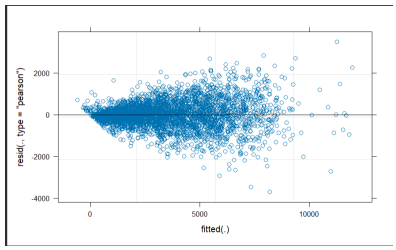
(e) Residuals vs. Fitted Model 3



(f) QQ Plot Model 3



(g) Residuals vs. Fitted Model 4



(h) QQ Plot Model 4

Figure 3: Residual and QQ plots for Models 1, 2, 3, and 4.