INTACT: A method for integration of longitudinal physical activity data from multiple sources

Jingru Zhang*[1], Erjia Cui[2], Hongzhe Li[3], and Haochang Shou[3]

[1]School of Data Science, Fudan University, Shanghai, 200433, China
[2]Division of Biostatisitcs and Health Data Science, University of Minnesota, Minneapolis, MN 55414
[3]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104

Abstract

Wearable devices and digital phenotyping are increasingly used in observational and interventional studies to assess physical activity. However, integrating and comparing data across studies and cohorts remains challenging due to variability in device types, acquisition protocols, and preprocessing methods. A key challenge is removing unwanted study- or device-specific effects while preserving meaningful biological signals. These difficulties are exacerbated by the longitudinal and within-day correlations inherent in high-resolution time series data collected from wearable sensors. To address this, we propose INTACT (INtegration of Time series data from weArable sensors for physiCal acTivity), a novel method for harmonizing physical activity intensity time series from accelerometers. INTACT models shared information through common eigenvalues and eigenfunctions while allowing for domain-specific scale and rotation adjustments. We demonstrate the effectiveness of INTACT on two cohorts from the National Health and Nutrition Examination Survey (NHANES), where physical activity measures were collected using different generations of accelerometers and processed into different units. Our results show that INTACT outperforms existing methods in mitigating domain effects while preserving biological signals, enabling more reliable cross-study comparisons of physical activity patterns.

Key words: Domain effects; Harmonization; NHANES data; Physical activity; Wearable device.

1 Introduction

The proliferation of novel wearable sensor devices has enabled continuous monitoring of various physiological and biological signals over extended periods. However, as the wearable technology landscape rapidly evolves—with new devices and frequent updates to existing algorithms (Harrison et al., 2015)—significant variability arises in how researchers select devices and acquire data for their studies, making cross-study data integration and comparison highly challenging. For example, Shah et al. (2017) reported substantial discrepancies in hourly step counts and daily heart rate summaries when participants wore two consumer-grade devices simultaneously. While they proposed linear models to adjust data from one device to another, such approaches fail to address more complex data structures, such as minute-level physical activity intensity measures collected from different studies without matched sample covariates.

*jr_zhang@fudan.edu.cn

Despite the growing reliance on wearable sensors, little attention has been given to the challenges posed by device heterogeneity, and existing methods for integrating and harmonizing device-specific data remain highly limited.

Many integration methods have been developed in other fields to address domain and batch effects—systematic differences between datasets arising from external factors unrelated to the variables of interest. In this paper, we use the term "domain" broadly to refer to a domain, site, device, or batch that contains homogeneous data. For example, Benito et al. (2004) applied distance-weighted discrimination to eliminate batch effects in microarray data, while Johnson et al. (2007) proposed an integration approach that removes domain-specific means and scales in linear models. Similar frameworks have been extended to multimodal neuroimaging to mitigate systematic differences due to scanner type, site, or acquisition protocols—such as longitudinal ComBat for multi-scanner imaging data (Beer et al., 2020) and CovBat for harmonizing covariance structures in multimodal imaging features (Chen et al., 2022). Additionally, methods for integrating single-cell data across multiple sites have been developed, including canonical correlation analysis (Butler et al., 2018), mutual nearest neighbors (Hie et al., 2019), and non-negative matrix factorization (Welch et al., 2019; Peng et al., 2021).

However, most existing methods are ill-suited for integrating physical activity data in two key ways: (1) they primarily focus on univariate and cross-sectional measures rather than time-dependent signals, and (2) they often treat each dimension separately, adjusting for differences in mean and variance without capturing disparities in the joint distribution of the data.

To address the inherent longitudinal and within-day correlations in physical activity data, we propose INtegration of Time-series data from weArable sensors for physiCal acTivity (INTACT), a framework designed to harmonize data across studies while accounting for domain-specific variations. INTACT addresses the challenge of repeated measures over multiple days by separately modeling between-subject and within-subject variations and applying domain effect corrections accordingly. It assumes a common eigenspace shared across datasets while allowing for domain-specific shifts, rotations, and scale differences. This shared eigenspace captures the true underlying eigen-directions and magnitudes that remain consistent across datasets, effectively distinguishing biological signals from unwanted variability introduced by differences in data acquisition and preprocessing. By leveraging these shared components, INTACT provides a robust integration framework that mitigates domain effects while preserving essential biological signals, enabling more reliable cross-study comparisons of physical activity patterns.

INTACT respects the inherent characteristics of epoch-level time series data from wearable devices, explicitly modeling repeated observations while preserving the original data distribution. By promoting an interpretable transformation, INTACT mitigates domain effects without compromising the integrity of biological signals. For efficient implementation, we derive an algorithm to estimate model parameters using an iterative eigendecomposition framework.

We evaluate the performance of INTACT by integrating physical activity data from the National Health and Nutrition Examination Survey (NHANES), collected using different devices across two study waves: Actigraph AM-7164 (NHANES 2003–2004) and Actigraph GT3X+ (NHANES 2013–2014). After applying our harmonization procedure, the two datasets exhibit greater consistency in activity intensity readings, reveal a more pronounced age effect on activity patterns, and show a stronger association between physical activity and mortality. Our results demonstrate that by effectively integrating shared information across heterogeneous data sources, INTACT enhances feature extraction and improves the robustness of physical activity analyses in diverse settings.

The remainder of this paper is organized as follows: Section 2 introduces our method INTACT for integrating longitudinal physical activity intensity time series data. Section 3 evaluates the performance of this integration method through simulations. In Section 4, we demonstrate INTACT using NHANES data. Finally, Section 5 concludes with a discussion.

2 Methodology for data integration

To effectively integrate longitudinal physical activity data from different domains, we develop INTACT, a method that accounts for domain-specific variations while preserving common biological features. First, we introduce a decomposition approach that partitions observed longitudinal activity data into shared and domain-specific components. This decomposition enables data harmonization by removing domain-specific effects while retaining biologically meaningful features (Section 2.1). Next, we outline the estimation procedure and implementation steps (Section 2.2) to ensure that the harmonized data are properly scaled and structured for downstream analysis.

2.1 Model for integrating longitudinal physical activity data

We define a dataset within a domain as data collected under the same conditions—such as within the same study or site using the same type of device. Suppose our data consist of epoch-level physical activity intensity measures from M different domains. Under domain a, there are n_a subjects, and for subject i, the activity data are collected over r_i days. The total number of observations is $N_a = \sum_{i=1}^{n_a} r_i$. Denote by $Y_{ij}^{(a)}(t)$ the log-transformed activity intensity value on day j for subject i from domain a at timestamp t. The original data $Y_{ij}^{(a)}(t)$ are scaled by dividing them by the square root of the L_2-norm of the sample variance matrix.

We then adopt the following regression model for the scaled data $\tilde{Y}_{ij}^{(a)}(t)$:

$$\tilde{Y}_{ij}^{(a)}(t) = \alpha(t) + \gamma^{(a)}(t) + Z_{ij}^T\beta(t) + \eta_{ij}^{(a)}(t),\ a = 1, \ldots, M,\ i = 1, \ldots, n_a,\ j = 1, \ldots, r_i, \tag{1}$$

where we explicitly model and correct the domain-specific mean shift, and we include the domain-specific and subject-specific variability. Here, $\alpha(t)$ is the overall mean activity at time t across M different domains, $\gamma^{(a)}$ is the domain-specific mean shift, Z_{ij} is a vector of biological covariates such as age and sex with the fixed effect coefficients $\beta(t)$, and $\eta_{ij}^{(a)}(t)$ is the domain-specific mean-zero random term that describes the variability of $\tilde{Y}_{ij}^{(a)}(t)$ and needs to be further decomposed. We decompose $\eta_{ij}^{(a)}(t)$ via the structured functional principal component model as considered in Shou et al. (2015)

$$\eta_{ij}^{(a)}(t) = X_i^{(a)}(t) + W_{ij}^{(a)}(t) + \epsilon_{ij}^{(a)}(t), \tag{2}$$

where $X_i^{(a)}(t)$ and $W_{ij}^{(a)}(t)$ are uncorrelated mean-zero subject-specific and day-specific latent random processes with covariance operators $S^{X(a)}(s,t)$ and $S^{W(a)}(s,t)$, respectively, and $\epsilon_{ij}^{(a)}(t)$ represents the mean-zero random noise that accounts for the independent time-specific measure errors at each time point. Using the Karhunen–Loéve expansion for $X_i^{(a)}(t)$ and $W_{ij}^{(a)}(t)$ in (2), we obtain

$$X_i^{(a)}(t) \approx \sum_{k=1}^{K_X} \sqrt{\lambda_k^{X(a)}}\psi_k^{X(a)}(t)\xi_{ik}^{X(a)}, \tag{3}$$

$$W_{ij}^{(a)}(t) \approx \sum_{k=1}^{K_W} \sqrt{\lambda_k^{W(a)}}\psi_k^{W(a)}(t)\xi_{ijk}^{W(a)}, \tag{4}$$

where $\psi_k^{X(a)}(t)$ and $\psi_k^{W(a)}(t)$ are the eigenfunctions of $S^{X(a)}(s,t)$ and $S^{W(a)}(s,t)$, $\lambda_k^{X(a)}$ and $\lambda_k^{W(a)}$ are the eigenvalues of $S^{X(a)}(s,t)$ and $S^{W(a)}(s,t)$, $\xi_{ik}^{X(a)}$ and $\xi_{ijk}^{W(a)}$ are the scores which are mutually independent random variables with mean 0 and variance 1, and K_X and K_W represent the number of principal components of interest for X and W, respectively.

Let $\Psi^{X(a)}(t) = (\psi_1^{X(a)}(t), \ldots, \psi_{K_X}^{X(a)}(t))$, $\Delta^{X(a)} = \text{diag}(\lambda_1^{X(a)}, \ldots, \lambda_{K_X}^{X(a)})$, and $\xi_i^{X(a)} = (\xi_{i1}^{X(a)}, \ldots, \xi_{iK_X}^{X(a)})^T$. Let $\Psi^{W(a)}(t)$, $\Delta^{W(a)}$, and $\xi_{ij}^{W(a)}$ be the corresponding terms for the W part. We rewrite (3) and (4) as

$$X_i^{(a)}(t) \approx \Psi^{X(a)}(t)(\Delta^{X(a)})^{1/2}\xi_i^{X(a)},$$
$$W_{ij}^{(a)}(t) \approx \Psi^{W(a)}(t)(\Delta^{W(a)})^{1/2}\xi_{ij}^{W(a)},$$

and further decompose the eigenfunctions and eigenvalues as common shared and domain-specific parts via

$$X_i^{(a)}(t) \approx \Phi^X(t)U^{X(a)}(\Lambda^X)^{1/2}(D^{X(a)})^{1/2}\xi_i^{X(a)}, \tag{5}$$
$$W_{ij}^{(a)}(t) \approx \Phi^W(t)U^{W(a)}(\Lambda^W)^{1/2}(D^{W(a)})^{1/2}\xi_{ij}^{W(a)},$$

where $\Phi^X(t)$ and $\Phi^W(t)$ are common eigenfunctions shared across domains, satisfying $\int(\Phi^X(t))^T\Phi^X(t)dt = I_{K_X}$ and $\int(\Phi^W(t))^T\Phi^W(t)dt = I_{K_W}$, $U^{X(a)}(t)$ and $U^{W(a)}(t)$ are domain-specific rotations, satisfying $(U^{X(a)})^TU^{X(a)} = I_{K_X}$ and $(U^{W(a)})^TU^{W(a)} = I_{K_W}$, Λ^X and $D^{X(a)}$ are $K_X \times K_X$ diagonal matrices, Λ^W and $D^{W(a)}$ are $K_W \times K_W$ diagonal matrices. The diagonals of the matrices Λ^X and Λ^W represent the common eigenvalues, and the diagonals of the matrices $D^{X(a)}$ and $D^{W(a)}$ describe the domain-specific scales.

The set of unknown parameters is estimated by minimizing the objective function provided in Section 2.2. Finally, to remove domain-specific effects while preserving common biological features, we propose the following two forms of harmonized data:

- INTACT:

$$\hat{Y}_{ij}^{\text{intact}(a)}(t) = c_0\Big(\hat{\alpha}(t) + Z_{ij}^T\hat{\beta}(t) + c_X^{(a)}\hat{\Phi}^X(t)(\hat{\Lambda}^X)^{1/2}\hat{\xi}_i^{X(a)} + c_W^{(a)}\hat{\Phi}^W(t)(\hat{\Lambda}^W)^{1/2}\hat{\xi}_{ij}^{W(a)} + c_\epsilon^{(a)}\hat{\epsilon}_{ij}^{(a)}(t)\Big),$$

- INTACT$_0$:

$$\hat{Y}_{ij}^{\text{intact}_0(a)}(t) = c_0\Big(\hat{\alpha}(t) + Z_{ij}^T\hat{\beta}(t) + c_X^{(a)}\hat{\Phi}^X(t)(\hat{\Lambda}^X)^{1/2}\hat{\xi}_i^{X(a)} + c_W^{(a)}\hat{\Phi}^W(t)(\hat{\Lambda}^W)^{1/2}\hat{\xi}_{ij}^{W(a)}\Big),$$

where c_0, $c_X^{(a)}$, $c_W^{(a)}$, and $c_\epsilon^{(a)}$ are set to match the original variance scale of all data (see details in Section 2.2).

The noise term $\epsilon_{ij}^{(a)}(t)$ is harmonized and incorporated in $\hat{Y}_{ij}^{\text{intact}(a)}(t)$, whereas it is omitted in $\hat{Y}_{ij}^{\text{intact}_0(a)}(t)$, resulting in a noise-free, harmonized observation for $\hat{Y}_{ij}^{\text{intact}_0(a)}(t)$. As $\hat{Y}_{ij}^{\text{intact}(a)}(t)$ retains residual noise, it more closely aligns with the original data distribution, while the smoothed form $\hat{Y}_{ij}^{\text{intact}_0(a)}(t)$ is suitable for downstream analyses that require PCA feature extraction.

2.2 Estimation and implementation of the integration method

To extract shared information for longitudinal data across different domains, we deal with X and W parts respectively. However, since these two parts cannot be directly observed and only $Y_{ij}^{(a)}(t)$ (and consequently $\eta_{ij}^{(a)}(t)$) is available, our method relies on the sample covariance of the X and W parts, which is derived from $\eta_{ij}^{(a)}(t)$ as described in Shou et al. (2015) and detailed in Web Appendix A.

In the following, we only present the estimation method for the shared features of the X part. The estimation for the W part is similar and thus omitted. For a matrix A, let A_k be the kth column of A and A_{kk} be its (k,k) element. Let e_k be the kth column of the identity matrix I_{K_X}. Since (5) implies

$$S^{X(a)}(s,t) \approx \Phi^X(s) U^{X(a)} \Lambda^X D^{X(a)} (U^{X(a)})^T (\Phi^X(t))^T,$$

we consider the following objective to estimate the set of unknown parameters $\{\Phi^X(t), \Lambda^X, \{U^{X(a)}, D^{X(a)}\}_{a=1}^M\}$:

$$\begin{aligned} \text{minimize} \quad & \sum_{a=1}^{M} \int (\Phi^X(s) U^{X(a)} \Lambda^X D^{X(a)} (U^{X(a)})^T (\Phi^X(t))^T - \hat{S}^{X(a)}(s,t))^2 ds dt \\ & + \sum_{a=1}^{M} \sum_{k=1}^{K_X} \left\{ \lambda_k \| U_k^{X(a)} - e_k \|_2^2 + \nu_k (D_{kk}^{X(a)} - b^{(a)})^2 \right\} + \omega \sum_{k=1}^{K_X} \int \partial^2 \Phi_k^X(t)^2 dt, \qquad (6) \\ \text{subject to} \quad & \sum_{a=1}^{M} b^{(a)} = M,\ b^{(a)} \geq 0, \end{aligned}$$

where $\hat{S}^{X(a)}(s,t)$ is the sample covariance matrix of $X_i^{(a)}(t)$, Λ^X, $\{D^{X(a)}\}_{a=1}^M$ are nonnegative diagonal, and $\Phi^X(t)$, $\{U^{X(a)}\}_{a=1}^M$ are orthonormal. Here, in (6) the penalty term $\lambda_k \| U_k^{X(a)} - e_k \|_2^2$ encourages $U^{X(a)}$ to be close to I_{K_X}, and $\nu_k (D_{kk}^{X(a)} - b^{(a)})^2$ ensures that $D^{X(a)}$ is not far away from $b^{(a)} I_{K_X}$. We set the tuning parameters such that $\lambda_1 > \cdots > \lambda_{K_X}$ and $\nu_1 > \cdots > \nu_{K_X}$, promoting consistency in eigenvectors and scaling for the initial directions, while allowing flexibility in the subsequent directions. The last term in (6) is a smoothing penalty with ω representing a constant smoothing parameter. We impose the restriction that $\sum_{a=1}^M b^{(a)} = M,\ b^{(a)} \geq 0$ to guarantee the model identifiability, since $\Lambda^X D^{X(a)}$ are in product form.

In practice, we expand the functions in terms of known orthogonal basis functions when solving the above optimization problem. It is sufficient to adopt proportional tuning parameters λ_k and ν_k to simplify the tuning procedure (see details in Web Appendix C). We set the number of latent factors K_X and K_W to 2 in our numerical studies, and our results are robust when they are chosen from a reasonably wide range.

Specifically, our integration procedure INTACT is summarized as follows.

Step 1 (Preprocessing). We vectorize the functional outcome at the discrete sampling points $\mathcal{T} = \{t_1, \ldots, t_p\}$ and define $\boldsymbol{Y}^{(a)} = (\boldsymbol{Y}_{11}^{(a)}, \ldots, \boldsymbol{Y}_{1r_1}^{(a)}, \ldots, \boldsymbol{Y}_{n_a 1}^{(a)}, \ldots, \boldsymbol{Y}_{n_a r_{n_a}}^{(a)})^T \in \mathbb{R}^{N_a \times p}$, where $\boldsymbol{Y}_{ij}^{(a)} = (Y_{ij}^{(a)}(t_1), \ldots, Y_{ij}^{(a)}(t_p))^T$. We scale the observations as $\tilde{Y}_{ij}^{(a)}(t) = Y_{ij}^{(a)}(t) / \sqrt{\| \boldsymbol{S}_Y^{(a)} \|_2}$, where $\boldsymbol{S}_Y^{(a)}$ is the sample covariance matrix, defined as $\boldsymbol{S}_Y^{(a)} = (\boldsymbol{Y}^{(a)})^T \boldsymbol{Y}^{(a)} / N_a$. Similarly, we define $\tilde{\boldsymbol{Y}}^{(a)}$ in the same manner as $\boldsymbol{Y}^{(a)}$, for all $a = 1, \ldots, M$. Let the scaling factor $c_0 = \sqrt{\| \boldsymbol{S}_0 \|_2 / \| \tilde{\boldsymbol{S}}_0 \|_2}$, where $\boldsymbol{S}_0$ and $\tilde{\boldsymbol{S}}_0$ are the sample covariance matrices of $\{\boldsymbol{Y}_{ij}^{(1)}, \ldots, \boldsymbol{Y}_{ij}^{(M)}\}$ and $\{\tilde{\boldsymbol{Y}}_{ij}^{(1)}, \ldots, \tilde{\boldsymbol{Y}}_{ij}^{(M)}\}$, respectively.

Step 2 (Estimate $\eta_{ij}^{(a)}(t)$ in (1)). For each t, estimate $\alpha(t), \gamma^{(a)}(t), \beta(t)$ using REML for linear mixed effects models with subject-specific random intercept (`lmer` function in the R package `lme4`) and obtain estimates $\hat{\alpha}(t), \hat{\gamma}^{(a)}(t), \hat{\beta}(t)$. Let

$$\hat{\eta}_{ij}^{(a)}(t) = \tilde{Y}_{ij}^{(a)}(t) - \hat{\alpha}(t) - \hat{\gamma}^{(a)}(t) - Z_{ij}^T \hat{\beta}(t). \qquad (7)$$

Apply functional PCA (`mfpca.face` function in the R package `refund` (Cui et al., 2023)) to $\hat{\eta}_{ij}^{(a)}(t)$ and get an estimate of noise $\epsilon_{ij}^{(a)}(t)$ denoted by $\hat{\epsilon}_{ij}^{(a)}(t)$.

Step 3 (Estimate common features). Let $\hat{\eta}_{ij}^{*(a)}(t) = \hat{\eta}_{ij}^{(a)}(t) - \hat{\epsilon}_{ij}^{(a)}(t)$. Use $\hat{\eta}_{ij}^{*(a)}(t)$ to obtain sample covariance of X and W denoted by $\hat{S}^{X(a)}(s,t)$ and $\hat{S}^{W(a)}(s,t)$ (Web Appendix A). Perform eigen-decomposition on $\hat{S}^{X(a)}(s,t)$ and $\hat{S}^{W(a)}(s,t)$ to obtain unconstrained estimates of $\Psi^{X(a)}(t)$, $\Delta^{X(a)}$, $\Psi^{W(a)}(t)$, and $\Delta^{W(a)}$, denoted by $\hat{\Psi}^{X(a)}(t)$, $\hat{\Delta}^{X(a)}$, $\hat{\Psi}^{W(a)}(t)$, and $\hat{\Delta}^{W(a)}$, respectively. Regress $\hat{\eta}_{ij}^{(a)}(t)$ on $\hat{\Psi}^{X(a)}(t)(\hat{\Delta}^{X(a)})^{1/2}$ and $\hat{\Psi}^{W(a)}(t)(\hat{\Delta}^{W(a)})^{1/2}$ to estimate scores $\xi_i^{X(a)}$ and $\xi_{ij}^{W(a)}$, denoted by $\hat{\xi}_i^{X(a)}$ and $\hat{\xi}_{ij}^{W(a)}$, respectively (Web Appendix B). Apply the proposed method (6) to $\hat{S}^{X(a)}(s,t)$ and $\hat{S}^{W(a)}(s,t)$ to obtain estimates $\hat{\Phi}^X(t)$, $\hat{\Lambda}^X$, $\hat{\Phi}^W(t)$, and $\hat{\Lambda}^W$.

Step 4 (Data integration). Let

$$\begin{aligned}
\hat{\eta}_{ij}^{\text{intact}(a)}(t) =& c_X^{(a)}\hat{\Phi}^X(t)(\hat{\Lambda}^X)^{1/2}\hat{\xi}_i^{X(a)} + c_W^{(a)}\hat{\Phi}^W(t)(\hat{\Lambda}^W)^{1/2}\hat{\xi}_{ij}^{W(a)} + c_\epsilon^{(a)}\hat{\epsilon}_{ij}^{(a)}(t),\\
\hat{\eta}_{ij}^{\text{intact}_0(a)}(t) =& c_X^{(a)}\hat{\Phi}^X(t)(\hat{\Lambda}^X)^{1/2}\hat{\xi}_i^{X(a)} + c_W^{(a)}\hat{\Phi}^W(t)(\hat{\Lambda}^W)^{1/2}\hat{\xi}_{ij}^{W(a)},
\end{aligned}$$

where

$$(c_X^{(a)})^2 = \frac{\|\sum_{a=1}^M \hat{\boldsymbol{S}}^{X(a)}\|_2/M}{\|var(\hat{\boldsymbol{\Phi}}^X(\hat{\Lambda}^X)^{1/2}\hat{\xi}_i^{X(a)})\|_2},\ (c_W^{(a)})^2 = \frac{\|\sum_{a=1}^M \hat{\boldsymbol{S}}^{W(a)}\|_2/M}{\|var(\hat{\boldsymbol{\Phi}}^W(\hat{\Lambda}^W)^{1/2}\hat{\xi}_{ij}^{W(a)})\|_2},$$

$$(c_\epsilon^{(a)})^2 = \frac{\|\sum_{a=1}^M \hat{\boldsymbol{S}}^{\epsilon(a)}\|_2/M}{\|\hat{\boldsymbol{S}}^{\epsilon(a)}\|_2},$$

and $\hat{\boldsymbol{S}}^{X(a)}$, $\hat{\boldsymbol{S}}^{W(a)}$, $\hat{\boldsymbol{S}}^{\epsilon(a)}$, $\hat{\boldsymbol{\Phi}}^X$, $\hat{\boldsymbol{\Phi}}^W$ are the discrete versions of $\hat{S}^{X(a)}(s,t)$, $\hat{S}^{W(a)}(s,t)$, $\hat{S}^{\epsilon(a)}(s,t)$, $\hat{\Phi}^X(t)$, $\hat{\Phi}^W(t)$, respectively. We remark that for the two types of the integration of η, $\hat{\eta}_{ij}^{\text{intact}(a)}(t)$ includes the noise term $\hat{\epsilon}_{ij}^{(a)}(t)$, while it is removed in $\hat{\eta}_{ij}^{\text{intact}_0(a)}(t)$.

We then add the overall mean and covariates effects estimated in Step 2 to obtain the integrated data

$$\begin{aligned}
&\text{(INTACT)} \quad \hat{Y}_{ij}^{\text{intact}(a)}(t) = c_0(\hat{\alpha}(t) + Z_{ij}^T\hat{\beta}(t) + \hat{\eta}_{ij}^{\text{intact}(a)}(t)),\\
&\text{(INTACT}_0\text{)} \quad \hat{Y}_{ij}^{\text{intact}_0(a)}(t) = c_0(\hat{\alpha}(t) + Z_{ij}^T\hat{\beta}(t) + \hat{\eta}_{ij}^{\text{intact}_0(a)}(t)).
\end{aligned}$$

3 Simulations

In the simulation study, data are generated according to the following model

$$\begin{aligned}
&Y_{ij}^{(a)}(t) = \alpha(t) + \gamma^{(a)}(t) + Z_{ij}^T\beta(t) + \eta_{ij}^{(a)}(t),\\
&a = 1,\ldots,M,\ i = 1,\ldots,n_a,\ j = 1,\ldots,r_i,
\end{aligned}$$

where

$$\begin{aligned}
\eta_{ij}^{(a)}(t) =& X_i^{(a)}(t) + W_{ij}^{(a)}(t) + \epsilon_{ij}^{(a)}(t)\\
=& \Phi^X(t)U^{X(a)}(\Lambda^X)^{1/2}(D^{X(a)})^{1/2}\xi_i^{X(a)} + \Phi^W(t)U^{W(a)}(\Lambda^W)^{1/2}(D^{W(a)})^{1/2}\xi_{ij}^{W(a)} + \epsilon_{ij}^{(a)}(t).
\end{aligned}$$

We set $\alpha(t) = 0$, $\gamma^{(a)}(t) = 0$, $M = 2$, $n_a = 80$, $r_i \in \{3,4,5,6,7\}$, and $t \in \{1,\ldots,144\}$. Let $Z_{ij} \overset{iid}{\sim} N(3.5, 0.75^2)$ and $\beta(t) = 0.3(t-1)/143$. Let $\Phi^X(t) = (\phi_1^X(t), \phi_2^X(t))$ and $\Phi^W(t) = (\phi_1^W(t), \phi_2^W(t))$. The shape of the common features is shown in Figure 1A. Let

$$U^{X(1)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ U^{X(2)} = \begin{pmatrix} 0.97 & \sqrt{1-0.97^2} \\ -\sqrt{1-0.97^2} & 0.97 \end{pmatrix},$$

$$\Lambda^X = \begin{pmatrix} 2.5 & 0 \\ 0 & 1 \end{pmatrix}, \; D^{X(1)} = \begin{pmatrix} 12 & 0 \\ 0 & 13 \end{pmatrix}, \; D^{X(2)} = \begin{pmatrix} 3 & 0 \\ 0 & 3.2 \end{pmatrix},$$

$$U^{W(1)} = \begin{pmatrix} 0.95 & \sqrt{1-0.95^2} \\ -\sqrt{1-0.95^2} & 0.95 \end{pmatrix}, \; U^{W(2)} = \begin{pmatrix} 0.9 & \sqrt{1-0.9^2} \\ -\sqrt{1-0.9^2} & 0.9 \end{pmatrix},$$

$$\Lambda^W = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \; D^{W(1)} = \begin{pmatrix} 12 & 0 \\ 0 & 11 \end{pmatrix}, \; D^{W(2)} = \begin{pmatrix} 3 & 0 \\ 0 & 3.1 \end{pmatrix}.$$

Let $\xi_i^{X(a)} \overset{iid}{\sim} N(0, I_2)$, $\xi_{ij}^{W(a)} \overset{iid}{\sim} N(0, I_2)$, and $\epsilon_{ij}^{(a)} \overset{iid}{\sim} N(0, 0.5^2)$, $a = 1, 2$.

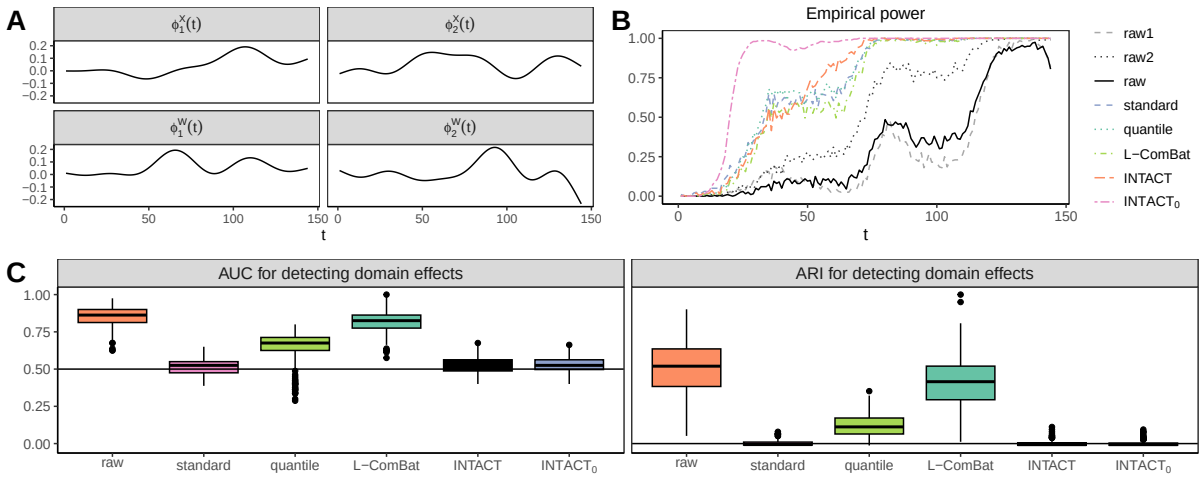

Figure 1: (A–C) Simulation settings and results. The abbreviations used in this figure are as follows: domain 1 data (raw1), domain 2 data (raw2), combined raw data (raw), simple standardization (standard), quantile normalization (quantile), longitudinal ComBat (L-ComBat), and proposed methods (INTACT and INTACT$_0$). (A) Shape of common features. Top: $\phi_1^X(t)$ and $\phi_2^X(t)$; bottom: $\phi_1^W(t)$ and $\phi_2^W(t)$. (B) Empirical power for testing $H_0 : \beta(t) = 0$. (C) Detection of domain effects. Left: box plots of AUC values; Right: box plots of ARI values.

To make the original data comparable with the proposed method, we scale $Y_{ij}^{(a)}(t)$ by dividing them by the square root of L_2-norm of the sample variance matrix (Step 1 in Section 2.2). We take the scaled data $\tilde{Y}_{ij}^{(a)}(t)$ as the baseline and call them raw data for fairness of comparison. We apply the simple standardization, the quantile normalization (Amaratunga and Cabrera, 2001; Bolstad et al., 2003), and the longitudinal ComBat method (Beer et al., 2020) to the raw data for comparison. Here, the simple standardization and the quantile normalization approaches are used at each time point, and details can be found in Web Appendix D. These competing methods apply pointwise harmonization to the raw data, aiming to harmonize the marginal distribution at each time point across different domains. However, they do not ensure harmonization of the overall temporal structure. Furthermore, simple standardization and quantile normalization treat all measurements as independent observations, disregarding the dependency among repeated measures within subjects. As a result, these methods are less suitable for the longitudinal, within-day correlated data considered in this study.

Under the above simulation setting, we generate 200 datasets. We first examine whether data from different domains are harmonized well. For each dataset, we randomly draw one observation (day) from the repeated observations and obtain a resampled dataset. To avoid the possible confounding of domain effects with other covariate effects, we subtract the fixed part $\hat{\alpha}(t) + Z_{ij}^T \hat{\beta}(t)$ from harmonized data.

For each resampled dataset, we randomly split the sample into 50% training set and 50% testing set and train a random forest model on the training set to classify the domain label from the residualized activity intensity measures from a day, and assess predictive performance on the testing set. Our random forest algorithm is implemented via the R package `randomForest` with default parameters. Lower area under ROC curve (AUC) and adjusted Rand indice (ARI) for recognizing data from different domains indicate improved harmonization. We see from Figure 1C that the simple standardization, INTACT and INTACT_0 substantially reduce the chance of detecting different domains compared to unharmonized data and other competing methods.

We then examine the detection of biological signals by testing $H_0 : \beta(t) = 0$ for each t. We apply the empirical likelihood ratio test (Wang et al., 2010) to the domain 1 data, the domain 2 data, the raw data, and data integrated by the simple standardization, the quantile normalization, the longitudinal ComBat, and the proposed method. Figure 1B presents the empirical power of testing $H_0 : \beta(t) = 0$. We see the proposed denoised harmonized data $\hat{Y}_{ij}^{\text{intact}_0(a)}$ obtain much higher power and exhibit more significant biological signal under almost the whole interval. When $t \leq 50$, the proposed integrated data $\hat{Y}_{ij}^{\text{intact}(a)}$ performs similarly to the simple standardization, the quantile normalization and the longitudinal ComBat method; however, when $t > 50$, $\hat{Y}_{ij}^{\text{intact}(a)}$ shows superior performance.

4 Application to NHANES data

We evaluate our method using NHANES 2003–2004 and NHANES 2013–2014 physical activity data. The NHANES 2003–2004 dataset was collected using the Actigraph AM-7164, a hip-worn accelerometer, while the NHANES 2013–2014 dataset used the Actigraph GT3X+, a wrist-worn accelerometer. The NHANES 2003–2004 dataset includes 7,176 participants aged 6 years and older, and the NHANES 2013–2014 dataset includes 7,776 participants aged 3 years and older. In both studies, participants wore physical activity monitors for seven consecutive days. Data were extracted from the NHANES website, and we use an analyzable format available at `https://functionaldataanalysis.org/dataset_nhanes.html` (Crainiceanu et al., 2024).

A key challenge in integrating these datasets arises from differences in measurement units. The 2003–2004 data recorded physical activity counts, with a mean of 222.2, whereas the 2013–2014 data used the Monitor-Independent Movement Summary (MIMS) metric, which captures activity intensity on a smaller scale, with a mean of 8.863. Although MIMS is designed to be device-agnostic and is supported by open-source software, it is limited to tri-axial data and cannot be applied to the 2003–2004 dataset, which was collected using a uni-axial accelerometer.

In this study, we aim to integrate longitudinal minute-by-minute physical activity data from the two NHANES devices to enhance the detection of biologically meaningful signals. Specifically, we focus on participants aged 50 years and older, with physical activity data linked to covariates $Z_{ij} = (\text{gender, age, income, BMI})^T$. To ensure data quality, we exclude eight participants from NHANES 2003–2004 due to extreme physical activity records and include only days with at least 10 hours of estimated wear time. Activity intensities are summarized into a 1,440-dimensional vector per day. The NHANES physical activity data are then linked to 2019 mortality outcomes using the Public-use Linked Mortality Files, accessible at `https://www.cdc.gov/nchs/data-linkage/mortality-public.htm`.

After preprocessing, the final dataset includes: 1,870 participants (11,125 total observation days) from NHANES 2003–2004, and 1,986 participants (12,961 total observation days) from NHANES 2013–2014 To integrate the datasets, we apply a log transformation to activity intensities and evaluate multiple harmonization methods, including INTACT, its simplified variant INTACT_0, simple standardization, quantile normalization, and longitudinal ComBat.

4.1 Analysis of discrepancies between two covariance matrices

Although both the competing harmonization methods and INTACT (and INTACT$_0$) are generally effective at reducing mean differences between domains, they may not effectively address covariance discrepancies. Here, we examine the differences between the covariance matrices of data collected by the two devices. We resample the data by drawing one observation (day) from repeated observations for each participant in two domains. We perform the procedure $K = 200$ times and denote the new datasets by $D_k^{(a)},\ a = 1, 2,\ k = 1, \ldots, K$. Let

$$Var_1 = \frac{1}{K}\sum_{k=1}^{K} var(D_k^{(1)}), \quad Var_2 = \frac{1}{K}\sum_{k=1}^{K} var(D_k^{(2)}).$$

Figure 2A shows the heat map for $Var_1 - Var_2$. We observe almost no difference between the two covariance matrices for the data integrated by INTACT and INTACT$_0$. By comparison, the raw data present the covariance difference in a much larger scale. The performances of the simple standardization, the quantile normalization, and the longitudinal ComBat method are similar, and all of them show that there still exist covariance differences.

4.2 Detection of domain effects

We resample 120 subjects from each domain for harmonized data, and draw one observation (day) from repeated observations for each subject. To avoid the possibility that domain effects could be detected through other covariates, we subtract the fixed-effects $\hat{\alpha}(t) + Z_{ij}^T\hat{\beta}(t)$ from harmonized data. For each resampled dataset, we randomly choose 50% sample into the training set and train a random forest model to recognize observations from different domains. We use the rest 50% sample to assess predictive performance.

We carry out the procedure 200 times, and draw box plots of AUC and ARI values in Figure 2B. We see that INTACT, INTACT$_0$, and the quantile normalization method have much lower AUC and ARI values, which indicates improved harmonization.

4.3 Age effects in physical activity

Many studies suggest that age can impact circadian activity curves (Xiao et al., 2015; Wrobel et al., 2021). To examine whether harmonized data reveal a stronger association between age and physical activity patterns, we test the null hypothesis of no age effect on activity intensity. We resample 120 subjects from each domain and apply a pointwise empirical likelihood ratio test (Wang et al., 2010) to the linear mixed-effects model. The procedure is repeated 200 times.

Figure 2C shows the empirical power results. When using resampled data from NHANES 2003-2004 or NHANES 2013-2014, the null hypothesis of no age effect cannot be rejected, and the empirical power is nearly zero as indicated by the dark gray dashed and dotted curves. However, significant age effects are observed when either the combined raw data from both domains or the harmonized data are used. Among these, INTACT and INTACT$_0$ demonstrate the highest power compared to existing approaches across nearly the entire time interval.

4.4 Association between mortality and physical activity

Recent studies have increasingly focused on the relationship between mortality and physical activity patterns (Cui et al., 2021). To examine this association, we apply a Cox model to hourly average log-transformed physical activity intensities, adjusted for gender, age, income, and BMI, using a 5-year censoring period. The model is implemented using the `coxph()` function from the R package `survival`.

Table 1: The average 10-fold cross-validated Harrell's C-index when applying the Cox model to the 12-dimensional activity vector along with gender, age, income, and BMI

	domain 1	domain 2	all
raw	0.789	0.765	0.774
standard	0.789	0.765	0.780
quantile	0.757	0.764	0.766
L-ComBat	0.788	0.765	0.780
INTACT	0.794	0.767	0.781
INTACT_0	0.792	0.775	0.784

For participants with multiple days of measurements, we average the log-transformed activity intensities across days.

We first apply separate Cox models to the activity intensity at each hour. Figure 2D shows the hazard ratios and their corresponding Bonferroni-corrected 99.79%(=1-0.05/24) confidence intervals for each hour-level physical activity intensity. All methods reveal significant results from approximately 9:00 to 20:00, with hazard ratios below 1, indicating that daytime physical activity significantly reduces mortality risk. From about 22:00 to 7:00, hazard ratios for INTACT and INTACT_0 exceed 1, while those for other methods hover around or exceed 1. Only the proposed methods show significant results during this period, suggesting that nighttime activity may negatively impact lifespan. INTACT and INTACT_0 also work well for minute-level data, with similar results provided in the Web Appendix E.

We further analyze the relationship between overall activity patterns and mortality by averaging activity intensities over 2-hour intervals and summarizing the activity trend as a 12-dimensional vector. The Cox model is then applied to this activity vector along with gender, age, income, and BMI. Table 1 presents the mean 10-fold cross-validated Harrell's C-index. The results show that INTACT and INTACT_0 achieve higher C-indexes than other methods, indicating better predictive performance.

4.5 Downstream analysis of summary statistics following data integration

Since most of the studies collecting wearable device data would derive summary activity markers for downstream analysis, we examine whether the harmonized epoch-level physical activity intensity measures also result in more consistent scalar-level summary statistics between subjects in NHANES 2003-2004 and NHANES 2013-2014. We randomly draw 1000 subjects from NHANES 2003-2004, and match them with 1000 subjects in NHANES 2013-2014 according to gender, age, income, and BMI. We adopt the `pairmatch()` function from the R package `optmatch` (Hansen, 2007), using distances computed by `match_on()` based on logit propensity scores. For each subject, we draw one observation from repeated observations and calculate the summary statistics: total activity intensities (TAC), total log-transformed activity intensities (TLAC), average of activity intensities during the most active 10-hour period (M10), time of M10 (M10t), average of activity intensities during the least active 5-hour period (L5), and time of L5 (L5t). Then we compute the absolute value of mean and variance differences in each statistic between the two domains.

We perform the procedure 200 times and draw box plots for the differences in summary statistics in log10 scale in Figure 2E for better visualization. It shows the presence of differences in the summary statistics for the raw data. The performances of the simple standardization and the longitudinal ComBat methods are similar. Compared to the raw data, the data integrated by these two methods show reduced mean differences in TLAC, L5, and L5t, but increased mean differences in TAC and M10, as well as increased variance differences in all statistics except TLAC. After applying quantile normalization, the mean differences in TLAC, L5, and L5t decrease, and the variance difference in L5 decreases, while

Table 2: The average 10-fold cross-validated Harrell's C-index when applying the Cox model to the activity summary statistic vector (TAC, TLAC, M10, M10t, L5, L5t)

	domain 1	domain 2	all
raw	0.739	0.730	0.730
standard	0.720	0.723	0.713
quantile	0.611	0.714	0.667
L-ComBat	0.721	0.723	0.713
INTACT	0.749	0.738	0.734
INTACT_0	0.752	0.742	0.732

the other results remain similar to the raw data. In contrast, INTACT and INTACT_0 result in relatively low differences in both mean and variance. The data integrated using INTACT exhibit no greater mean differences than the raw data and have nearly the second smallest variance differences across all summary statistics. The denoised integration, INTACT_0, yields the smallest differences in both mean and variance in all summary statistics, indicating improved harmonization.

We also examine the association between mortality and the physical activity summary statistics vector (TAC, TLAC, M10, M10t, L5, L5t). To do this, we apply the Cox model to each matched dataset and calculate the mean 10-fold cross-validated Harrell's C-index. Table 2 presents the average Harrell's C-index across 200 matched datasets. INTACT and INTACT_0 consistently achieve higher C-index values than other methods, whether using data from only domain 1 (NHANES 2003-2004), only domain 2 (NHANES 2013-2014), or combining data from both domains.

5 Discussion

As wearable devices continue to proliferate in both consumer markets and research settings, differences in device features, accuracy, and advantages have raised analytical challenges (Scalise and Cosoli, 2018; Vooris et al., 2019; Cosoli et al., 2020). A critical unresolved challenge remains: how to effectively integrate physical activity data from different wearables to enable meaningful cross-study comparisons.

In this paper, we propose a novel data integration method specifically designed to accommodate the longitudinal and within-day correlations inherent in physical activity data. Our approach leverages a functional mixed-effects model, which accounts for the structure of the data by separately modeling subject-specific and visit-specific processes. This framework ensures that biological signals are preserved while mitigating unwanted domain effects, providing a robust solution for harmonizing physical activity data across studies.

Many existing approaches rely on pointwise harmonization, which aligns the marginal distribution at each time point but fails to harmonize the joint distribution across domains. In contrast, INTACT (and INTACT_0) focuses on the principal directions of different datasets to achieve a more consistent joint distribution. Our method assumes a shared eigenspace across datasets while allowing for domain-specific rotations and scales. By assigning different weights to penalty terms, INTACT (and INTACT_0) ensures consistency in eigenvectors and scales for the initial principal directions while maintaining flexibility for subsequent directions, effectively balancing harmonization and adaptability.

The effectiveness of INTACT (and INTACT_0) has been demonstrated through simulations and real-world datasets collected using different device versions within the same study. In the NHANES application, some activity intensity data may have questionable reliability. While we did not perform a detailed analysis to address this issue, our method—which relies on covariance matrix estimation—could be adapted to account for data reliability by incorporating only high-confidence observations in the covari-

ance estimation process. We leave this extension for future research.

Another future research direction is to generalize this method to data from different device brands, studies, and sample characteristics, for example, diseased vs. healthy populations or children vs. adults. Additionally, our model requires data with the same time resolution (e.g., minute-level). While we can often align and interpolate data collected at different resolutions (e.g., UK Biobank data at 5-second intervals vs. minute intervals), the impact of such preprocessing steps has not been fully evaluated. Investigating this aspect could be an interesting topic for future research.

APPENDIX

A Covariance estimation

We employ the method of moments (MoM) introduced in Shou et al. (2015) to derive sample covariance of X and W in the model

$$\hat{\eta}_{ij}^{*(a)}(t) = X_i^{(a)}(t) + W_{ij}^{(a)}(t) \tag{A.1}$$

$$= \Psi^{X(a)}(t)(\Delta^{X(a)})^{1/2}\xi_i^{X(a)} + \Psi^{W(a)}(t)(\Delta^{W(a)})^{1/2}\xi_{ij}^{W(a)}. \tag{A.2}$$

From (A.1), we have

$$E\{\hat{\eta}_{ij}^{*(a)}(t) - \hat{\eta}_{kl}^{*(a)}(t)\}\{\hat{\eta}_{ij}^{*(a)}(s) - \hat{\eta}_{kl}^{*(a)}(s)\}^T = \begin{cases} 2S^{W(a)}(t,s) & \text{if } i = k, j \neq l \\ 2\{S^{X(a)}(t,s) + S^{W(a)}(t,s)\} & \text{if } i \neq k. \end{cases}$$

Let $\hat{\boldsymbol{\eta}}^{*(a)} = (\hat{\boldsymbol{\eta}}_{11}^{*(a)}, \ldots, \hat{\boldsymbol{\eta}}_{1r_1}^{*(a)}, \ldots, \hat{\boldsymbol{\eta}}_{n_a1}^{*(a)}, \ldots, \hat{\boldsymbol{\eta}}_{n_ar_{n_a}}^{*(a)})^T \in \mathbb{R}^{N_a \times p}$, where $\hat{\boldsymbol{\eta}}_{ij}^{*(a)} = (\hat{\eta}_{ij}^{*(a)}(t_1), \ldots, \hat{\eta}_{ij}^{*(a)}(t_p))^T$. Let $k_1 = \sum_i r_i^2$, $A_{N_a \times N_a} = \text{diag}\{A_1, \ldots, A_{n_a}\}$ with $A_i = r_i I_{r_i}$, and $E_{n_a \times N_a} = \text{diag}\{1_{r_1}^T, \ldots, 1_{r_{n_a}}^T\}$. If $H^{W(a)} = 2S^{W(a)}$ and $H^{XW(a)} = 2\{S^{X(a)} + S^{W(a)}\}$, then we have the following explicit MoM estimators

$$\begin{aligned}
\hat{H}^{W(a)} &= \frac{1}{k_1 - N_a} \sum_i \sum_{j \neq l} (\hat{\boldsymbol{\eta}}_{ij}^{*(a)} - \hat{\boldsymbol{\eta}}_{il}^{*(a)})(\hat{\boldsymbol{\eta}}_{ij}^{*(a)} - \hat{\boldsymbol{\eta}}_{il}^{*(a)})^T \\
&= \frac{2}{k_1 - N_a} \hat{\boldsymbol{\eta}}^{*(a)}(A - E^T E)(\hat{\boldsymbol{\eta}}^{*(a)})^T, \\
\hat{H}^{XW(a)} &= \frac{1}{N_a^2 - k_1} \sum_{i \neq k} \sum_{jl} (\hat{\boldsymbol{\eta}}_{ij}^{*(a)} - \hat{\boldsymbol{\eta}}_{kl}^{*(a)})(\hat{\boldsymbol{\eta}}_{ij}^{*(a)} - \hat{\boldsymbol{\eta}}_{kl}^{*(a)})^T \\
&= \frac{2}{N_a^2 - k_1} \hat{\boldsymbol{\eta}}^{*(a)}(N_a I - 11^T - A + E^T E)(\hat{\boldsymbol{\eta}}^{*(a)})^T.
\end{aligned}$$

Hence, we obtain $\hat{S}^{W(a)} = \hat{H}^{W(a)}/2$ and $\hat{S}^{X(a)} = (\hat{H}^{XW(a)} - \hat{H}^{W(a)})/2$.

B Principal score estimation

To estimate scores in (A.2), we first adopt structured functional PCA (Shou et al., 2015) to obtain estimates of eigenfunctions denoted by $\hat{\Psi}^{X(a)}(t)$, $\hat{\Psi}^{W(a)}(t)$, and estimates of eigenvalues denoted by $\hat{\Delta}^{X(a)}$,

$\hat{\Delta}^{W(a)}$. Since $\xi_i^{X(a)}$ and $\xi_{ij}^{W(a)}$ are mutually independent random variables with mean 0 and variance I, (A.2) implies

$$E_{\xi_{ij}^{W(a)}}[\hat{\eta}_{ij}^{*(a)}] = \Psi^{X(a)}(\Delta^{X(a)})^{1/2}\xi_i^{X(a)}.$$

Thus, we have the following estimate of $\xi_i^{X(a)}$

$$\hat{\xi}_i^{X(a)} = (\hat{\Delta}^{X(a)})^{-1/2}(\hat{\Psi}^{X(a)})^T(\sum_{j=1}^{r_i}\hat{\eta}_{ij}^{*(a)}/r_i). \tag{B.3}$$

Plugging (B.3) into (A.2), we estimate $\xi_{ij}^{W(a)}$ by using

$$\hat{\xi}_{ij}^{W(a)} = (\hat{\Delta}^{W(a)})^{-1/2}(\hat{\Psi}^{W(a)})^T(\hat{\eta}_{ij}^{*(a)} - \hat{\Psi}^{X(a)}(\hat{\Delta}^{X(a)})^{1/2}\hat{\xi}_i^{X(a)}).$$

C Common feature estimation

We expand the functions $X_i^{(a)}(t)$ and $\Phi^X(t)$ in (5) in terms of known orthogonal basis functions $B(t) = (B_1(t), \ldots, B_q(t))$, i.e.,

$$X_i^{(a)}(t) = B(t)x_i^{(a)}, \quad \Phi^X(t) = B(t)\Phi^X, \tag{C.4}$$

where $x_i^{(a)} \in \mathbb{R}^q$ and $\Phi^X \in \mathbb{R}^{q\times K_X}$. Plugging (C.4) into the problem (6), we obtain the following equivalent problem

$$\begin{aligned}
\text{minimize} \quad & \sum_{a=1}^{M}\left\|\hat{s}^{X(a)} - \Phi^X U^{X(a)}\Lambda^X D^{X(a)}(U^{X(a)})^T(\Phi^X)^T\right\|_F^2 \\
& + \sum_{a=1}^{M}\sum_{k=1}^{K_X}\left\{\lambda_k\|U_k^{X(a)} - e_k\|_2^2 + \nu_k(D_{kk}^{X(a)} - b^{(a)})^2\right\} + \omega\text{tr}((\Phi^X)^T B_0\Phi^X) \\
\text{subject to} \quad & 1^T b = M,\ b \geq 0,
\end{aligned} \tag{C.5}$$

where $\Lambda^X,\ \{D^{X(a)}\}_{a=1}^M$ are nonnegative diagonal, $\Phi^X,\ \{U^{X(a)}\}_{a=1}^M$ are orthonormal, $\hat{s}^{X(a)}$ is the sample variance matrix of $x_i^{(a)}$, $B_0 = \int B''(t)^T B''(t)dt$, and $b = (b^{(1)}, \ldots, b^{(M)})^T$.

Here, we omit the upper index X and rewrite the objective function (C.5) as

$$\begin{aligned}
& L(\Phi, \Lambda, \{D^{(a)}, U^{(a)}\}_{a=1}^M, b) \\
= & \sum_{a=1}^{M}\left\{-\langle\hat{s}^{(a)}, \Phi U^{(a)}\Lambda D^{(a)}(U^{(a)})^T\Phi^T\rangle + \frac{1}{2}\|\Lambda D^{(a)}\|_F^2\right\} + \sum_{a=1}^{M}\sum_{k=1}^{K_X}\left\{\frac{\lambda_k}{2}\|U_k^{(a)} - e_k\|_2^2\right. \\
& \left. + \frac{\nu_k}{2}(D_{kk}^{(a)} - b^{(a)})^2\right\} + \frac{\omega}{2}\text{tr}(\Phi^T B_0\Phi).
\end{aligned}$$

We define the following approximations to the objective function:

$$\begin{aligned}
& \hat{L}_1(\Phi; \hat{\Phi}, \hat{\Lambda}, \{\hat{D}^{(a)}, \hat{U}^{(a)}\}_{a=1}^M, \hat{b}) \\
= & \sum_{a=1}^{M}\left\{-\langle\hat{s}^{(a)}, \hat{\Phi}\hat{U}^{(a)}\hat{\Lambda}\hat{D}^{(a)}(\hat{U}^{(a)})^T\hat{\Phi}^T\rangle - 2\langle\hat{s}^{(a)}\hat{\Phi}\hat{U}^{(a)}\hat{\Lambda}\hat{D}^{(a)}(\hat{U}^{(a)})^T, \Phi - \hat{\Phi}\rangle + \frac{1}{2}\|\hat{\Lambda}\hat{D}^{(a)}\|_F^2\right\} \\
& + \sum_{a=1}^{M}\sum_{k=1}^{K_X}\left\{\frac{\lambda_k}{2}\|\hat{U}_k^{(a)} - e_k\|_2^2 + \frac{\nu_k}{2}(\hat{D}_{kk}^{(a)} - \hat{b}^{(a)})^2\right\} + \langle\omega B_0\hat{\Phi}, \Phi - \hat{\Phi}\rangle + \frac{\omega}{2}\text{tr}(\hat{\Phi}^T B_0\hat{\Phi}),
\end{aligned}$$

$$
\begin{aligned}
&\hat{L}_2(\{U^{(a)}\}_{a=1}^M; \hat{\Phi}, \hat{\Lambda}, \{\hat{D}^{(a)}, \hat{U}^{(a)}\}_{a=1}^M, \hat{b}) \\
=&\sum_{a=1}^M \left\{ -\langle \hat{s}^{(a)}, \hat{\Phi}\hat{U}^{(a)}\hat{\Lambda}\hat{D}^{(a)}(\hat{U}^{(a)})^T\hat{\Phi}^T\rangle - 2\langle \hat{\Phi}^T\hat{s}^{(a)}\hat{\Phi}\hat{U}^{(a)}\hat{\Lambda}\hat{D}^{(a)}, U^{(a)} - \hat{U}^{(a)}\rangle + \frac{1}{2}\|\hat{\Lambda}\hat{D}^{(a)}\|_F^2 \right\} \\
&+ \sum_{a=1}^M \sum_{k=1}^{K_X} \left\{ \frac{\lambda_k}{2}\|U_k^{(a)} - e_k\|_2^2 + \frac{\nu_k}{2}(\hat{D}_{kk}^{(a)} - \hat{b}^{(a)})^2 \right\} + \frac{\omega}{2}\mathrm{tr}(\hat{\Phi}^T B_0 \hat{\Phi}).
\end{aligned}
$$

Algorithm 1 summarizes the procedure for solving the problem (C.5).

Algorithm 1 Common feature estimation under the model (C.5)

Input: $\hat{s}^{X(a)}$, $a = 1, \ldots, M$, and hyperparameter ρ
Output: $\Phi^K,\ \Lambda^K,\ \{D^{(a)K}, U^{(a)K}\}_{a=1}^M, b^K$
1. Initialize parameters $\Phi^0,\ \Lambda^0,\ \{D^{(a)0}, U^{(a)0}\}_{a=1}^M, b^0$ using Algorithm 2
2. For $\tau = 0$ to $\tau = K - 1$
   $\Phi^{\tau+1} = \arg\min_{\Phi^T\Phi=I} \hat{L}_1(\Phi; \Phi^\tau, \Lambda^\tau, \{D^{(a)\tau}, U^{(a)\tau}\}_{a=1}^M, b^\tau) + \frac{\rho}{2}\|\Phi - \Phi^\tau\|_F^2$
   $\Lambda^{\tau+1} = \arg\min_{\Lambda\succeq 0} L(\Phi^{\tau+1}, \Lambda, \{D^{(a)\tau}, U^{(a)\tau}\}_{j=1}^M, b^\tau) + \frac{\rho}{2}\|\Lambda - \Lambda^\tau\|_F^2$
   $\{D^{(a)\tau+1}\}_{a=1}^M = \arg\min_{D^{(a)}\succeq 0} L(\Phi^{\tau+1}, \Lambda^{\tau+1}, \{D^{(a)}, U^{(a)\tau}\}_{a=1}^M, b^\tau) + \frac{\rho}{2}\sum_{a=1}^M \|D^{(a)} - D^{(a)\tau}\|_F^2$
   $b^{\tau+1} = \arg\min_{1^T b = M,\ b\geq 0} L(\Phi^{\tau+1}, \Lambda^{\tau+1}, \{D^{(a)\tau+1}, U^{(a)\tau}\}_{a=1}^M, b) + \frac{\rho}{2}\sum_{a=1}^M (b^{(a)} - b^{(a)\tau})^2$
   $\{U^{(a)\tau+1}\}_{a=1}^M = \arg\min_{(U^{(a)})^T U^{(a)}=I} \hat{L}_2(\{U^{(a)}\}_{a=1}^M; \Phi^{\tau+1}, \Lambda^{\tau+1}, \{D^{(a)\tau+1}, U^{(a)\tau}\}_{a=1}^M, b^{\tau+1})$
   $\qquad + \frac{\rho}{2}\sum_{j=1}^M \|U^{(a)} - U^{(a)\tau}\|_F^2$

In Φ-update of Algorithm 1, the subproblem to be solved is

$$
\Phi^{\tau+1} = \arg\min_{\Phi^T\Phi=I} -2\langle A, \Phi\rangle = \arg\min_{\Phi^T\Phi=I} \|\Phi - A\|_F^2,
$$

where $A = \sum_{a=1}^M \hat{s}^{(a)}\Phi^\tau U^{(a)\tau}\Lambda^\tau D^{(a)\tau}(U^{(a)\tau})^T - \omega B_0\Phi^\tau/2 + \Phi^\tau\rho/2$. Let $A = \tilde{Q}\tilde{D}\tilde{P}^T$ be the SVD decomposition of A. Then the solution is $\Phi^{\tau+1} = \tilde{Q}\tilde{P}^T$.

The subproblem for Λ-update of Algorithm 1 is

$$
\begin{aligned}
\Lambda^{\tau+1} =& \arg\min_{\Lambda\succeq 0} \sum_{a=1}^M \left\{ -\langle \hat{s}^{(a)}, \Phi^{\tau+1}U^{(a)\tau}\Lambda D^{(a)\tau}(U^{(a)\tau})^T(\Phi^{\tau+1})^T\rangle + \frac{1}{2}\|\Lambda D^{(a)\tau}\|_F^2 \right\} \\
&+ \frac{\rho}{2}\|\Lambda - \Lambda^\tau\|_F^2 \\
=& \arg\min_{\Lambda\succeq 0} -\mathrm{tr}(\Lambda B) + \frac{1}{2}tr\left(\Lambda^2 E\right) \\
=& \max(\mathrm{diag}(B_{11}/E_{11}, \ldots, B_{K_XK_X}/E_{K_XK_X}), 0),
\end{aligned}
$$

where $B = \sum_{a=1}^M D^{(a)\tau}(U^{(a)\tau})^T(\Phi^{\tau+1})^T\hat{s}^{(a)}\Phi^{\tau+1}U^{(a)\tau} + \rho\Lambda^\tau$ and $E = \sum_{a=1}^M (D^{(a)\tau})^2 + \rho I$.

It can be proved that D-update of Algorithm 1 consists of M decoupled subproblems:

$$\begin{aligned}
D^{(a)\tau+1} &= \arg\min_{D^{(a)}\succeq 0} -\langle \hat{s}^{(a)}, \Phi^{\tau+1}U^{(a)\tau}\Lambda^{\tau+1}D^{(a)}(U^{(a)\tau})^T(\Phi^{\tau+1})^T\rangle + \frac{1}{2}\|\Lambda^{\tau+1}D^{(a)}\|_F^2 \\
&\quad + \sum_{k=1}^{K_X}\frac{\nu_k}{2}(D_{kk}^{(a)} - b^{(a)\tau})^2 + \frac{\rho}{2}\|D^{(a)} - D^{(a)\tau}\|_F^2 \\
&= \arg\min_{D^{(a)}\succeq 0} -\mathrm{tr}(D^{(a)}F) + \frac{1}{2}\mathrm{tr}((D^{(a)})^2G) \\
&= \max(\mathrm{diag}(F_{11}/G_{11}, \ldots, F_{K_XK_X}/G_{K_XK_X}), 0), \quad a = 1, \ldots, M,
\end{aligned}$$

where $F = (U^{(a)\tau})^T(\Phi^{\tau+1})^T\hat{s}^{(a)}\Phi^{\tau+1}U^{(a)\tau}\Lambda^{\tau+1} + b^{(a)\tau}\mathrm{diag}(\nu_1, \ldots, \nu_{K_X}) + \rho D^{(a)\tau}$ and $G = (\Lambda^{\tau+1})^2 + \mathrm{diag}(\nu_1, \ldots, \nu_{K_X}) + \rho I$.

Let $\nu = \mathrm{diag}(\nu_1, \ldots, \nu_{K_X})$, $F_b = (\sum_{k=1}^{K_X}\nu_k)I$ and $\psi = (\mathrm{tr}(\nu D^{(1)\tau+1}), \ldots, \mathrm{tr}(\nu D^{(M)\tau+1}))^T$. The subproblem for b-update of Algorithm 1 is

$$b^{\tau+1} = \arg\min_{1^Tb=M,\ b\geq 0}\frac{1}{2}b^T(F_b + \rho\mathbf{I})b - (\psi + \rho b^\tau)^Tb,$$

which is a quadratic programming problem.

U-update of Algorithm 1 also consists of M decoupled subproblems:

$$\begin{aligned}
U^{(a)\tau+1} &= \arg\min_{(U^{(a)})^TU^{(a)}=I} -2\langle(\Phi^{\tau+1})^T\hat{s}^{(a)}\Phi^{\tau+1}U^{(a)\tau}\Lambda^{\tau+1}D^{(a)\tau+1}, U^{(a)}\rangle + \sum_{k=1}^{K_X}\frac{\lambda_k}{2}\|U_k^{(a)} - e_k\|_2^2 \\
&\quad + \frac{\rho}{2}\|U^{(a)} - U^{(a)\tau}\|_F^2 \\
&= \arg\min_{(U^{(a)})^TU^{(a)}=I} -2\langle H, U^{(a)}\rangle = \arg\min_{(U^{(a)})^TU^{(a)}=I}\|H - U^{(a)}\|_F^2, \quad a = 1, \ldots, M,
\end{aligned}$$

where $H = (\Phi^{\tau+1})^T\hat{s}^{(a)}\Phi^{\tau+1}U^{(a)\tau}\Lambda^{\tau+1}D^{(a)\tau+1} + U^{(a)\tau}\rho/2 + \mathrm{diag}(\lambda_1, \ldots, \lambda_{K_X})/2$. Let $H = \tilde{Q}\tilde{D}\tilde{P}^T$ be the SVD decomposition of H. Then the solution is $U^{(a)\tau+1} = \tilde{Q}\tilde{P}^T$.

Algorithm 2 Parameter initialization for the model (C.5)

Input: $\hat{s}^{X(a)}$, $a = 1, \ldots, M$
Output: Intialized parameters $\Phi^0, \Lambda^0, \{D^{(a)0}, U^{(a)0}\}_{a=1}^M, b^0$
1. $\bar{s} = \sum_{a=1}^M \hat{s}^{X(a)}/M$
2. Let $\bar{s} = \tilde{\Phi}\tilde{\Lambda}(\tilde{\Phi})^T$ be the eigendecomposition of $\bar{s}$. Initialize $\Phi^0 = \tilde{\Phi}$.
3. For $a = 1$ to $a = M$
   Let $\hat{s}^{X(a)} = \tilde{\Phi}^{(a)}\tilde{\Lambda}^{(a)}(\tilde{\Phi}^{(a)})^T$ be the eigendecomposition of $\hat{s}^{X(a)}$.
   Let $\tilde{\Phi}^T\tilde{\Phi}^{(a)} = \tilde{Q}^{(a)}\tilde{D}^{(a)}(\tilde{P}^{(a)})^T$ be the SVD decomposition of $\tilde{\Phi}^T\tilde{\Phi}^{(a)}$.
   Let $\mathrm{diag}(\tilde{D}^{(a)0}) = \mathrm{diag}(\tilde{\Lambda}^{(a)})/\mathrm{diag}(\tilde{\Lambda})$.
   Initialize $U^{(a)0} = \tilde{Q}^{(a)}(\tilde{P}^{(a)})^T$.
4. Let $c = \sum_{a=1}^M \tilde{D}_{11}^{(a)0}/M$.
5. Initialize $D^{(a)0} = \tilde{D}^{(a)0}/c$, $a = 1, \ldots, M$, $b^0 = (D_{11}^{(1)0}, \ldots, D_{11}^{(M)0})^T$.
6. Initialize $\Lambda^0 = c\tilde{\Lambda}$.

In practice, we use orthogonal polynomials of degree 15 as the basis functions. We adopt proportional tuning parameters λ_k and ν_k by setting $\lambda_k = \lambda_0\Lambda_{kk}^0$ and $\nu_k = \nu_0\Lambda_{kk}^0$, where Λ_{kk}^0 is the (k,k)-element of the initial Λ^0.

D Simple standardization and quantile normalization approaches

To make the simple standardization and the quantile normalization approaches comparable with INTACT and INTACT$_0$, we apply them to the raw data at each timepoint.

D.1 Simple standardization

For each t, denote by $\mathcal{A}^{(a)}(t)$ the set of $\hat{\eta}_{ij}^{(a)}$ (7) in domain (a), i.e., $\mathcal{A}^{(a)}(t) = \{\hat{\eta}_{ij}^{(a)},\ i = 1, \ldots, n_a,\ j = 1, \ldots, r_i\}$. Denote by $e(t)$ the variance of data in $\cup_a \mathcal{A}^{(a)}(t)$, $e^{(a)}(t)$ the variance of data in $\mathcal{A}^{(a)}(t)$. The simple standardized data are defined by

$$\hat{Y}_{ij}^{s(a)}(t) = \hat{\alpha}(t) + Z_{ij}^T \hat{\beta}(t) + \hat{\eta}_{ij}^{s(a)}(t),$$

where $\hat{\eta}_{ij}^{s(a)}(t) = \hat{\eta}_{ij}^{(a)}(t)\sqrt{e(t)/e^{(a)}(t)}$.

D.2 Quantile normalization

For each t, we use the normalize function from the R package broman on $(\mathcal{A}^{(1)}(t), \ldots, \mathcal{A}^{(M)}(t))$. If the lengths of $\mathcal{A}^{(1)}(t), \ldots, \mathcal{A}^{(M)}(t)$ differ, we first apply linear interpolation to equalize their lengths before passing them into the function. After normalization, we then linearly interpolate the output to restore it to its original length and denote by $\hat{\eta}_{ij}^{q(a)}(t)$. The quantile normalized data are defined by

$$\hat{Y}_{ij}^{q(a)}(t) = \hat{\alpha}(t) + Z_{ij}^T \hat{\beta}(t) + \hat{\eta}_{ij}^{q(a)}(t).$$

E Association between mortality and minute-level physical activity

To examine the association between mortality and minute-level physical activity, we apply a Cox model to log-transformed minute-level physical activity intensities, adjusted for gender, age, income, and BMI, using a 5-year censoring period. The model is implemented using the `coxph()` function from the R package `survival`. For participants with multiple days of measurements, we average the log-transformed activity intensities across days. We apply separate Cox models to the activity intensity at each minute. Figure 3 shows the hazard ratios and their corresponding Bonferroni-corrected 99.997%(=1-0.05/1440) confidence intervals for each minute-level physical activity intensity. All methods reveal significant results from approximately 8:00 to 20:00, with hazard ratios below 1, indicating that daytime physical activity significantly reduces mortality risk. From about 22:00 to 6:00, hazard ratios for INTACT and INTACT$_0$ exceed 1, while those for other methods hover around or exceed 1. Only INTACT and INTACT$_0$ show significant results during this period, suggesting that nighttime activity may negatively impact lifespan.

References

Amaratunga, D. and J. Cabrera (2001). Analysis of data from viral dna microchips. *Journal of the American Statistical Association 96*(456), 1161–1170.

Beer, J. C., N. J. Tustison, P. A. Cook, C. Davatzikos, Y. I. Sheline, R. T. Shinohara, K. A. Linn, A. D. N. Initiative, et al. (2020). Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage 220*, 117129.

Benito, M., J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron (2004). Adjustment of systematic microarray data biases. *Bioinformatics 20*(1), 105–114.

Bolstad, B. M., R. A. Irizarry, M. Åstrand, and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19*(2), 185–193.

Butler, A., P. Hoffman, P. Smibert, E. Papalexi, and R. Satija (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology 36*(5), 411–420.

Chen, A. A., J. C. Beer, N. J. Tustison, P. A. Cook, R. T. Shinohara, H. Shou, and A. D. N. Initiative (2022). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human brain mapping 43*(4), 1179–1195.

Cosoli, G., S. Spinsante, and L. Scalise (2020). Wrist-worn and chest-strap wearable devices: Systematic review on accuracy and metrological characteristics. *Measurement 159*, 107789.

Crainiceanu, C. M., J. Goldsmith, A. Leroux, and E. Cui (2024). *Functional Data Analysis with R* (1st ed.). Chapman and Hall/CRC.

Cui, E., C. M. Crainiceanu, and A. Leroux (2021). Additive functional cox model. *Journal of Computational and Graphical Statistics 30*(3), 780–793.

Cui, E., R. Li, C. M. Crainiceanu, and L. Xiao (2023). Fast multilevel functional principal component analysis. *Journal of Computational and Graphical Statistics 32*(2), 366–377.

Hansen, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *New Functions for Multivariate Analysis 7*(2), 18–24.

Harrison, D., P. Marshall, N. Bianchi-Berthouze, and J. Bird (2015). Activity tracking: barriers, workarounds and customisation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 617–621.

Hie, B., B. Bryson, and B. Berger (2019). Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology 37*(6), 685–691.

Johnson, W. E., C. Li, and A. Rabinovic (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics 8*(1), 118–127.

Peng, M., Y. Li, B. Wamsley, Y. Wei, and K. Roeder (2021). Integration and transfer learning of single-cell transcriptomes via cfit. *Proceedings of the National Academy of Sciences 118*(10), e2024383118.

Scalise, L. and G. Cosoli (2018). Wearables for health and fitness: Measurement characteristics and accuracy. In *2018 IEEE international instrumentation and measurement technology conference (I2MTC)*, pp. 1–6. IEEE.

Shah, Y., J. Dunn, E. Huebner, and S. Landry (2017). Wearables data integration: Data-driven modeling to adjust for differences in jawbone and fitbit estimations of steps, calories, and resting heart-rate. *Computers in Industry 86*, 72–81.

Shou, H., V. Zipunnikov, C. M. Crainiceanu, and S. Greven (2015). Structured functional principal component analysis. *Biometrics 71*(1), 247–257.

Vooris, R., M. Blaszka, and S. Purrington (2019). Understanding the wearable fitness tracker revolution. *International Journal of the Sociology of Leisure 2*, 421–437.

Wang, S., L. Qian, and R. J. Carroll (2010). Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika 97*(1), 79–93.

Welch, J. D., V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell 177*(7), 1873–1887.

Wrobel, J., J. Muschelli, and A. Leroux (2021). Diurnal physical activity patterns across ages in a large uk based cohort: the uk biobank study. *Sensors 21*(4), 1545.

Xiao, L., L. Huang, J. A. Schrack, L. Ferrucci, V. Zipunnikov, and C. M. Crainiceanu (2015). Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics 16*(2), 352–367.

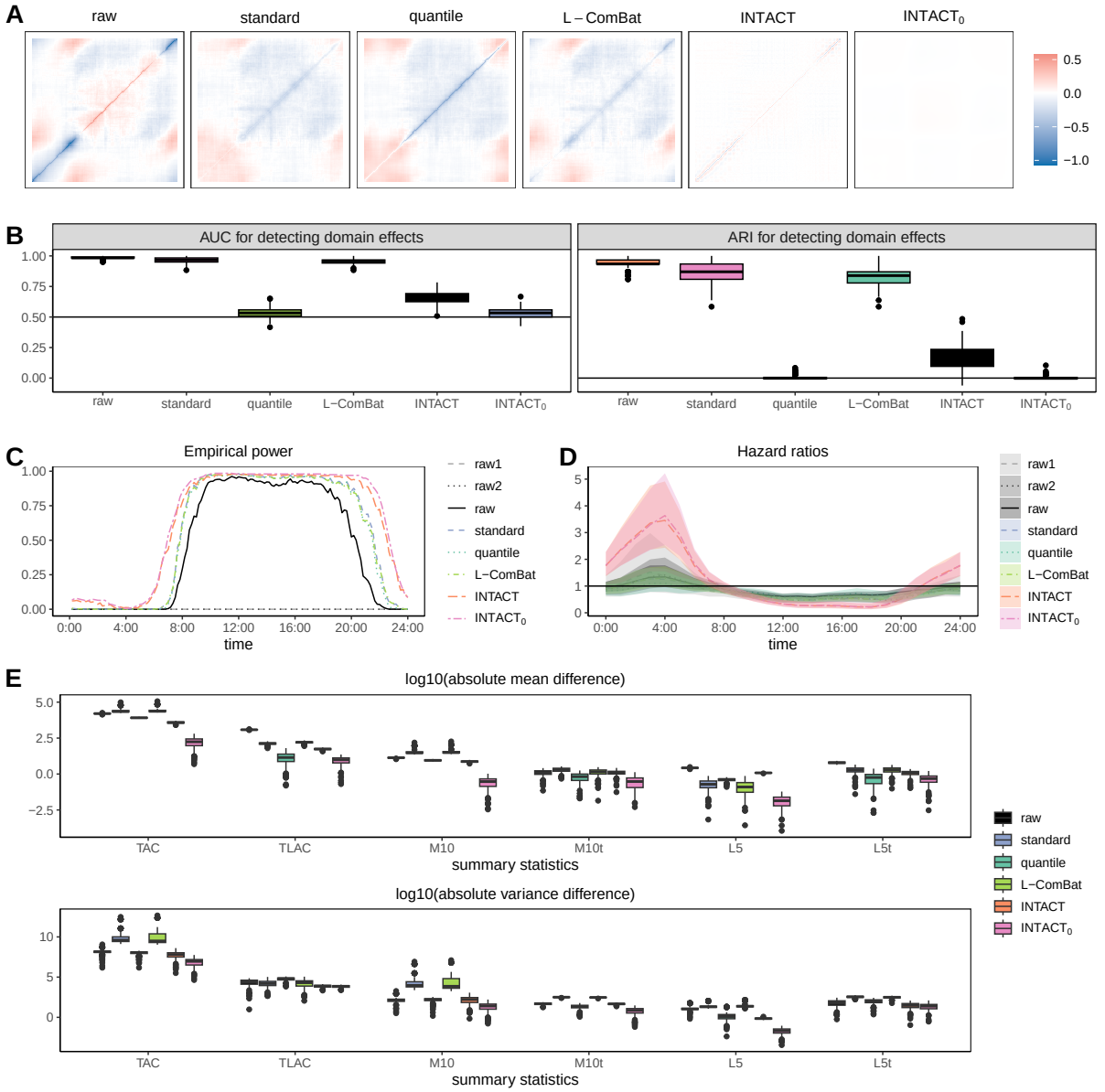

Figure 2: (A–E) Integration results of NHANES physical activity data. Abbreviations used in this figure are as follows: raw data (raw), simple standardization (standard), quantile normalization (quantile), longitudinal ComBat (L-ComBat), proposed integration methods (INTACT and INTACT$_0$), NHANES 2003–2004 data (raw1), and NHANES 2013–2014 data (raw2). (A) Heat map of $Var_1 - Var_2$. (B) Detection of domain effects. Left: box plots of AUC values; Right: box plots of ARI values. (C) Empirical power for testing the null hypothesis of no age effect, based on resampling 120 subjects per domain. (D) Hazard ratios and corresponding Bonferroni-corrected confidence intervals from Cox model analysis. The black solid line indicates a hazard ratio of 1. (E) Box plots of absolute mean and variance differences in summary statistics between two domains.

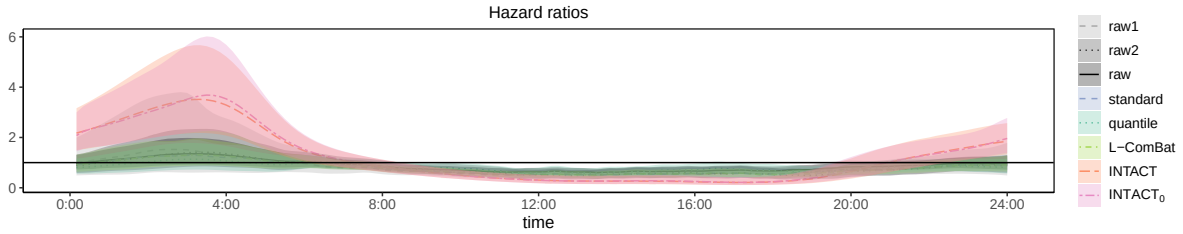

Figure 3: Hazard ratios and corresponding Bonferroni-corrected confidence intervals from Cox model analysis. The black solid line indicates a hazard ratio of 1.