Quantifying the difference in activity between diabetics and healthy controls using activity data measured by accelerometers

by

Chih-Kai Chang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

May, 2018

Abstract

The National Health and Nutrition Examination Survey (NHANES) is a combination of multiple cross-sectional studies designed to be representative for the US population. In addition to standard demographic, environmental and behavioral variables the 2003-2004 and 2005-2006 NHANES samples recorded activity data using a hip accelerometer, at the minute level and up to 7 days a week for each individual. The primary goal of this thesis is to understand how activity patterns are associated with health. The secondary goal of the thesis is to quantify the subject-specific patterns of physical activity transitions and study their distribution in the population.

Primary Reader: Ciprian Crainiceanu, PhD

Secondary Reader: Reader: Jacek K. Urbanek, PhD

Acknowledgments

This document is the result of two years of study and work as research assistant at the Johns Hopkins Bloomberg School of Public Health. In the journey of pursuing a Master of science degree in biostatistics, I have grown mentally as a person and learned substantially as a statistician. I would like to thank those who instructed me and guided me during these two years. First, my thesis advisor Ciprian Crainiceanu has been a wonderful professor and mentor who instructed me and provided me with opportunities I am very grateful for. In addition, Jacek K. Urbanek have been a considerable role in helping to the process of this thesis. This thesis would not have been possible without him. Furthermore, I would like to thank all of the members in Wearable Implement Technology group. They helped me solve multiple statistical and computational problems that proved crucial to addressing the goals of my research. I would also like to thank all the excellent faculty in the Department of Biostatistics at Johns Hopkins University who taught me along this journey. Last, a very special thanks to my parents who have always supported me financially and morally.

Contents

CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Diabetes is a chronic metabolic disorder in which the amount of insulin produced by the body falls below the normal range. According to the published literature (Healy et. al,2008, Jeon et. al 2007),there is an association between physical activity and diabetes. Some of these papers (Colberg et. al,2016, Sigal et. al 2006) suggest that doing moderate intensity physical activity, such as brisk walking, can substantially reduce the risk of type 2 diabetes. However, quantifying the association between diabetes and objectively measured physical activity and what type and volume physical activity are more strongly associated with reducing risk of developing diabetes and are an area of ongoing research (American Diabetes Association, 2004). Understanding these associations could substantially impact the range of interven-

tions and provide specific hypothesis about the types of interventions that could be most effective. Some evidence exists that increasing activity levels for individuals with consistently high levels of activity may be associated with a lower risk of developing diabetes (Aune et. al, 2015).

Given that lack of physical activity is believed to be in the causal pathway of developing diabetes, understanding the patterns of objectively measured physical activity might provide novel insights into these associations above and beyond the effect of the total volume of activity. The National Health and Nutrition Examination Survey (NHANES) is a collection of cross-sectional studies designed to assess the health status of people in the United States. To assess the pattern of physical activity transition, we will used the physical activity counts information and diabetes status information in NHANES.

1.2 Data

The NHANES data is available from the Center for Disease Control website and categorized into several areas: demographics, dietary, examination, laboratory, questionnaire, and limited access respectively. Throughout this study, individuals wore accelerometers on hip. The accelerometry data can be downloaded from the "Physical Activity Monitor" subcategory. Regarding the data structure of activity counts, we used 1440+ format to analyze (Leroux et. al, 2017). It recorded accelerometry data

at the minute level. Here 1440 corresponds to the number of minutes in a day and "+" indicates that data may also contain quality and wear/non-wear information. According to a published paper (Leroux et. al, 2017?), this format can reduce the file size and streamlines the process of identifying non-wear time and activity counts.

We used NHANES 2003-2004 and 2005-2006 with a total of approximately 5000 subjects completing the health examination component of the survey every year. These two studies have information for 50231 and 52185 subject-days (rows) respectively, including accelerometry information (1440+ format) and hundreds of predictors (columns) such as age, weight, height, medical history (demographics covariates and health status). In addition to the activity measurements, the data contains diabetes status for 6517 (approximately 65.2%) of the population. This information was coded "Yes" after an individual was diagnosed with diabetes by a doctor or health professional. This information will be referred to as an indicator to determine whether this individual has diabetes. The survey is unique in that it combines interviews and physical examinations. The number of individuals with diabetes in the NHANES 2003-2004 and NHANES 2005-2006 was 1149 (11.5%). Of course, the percent of individuals with diabetes varies strongly with age and below we provide a table of number and percent of individuals by study and age group who have diabetes.

Table 1.1: Number and percentages of diabetics by age group

Number of diabetics (percentages)	6-10 y.o. 0 (0 %)	11-16 y.o. 0 (0%)	17-22 y.o. 2 (1.2%)	23-28 y.o. 5 (1.1%)	29-34 y.o. 25 (5.3%)
	35-40 y.o. 53 (12.9%)	41-46 y.o. 67 (15.1%)	47-52 y.o. 96 (23.0%)	53-58 y.o. 111 (31.8%)	59-64 y.o. 203 (41.4%)
	65-70 y.o. 212 (47.1%)	71-76 y.o. 161 (44.4%)	77-84 y.o. 143 (38.4%)		

Health status and physical activity information obtained from accelerometers is missing for some individuals and these data are omitted from our analyses. More precisely, there were 6517 individuals with diabetes status in NHANES 2003-2004 and NHANES 2005-2006. Furthermore, even if some subjects had activity data, the quality of the data was not considered to be adequate and some days needed to be excluded (according to the NHANES protocol). After these exclusions and data deletions we were left with a reduced data set. Below we provide a table indicating the number of days of accelerometry per subject for each of the two studies. A zero indicates that there were 0 days of accelerometry for that particular subject (the subject did not have any accelerometry data) and so on. For example, there were 1028 (8.4%) individuals who had at least 3 days of accelerometry data.

Table 1.2: The number of days of accelerometry per subject

	Number of Days						
	1	2	3	4	5	6	7
# subjects	777 (6.0%)	824 (6.4%)	1082 (8.4%)	1381 (10.9%)	1869 (14.6%)	2846 (22.2%)	4023 (31.4%)

Also, We assume that the final numbers in our dataset will indicate the number of people with accelerometry among subjects who have diabetes health status.

1.3 Objectives

The main objectives of this analysis is to investigate the pattern of physical activity of individuals with diabetes and those without diabetes. Since we already know physical activity is related to the risk of getting diabetes, we focus on the pattern of physical activity transition (PAT) in this study.

The analysis that approach our objective is organized as follows: Exploratory data analysis (Chapter 2), analysis of matched case-control study (Chapter 3), analysis of difference in transition applied survey weight GLM (Nordberg, 1989) (Chapter 4), and conclusion (Chapter 5).

Chapter 2

Data Structure/EDA

2.1 Activity count

As mentioned previously, the physical activity information used in this analysis provides a single 'activity count' for every minute of the day. This minute level count is determined using an algorithm that summarized the raw data from the accelerometer over a sixty-second window. In NHANES each individual wore an uniaxial ActiGraph GT1M accelerometer at the hip for seven consecutive days. While some papers (Tudor-Locke, 2012, Bai et al., 2014; Schrack et al., 2013) suggested that the activity count magnitude can vary by device location, in NHANES the location of the device was quite consistent and we assume that measurements are comparable across individuals. For the purpose of this thesis we will just assume that counts are comparable and on the same scale for all individuals and not test whether the

assumption is correct. For preprocessing of the data we used previously developed R functions for NHANES 20032006 (Dane et. al, 2014).

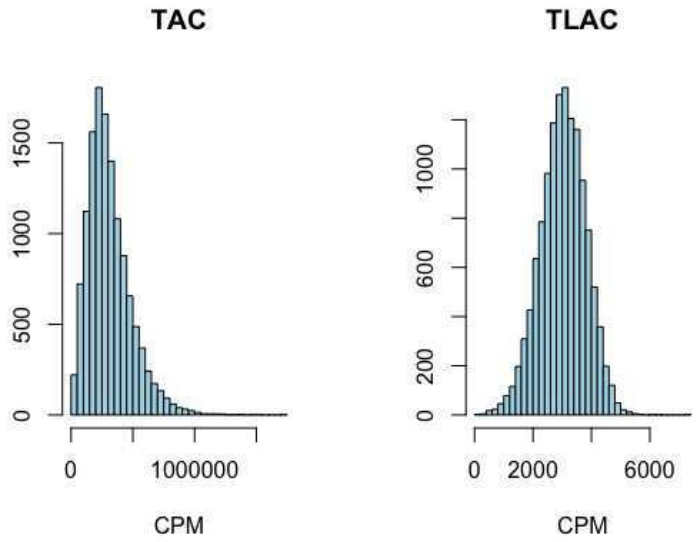
One important feature of the marginal distribution of the activity count data is that it is highly skewed. Working with the data on the original scale can thus substantially affect inference as a few outliers may lead to substantial changes in the inference. Here we follow the advice in (Schrack et. al, 2013) and apply the transformation: counts $\rightarrow \log(1 + \text{counts})$ at the minute level to reduce skewness, transform 0 to 0 and maintain the positivity of measurements.

Applying the log transformation to the activity count data results in a much more symmetric distribution of daily Total activity counts. Figure 2.1 illustrates the substantial change in the distribution of daily total activity counts after the log transformation. The histogram on the right, daily total log activity counts (TLAC) exhibits substantially less skeweness (0.045) than the histogram on the left, daily total activity counts (TAC). This is very important in practice, especially when we compare various groups using, for example, t-tests or regression analysis and the groups are moderate or small in size. To be precise, in our notation for TAC and TLAC, let y_{ijk} be minute level activity count (as obtained from the Actigraph proprietary algorithm) for subject i, on day j and minute k. The formulas for TAC and TLAC are :

$$TAC_i = \frac{1}{\text{valid days for subject i}} \sum_{k=1}^{1440} \sum_{j=\text{valid day}} y_{ijk}$$

$$TLAC_i = \frac{1}{\text{valid days for subject i}} \sum_{k=1}^{1440} \sum_{j=\text{valid day}} \log(1 + y_{ijk})$$

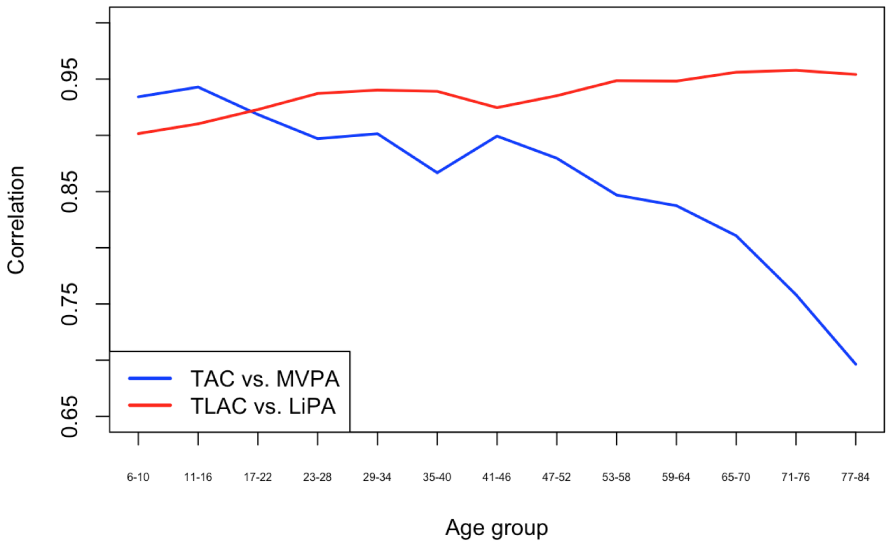
Figure 2.1: Comparison of Transformed Daily Daily Total (log) Counts



Another important reason for using TLAC instead of TAC was highlighted by Varma et. al, 2017 who showed that TLAC is highly correlated with the Light-Intensity Physical Activity (LIPA), the major contributor to total accumulated physical activity across the day. In contrast, TAC is more correlated with moderate-to-vigorous activity (MVPA). Another important characteristic of TLAC is that its correlation with LIPA is maintained across the age span, whereas the association between TAC and MVPA becomes less strong as age increases. Indeed, Figure 2.2 shows that the correlation between TAC and MVPA decreases when age increases (correlation coefficient decreases from 0.93 to 0.70), while the correlation between

TLAC and LIPA remains stable across the age groups and remains higher than 0.90. Therefore, in this thesis we focus on TLAC intead of TAC. In general, one needs to be particularly careful with TAC and use appropriate tools for highly skewed data.

Figure 2.2: TAC v.s. MVPA and TLAC v.s. LIPA



A major problem in accelerometry data is the missing observations, which can be due to invalid days, device malfunction, or non-wear. Here we will assume that missing data are non-informative, though more analysis will be required to verify whether this assumption affects our analyses. A thorough sensitivity analysis may help, but it is beyond the scope of this analysis.

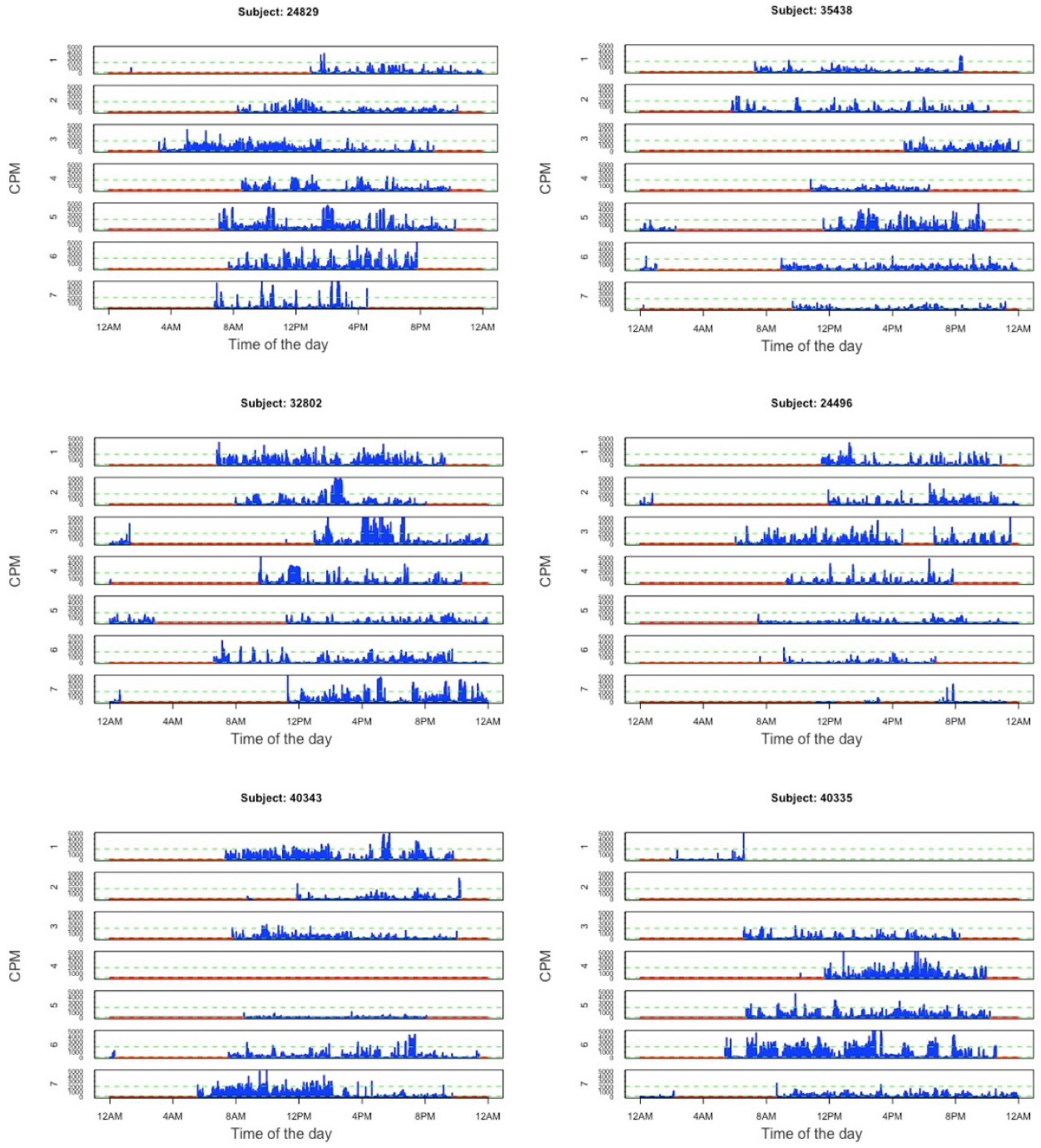
2.2 Visualizing Activity

The first step to analyzing complex activity data is to visualize it. We first display the activity counts at the minute level for each subject to provide the necessary intuition.

Activity counts can be seen as a time series throughout the day, where the x-axis is labeled in one-minute time increments and a particular time of the day (say midnight) is considered as the beginning and end of the day. This can be changed, but throughout the document we will use the midnight-to-midnight representation of the data. To do this we randomly selected six individuals and plotted their individual daily trajectories in Figure 2.3. In the figure, each panel corresponds to single subject and each sub-panel represents a different day. Within each panel, blue indicates the magnitude of the activity count at every minute while red color indicates that the subject did not wear a device (as estimated from the NHANES-specific algorithms for missing data). Areas where neither blue or red is shown, indicate that no data are available for those particular intervals. The two horizontal green dash lines in each panel are thresholds that indicate the boundary between inactivity and LIPA and between LIPA and MVPA. These thresholds are equal to 2019 and 100 and they are the standard values used in NHANES.

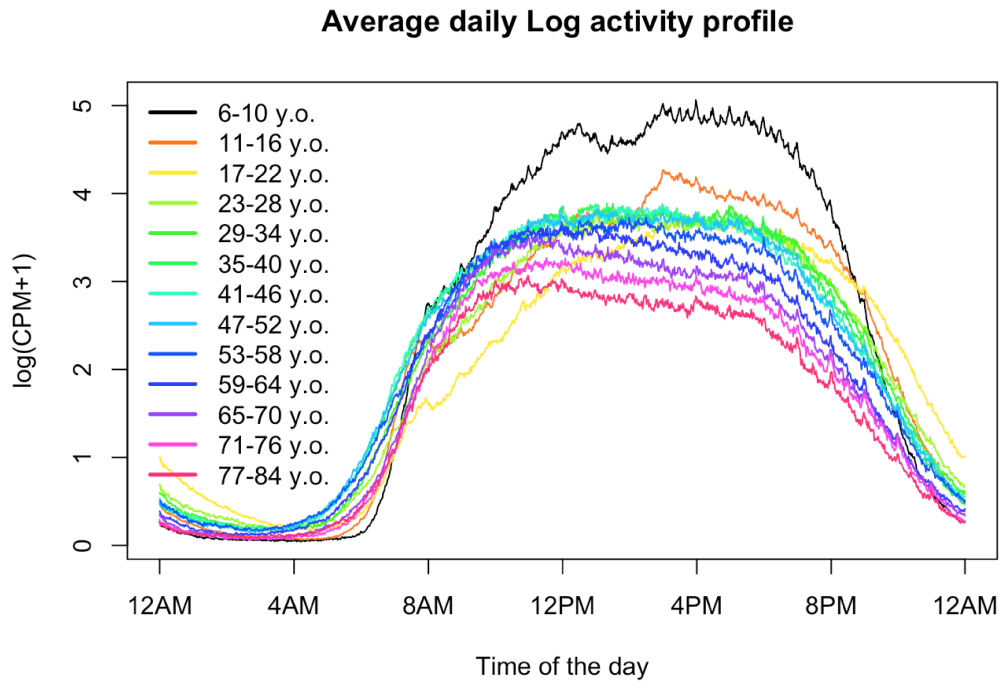
These graphs indicate that there is substantial within- and between-subject variability and that some days should be excluded from the analysis. We excluded data that were either of low quality (quality flag = 0) or entire days that had more than 10 hours/day of estimated non-wear time. Another visible characteristic of the data is that most activity seems to be concentrated between 8am and 11pm, though some of this may be due to the NHANES wear protocol. Indeed, subjects were instructed to not wear the device during sleep, which accounts for the long periods during the night that were identified as non-wear (shown in red). The thresholds (green dash lines) can be used to categorize and summarized into different activity levels. For example, we can obtain time and volume of activity in MVPA and LIPA, respectively.

Figure 2.3: Activity counts distribution



Due to the large sample size of NHANES, it is impractical to visualize all days and all subjects. However, we could easily visualize the within-age-group average at the minute level. Figure 2.4 displays these averages in 5-year age groups starting with the 6 to 10 year olds and ending with the 77 to 84 year olds. This figure indicates that most activity within a day happens between 8am and 11pm. With the exception of children and teens we see a continuous decrease in overall activity across the lifespan and the time of day. The shape of the mean function within age groups stays relatively similar, especially after age 23 and the decrease in activity is more accentuated in older age and accelerated in the second part of the day (3pm to 8pm). We will not be using these daily age group profiles in our analyses, but they provide valuable insight and indicate that, indeed, objective measures of activity based on accelerometers behave reasonably, at least in terms of age.

Figure 2.4: Average Daily TLAC



Chapter 3

Matched Case-Control Study

3.1 Quantify Activity Transition (PAT)

Our overall goal is to study the associations between diabetes status and within-day transition patterns among various types of activity intensities. More precisely, the original counts data collected at every minute is first transformed into four categories: "non-wear", "sedentary", "low intensity physical activity (LIPA)" and "Moderate to Vigorous Physival Activity (MVPA)" using published thresholds (Troiano et. al,2008). Thus, the original time series of counts is transformed into a time series of activity intensities. We estimate the subject-specific transition probabilities between these activity intensity stages and investigate whether these transitions are different by disease group. This is done using matched case-controls and regression modeling approaches.

3.1.1 Categorized Physical Activity Data

To categorize PA into sedentary, light intense physical activity (LIPA), and moderate to vigorous activity (MVPA) for every minute $k = 1, \ldots, 1440$ we define

$$\text{flag}_k = \begin{cases} 1 & \text{if subject wears the device at minute } k \\ 0 & \text{if subject does not wear the device at minute } k \end{cases} \tag{3.1}$$

and the state-time series is obtained as

$$\text{State}_k = \begin{cases} \text{Sedentary} & \text{if } \text{AC}_k \leq 100 \text{ and } \text{flag}_\text{k} = 1 \\ \text{LIPA} & \text{if } 100 < \text{AC}_k \leq 2019 \text{ and } \text{flag}_k = 1 \\ \text{MVPA} & \text{if } \text{AC}_k > 2019 \text{ and } \text{flag}_k = 1 \\ \text{NA} & \text{if } \text{flag}_k = 0 \end{cases} \tag{3.2}$$

Here we denoted by AC_k the activity count at minute k of the day. Before we used the notation y_{ijk} to indicate the minute k of the day j for subject i. Here we have dropped the i and j indexes to make definitions easier to read, but this procedure is applied to every subject and every day at every minute. Thus, the index k is a specific minute time point in a day, from midnight 00:00 ($k = 1$) to 23:59 ($k = 1440$). The flag_k is an indicator variable for wear, non-wear status. The State_k variable is a derived variable indicating at every minute the intensity of physical activity with 3 levels of intensity (Sedentary, LIPA, and MVPA) and one level that is not defined due to non-wear (NA). After categorizing physical activity into specific states, we

can extract transition probabilities between the various states during the day. These numbers can then be averaged across days to provide a subject-specific number or they can be analyzed at the day level using a repeat measures approach. Here we choose the simpler approach of averaging at the subject level

3.1.2 Average Daily Activity Transition

For every subject, we calculated the average daily number of physical activity transitions among three different states in a day, Sedentary (labeled S), LIPA (labeled L), and MVPA (labeled M), respectively. We excluded invalid days and calculated the average daily number of transitions for each individual.

A simple example to better understand what we are doing can be seen if the set of transitions is

$$S - S - S - S - L - S - L - S - S - M - S$$

. Therefore, there are nine different type of physical activity transitions events: $S-S$, $S-L$, $S-M$, $L-S$, $L-L$, $L-M$, $M-S$, $M-L$, $M-M$, respectively. The number of physical activity transitions are $nPAT(S-S) = 4$, $nPAT(S-L) = 2$, $nPAT(L-S) = 2$, respectively.

3.1.3 Normalization of Physical Activity Transitions

We will be focusing on analyzing and comparing all nine types of transition between the three states, though we will need to normalize data first. More precisely, we want to use data that are more comparable across subjects and days. For example, some subjects may have more transitions simply because they had more wear time. Therefore, it would make more sense to compare the number of events relative to a reference set. That is, we have to consider the interpretation of each method, and the meaning behind it. We will use three different types of normalizion methods.

The first one is the number of physical activity transitions (nPAT) relative to the **total active count (TAcC)** per day. TAcC is defined as the number of activity counts greater than zero when the subject wore the device. More precisely, we define Active Counts at every minute k as

$$\mathrm{AcC}_k = \begin{cases} 1 & \text{if } \mathrm{AC}_k \geq 1 \text{ and } \mathrm{flag}_k = 1 \\ 0 & \text{if } \mathrm{AC}_k = 0 \text{ and } \mathrm{flag}_k = 1 \\ \mathrm{NA} & \text{if } \mathrm{flag}_i = 0 \end{cases} \tag{3.3}$$

Once the Active Counts are defined, we define the Total Active Counts as

$$\mathrm{TAcC} = \sum_{k=1}^{1440} \mathrm{AcC}_k \ ,$$

where, again, we conveniently ignored the indexing by subject and day. Across days we simply add the TAcC to obtain a subject specific TAcC.

To better understand this assume that the data for 10 minutes is

$$S - S - S - S - L - S - L - S - S - M - S$$

and the corresponding number of counts were

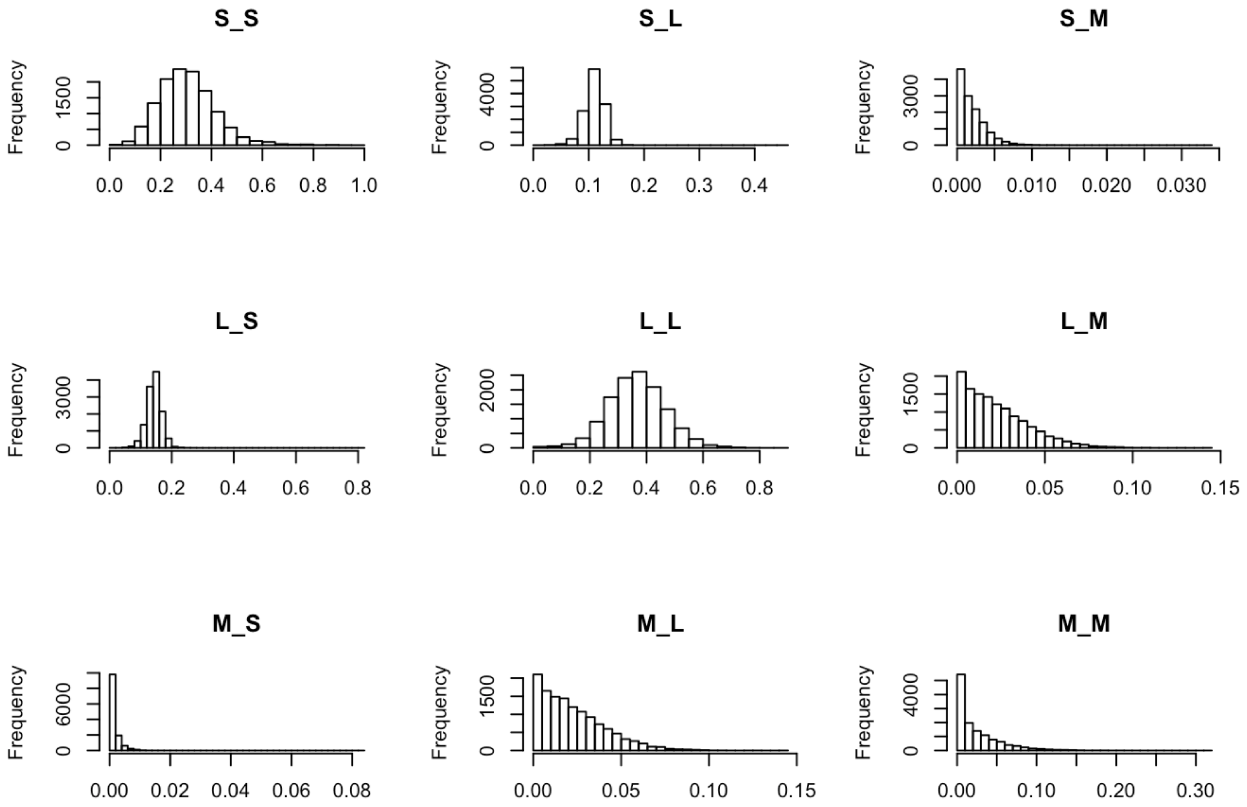
$$0, 90, 0, 25, 500, 0, 340, 0, 35, 3300, 27 \ .$$

In this case the number of transitions from S to L is equal to $nPAT(S - L) = 2$, while the number of Active Counts, that is the number of minutes that have a count larger than 0 is AcC=7. Thus, we define the normalized proportion of transitions relative to the number of active minutes as

$$\mathrm{P}_{\mathrm{TAcC}}(S - L) = \frac{2}{7} = 0.286 \ .$$

This normalized quantity can be interpreted as the frequency of physical activity transitions from one state to another, given that the subject was active (AC¿0). In this example, we can say that given that the subject is active, the physical activity transition from sedentary to LiPA is 0.2. Figure 3.1 displays the histograms of the proportion of transitions normalized by TAcC across subjects. For activity transition that involve MVPA (panels in the last row and column), the distribution of normalized transition probabilities is highly skewed. For the other types of transitions the distributions are more symmetric. Part of this difference could be explained by the fact that some individuals simply have very few periods when they are in MVPA.

Figure 3.1: Data Normalized by TAcT



The second method used for normalization is using the daily **total wear time (TWT)**. At every minute, we define that a subject wears the device if the wear flag at that particular minute equals to one. Using the same logic as for TAcC, we calculate the number of minutes a subject was wearing a device in a day and defined it as total wear time (TWT). More precisely,

$$\text{TWT} = \sum_{k=1}^{1440} \text{flag}_k \ .$$

Continuing with our example, assume that the device was worn at each of the 11. Then nPAT$(S-L)=2$ and TWT $=11$. Therefore,

$$P_{TWT}(S-L)=\frac{2}{11}=0.182\ .$$

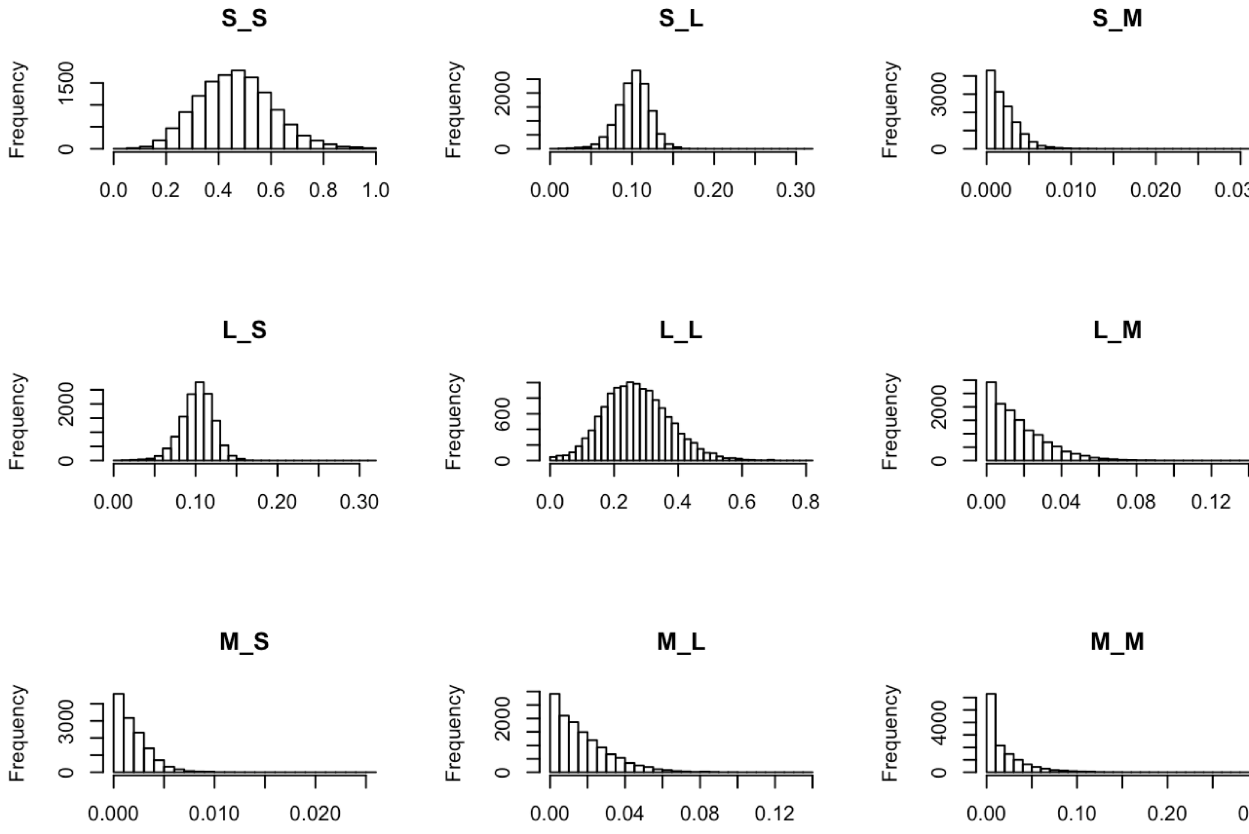
This variable can be interpreted as the frequency of physical activity transition from one state to another given that the subject was wearing the device. Figure 3.2 displays the histogram of frequency of transitions relative to total wear time across subjects. As in the previous case, the distributions for transitions that involve MVPA are highly skewed (panels in the first and last column). The distributions for all other transitions are more symmetric. While the distributions do not change substantially in shape, the actual numbers are different in Figure 3.1 versus Figure 3.2. Our example should explain exactly why this is the case.

The last normalization procedure for the number of physical activity transition (nPAT) is done by dividing by the daily **total time spent in a certain state** (time at risk). The first step is to calculate the average daily time in a particular activity state for each individual (e.g. Sedentary). In our example, there are 2 S-L transitions and the subject was Sedentary in 8 out the 11 minutes. In this case the proportion of $S-L$ transitions normalized by the time at risk is $2/8=0.25$.

In general, we define the

$$P_S(S-L)=\frac{\text{nPAT(S}-\text{L)}}{\sum_k I\{\text{State}_k=S\}}\ ,$$

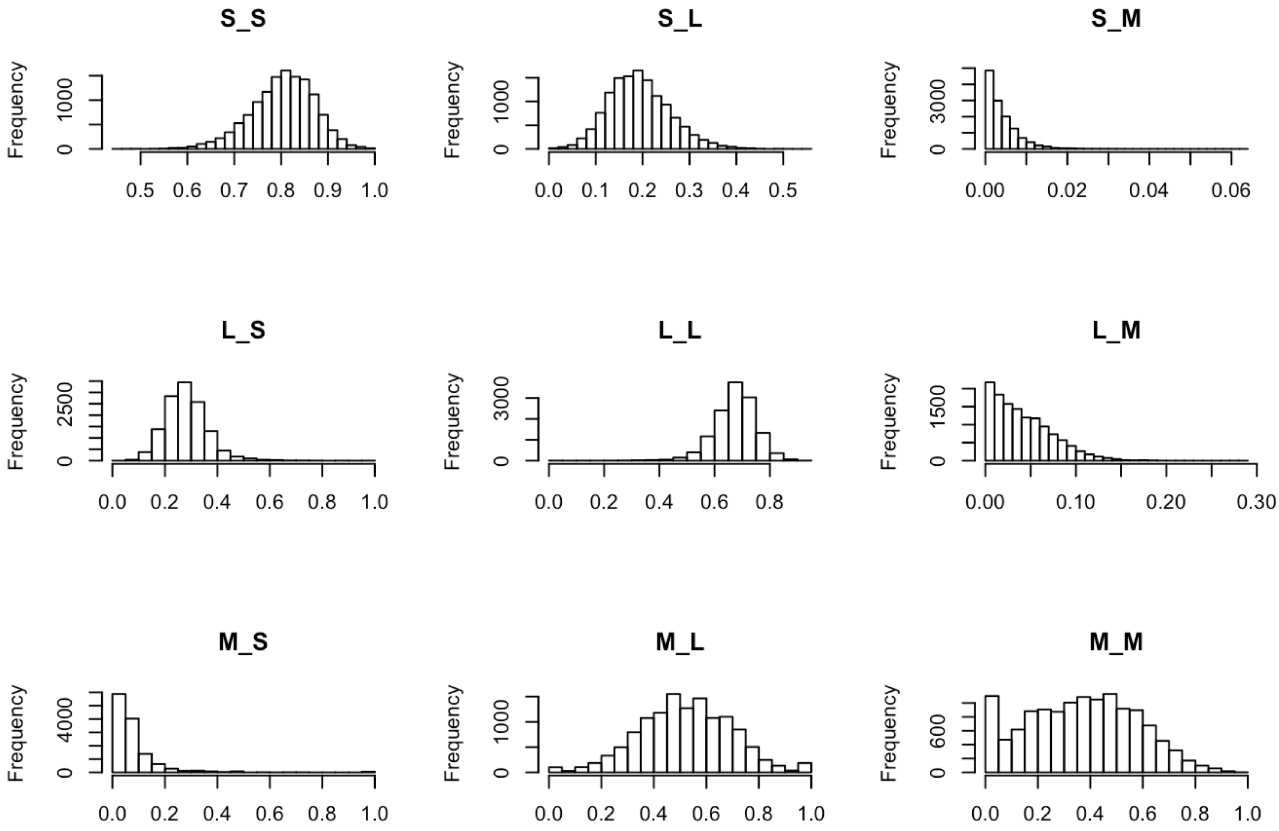
Figure 3.2: Data Normalized by TWT



where, recall, nPAT(S-L) is the number of S to L transitions and $I\{\text{State}_k = S\}$ is the indicator that the State_k is equal to s at minute k.

Figure 3.3 provides the nprmalized transition probabilities bythe time at risk. Again, the distributions that contain MVPA tend to be highly skewed whereas all other transitions are symmetric.

Figure 3.3: Data normalized by Time at risk



3.2 Diabetes v.s. Non-Diabetes

After quantifying the physical activity profile for each individual, we conducted a matched comparison of individuals with diabetes versus individuals without diabetes. The covariates we used to match are age, BMI and gender. The margin for matching age is one year and the margin of matching BMI is one BMI unit. We have conducted three types of matching, one to one (labeled 1-1), one to two (labeled 1-2), and one to all (labeled 1-all). For 1-2 matching, we matched each individual with diabetes

with two individuals without diabetes because there are many more individuals with diabetes than without. For 1-all matching we matched an individual with diabetes to all of the individuals without diabetes within the pre-defined margin for age and BMI and of the same gender.

3.2.1 Compare covariates

There are total 943 subjects in diabetes group and total 1799 subjects in control group. After matching, for 1-1 matching ,there are both 409 subjects in diabetes and control groups. For 1-2 matching, the sample size is 214 in the diabetes group and 428 in control groups. As for 1-all matching, there are 403 subjects in diabetes group and 451 in the control group.

3.2.2 Compare transition probability

To compare whether the physical activity transition probabilities is different between the diabetes and non-diabetes groups, we obtained the bootstrap confidence intervals for the the absolute median difference, relative risk, and log odds ratio of the normalized transition proportions. Our null hypothesis is that there is no difference between the two groups and set type one error $\alpha = 0.05$.

Absolute Bootstrap median difference: $|(P(Diabetes) - P(Non - Diabetes)) * 100\%|$

Relative Risk : $P(D)/P(ND)$

Log OR : $log[P(D)/1-P(D)]/[P(ND)/1-P(ND)]$

Here $P(D)$ and $P(ND)$ are generic notation for the normalized transition frequency for the diabetes and non-diabetes groups, respectively. We have not indicated exactly what transition we are focusing on because we will consider all of them. In the following section, we will provide the results for each normalization procedure and matching method.

3.2.2.1 Normalized by Total Wear Time

1-1 Matching

For one to one matching, given that subject worn a device , individuals without diabetes tends to have more physical activity transition from LIPA to MVPA and from MVPA to LIPA than individuals with diabetes. The absolute median difference for physical activity transition from LIPA to MVPA is 0.0028 with 95% confidence interval (CI) [0.0010, 0.0040]. The probability of transition form LIPA to MVPA for non-diabetic individuals is 48% higher (relative risk= 0.62 with 95% CI [0.53, 0.82]) than that for diabetic individuals . The log-odds ratio for this type of transition is -0.46 [95% CI: -0.20, -0.69]

Secondly, the probability of transition form MVPA to LIPA for non-diabetic individuals is 39% higher (relative risk = 0.61 with 95% CI [0.51, 0.83]) than that for diabetic individuals , and its absolute median difference between two group is 0.0029 with 95% CI [0.0011, 0.0039] and its log-odds ratio is -0.45 with 95% CI [-0.20,

-0.67].

Furthermore, the probability of transition form MVPA to sedentary for non-diabetic individuals is 24% higher (relative risk= 0.76 with 95% CI [0.62, 0.90]) than that for diabetic individuals and the log-odds ratio for this type of transition is -0.28 [95% CI: -0.47, -0.10].

In addition, transition form sedentary to MVPA is also found to be significantly different between diabetes and non-diabetes groups. The probability of transition form sedentary to MVPA for non-diabetic individuals is 28% higher (relative risk = 0.72 with 95% CI [0.58, 0.96]) than that for diabetic individuals , and its absolute median difference between two group is 0.0002 with 95% CI [0.0000, 0.0004], and its log-odds ration is -0.30 with 95% CI [-0.55, -0.04].

Thirdly, we also found that given that subject worn a device, individuals without diabetes tends to stay longer in MVPA than those with diabetes. The absolute median difference is 0.0021 with 95% confidence interval [0.0007, 0.0042], and the relative risk is 0.48 with 95% CI [0.25, 0.76], which means non-diabetic individuals have 52% higher probability in transition from MVPA to MVPA than non-diabetic individuals.

Last, given that subject worn a device, individuals without diabetes tends to stay longer in LIPA than those with diabetes. The absolute median difference is 0.030 with 95% confidence interval [0.005, 0.050], and the relative risk is 0.89 with 95% CI [0.82, 0.98], which means non-diabetic individuals have 11% higher probability in transition from LIPA to LIPA than non-diabetic individuals.

As for other type of physical activity transition, there is no significant difference between diabetes and non-diabetes groups.

In conclusion , individuals without diabetes tends to have more physical activity transition frequencies between LIPA and MVPA than individuals with diabetes given a subject worn a device.

1-2 Matching

For 1-2 matching, again results are consistent with those for 1-1 matching, though another transition was found to be statistically different between the diabetes and non-diabetes group. Transition from sedentary to sedentary becomes significant. The probability of this transition for diabetic individual is 12% higher (relative risk is 1.12 with 95% CI [1.04, 1.18]) than non-diabetic individual and its absolute median difference is 0.0003 [95% CI: 0.0001, 0.0005]. Others transitions are still not significant but p-value for test the difference between two group become smaller.

1-All Matching

For 1-all matching, all of the physical activity transition frequencies were found to be statistically significantly different.

3.2.2.2 Normalized by Total Active Time

1-1 Matching

We now focus on comparing the transition frequencies normalized by Total Active Time (TAcT). For 1-1 matching, individuals without diabetes have more physical

activity transition (PAT) from LIPA to MVPA and from MVPA to LIPA than individuals with diabetes. The absolute median probability difference for physical activity transition from LIPA to MVPA is 0.0030 with a 95% confidence interval (CI) [0.0013, 0.0053]. Its relative risk and log-odds ratio are 0.69 [95% CI:0.54, 0.84], -0.39 [95% CI: -0.63, -0.17] respectively.

The absolute median difference in the probability of transition from MVPA to LIPA between two group is 0.035 with a 95% CI [0.0017, 0.0052]. Its relative risk is 0.66 [95% CI 0.53, 0.82]. As for its log-odds ratio, it is -0.42 with a 95% confidence interval [-0.63, -0.21].

Another finding is that individuals with diabetes have more physical activity transition (PAT) from LIPA to sedentary and sedentary to LIPA than individuals without diabetes. The absolute median difference for physical activity transition from LIPA to sedentary individuals is 0.0071 with 95% confidence interval [0.0035, 0.0104]. The relative is 1.05 with 95% CI [1.02, 1.07]. This means the probability of transition form LIPA to sedentary for diabetic individuals is 5% higher than that for non-diabetic individuals. Its log odds ratio is 0.06 with 95% CI [0.03, 0.08].

In addition, the absolute median difference for transition from sedentary to LiPA is 0.0039 with 95% CI [0.0014, 0.0074]. The probability for that transition for diabetic individuals is 7% higher than that for non-diabetic individuals (relative risk = 1.07 with 95% CI [1.01, 1.15]), and its log odds ratio is 0.11 [95% CI 0.06, 0.17].

A third finding was that individuals without diabetes have more physical activity

transition (PAT) from MVPA to MVPA than individuals with diabetes. The absolute median difference is 0.0034 with 95% confidence interval (CI) [0.0010, 0.0054], and the relative risk is 0.44 with 95% CI [0.28, 0.76], which means the probability for transition from MVPA to MVPA in non-diabetic individuals is 56% higher than diabetic individual.

Fourthly, individuals without diabetes have more physical activity transition (PAT) from LIPA to LIPA than individuals with diabetes. The absolute median difference is 0.039 with 95% confidence interval (CI) [0.0061, 0.048], and the relative risk is 0.91 with 95% CI [0.88, 0.98] individual.

To sum up, non-diabetic individuals tend to have more transition between LIPA and MVPA, fewer transitions between sedentary and LIPA. Diabetes status seems to affect more strongly the LIPA to MVPA than the sedentary to LiPA transitions.

With this matching there were no statistically significant differences in physical transition frequencies from sedentary to sedentary, LIPA to LIPA, sedentary to MVPA, and MVPA to sedentary.

1-2 Matching

For 1-2 matching, results are consistent with those for 1-1 matching.

1-All Matching

For 1-all matching, results are consistent with those for 1-1 matching.

3.2.2.3 Normalized by Time spent in Certain State

1-1 Matching

In this section, we now draw our attention on comparing the transition frequencies normalized by total time spent in a certain state. For one to one matching, given that subject spent time in LIPA, individuals without diabetes tends to have more physical activity transition from LIPA to MVPA than individuals with diabetes. The absolute median difference for transition from LIPA to MVPA is 0.0071 with 95% confidence interval (CI) [0.0030, 0.010]. The probability of transition form LiPA to MVPA for non-diabetic individuals is 37% higher (relative risk= 0.63 with 95% CI [0.53, 0.82]) than that for diabetic individuals and its log odds ratio is -0.43 [95% CI: -0.65, -0.20].

Secondly, given that subject spent time in LIPA, individuals with diabetes tends to have more physical activity transition from LIPA to sedentary than individuals with diabetes. The absolute median difference is 0.0265 [95% CI : 0.0109, 0.0452]. Besides, the relative risk for transition from LiPA to sedentary is 1.10 [95% CI : 1.04, 1.17] and its log-odds ratio is 0.14 [95% CI : 0.08, 0.20]. In addition, given that subject spent time in LIPA, individuals with diabetes tends to stay in LIPA longer than individuals with diabetes. The absolute median difference is 0.0190 [95% CI : 0.0034, 0.0336]. Besides, the relative risk for transition from LIPA to LIPA is 0.97 [95% CI : 0.95, 0.99].

Thirdly, we also found that given that subject was MVPA, individuals without

diabetes tends to stay longer in MVPA than those with diabetes. The absolute median difference is 0.0597 with 95% confidence interval (CI) [0.0187, 0.0960], and the relative risk is 0.79 with 95% CI [0.69, 0.93], which means the probability of transition from MVPA to MVPA for non-diabetic individuals is 21% higher than diabetic individuals. The log-odds ratio for this type of transition is -0.31 with 95% CI $[-0.45, -0.15]$.

To sum up, given a subject was in LIPA, individuals without diabetes have more transition to MVPA, while individuals with diabetes tend to transit to sedentary. Once an individual stays in MVPA. An non-diabetic individual tends to stay in MVPA 21% longer than diabetic individual. Besides, once an individual stays in LIPA. An non-diabetic individual tends to stay in MVPA 3% longer than diabetic individual.

1-2 Matching

Again, for 1-2 matching, results are consistent with those for 1-1 matching. There is, however, a new transition become statistically significant. Given an individual is in sedentary, transition from sedentary to sedentary also become significant with absolute median difference 0.0113 [95% CI:0.0003, 0.0264], relative risk 1.01 [95% CI: 1.00, 1.03], and log-odds ratio 0.08 [95% CI: 0.06 , 0.10]

1-All Matching

For 1-all matching, all of the physical activity transition frequencies were found to be statistically significantly different. The two exceptions were the transition from MVPA to sedentary and MVPA to LIPA.

Information for the plot Figures 3.4, 3.5, and 3.6, summarize the findings presented in this section for 1-1, and 1-2 matching in different normalized method, respectively. All figures contain three different color of arrows: gray, blue and red. Gray arrows indicate that for that particular transition frequency there was not a statistically significant difference between the two groups. Red arrows indicate that for that particular transition frequency the mean of the non-diabetic group was found to be higher than for the non-diabetic group. Blue arrows indicate that for that particular transition frequency the mean of the non-diabetic group was found to be lower than for the non-diabetic group. We also presented three value near an arrow. They are median difference, relative risk, and log odds ratio, respectively together with their 95% confidence intervals (details on the top right side of the plot).

Figure 3.4: Compare Transition Normalized by Total Wear Time

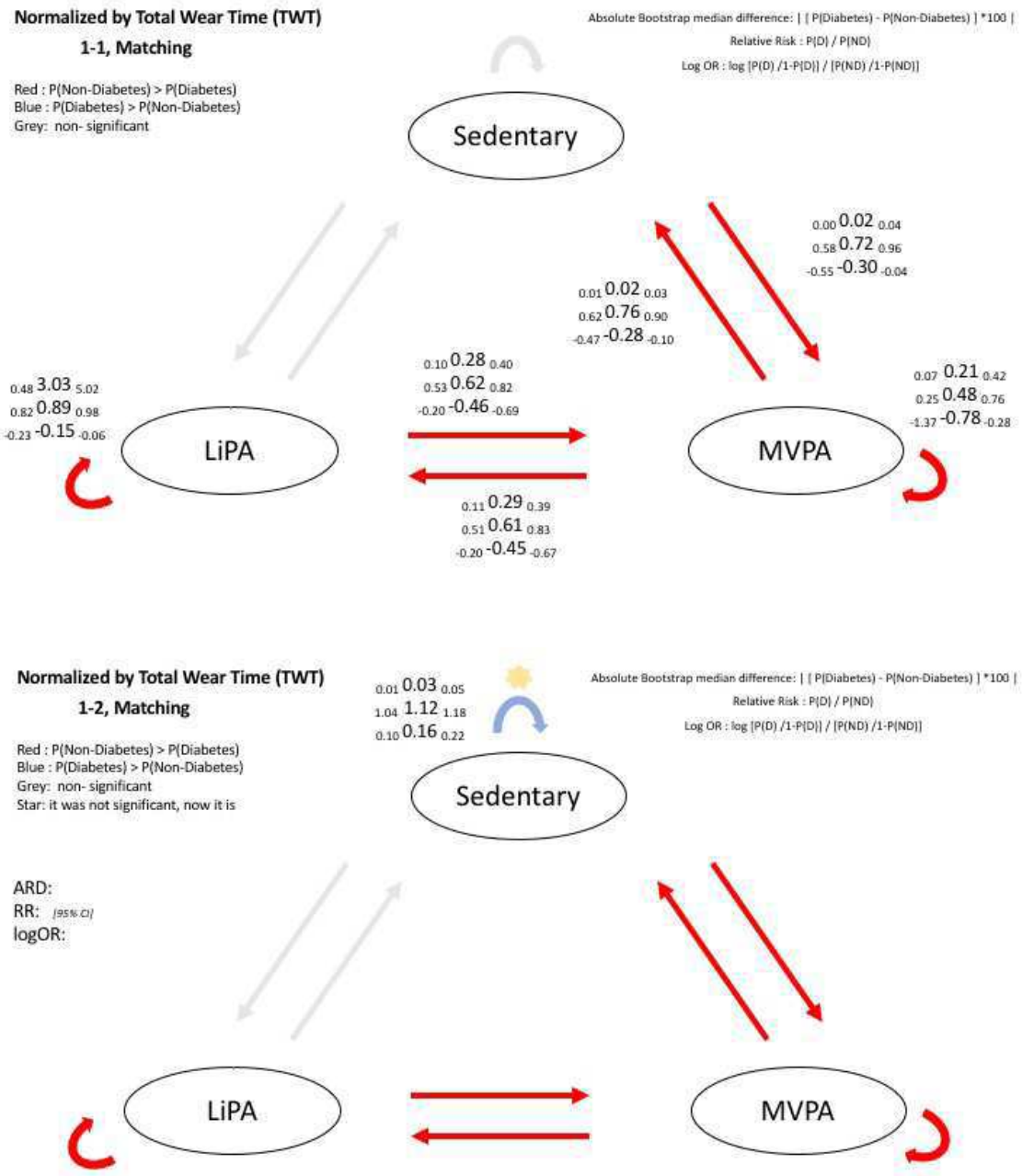
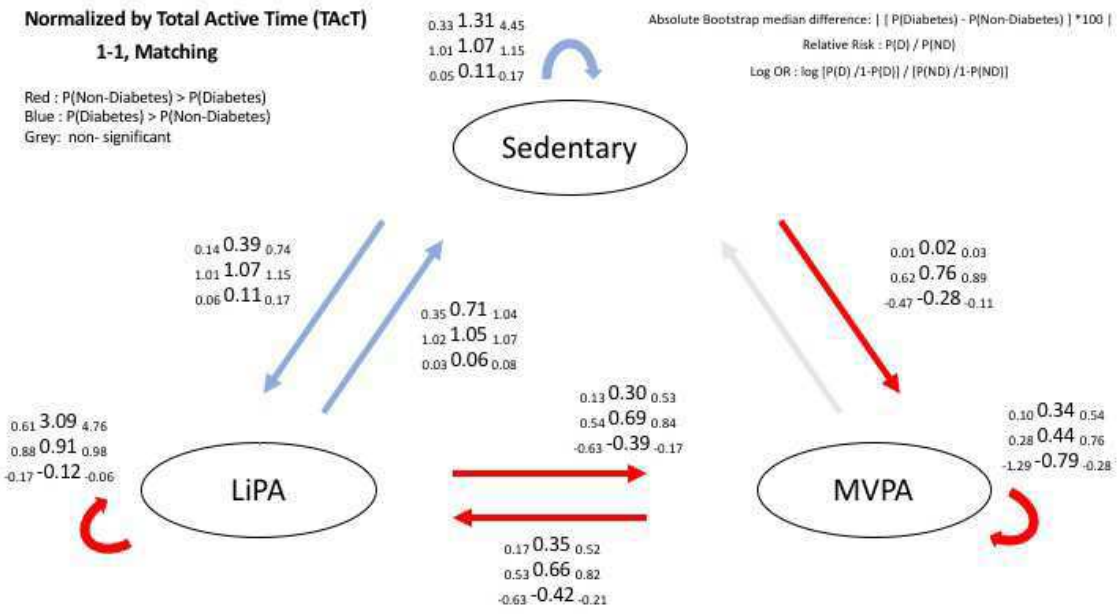


Figure 3.5: Compare Transition Normalized by TAcT



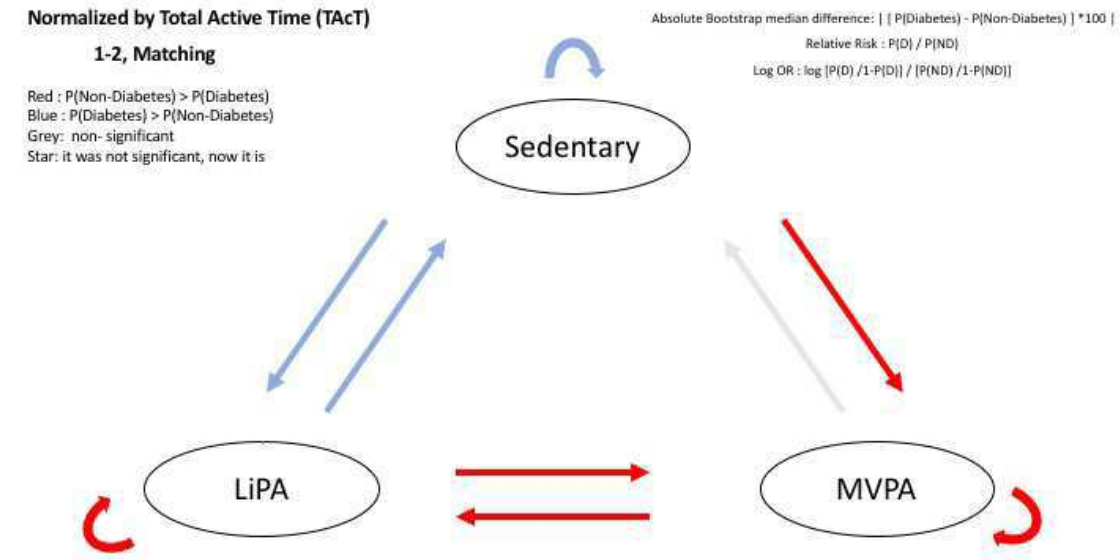
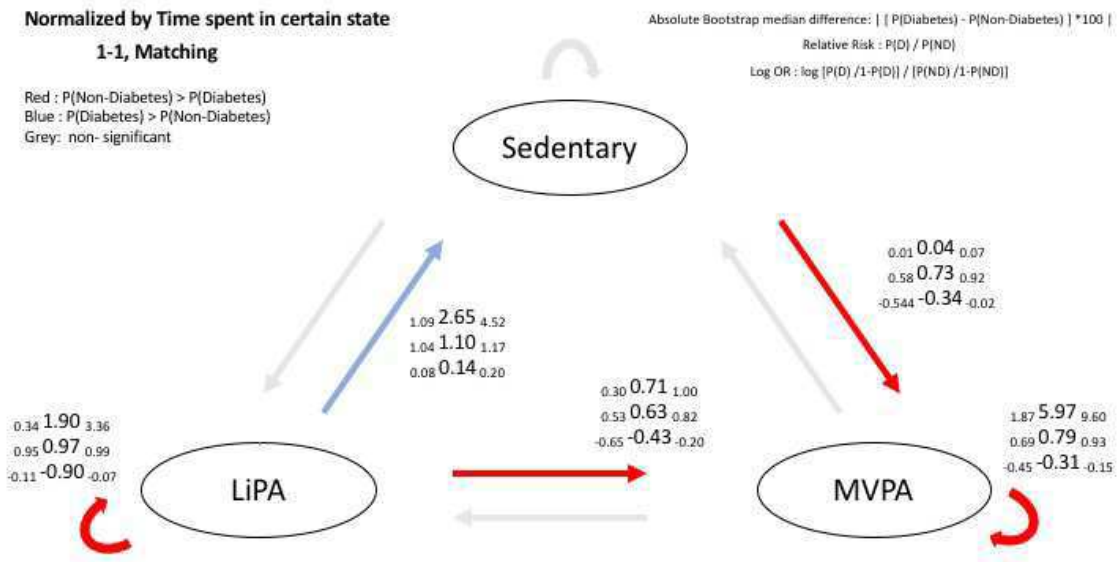
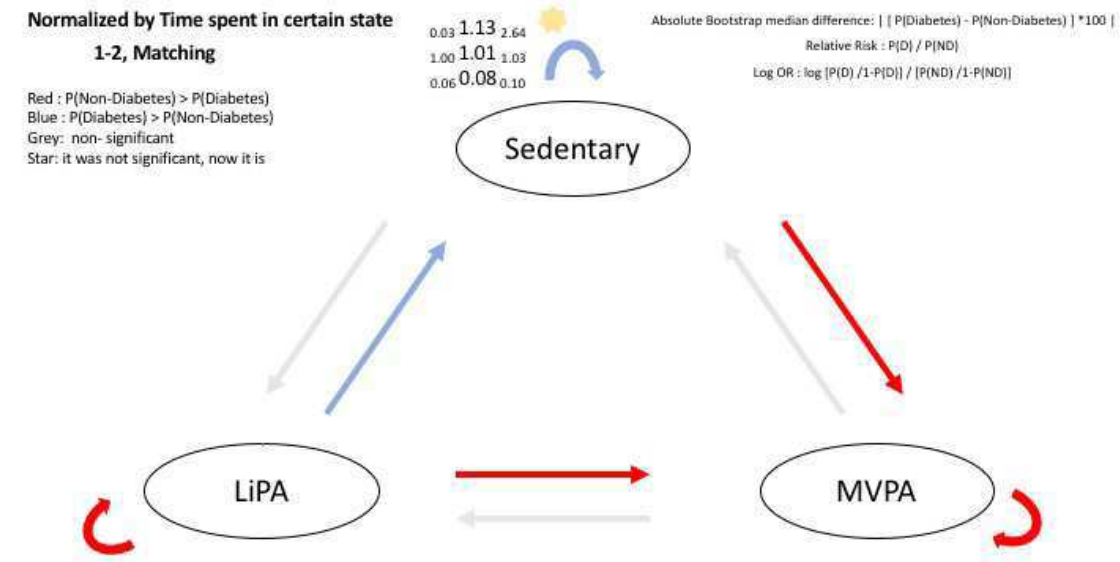


Figure 3.6: Compare Transition Normalized by Time spent in certain state





Chapter 4

Conclusions

As wearable devices become more less expensive and accurate, the use of actigraphy devices is likely to rapidly increase. There are the wealth of research and knowledge supporting that physical activity is associated with human health, and the risk of many chronic diseases. It is crucial that statistical methods are developed to make inference from this complex data. In this document, we investigated the association of activity transition and diabetes to help understand aspects of risk of getting diabetes and prevent it.

4.1 The impact of Activity Transition

From the analysis described in previous chapters, it can be concluded that indeed, the pattern of activity transition does significantly affect the risk of getting diabetes.

However, most of transitions is not predictive for the risk of getting diabetes after adjust total log activity counts (TLAC). There is still moderate support in the data for the belief that transition from sedentary to moderate to vigorous activity (MVPA) can predict the risk of diabetes. While this support is based on match-control study, activity transition patterns can provide us more information for predicting the risk of diabetes. Other than total active counts, activity transition might take an important role in human health.

4.2 Future Study

Considering the relatively strong predictive power for some transitions that respond negatively to the risk of getting diabetes, more work needs to be done to determine if it is still significant after adjusting for all kind of transitions in a model together and whether the result still hold if we only focus on a specific time period.

Chapter 5

Bibliography

Amy Luke, Lara R Dugas, Ramon A Durazo-Arvizu, Guichan Cao, and Richard S Cooper. Assessing Physical Activity and its Relationship to Cardiovascular Risk Factors: NHANES 2003-2006. BMC Public Health. 2011; 11: 387.

Auguie, B (2016). gridExtra: Functions for "Grid" Graphics. R package version 2.3. https://cran.r-project.org/web/packages/gridExtra/index.html

American Diabetes Association. Physical Activity/Exercise and Diabetes. Diabetes Care Jan 2004, 27 (suppl 1) s58-s62.

Aune, Dagfinn and Norat, Teresa and Leitzmann, Michael and Tonstad, Serena and Vatten, Lars Johan. Physical activity and the risk of type 2 diabetes: a systematic

review and dose–response meta-analysis. European Journal of Epidemiology 2015 (30), p 529-542.

Bai J, He B, Shou H, Zipunnikov V, Glass TA, Crainiceanu CM. Normalization and extraction of interpretable metrics from raw accelerometry data. Biostatistics. 2014 Jan;15(1):102-116.

Christie Y. Jeon, R. Peter Lokken, Frank B. Hu, Rob M. van Dam. Physical Activity of Moderate Intensity and Risk of Type 2 Diabetes. Diabetes Care Mar 2007, 30 (3) 744-752.

Colberg, Sheri R. and Sigal, Ronald J. and Yardley, Jane E. and Riddell, Michael C. and Dunstan, David W. and Dempsey, Paddy C. and Horton, Edward S. and Castorino, Kristin and Tate, Deborah F. Physical Activity/Exercise and Diabetes: A Position Statement of the American Diabetes Association. Diabetes Care (39). p.2065–2079.

Dane R. Van Domelen and W. Stephen Pittard. Flexible R Functions for Processing Accelerometer Data, with Emphasis on NHANES 2003-2006. The R Journal Vol. 6/2, December 2014/ p 52-62.

Darrell J. Gaskin and Roland J. Thorpe Jr and Emma E. McGinty and Kelly Bower and Charles Rohde and J. Hunter Young and Thomas A. LaVeist and Lisa Dubay. Disparities in Diabetes: The Nexus of Race, Poverty, and Place. American Journal of Public Health, 2014 (104). P. 2147-2155.

Efron, B. Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7 (1979), no. 1, 1–26. doi:10.1214/aos/1176344552. https://projecteuclid.org/euclid.aos/1176344552

Genevieve N. Healy, Katrien Wijndaele, David W. Dunstan, Jonathan E. Shaw, JoSalmon, Paul Z. Zimmet, Neville Owen. Objectively Measured Sedentary Time, Physical Activity, and Metabolic Risk. Diabetes Care Feb 2008, 31 (2) 369-371.

Hadley Wickham. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package. Version 2.2.1 https://cran.r-project.org/web/packages/ggplot2/index.htm

Hadley Wickham. Flexibly Reshape Data. R package. Version 0.8.7. https://cran.r-project.org/web/packages/reshape/reshape.pdf.

Jeon, Christie Y. and Lokken, R. Peter and Hu, Frank B. and van Dam, Rob M. Physical Activity of Moderate Intensity and Risk of Type 2 Diabetes. Diabetes Care (2007). V. 30 P. 744–752

Karas, Marta and Bai, Jiawei and Strczkiewicz, Marcin and Harezlak, Jaroslaw and Glynn, Nancy W. and Harris, Tamara and Zipunnikov, Vadim and Crainiceanu, Ciprian and Urbanek, Jacek K. Accelerometry data in health research: challenges and opportunities. Review and examples. bioRxiv (2018)

Nelder, J. A., and R. W. M. Wedderburn. "Generalized Linear Models." Journal of the Royal Statistical Society. Series A (General) 135, no. 3 (1972): 370-84.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. Journal of Official Statistics, 5(3), 223. Retrieved from https://search.proquest.com/docview/1266808183?a

Randall Pruim. Data from the US National Health and Nutrition Examination Study. R package version 2.1.0. https://cran.r-project.org/web/packages/NHANES/NHANES.pdf

Scott J Strath, PhD, Robert G Holleman, MPH, Caroline R Richardson, MD, David L Ronis, PhD, and Ann M Swartz, PhD. Objective Physical Activity Accumulation in Bouts and Nonbouts and Relation to Markers of Obesity in US Adults. Prev Chronic Dis. 2008 Oct; 5(4): A131.

Schrack JA, Zipunnikov V, Goldsmith J, Bai J, Simonsick EM4, Crainiceanu C,

Ferrucci L. Assessing the "physical cliff": detailed quantification of age-related differences in daily patterns of physical activity. J Gerontol A Biol Sci Med Sci. 2014 Aug;69(8):973-979.

Steeves JA, Murphy RA, Crainiceanu CM, Zipunnikov V, Van Domelen DR, Harris TB. Daily Patterns of Physical Activity by Type 2 Diabetes Definition: Comparing Diabetes, Prediabetes, and Participants with Normal Glucose Levels in NHANES 2003-2006. Prev Med Rep. 2015;2:152-157.

Troiano RP, Berrigan D, Dodd KW, Msse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. 2008 Jan;40(1):181-8.

Tudor-Locke C, Camhi SM, Troiano RP. A catalog of rules, variables, and definitions applied to accelerometer data in the National Health and Nutrition Examination Survey, 2003-2006. Prev Chronic Dis. 2012;9:E113.

Vijay R. Varma PhD, Debangan Dey , Andrew Leroux , Junrui Di , Jacek Urbanek PhD, Luo Xiao PhD, Vadim Zipunnikov PhD. Total volume of physical activity: TAC, TLAC or TAC(). Preventive medicine, 2017 (106).

Varma VR, Dey D, Leroux A, Di J, Urbanek J, Xiao L, et al. Re-evaluating the effect of age on physical activity over the lifespan. Preventive medicine. 2017 Aug;101:102-8.

Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J. Active Smoking and the Risk of Type 2 DiabetesA Systematic Review and Meta-analysis. JAMA. 2007;298(22):2654-2664.

Wolff-Hughes DL, Fitzhugh EC, Bassett DR, Churilla JR. Waist-Worn Actigraphy: Population-Referenced Percentiles for Total Activity Counts in U.S. Adults. Journal of physical activity health. 2015 Apr;12(4):447-53.

Vita

Chih-Kai Chang was born in Kaohsiung, Taiwan on August 28, 1992. He pursed undergraduate studies in industrial engineering and minor in economics at National Tsing Hua University, attaining his B.S. degree in 2015. In the fall of 2016, Chih-Kai enrolled in Johns Hopkins University as graduate student in biostatistics department.