Modeling Temporal Variation in Physical Activity Using Functional Principal Components Analysis

Selene Yue Xu, et al. *[full author details at the end of the article]*

Received: 14 December 2017 / Revised: 26 February 2019 / Accepted: 2 April 2019 /
Published online: 12 April 2019
© The Author(s) 2019

Abstract

Accelerometers are person-worn sensors that provide objective measurements of movement based on minute-level activity counts, thus providing a rich framework for assessing physical activity patterns. New statistical approaches and computational tools are needed to exploit these densely sampled time-series data. We implement a functional principal component mixed model approach to ascertain temporal activity patterns in 578 overweight women (60% cancer survivors) and summarize individual patterns with unique personalized principal component scores. We then test if these patterns are associated with health by performing multiple regression of health outcomes (including biomarkers, namely, insulin, C-reactive protein, and quality of life) on activity patterns represented by these scores. Our model elucidates the most important patterns/modes of variation in physical activities. Results show that health outcomes including biomarkers and quality of life are strongly associated with the total volume, as well as temporal variation in activity. In addition, associations between physical activity and health outcomes are not modified by cancer status. Our findings suggest that employing a multilevel functional principal component analysis approach can elicit important temporal patterns in physical activity. It further allows us to study the relationship between health outcomes and activity patterns, and thus could be a valuable modeling approach in behavioral research.

Keywords Accelerometer · Functional method · Principal component analysis · Physical activity patterns

SYX, SG, and JK were partially supported by a Grant from the UCSD Institute for Public Health; SG, JK, REP, CLR, DDS, and LN were partially supported by funding from the National Cancer Institute (U54 CA155435-01); LN was partially supported by funding from the National Institute of Aging (PO1AG052352). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

✉ Loki Natarajan
lnatarajan@ucsd.edu

Extended author information available on the last page of the article

1 Introduction

Physical inactivity and sedentary behavior are recognized risk factors for many chronic diseases [2,3,25,26], driving research on levels of physical activity needed to maintain a healthy lifestyle and prevent disease. Accurate measurement of physical activity is critical for designing and implementing interventions aimed at modifying this important behavior. Accelerometers, person-worn sensors that measure physical activity based on activity counts derived from accelerations, have become the norm in physical activity measurement. Use of accelerometers has many advantages, the most commonly cited of which is that they may be less prone to the biases associated with self-report [18,22]. A less recognized advantage is that accelerometers record minute-level (or even second-level) activity, and hence provide a rich and objective framework for analyzing physical activity patterns.

The majority of research utilizing accelerometers has focused on aggregate or summary statistics such as daily average activity or weekly percentage of sedentary/vigorous activity. While aggregate levels may provide a measure of average habitual physical activity, there is also much to be gleaned from a more thorough analysis of temporal activity patterns throughout the day. Analysis of these patterns would potentially allow a more refined understanding of the implications of the circadian rhythms of exercise for health, and possibly identify critical windows-of-opportunity for intervening to increase physical activity or reduce sedentary behavior. To date, the analysis of temporal patterns of physical activity has generally divided the day into, potentially, arbitrary time segments which may miss important aspects of the temporal pattern. More recently functional analysis techniques such as wavelet-based or principal component functional models [14,15,24,27] have been proposed to study the full spectrum of accelerometer data. These approaches model minute-level information, rather than reducing the data to daily or weekly summaries, thus exploiting these densely sampled measures.

The use of functional data analysis is gradually becoming more common in the analysis of activity patterns in children and adolescents [9,11,23], where a day structured around school hours makes assessment of circadian activity patterns more tangible. Among adults, a recent study by [28] applied the functional principal component analysis technique to a cohort of older men to examine associations between diurnal patterns of accelerometry measured physical activity with sleep, cognitive function, and mortality. The study used accelerometry measures from several days to calculate an “average day” of activity pattern at the minute level which was then subjected to the functional principal components analysis. The study found specific patterns of daily activity to be associated with worse sleep, decreased cognition, and greater mortality. Methods have also been proposed to accommodate and analyze specific experimental designs [13].

In this article, we implemented a functional principal components mixed model [6] which uses dimension reduction via principal components to model the minute-level activity counts. Furthermore, unlike the approach in [28] which averaged an individual’s daily activity records, the use of mixed models allows inclusion of repeated daily records in the model. We used this model to characterize subject-to-subject variation in activity patterns throughout the day, and examine associations between these activity

patterns and health outcomes. We leveraged data from two large randomized trials: MENU a randomized diet intervention among 245 overweight women and Reach for Health a trial examining the use of metformin and a lifestyle intervention in 333 overweight, postmenopausal, breast cancer survivors. Use of these two trials provided us with accelerometer data on 578 overweight women, a rich array of health outcomes including biomarkers and quality of life. Further, because 60% of the study sample were cancer survivors, we were able to test whether activity patterns differed by cancer status, and whether cancer status moderated the associations between physical activity and health outcomes.

2 Study Sample

Our study sample comprised 578 overweight women who participated in the NIH-funded Transdisciplinary Research on Energetics and Cancer (TREC) Study at UCSD (2011–2017). The primary aim of these studies was to enhance knowledge on the role of obesity and energetics on cancer risk [16]. To this end, two randomized controlled weight-loss interventions were conducted: the MENU trial ($N = 245$) in overweight otherwise healthy women, and the Reach for Health trial ($N = 333$) in postmenopausal overweight breast cancer survivors. In both trials, physical activity was assessed via 7-day accelerometry. We used baseline data from these trials in the current project. Details on the study design and protocols of these studies have been previously published and are briefly described below.

The MENU trial was a 12-month behavioral intervention study among 245 overweight non-diabetic women to investigate the role of dietary macronutrient composition on weight loss [12,21]. Participants were randomized to one of three diets: a lower fat (20% of energy) and higher carbohydrate (65% of energy) diet; a lower carbohydrate (45% energy) and higher monounsaturated fat (35% energy) diet; or a walnut-rich (35% fat) and lower carbohydrate (45%) diet. The Reach for Health (RfH) Study was a 2×2 six-month randomized trial of 333 overweight, postmenopausal early-stage breast cancer survivors, aiming to test the impact of metformin treatment alone, a lifestyle-based intervention alone, both or neither on weight loss and biomarkers associated with cancer risk [17].

Using similar data collection procedures, both trials collected information on demographic, lifestyle, and psychosocial factors. Height and weight were measured at clinic visits and used to calculate BMI. Physical activity and sedentary behavior were obtained via a triaxial accelerometer, the GT3X Actigraph monitors (ActiGraph, LLC; Pensacola, FL), which is set to collect data at 30 Hz [1]. Self-report questionnaires were used to elicit demographic information, and the validated SF-36 scale was used to ascertain quality-of-life scores [4]. Fasting blood samples were drawn at clinic visits and used to assay glucoregulatory and inflammatory markers. Details on tumor characteristics and treatment were abstracted from medical records in the RfH Study.

Fig. 1 A typical day of accelerometer data

3 Statistical Framework

3.1 Overview

Accelerometer-based activity counts form irregular functions characterized by peaks of varying frequencies and locations. Output for a single day is displayed in Fig. 1. On this day, this participant wore the accelerometer for 863 min from 8:22 am to 10:44 pm. 591 min were spent in sedentary time (i.e., activity counts < 100 [10]) and 7 min in moderate to vigorous physical activities (i.e., activity counts > 1952 [10]). To handle these high-dimensional time-series data, we used principal component methods to reduce dimension. Also, our studies collected accelerometer measures on multiple days for each participant. To account for the hierarchical nature of these data (days within subjects), we used multilevel functional mixed models. Specifically, we used a two-level functional principal component analysis model as proposed by [6]. We briefly describe the mathematical underpinnings of this approach below; details can be found in [6].

3.2 The Model

As a first step, we will decompose each daily activity record as follows:

$$X_{ij}(t) = \mu(t) + Z_i(t) + W_{ij}(t),$$

where $X_{ij}(t)$ is the accelerometer count function measured at time t on the j th day for subject i . $\mu(t)$ represents the overall population mean acceleration function at time t .

$Z_i(t)$ is the subject-specific deviation from the overall mean function. $W_{ij}(t)$ is the residual subject- and day-specific deviation from the subject-mean function. Note, $Z_i(t)$ represents the subject level activity curve that will enable us to characterize subject-to-subject variation in activity. We call $Z_i(t)$ level 1 functions, and $W_{ij}(t)$ level 2 functions. This model specifies the 2-level hierarchical structure.

Next, to account for the minute-level high-dimensional data, we will further decompose level 1 and level 2 functions through the Karhunen–Loeve expansion [19]

$$Z_i(t) = \sum_k \xi_{ik} \phi_k^{(1)}(t)$$

$$W_{ij}(t) = \sum_m \beta_{ijm} \phi_m^{(2)}(t),$$

where ξ_{ik} and β_{ijm} are level 1 and level 2 principal component scores, and $\phi_k^{(1)}(t)$ and $\phi_m^{(2)}(t)$ are level 1 and level 2 eigenfunctions. Substituting these into our original model, we obtain

$$X_{ij}(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k^{(1)}(t) + \sum_{m=1}^{\infty} \beta_{ijm} \phi_m^{(2)}(t),$$

where $\mu(t)$, $\phi_k^{(1)}(t)$, and $\phi_m^{(2)}(t)$ are fixed functional effects, and ξ_{ik} and β_{ijm} are random variables with mean zero. The principal component scores, ξ_{ik} , can be used to distinguish temporal activity patterns between individuals, and are hence the key focus of the current work. The model is built under the following assumptions [6]:

- A.1 $E(\xi_{ik}) = 0$, $var(\xi_{ik}) = \lambda_k^{(1)}$, for any i , $k_1 \neq k_2$, $E(\xi_{ik_1} \xi_{ik_2}) = 0$;
- A.2 $\{\phi_k^{(1)}(t) : k = 1, 2, \dots\}$ is an orthonormal basis of $L^2[0, 1]$;
- A.3 $E(\beta_{ijm}) = 0$, $var(\beta_{ijm}) = \lambda_m^{(2)}$, for any i, j , $m_1 \neq m_2$, $E(\beta_{ijm_1} \beta_{ijm_2}) = 0$;
- A.4 $\{\phi_m^{(2)}(t) : m = 1, 2, \dots\}$ is an orthonormal basis of $L^2[0, 1]$;
- A.5 $\{\xi_{ik} : k = 1, 2, \dots\}$ are uncorrelated with $\{\beta_{ijm} : m = 1, 2, \dots\}$.

We will not delve into the details of estimating these scores here except to note that we used eigenfunction decompositions and method of moment methods as detailed in [6].

3.3 Choosing the Number of Principal Components

Selecting the number of principal components involves a trade-off between explaining variation in the data, while reducing artifactual patterns. We want to find the optimal balance between under-fitting and over-fitting. Usually, this number is chosen to ensure that a “sufficient” amount of the level 1 (and level 2) variation in the dataset is explained by the selected principal components. What constitutes “sufficient” varies by application and context. In our application we will choose enough level 1 components to explain 50% of the between-subject variation.

3.4 Derived Features

As we will demonstrate, each level-1 eigenfunction, $\phi_k^{(1)}(t)$, for $k = 1, 2, \dots$ can be used to distinguish a temporal activity pattern. For each eigenfunction, the model also provides level 1 principal component scores ξ_{ik} , $k = 1, 2, \dots$ for each subject i , which quantify the extent to which subject i subscribes to the temporal pattern associated with the corresponding eigenfunction. Thus these scores can be used to test if variation in a physical activity pattern defined by a particular eigenfunction is associated with health outcomes.

Another potentially useful statistic is the L_2 norm of the level 1 principal component scores, which for subject i is defined as $\sqrt{\sum_k \xi_{ik}^2}$. We hypothesized that a higher L_2 norm would be associated with a healthier profile. This is based on evidence [7,8] that it is better to exercise multiple times in a day in short bouts of time, rather than to exercise just once for a long stretch of time. There has been growing literature investigating how patterns of sedentary and active time accumulation are associated with all-cause mortality independently of total sedentary and total active times. Studies show that a larger proportion of longer sedentary bouts are positively associated with mortality, suggestive that physical activity guidelines should target not only reducing but also interrupting sedentary time to reduce mortality risk [7,8].

A measure that could distinguish short bouts of exercise versus a single long exercise bout is the total variation of the accelerometer profile, defined as the integral of the absolute value of the derivative of the activity functions, i.e., $\int |X'_{ij}(t)|dt$. We hypothesized that the L_2 norm of the level 1 PC scores would also capture the total variation of daily activity, based on the following rationale. Recall the subject-specific deviation function $Z_i(t) = \sum_k \xi_{ik} \phi_k^{(1)}(t)$, and suppose that the k -th principal component score, ξ_{ik} has a large (absolute) value. This indicates that the corresponding eigenfunction, $\phi_k^{(1)}(t)$, gets multiplied by a coefficient of large magnitude. Thus if the eigenfunction $\phi_k^{(1)}(t)$ represented a pattern with highs and lows at different times of the day, these will get magnified by the large ξ_{ik} . Now, one potential way for the L_2 measure to have a large value is if there are several level 1 principal component scores with moderate-large values. This would result in multiple magnified eigenfunctions getting superimposed to produce greater and more frequent oscillations in the activity profile.

To visually illustrate this hypothesis, we simulated two daily profiles with the same average physical activity but different L_2 norms. The comparison is shown in Fig. 2. The blue curve represents a profile with $\xi_1 = 1000$ and $\xi_k = 0$ for $k > 1$. The green curve represents a profile with $\xi_1 = 1000$, $\xi_2 = 500$, $\xi_3 = 500$, $\xi_4 = -3300$, and $\xi_k = 0$ for $k > 4$. The two daily activity profiles have the same average activity but the green profile clearly has larger L_2 norm than the blue profile. To analytically justify the L_2 norm, we calculated the total variation of each daily profile in our sample using the above definition, and compared the total variation to the L_2 norm. The correlation between the L_2 norm of the level 1 PC scores and total variation of daily profiles was 0.70 (p -value < 0.005), suggesting that the L_2 norm provides a reasonable assessment of total variation. In fact, if we define total variation only based on level 1 PC functions

Fig. 2 Compare two simulated daily activity profiles with the same average activity but different L_2 norms

Fig. 3 Compare a randomly chosen daily profile of the person with the largest L_2 norm of the level-1 PC scores to a randomly chosen daily profile of the person with the smallest L_2 norm of the level-1 PC scores

and ignore the residual function $W_{ij}(t)$, i.e., total variation $= \int |\mu'(t) + Z'_i(t)|dt$, the correlation between the L_2 norm of the level 1 PC scores and total variation defined this way was even higher at 0.83 (p value < 0.005). These results motivated our use of the L_2 norm as it is more easily derived than total variation. Finally as an illustration of the usefulness of these L_2 norms in our application, we plotted the activity profile of the two participants with the largest and smallest L_2 norms in Fig. 3.

We emphasize that the L_2 measure is a level 1 measure, and there may be many ways to attain a large L_2 -value, of which the above interpretation is but one.

4 Statistical Analysis

4.1 Data Processing

As mentioned in Sect. 2, objective physical activity in our study was assessed via the GT3X Actigraph, which is a triaxial lightweight accelerometer approximately 2"X2"X1" in size. The Actigraph GT3X+ monitor was set to collect acceleration data at 30 Hz. The ActiLife program applied a band-pass filter to remove non-human acceleration signal from the data and then summarized the signal to counts per minute using a proprietary algorithm [1]. The magnitude of the count is related to intensity of the activity [1]. The device has been validated and calibrated for use in both controlled and field conditions [1]. Participants were provided with written protocols for best positioning of the device and instructed to wear it on the hip for 7 days during all waking hours, except for when in contact with water. Non-wear time was identified via pre-defined algorithms of consecutive zero counts using standard protocols [5] and labeled as missing data.

To account for varying start and stop wear times and to maximize information, we retained days with 12 or more hours of consecutive wear time, and only used these “complete profiles” in our analysis. The set of complete profiles consisted of 570 participants, out of the 578 originally recruited, so that < 1% of the subject sample was discarded. Out of the 4475 original daily profiles, 1062 (24%) were discarded, leaving a total of 3413 daily records for our analysis. The number of days per participant in the resulting data set ranges from 1 to 14 and has a mean (SD) of 6.0 (1.8). Lastly, we aligned all daily records by the first minute of device wear, thus allowing each participant to start her daily activity at different times of the day. On average, participants started wearing the accelerometers at approximately 7:30 am (SD 85 min). In the following analysis, we used the first 720 points (i.e., 12 h) of each record.

4.2 Temporal Activity Patterns and Health Outcomes

We derived the principal component eigenfunctions and scores, as described in the previous sections. We first compared each activity pattern between the cancer (RfH) and non-cancer (MENU) samples by performing two-sample *t* tests on the level 1 principal component scores associated with this activity pattern. Next, to examine associations between temporal activity patterns and health outcomes, we developed multiple regression models of health outcomes on activity patterns represented by the level 1 principal component scores. We adjusted for age at study entry, education, BMI, smoking status, and cancer status. Additionally we tested the interaction between cancer status and activity patterns, in particular, whether having cancer changes the relationship between health outcomes and activity patterns. More specifically, we fit the following regression models:

- 1 health outcome = $\beta_0 + \beta_1 * PC1 + \beta_2 * PC2 + \beta_3 * PC3 + \beta_4 * PC4 +$
other covariates
- 2 health outcome = $\beta_0 + \beta_1 * L_2 +$ other covariates.

Here $PC1$, $PC2$, $PC3$, and $PC4$ refer to the first, second, third, and fourth level 1 PC scores. L_2 refers to the L_2 norm of the level 1 PC scores. Other covariates included were age, education (1 if college and above; 0 otherwise), BMI, smoke (1 if ever smoked; 0 if never smoked), and cancer status (1 if yes; 0 if no). Health outcomes included insulin, C-reactive protein (CRP), homeostatic model assessment (HOMA), insulin status (1: $HOMA \geq 3$, i.e., insulin resistant (IR), 0: $HOMA < 3$, i.e., insulin sensitive (IS)), physical quality of life (QoL), and mental quality of life (QoL). We used linear regression models for the continuous outcomes (biomarkers, QoL) and logistic regression for insulin status. Biomarkers were log-transformed to better approximate the assumption of Gaussian residuals.

To study interactions between cancer and activity patterns, we fit the following models:

$$3 \text{ health outcome} = \beta_0 + PC1 * cancer + PC2 * cancer + PC3 * cancer + PC4 * cancer + \text{other covariates}$$

$$4 \text{ health outcome} = \beta_0 + L_2 * cancer + \text{other covariates}.$$

Here * represents interaction between two terms. The first two regression models investigate the relationships between activity patterns and health outcomes. The last two regression models test if these associations differ by cancer status. Regression coefficients with 95% confidence intervals and corresponding p values are presented. Importantly, $PC1$, $PC2$, $PC3$, $PC4$ were all included in the same model allowing us to examine the effect of one temporal pattern, while controlling for others. All analyses used the statistical software package R 3.3.2 [20].

5 Results

5.1 Study Sample Descriptives

Sample characteristics are provided in Table 1. The MENU Study randomized 245 overweight women to one of three diet arms. Women in the MENU Study were on average 50 years old ($SD = 10$) at study entry, age range 22–72 years. Mean (SD) years of education was 15.6(2.2). Race/ethnicity was self-reported as white non-Hispanic 74.3%, African-American 5.7%, Hispanic 17.1%, Asian 1.6%, Native American 0.4%, and mixed/other race 0.8%.

The RfH Study randomized 333 breast cancer survivors to one of four metformin (Y/N) $\times$ lifestyle intervention (Y/N) study arms. Women were on average 63 years old ($SD = 6.9$) at study entry, with a mean 2.7 ($SD = 2$) years from their breast cancer diagnosis; 51% had a college degree; the majority race was White (83%), with 4% African-American, 2% Asian, and 11% of mixed/other race; 11% reported Hispanic ethnicity.

The RfH cohort was older than the MENU participants, which is as expected since the RfH study recruited postmenopausal women. In both cohorts, more than half the sample had received a college degree. Importantly, biomarker and quality-of-life outcomes were similar between the two cohorts. Also, at study entry, women in the MENU and RfH studies had similar daily physical activity and sedentary time

Table 1 Characteristics of the two study cohorts (mean(SD) or percentages)

	RfH study	MENU study
Age	63.1 (6.9)	50.8 (10.0)
BMI	31.2 (5.0)	34.2 (3.4)
Education	51.7% college and above; 48.3% o.w	56.0% college and above; 44.0% o.w
Smoke	45.2% Yes; 54.7% No	27.6% Yes; 70.0% No; 2.5% don't know
Cancer	48% stage 1; 35% stage 2; 17% stage 3	None
Insulin ^a (pg/mL)	464 (335, 659)	484 (363, 726)
CRP ^a (10 ⁶ pg/mL)	3.1 (1.5, 6.5)	3.4 (1.5, 5.7)
HOMA ^a	2.89 (2.05, 4.14)	3.05 (2.25, 4.64)
Insulin Status ^b	53.8% IS; 46.2% IR	48.6% IS; 51.4% IR
Physical QOL	46.4 (8.8)	50.6 (7.2)
Mental QOL	51.2 (9.4)	50.4 (8.7)

^aMedian (1st quartile, 3rd quartile)^bIS insulin sensitive; IR insulin resistant**Table 2** Mean PC scores and cancer status

	Cancer (RfH study)	Controls (MENU study)	<i>t</i> -statistics	<i>p</i> -value
PC1	− 0.16	0.02	− 0.65	0.52
PC2	0.14	− 0.16	1.99	0.047
PC3	0.02	0.05	− 0.20	0.84
PC4	− 0.10	0.12	− 2.09	0.04
MVPA (min/d)	17.5	18.2	− 0.31	0.75
Sed Time (min/d)	451	441	1.18	0.24

estimates, spending approximately 18 min/d on moderate–vigorous activity and 7.5 hours being sedentary (Table 2)

5.2 Temporal Patterns

We fit the two-level hierarchical principal components model described in Sect. 3 to our study sample. Figure 4 illustrates the first four level 1 principal component (PC) eigenfunctions. As expected, the first level 1 PC curve represents an overall vertical shift of the mean activity curve, and explained 30% of the between-subject variation in the sample. This component captures total activity volume, so that a participant with a high score on this component is on average more physically active than one with a lower score. The second level 1 PC curve explained 10% of the variation, and emphasizes variations in the very early parts of the day. The third and fourth level 1 PC curve together explained ~ 12% and further parse timing of activity, reflecting variation contrasts between morning activity and activity throughout the rest of the day for the third PC, and variation in the middle and later parts of the day for the fourth PC.

Fig. 4 (MENU + RfH) The mean intensity curve and the effects of adding (red) and subtracting (blue) a suitable multiple of the first four level-1 PC curves (Color figure online)

Fig. 5 Smoothed heatmap of the correlation matrix between activity counts at any two time points during the day. X and Y axis represent time (min) since the beginning of the day (or since the subject puts on the accelerometer)

The PCfunctions and PC scores are estimated via the covariance matrix. The entry at the i th row and the j th column in this matrix is the empirical covariance value between the activity count at time i and that at time j . Figure 5 shows a smoothed heatmap of this covariance matrix after standardizing the rows and columns, so that it

Fig. 6 Compare the mean curves of the cancer population and the non-cancer population

is now a correlation matrix. We observed high correlation around the diagonal axis on this heatmap, as might be expected since activities that occur close in time tend to be similar in magnitude. These diagonal correlations were highest in the earliest and latest intervals, indicating highly consistent patterns at these times. Lowest correlations are observed between activities that are roughly 3 hours apart. Interestingly, we observed a slightly higher correlation between activities in the intervals 50–200 min and 500–650 min possibly due to a routine patterns at the start (pre-work) and end (post-work) of the day.

5.3 Physical Activity Patterns: Cancer Versus Non-cancer

We first investigated the differences in physical activity patterns between cancer survivors (RfH participants) and subjects without cancer (MENU participants). Figure 6 plots the mean activity function and Table 2 provides the summary statistics and two-sample *t*-test *p* values comparing the first four level 1 PC scores between the cancer and non-cancer groups. The plots show similar overall volume of activity as is borne out by the non-significant difference in the PC1 scores. However, there do appear to be nominal differences in temporal activity patterns between the cancer and non-cancer populations. In general, cancer survivors had significantly (at the 5%) higher PC2 scores and lower PC4 scores. The higher PC2 scores indicate that the cancer group tended to have lower morning activity, whereas the lower PC4 scores imply that this group tends to have higher mid-day activity and lower evening activity. These patterns are also observable in Fig. 6.

Given that participants in the RfH Study were older than MENU Study participants, we repeated the above analysis adjusting for age. The difference in PC2 scores by cancer status remained significant at the 5% level, although the PC4 scores were no longer different between groups.

5.4 Physical Activity and Health

5.4.1 Biomarkers

Table 3 shows the results of multiple linear regression models (as specified in the first and second model in Sect. 4) examining associations between physical activity patterns (i.e., PC1-PC4 scores and L_2 -score) and insulin, CRP, and HOMA, after adjusting for BMI, age, education, smoking, and cancer status.

BMI was the only covariate significantly associated with insulin level in both models. Each unit higher BMI was associated with roughly 0.04 higher log(insulin) level. In the first model, the PC1 component, a measure of physical activity volume, was strongly negatively associated with insulin level. None of the other PCs were associated with insulin. In the second model, the L_2 norm of level 1 PC scores showed significant negative association with insulin level. The interaction between cancer and PCs was non-significant (likelihood ratio test $p = 0.99$), neither was the interaction between cancer and L_2 norm (likelihood ratio test $p = 0.8$).

BMI and cancer status were significantly associated with CRP level in both models. Each unit higher BMI was associated with roughly 0.09 higher log(CRP) level; cancer survivors compared to the non-cancer group had on average 0.44 higher log(CRP). In the first model, the PC1 component was marginally negatively associated with CRP level ($p < 0.1$). None of the other PCs were associated with CRP. In the second model, the association between the L_2 norm of level 1 PC scores and CRP level was not statistically significant. The interaction between cancer and PCs was also non-significant (likelihood ratio test $p = 0.38$), neither was the interaction between cancer and L_2 norm (likelihood ratio test $p = 0.16$).

BMI was the only covariate significantly associated with HOMA level in both models. This result was also true when we discretized HOMA into insulin resistance status, with $\text{HOMA} > 3$ being insulin resistant (IR) and $\text{HOMA} \leq 3$ being insulin sensitive (IS), and performed logistic regression of this variable on the PC scores and covariates. Each unit higher BMI was associated with a 0.15 higher HOMA level, and 14% higher odds of being IR. In the first model, the PC1 component was strongly negatively associated with HOMA level (one unit increase in PC1 score was associated with 0.08 unit decrease in HOMA level); the PC3 component was marginally negatively associated with HOMA (one unit increase in PC3 score was associated with 0.11 unit decrease in HOMA level). In the logistic regression, both PC1 and PC3 were significantly negatively associated with insulin resistance status. Each unit higher in PC1 score was associated with a 7% lower odds of being IR, whereas each unit higher PC3 score was associated with a 12% lower odds of being IR. In the second model, the L_2 norm of level 1 PC scores showed significant negative association with HOMA: one unit increase in L_2 norms was associated with 0.13 unit decrease in HOMA level, and a 9% lower odds of being IR.

Table 3 Linear/logistic regression of biomarkers on the first four PC scores or the L_2 norm of PC scores (after adjusted for age, education, smoke, BMI, and cancer status)

Dependent variables	Independent variables	Coefficient estimate	SE	<i>p</i> -value	Confidence interval
Log Insulin	PC1	−0.019	0.006	0.002	(− 0.031, − 0.007)
	PC2	0.004	0.011	0.699	(− 0.018, 0.027)
	PC3	−0.020	0.012	0.103	(− 0.043, 0.004)
	PC4	0.028	0.017	0.110	(− 0.006, 0.062)
	BMI	0.038	0.005	0.000	(0.029, 0.047)
	Cancer	0.061	0.052	0.244	(− 0.042, 0.163)
Log CRP	PC1	−0.023	0.013	0.085	(− 0.048, 0.003)
	PC2	−0.034	0.024	0.168	(− 0.082, 0.014)
	PC3	−0.035	0.026	0.178	(− 0.085, 0.016)
	PC4	−0.017	0.037	0.648	(− 0.089, 0.056)
	BMI	0.092	0.010	0.000	(0.072, 0.112)
	Cancer	0.447	0.111	0.000	(0.228, 0.666)
HOMA	PC1	−0.081	0.028	0.004	(− 0.136, − 0.026)
	PC2	0.014	0.052	0.793	(− 0.089, 0.117)
	PC3	−0.107	0.055	0.053	(− 0.216, 0.001)
	PC4	0.092	0.079	0.244	(− 0.063, 0.248)
	BMI	0.150	0.022	0.000	(0.108, 0.193)
	Cancer	0.277	0.239	0.247	(− 0.192, 0.746)
InsulinStatus (1:IR, 0:IS)	PC1	−0.070	0.028	0.012	(0.882, 0.984)
	PC2	0.004	0.052	0.932	(0.908, 1.114)
	PC3	−0.126	0.055	0.022	(0.791, 0.981)
	PC4	0.057	0.081	0.480	(0.906, 1.243)
	BMI	0.132	0.023	0.000	(1.092, 1.196)
	Cancer	0.204	0.233	0.380	(0.779, 1.943)
Log insulin	L_2	−0.033	0.009	0.000	(− 0.050, − 0.015)
	BMI	0.038	0.005	0.000	(0.029, 0.047)
	Cancer	0.059	0.052	0.254	(− 0.043, 0.161)
Log CRP	L_2	−0.028	0.019	0.156	(− 0.066, 0.011)
	BMI	0.094	0.010	0.000	(0.074, 0.114)
	Cancer	0.432	0.111	0.000	(0.214, 0.651)
HOMA	L_2	−0.130	0.042	0.002	(− 0.211, − 0.048)
	BMI	0.151	0.022	0.000	(0.108, 0.193)
	Cancer	0.266	0.238	0.265	(− 0.202, 0.734)
InsulinStatus (1:IR, 0:IS)	L_2	−0.098	0.044	0.027	(0.828, 0.986)
	BMI	0.133	0.023	0.000	(1.093, 1.196)
	Cancer	0.193	0.231	0.403	(0.772, 1.915)

The parameter estimates for BMI and cancer status are also given. Note the PC scores and L_2 norm have been scaled down by dividing 1000

Table 4 Linear regression of quality of life on the first four PC scores or the L_2 norm of PC scores (after adjusted for age, education, smoke, BMI, and cancer status)

Dependent variables	Independent variables	Coefficient estimate	SE	<i>p</i> -value	Confidence interval
QOL physical	PC1	0.308	0.101	0.002	(0.110, 0.505)
	PC2	−0.111	0.188	0.555	(− 0.480, 0.258)
	PC3	0.047	0.198	0.812	(− 0.342, 0.436)
	PC4	−0.263	0.284	0.353	(− 0.820, 0.293)
	BMI	−0.505	0.078	0.000	(− 0.657, − 0.352)
	Cancer	−4.870	0.855	0.000	(− 6.550, − 3.190)
QOL mental	PC1	0.279	0.115	0.016	(0.053, 0.505)
	PC2	0.020	0.215	0.925	(− 0.403, 0.443)
	PC3	0.277	0.227	0.222	(− 0.168, 0.722)
	PC4	−0.694	0.325	0.033	(− 1.331, − 0.056)
	BMI	−0.011	0.089	0.900	(− 0.186, 0.164)
	Cancer	−1.658	0.980	0.091	(− 3.582, 0.267)
QOL physical	L_2	0.506	0.148	0.001	(0.215, 0.797)
	BMI	−0.503	0.077	0.000	(− 0.654, − 0.352)
	Cancer	−4.904	0.850	0.000	(− 6.573, − 3.235)
QOL mental	L_2	0.411	0.171	0.017	(0.075, 0.746)
	BMI	−0.014	0.089	0.875	(− 0.188, 0.160)
	Cancer	−1.604	0.979	0.102	(− 3.528, 0.319)

The parameter estimates for BMI and cancer status are also given. Note the PC scores here have been scaled down by dividing 1000. Similarly the PC scores used to compute the L_2 norm here have been scaled down by dividing 1000

5.4.2 Quality of Life

Table 4 shows the regression model results (as specified in the first and second model in Sect. 4) of the associations between physical activity and quality of life (physical and mental).

BMI and cancer status were significantly associated with physical QOL in both models. Each unit higher BMI was associated with roughly half a point lower physical QOL; cancer survivors compared to the non-cancer group had on average a five point lower physical QOL. In the first model, the PC1 component was strongly positively associated with physical QOL. None of the other PCs were associated with physical QOL. In the second model, the L_2 norm of level 1 PC scores showed significant positive association with physical QOL. The interaction between cancer and PCs was non-significant (likelihood ratio test $p = 0.88$), neither was the interaction between cancer and L_2 norm (likelihood ratio test $p = 0.42$).

Neither BMI nor cancer status was significantly associated with mental QOL in the models. In the first model, the PC1 component and the PC4 component (which measures variation in the middle and later parts of the day) were strongly associated with mental QOL. None of the other PCs were associated with mental QOL. In the

Table 5 Correlations (p value) between level 1 PC scores and traditional measures of physical activity including average physical activity per minute, number of minutes performing moderate to vigorous activity daily, and number of minutes in sedentary time daily

	Average physical activity	MVPA	Sedentary activity
PC1	0.75 (0.00)	0.68 (0.00)	− 0.17 (0.00)
PC2	0.05 (0.28)	0.00 (0.98)	− 0.07 (0.08)
PC3	0.16 (0.00)	0.03 (0.44)	− 0.17 (0.00)
PC4	0.04 (0.29)	0.07 (0.10)	0.01 (0.75)
L_2 norm of PC scores	0.71 (0.00)	0.67 (0.00)	− 0.18 (0.00)

second model, the L_2 norm of level 1 PC scores showed strong positive association with mental QOL. The interaction between cancer and PCs was non-significant (likelihood ratio test $p = 0.86$), neither was the interaction between cancer and L_2 norm (likelihood ratio test $p = 0.24$).

6 Discussion

In this work we have demonstrated the use of powerful functional principal component analysis to parse physical activity patterns from accelerometer data. Functional data analysis provides a sound mathematical framework for elucidating human physical activity patterns. Instead of looking at activity counts, we can now elicit temporal patterns and study variation in these patterns across individuals. The discussion on the first four level 1 PC functions demonstrates the types of patterns we were able to capture with this method. We could, of course, delve into more intricate patterns by looking beyond the first four PC functions. Importantly, from this model for each activity pattern, we derived a numerical score specific to each subject, thus establishing the foundation for individual-level health analysis.

Conventionally, public health researchers use summary statistics such as average activity count per minute, number of minutes in moderate to vigorous physical activities (MVPA) per day, and number of minutes in sedentary time per day, ignoring the full spectrum of activity profiles provided by accelerometer devices. The functional data model greatly enriches the framework to study physical activities. To illustrate some of the information that was obtained via our model but absent in the traditional physical activity measures, we refer to Table 5. The first level 1 PC scores, which, as noted earlier, measured the total volume/level of activity, was significantly associated with average activity and MVPA with correlations close to 0.7 or higher. Hence this measure does not provide much more information than the conventional measures of activity, in this sample. However, PC 2 to PC 4 scores exhibited low correlation with average activity, MVPA, and sedentary activity. As we discussed in Sect. 5, PC 3 was significantly associated with HOMA level/insulin status, and PC 4 was associated with mental quality of life. Such associations would have been missed in a study that only used standard summary statistics, such as MVPA or sedentary time, from accelerometer readings.

As an aside, in our sample, the L_2 -norm score was highly correlated with total physical activity (and MVPA). In particular, individuals with high L_2 -norm score, also usually had high levels of activity, so that it was not possible to examine the independent effects of the L_2 -metric. However, in general, it is easy to imagine two individuals with similar total activity, yet one has greater total variation in activity over the day by virtue of her being a “fidgeter,” while the other has low variation by virtue of her being more “still.” As noted in Sect. 3, in certain circumstances, the L_2 metric can effectively capture total variation, as illustrated in Fig. 2. We hasten to add, however, that the L_2 -norm is a level 1 measure, and this is only one implication of a large versus small L_2 -norm; there could be other interpretations as well.

Since our sample comprised two studies, one with overweight breast cancer survivors and the other with overweight healthy women, we had the unique opportunity to compare physical activity patterns based on cancer status. Both groups had similar total activity, as evidenced by the non-significant differences in PC1 scores, MVPA, and sedentary time (Table 2) between these groups. However, there was some evidence that the temporal activity patterns differed: after adjusting for age, PC2 scores differed by cancer status, and indicated lower activity in the early part of the day for the cancer group. This information may be useful for tailoring interventions for cancer survivors. Most importantly, having cancer did not moderate the relationship between activity patterns and health outcomes (insulin, CRP, HOMA, insulin status, physical quality of life, and mental quality of life). Thus, the putative beneficial impact of physical activity on health persists irrespective of cancer status, an important public health message for the large cancer survivor population in the world.

We recognize some limitations of the model, such as the interpretability of the more intricate/complex activity trends. Moreover, we have not incorporated the level 2 principal components. Further research can be conducted to investigate these components. We also used a rather simple curve registration method which aligns all daily profiles by the starting point and truncates at 12 h. More advanced time warping functions could be incorporated to better align the data for analysis. Another natural next step would be to extend this work from a two-level functional PCA model to a three-level model incorporating “waves” as the third-level variation. In particular, since we are using baseline data from clinical trials, we will also obtain longitudinal data collected in waves (baseline, follow-up, etc.) tracking changes over time. Thus it would be natural to study dynamic changes of temporal physical activity patterns for each participant. This approach could be extremely important for intervention studies.

In summary, the functional principal component analysis method has proven to be useful in analyzing accelerometer data. From a methodological perspective, while we primarily followed the framework proposed by [6], we introduced the L_2 -norm of scores, which could provide an additional summary metric of physical activity. We also showed how the covariation of activity could be used to identify periods of consistent behavior across days. From a public health perspective, we demonstrated how temporal activity patterns are associated with health. These findings further our understanding of health benefits of physical activity, by showing that timing of activity, in addition to total activity volume, can have beneficial effects. If replicated, these results could help us refine physical activity guidelines and recommendations for healthy living.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Bassett DR (2012) Device-based monitoring in physical activity and public health research. *Physiol Meas* 33:1769–1783
2. Bijnen FC, Caspersen CJ, Mosterd WL (1994) Physical inactivity as a risk factor for coronary heart disease: a WHO and International Society and Federation of Cardiology position statement. *Bull World Health Organ* 72:1–4
3. Blair SN (2009) Physical inactivity: the biggest public health problem of the 21st century. *Br J Sports Med* 43:1–2
4. Brazier JE, Harper R, Jones NM, O’Cathain A, Thomas KJ, Usherwood T et al (1992) Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ*. 305(6846):160–4 PMID: PMC1883187
5. Choi LI, Liu Z, Matthews CE, Buchowski MS (2011) Validation of accelerometer wear and non-wear time classification algorithm. *Med Sci Sports Exerc* 43(2):357–64. <https://doi.org/10.1249/MSS.0b013e3181ed61a3>
6. Di CZ, Crainiceanu CM, Caffo BS, Punjabi NM (2009) Multilevel functional principal component analysis. *Ann Appl Stat* 3(1):458 Mar 1
7. Di J, Leroux A, Urbanek J, Varadhan R, Spira A, Schrack J, Zipunnikov V (2017) Patterns of sedentary and active time accumulation are associated with mortality in US adults: The NHANES study. *bioRxiv*. Jan 1:182337
8. Diaz KM, Howard VJ, Hutto B, Colabianchi N, Vena JE, Safford MM, Blair SN, Hooker SP (2017) Patterns of sedentary behavior and mortality in US middle-aged and older adults: a national cohort study. *Ann Internal Med* 167(7):465–75 Oct 3
9. Fan R, Chen V, Xie Y, Yin L, Kim S, Albert PS, Simons-Morton B (2015) A functional data analysis approach for circadian patterns of activity of teenage girls. *J Circadian Rhythm* 13:3. <https://doi.org/10.5334/jcr.ac>
10. Freedson PS, Melanson E, Sirard J (1998) Calibration of the computer science and applications. *Inc Accelerometer Med Sci Sports Exerc* 30(5):777–781
11. Goldsmith J, Liu X, Jacobson J, Rundle A (2016) New insights into activity patterns in children, found using functional data analyses. *Med Sci Sports Exerc* 48(9):1723–1729. <https://doi.org/10.1249/MSS.0000000000000968>
12. Le T, Flatt SW, Natarajan L, Pakiz B, Quintana EL, Heath DD, Rana BK, Rock CL (2016) Effects of diet composition and insulin resistance status on plasma lipid levels in a weight loss intervention in women. *J Am Heart Assoc* 5(1):e002771. <https://doi.org/10.1161/JAHA.115.002771>
13. Li H, Keadle SK, Staudenmayer J, Assaad H, Huang J Z, Carroll R J (2015) Methods to assess an exercise intervention trial based on 3-level functional data. *Biostatistics*, kvx015
14. Morris JS, Arroyo C, Coull BA, Ryan LM, Herrick R, Gortmaker SL (2006) Using wavelet-based functional mixed models to characterize Population heterogeneity in accelerometer Profiles: a case study. *J Am Stat Assoc* 101(476):1352–1364
15. Morris JS, Carroll RJ (2006) Wavelet-based functional mixed models. *J R Stat Soc* 68(2):179–99 Apr 1
16. Patterson RE, Colditz GA, Hu FB, Schmitz KH, Ahima RS, Brownson RC, Carson KR, Chavarro JE, Chodosh LA, Gehlert S, Gill J, Glanz K, Haire-Joshu D, Herbst KL, Hoehner CM, Hovmand PS, Irwin ML, Jacobs LA, James AS, Jones LW, Kerr J, Kibel AS, King IB, Ligibel JA, Meyerhardt JA, Natarajan L, Neuhaus M, Olefsky JM, Proctor EK, Redline S, Rock CL, Rosner B, Sarwer DB, Schwartz JS, Sears DD, Sesso HD, Stampfer MJ, Subramanian SV, Taveras FM, Tchou J, Thompson B, Troxel AB, Wessling-Resnick M, Wolin KY, Thornquist MD (2013) The 2011–2016 Transdisciplinary Research on Energetics and Cancer (TREC) initiative: rationale and design. *Cancer Causes Control* 24(4):695–704. <https://doi.org/10.1007/s10552-013-0150-z>

17. Patterson RE, Marinac CR, Natarajan L, Hartman SJ, Cadmus-Bertram L, Flatt SW, Li H, Parker B, Oratowski-Coleman J, Villaseñor A, Godbole S, Kerr J (2015) Recruitment strategies, design, and participant characteristics in a trial of weight-loss and metformin in breast cancer survivors. *Contemp Clin Trials* 47:64–71. <https://doi.org/10.1016/j.cct.2015.12.009>
18. Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M (2008) A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act* 5:56
19. Ramsay JO, Silverman BW (2005) *Functional Data Analysis*. Springer, Berlin
20. R Core Team (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria <http://www.R-project.org/>
21. Rock CL, Flatt SW, Pakiz B, Quintana EL, Heath DD, Rana BK, Natarajan L (2016) Effects of diet composition on weight loss, metabolic factors and biomarkers in a 1-year weight loss intervention in obese women examined by baseline insulin resistance status. *Metabolism* 65(11):1605–1613. <https://doi.org/10.1016/j.metabol.2016.07.008>
22. Sallis JF, Saelens BE (2000) Assessment of physical activity by self-report: status, limitations, and future directions. *Res Q Exerc Sport* 71:S1–14
23. Sera F, Griffiths LJ, Dezateux C, Geraci M, Cortina-Borja M (2017) Using functional data analysis to understand daily activity levels and patterns in primary school-aged children: cross-sectional analysis of a UK-wide study. *PLoS ONE* 12(11):e0187677. <https://doi.org/10.1371/journal.pone.0187677>
24. Shou H, Zipunnikov V, Crainiceanu CM, Greven S (2015) Structured functional principal component analysis. *Biometrics* 71(1):247–257. <https://doi.org/10.1111/biom.12236>
25. Sullivan PW, Morroto EH, Ghushchyan V, Wyatt HR, Hill JO (2005) Obesity, inactivity, and the prevalence of diabetes and diabetes-related cardiovascular comorbidities in the U.S., 2000–2002. *Diabetes Care* 28:1599–1603
26. Wilmut EG, Edwardson CL, Achana FA, Davies MJ, Gorely T, Gray LJ, Khunti K, Yates T, Biddle SJ (2012) Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. *Diabetologia* 55(11):2895–2905. <https://doi.org/10.1007/s00125-012-2677-z>
27. Xiao L, Huang L, Schrack JA, Ferrucci L, Zipunnikov V, Crainiceanu CM (2015) Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics* 16(2):352–67. <https://doi.org/10.1093/biostatistics/kxu045>
28. Zeitzer JM, Blackwell T, Hoffman AR, Cummings S, Ancoli-Israel S, Stone K (2017) Osteoporotic Fractures in Men (MrOS) Study Research Group. Daily patterns of accelerometer activity predict changes in sleep, cognition, and mortality in older men. *J Gerontol Ser A* 73(5):682–687

Affiliations

Selene Yue Xu¹ · Sandahl Nelson^{2,3} · Jacqueline Kerr^{3,4,5} ·
 Suneeta Godbole⁵ · Eileen Johnson^{3,6} · Ruth E. Patterson^{3,4} ·
 Cheryl L. Rock^{3,4} · Dorothy D. Sears^{3,4} · Ian Abramson¹ · Loki Natarajan^{3,4}

¹ Department of Mathematics, UC San Diego, La Jolla, CA, USA

² Graduate School of Public Health, San Diego State University, San Diego, CA, USA

³ Department of Family Medicine and Public Health, UC San Diego, La Jolla, CA, USA

⁴ Moores UC San Diego Cancer Center, UC San Diego, La Jolla, CA, USA

⁵ Center for Wireless and Population Health Sciences, UC San Diego, La Jolla, CA, USA

⁶ Division of Epidemiology, School of Public Health, UC Berkeley, Berkeley, CA, USA