Decomposing changes of socioeconomic inequality in physical activity between 2005 and 2015 in the US using RIF regressions

Tong Liu

8634944

Major Paper presented to the

Department of Economics of the University of Ottawa

in partial fulfillment of the requirements of the M.A. Degree

Supervisor: Professor Paul Makdissi

ECO 6999

Ottawa, Ontario

December 2017

Abstract

This paper applies the RIF regression decomposition method to socioeconomic inequality in physical activity and extends the RIF method to health achievement index. Moreover, an Oaxaca-Blinder type decomposition is applied to the results of the RIF regression decompositions to estimate the relative contributions of various socioeconomic and demographic characteristics. Using the NHIS data on the general US population, we find positive evidence of an existing socioeconomic inequality which reduced in severity from 2005 to 2015. People with less social and human capital are found to experience more severe inequality on average.

Contents

1. Introduction

It is common knowledge that an adequate amount of physical activity is good for one's health. Indeed, the relationship between physical exercise and health is heavily investigated in the health literature. For cardiovascular diseases alone, adequate physical activity is suggested to help to maintain blood pressure and plasma viscosity, both of which can prevent the onset of cardiovascular diseases (MacAuley et al.,1996; Koenig, et al., 1997). It is also shown that inadequate physical exercise is associated with an increased risk of stroke (Sacco et al., 1998). More relevant to this paper, an adequate amount of leisure time physical activity is shown to have a preventive effect on coronary heart disease, hypertension, and diabetes in middle-aged men and women (Haapanen, et al., 1997). Regular leisure-time physical activity is also shown to help to maintain cognitive function and found to be associated with a reduced risk of dementia and Alzheimer's disease in the middle-aged population (Rovio et al., 2005). In their review of medical journals, Warburton et al. (2006) showed that "there is irrefutable evidence of the effectiveness of regular physical activity in the primary and secondary prevention of several chronic diseases" (p. 801) including cardiovascular disease, diabetes, cancer, hypertension, depression, osteoporosis, and premature death. Furthermore, an adequate amount of exercise of about 1000 kcal (4200 kJ) per week is shown to be associated with a 20%-30% reduction in all-cause mortality (Lee & Skerrett, 2001; Paffenbarger Jr, Hyde, Wing, & Hsieh, et al., 1986; Paffenbarger Jr et al., 1993).

Research on physical activity is also not uncommon in the field of health economics. According to Lechner (2009), a high level of sports activity is shown to have a significantly positive effect on labour market outcome and subjective measure of health. In their study of Russian children, Jahns et al. (2012) found a significant association between hours spent in moderate physical activity and a decrease in overweight prevalence. In a series of studies on Canadian adults, Sarma et al. (2014) and Sarma et al. (2015) found a significant association between the participation in leisure time physical activity and a reduction in the probability of being overweight, obese, and having diabetes, high blood pressure, and heart diseases. Moreover, Humphreys et al. (2014) found, using evidence from Canada, significant marginal effects of participation in physical activity on the probability of having chronic diseases and of having fair or poor self-reported health.

We can see there are plenty of studies confirming the causal relationship and statistical association between physical activity and health, which provides the incentive for the objective of this paper. The importance of physical activity as a determinant of health and economic well-being makes the inquiry into the inequality of physical activity meaningful. Reviewing the literature in health economics, there are, within the author's knowledge, few articles from the decomposition perspective in this topic. Nevertheless there are works done regarding physical activity inequality from the medical and health literature. In their systematic review of researches on the socioeconomic inequalities in physical activity using data on European countries and regions, Beenackers et al. (2012) found that, out of the 75 studies and 200 unique associations on the association between

total leisure-time physical activity and socioeconomic status, 68% of the associations and most of the studies confirm a positive association. Moreover, Cerin and Leslie (2008) found significant effects of educational attainment and annual household income on recreational walking and other leisure-time physical activities. These studies investigate the perceived inequality by focusing on socioeconomic differences on physical activity.

This paper aims to precisely measure the recent change in socioeconomic inequality in leisure time physical activity in the US adult population using health concentration and health achievement indexes and to investigate the factors which contributed to the change using RIF-OLS regression decomposition and the Oaxaca-Blinder decomposition. This angle of investigating the socioeconomic inequality in health behaviours is directly combining and applying some ideas from the inequality literature and the decomposition literature, as noted in the following paragraphs.

Firpo et al. (2009) first used influence function, a tool used in statistics and econometrics for robust estimation, to develop the recentered influence function (RIF) regression method to estimate the marginal impact of changing the distribution of explanatory variables on the quantiles of the distribution of the outcome variable. Following their development of the RIF regression method, Firpo et al. (2011) explored the identification requirements for the application of RIF regression by clearly outlining the necessary assumptions required to interpret the estimates as the intended result. Etile (2014), using data on French households, combined the RIF regression method with the Oaxaca-Blinder method to derive the composition and structural effects of education on the

change of inequality in BMI while also controlling for other possible explanatory variables in the regression. However, the previous studies all used absolute and univariate inequality indexes. Firpo et al. (2009) focused on the quantiles and Etile (2014) used deciles, quantiles, and the Gini index as measures of inequality.

This paper's closest reference is Heckley et al. (2016), which extended the RIF regression decomposition method to "bivariate rank dependent indices of socioeconomic inequality in health, including the concentration index" (p. 89). From the angle of the decomposition literature of bivariate rank dependent indices of inequality, Heckley et al. (2016) pointed out the shortcomings of the previous method used to decompose bivariate rank dependent indices and illustrated that the RIF regression decomposition method can be generally applied to a family of bivariate rank dependent indices.

This paper is, to the author's knowledge, the first empirical application of the RIF regression decomposition on physical activity inequality and the first to derive the RIF of health achievement index. Following Heckley et al. (2016) and Etile (2014), both RIF decomposition and the Oaxaca-Blinder type decomposition are carried out in this paper to estimate the marginal partial effects of various explanatory variables on physical activity inequality and the relative contributions of composition and structural effects to the overall effect.

This paper investigates the socioeconomic inequality of physical activity of the U.S. population of two time points, 2005 and 2015, and finds that there was a moderate

reduction in inequality and an increase in health achievement. Inequality among people of different income, education, age, and family size possibly caused the existing inequality. It is shown that people with less social and human capital often experience deeper inequality in physical activity.

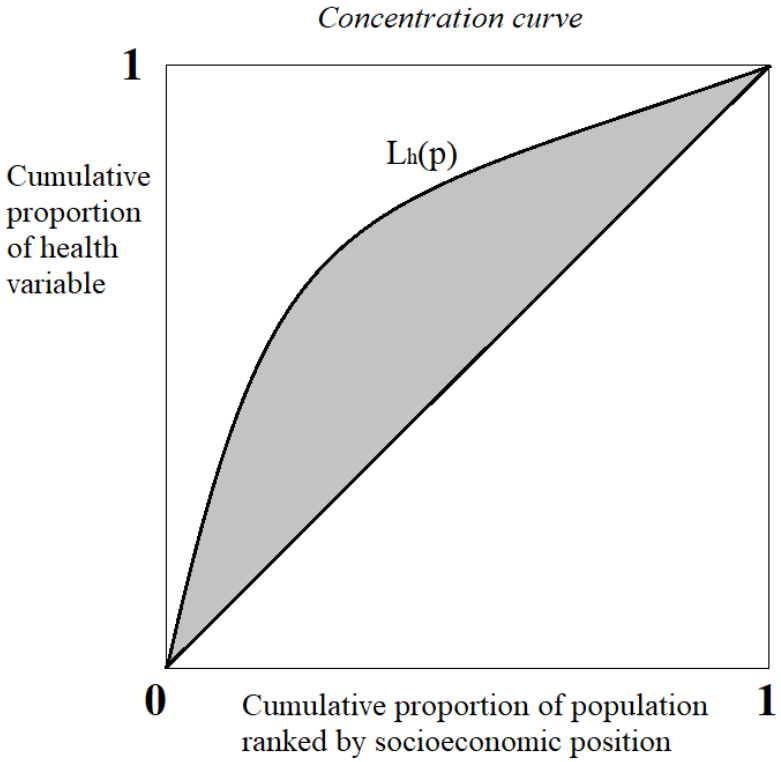
Section 2 introduces the methods and concepts used by this paper. Section 3 discusses the empirical estimation strategy and the identification issue, the objective and necessary assumptions, followed by section 4 which describes the data. Section 5 gives analyses of the data and the decomposition results. Section 6 concludes the paper with a discussion of the results and the limitations of the paper.

2. Preliminaries

2.1. Health Concentration Curve

The health concentration curve plots the cumulative proportion of the population ranked by income against the cumulative proportion of the objective variable, which gives a complete picture of how the objective variable varies across the full distribution of the ranked population (O'Donnell et al., 2008). An example is given in figure 1:

Figure 1



Source: Author

In figure 1, for any point on the $L_h(p)$ curve, L_h is the cumulative proportion of the health variable at the p cumulative proportion of population ranked by socioeconomic position. Both L_h and p have a range between 0 and 1. In the case of figure 1, the concentration curve is above the 45-degree line. If we consider the health variable to be a measure of illness, then this graph shows the poor on average experience more illness, meaning a propoor distribution of illness.

If the concentration curve is the same as the 45-degree line, there is no socioeconomic inequality in the distribution of the objective health variable in the sample population. In contrast, the greater the degree of inequality of the distribution of the health variable in

the population, the farther the concentration curve will lie from the 45-degree line (Wagstaff, 2002). If the concentration curve of a country at time T_1 lies everywhere closer to the 45-degree line than at time T_2, it is said that T_1's concentration curve dominates that of T_2, meaning a smaller degree of socioeconomic inequality in the health variable (O'Donnell et al., 2008).

Originally developed by Kakwani (1977) to be used together with the Lorenz curve to measure tax progressivity, the concentration curve and the related concentration index have been widely adopted in the health economics literature as a measure of inequality. For example, O'Donnell et al. (2007) used concentration curves of healthcare to examine if public healthcare spending in 11 Asian developing countries is propoor, and they find that most of the Asian developing countries have prorich public health care spending. Van Doorslaer et al. (1997) used income and self-assessed health to construct concentration curves and find that all of the developed countries they studied were prorich self-assessed health distribution, with particularly high inequality in the United States and the United Kingdom. Wagstaff (2000) used concentration curves to measure if child mortality is more unequally prevalent in poor families in one country than another.

In the actual construction of the curve, both individual-level data and grouped data can be used in the construction of the concentration curve (O'Donnell et al., 2008). In this paper's application of the concentration curve, the ratio of family income to the poverty line is used as the socioeconomic status indicator to rank individuals and physical activity is the objective health outcome.

2.2. Concentration Index

Definition of Concentration Index:

$$CI = 1 - 2\int_0^1 L_h(p)dp, \quad p \in [0,1] \quad L \in [0.1] \quad CI \in [-1,1] \qquad (1)$$

The concentration index (*CI*) is equal to one minus two times the area below the concentration curve $L_h(p)$. In our case of figure 1, the *CI* is equal to negative two times the shaded area. If the curve lies everywhere below the 45-degree line the *CI* value is positive. A negative *CI* value means a disproportionate distribution of the health variable to the individuals with lower socioeconomic ranks.

Another formula for the CI is:

$$CI = \frac{2}{\mu} cov(h, r) \qquad (2)$$

In equation 2, h is the objective health variable, μ is the sample mean of the health variable, r is the socioeconomic rank. Equation 2 indicates that the concentration index only depends on the relationship between the objective health variable and the rank of the socioeconomic status variable and not on the variation in the socioeconomic status variable itself (O'Donnell et al., 2008).

The sign of the concentration index indicates the direction of the association between the health variable and the socioeconomic status, "and its magnitude reflects both the strength of the relationship and the degree of variability in the health variable"

(O'Donnell et al., 2008, p. 96). Moreover, Koolman and van Doorslaer (2004) showed that multiplying the concentration index by 75 is equal to the percentage of the cumulative health variable needed to be linearly redistributed from the richer half to the poorer half of the sample population to arrive at a concentration index value of zero.

The concentration index gives a way to quantify the degree of socioeconomic inequality in a health variable (Kakwani et al., 1997; Wagstaff et al., 1989). The concentration index has been, for example, used to measure and compare the extent of socioeconomic inequality by Wagstaff (2000) in child mortality, by Wagstaff et al. (2003) in child malnutrition, by van Doorslaer et al. (1997) in adult health, and by O'Donnell et al. (2007) in health subsidies.

2.3. Achievement index (AI) of CI

Definition of the achievement index of CI:

$$AI = \mu(1 - CI) \tag{3}$$

The implicit value judgement underlying the concentration index is that the weight of the health variables of the individuals decreases proportionally as the individuals' cumulative socioeconomic ranks moves from 0 to 1 (Wagstaff, 2002). In the context of the concentration curve, this form of achievement, AI, can be thought of as a weighted average of the health outcome variable of the members of the sample population, where weights are proportional to the inverse of their cumulative socioeconomic rank (Wagstaff, 2002).

Think of an example opposite to the case in figure 2 with the concentration curve lying below the 45-degree line and let the health variable be a measure of good health. In this case, the value of (*1-CI*) is less than 1 as *CI* is positive. A deeper inequality against the socioeconomically poor will give a larger *CI*, a smaller (*1-CI*) value, and a smaller *AI* index, holding the mean, μ, constant. The resulting estimate will be a smaller good health achievement index in the distribution with deeper inequality. A deeper socioeconomic inequality against the socioeconomically poor will, in this example, make the value of *AI* further away from the mean of the good health variable. If the curve lies above the 45-degree line and the health outcome is a bad health indicator, it is straightforward to see that deeper inequality will give a larger ill health achievement index.

2.4. Recentered influence function

The influence function is a special form of directional derivative which is used to find the influence of a perturbation in a distribution on a statistic of that distribution. The IF is of a particular form that the perturbation distribution, denoted as δ_h, is equal to a cumulative distribution function that puts probability 1 at a particular value h:

$$\delta_h(l) = \begin{cases} 0, & \text{if } \ l < h \\ 1, & \text{if } \ l \geq h \end{cases} \tag{4}$$

Denote the original distribution as F_H, and the IF of the functional $v(F_H)$ evaluated at point h as IF(h;v). Define G_h as:

$$G_h = (1-\varepsilon)F_H + \varepsilon\delta_h, \quad \varepsilon \in [0,1] \tag{5}$$

G_h is a distribution ε away from F_H in the direction of δ_h. IF(h;v) is then defined as:

$$IF(h;v) = \frac{\partial v(G_h)}{\partial \varepsilon}|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{v(G_h) - v(F_H)}{\varepsilon} \tag{6}$$

if the limit is defined for every point in h. The IF gives the influence of any individual in the sample population on the objective functional $v(F_H)$ (Wilcox, 2005). Calculating the IF will yield an IF value for each individual in the sample population (Heckley et al., 2016). Now we can show the RIF is defined as:

$$RIF(h; v) = v(F_H) + IF(h; v) \tag{7}$$

The RIF can be seen as a transformation of the IF, which is obtained by adding back the original functional to the IF (Heckley et al., 2016). The expectation of the RIF is equal to the original statistic $v(F_H)$ (Firpo et al., 2009) as the expectation of the IF is zero (Monti, 1991):

$$\begin{aligned} E[RIF(h; v)] &= E[IF(h; v)] + E[v(F_H)] \\ &= \int IF(h; v, F_H) \cdot dF_H(h) + E[v(F_H)] \\ &= 0 + E[v(F_H)] = E[v(F_H)] \end{aligned}$$

Now we derive the RIF in the case of a rank dependent index, I :

$$G_{h,F_Y(y)} = (1 - \varepsilon)F_{H,F_Y} + \varepsilon\delta_{h,F_Y(y)}, \quad \varepsilon \in [0, 1] \tag{8}$$

$$\delta_{h,F_Y(y)}(l, r) = \begin{cases} 0, & \text{if } \ l < h \text{ or } r < F_Y(y) \\ 1, & \text{if } \ l \geq h \text{ or } r \geq F_Y(y) \end{cases} \tag{9}$$

Here in equation 8, $G_{h,\,F}$ is a bivariate distribution by a perturbation by $F_{H,F}$ in both h and F. In equation 9, $\delta_{h,F}$ is a joint cumulative distribution function that gives mass 1 to (h, $F_Y(y)$) jointly, where l and r are draws from H and F_Y respectively. Now the bivariate IF of the functional v^I at point (h, $F_Y(y)$) is defined as:

$$IF(h, F_Y(y); v^I) = \frac{\partial v^I(G_{h,F_Y(y)})}{\partial \varepsilon}|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{v^I(G_{h,F_Y(y)}) - v(F_{H,F_Y})}{\varepsilon} \tag{10}$$

given that this limit function is defined for every point (h, $F_Y(y)$). The RIF of the rank dependent index, I, is now defined as:

$$RIF(h, F_Y(y); v^I) = v^I(F_{H,F_Y}) + IF(h, F_Y(y); v^I) \tag{11}$$

2.5. Oaxaca-Blinder Decomposition

The Oaxaca-Blinder decomposition decomposes a change or a gap in the mean outcomes between two groups which can be two time points or two countries, assuming a linear structural model:

$$Y_g = X_g \beta_g + v_g, \quad \text{for } g = A, B \tag{12}$$

where $E[v_g|X]=0$. A and B indicate the two groups. The model is assumed to be linear in observable characteristics, X, and unobservable characteristic, v. Let $D_g=1$ be an indication of group g membership, and an assumption of conditional independence/ignorability is required to restrict the conditional distribution of the unobservables (Firpo et al., 2011): for g = A, B, and (D_g, X, v) having a joint distribution, v is independent of D_g given $X=x$ for all x in X. Ignorability ensures there is no confounding effect of v on X, which identifies the structural effect. The zero conditional mean ensures the expected value of v equals zero. The linearity assumption, with the previous two assumptions, achieves the identifications of structural and composition effects in the decomposition of the gap. Now the gap in the mean outcomes, Δ_O, can be written as

$$
\begin{aligned}
\Delta_O &= E[Y_B] - E[Y_A] \\
&= E[Y_B|D_B = 1] - E[Y_A|D_B = 0] \\
&= E[E(Y_B|X, D_B = 1)|D_B = 1] - E[E(Y_A|X, D_B = 0)|D_B = 0] \\
&= (E[X|D_B = 1]\beta_B + E[v_B|D_B = 1]) - \\
&\quad (E[X|D_B = 0]\beta_A + E[v_A|D_B = 0]) \\
&= E[X|D_B = 1]\beta_B - E[X|D_B = 0]\beta_A \\
&= (E[X|D_B = 1]\beta_B - E[X|D_B = 1]\beta_A) + \\
&\quad (E[X|D_B = 1]\beta_A - E[X|D_B = 0]\beta_A) \\
&= \Delta_S + \Delta_X
\end{aligned} \tag{13}
$$

The empirical estimation is then carried out as:

$$
\begin{aligned}
\widehat{\Delta_O} &= \overline{X}_B\widehat{\beta}_B - \overline{X}_B\widehat{\beta}_A + \overline{X}_B\widehat{\beta}_A - \overline{X}_A\widehat{\beta}_A \\
&= \overline{X}_B(\widehat{\beta}_B - \widehat{\beta}_A) + (\overline{X}_B - \overline{X}_A)\widehat{\beta}_A \\
&= \widehat{\Delta_S} + \widehat{\Delta_X}
\end{aligned} \tag{14}
$$

This decomposition method, developed by Oaxaca (1973) and Blinder (1973) in their seminal papers, aims to examine how much of the gap can be attributed to the differences in observable characteristics (Elder et al., 2010). The structural effect measures the component of the overall change attributable to the differences in the returns to the observable characteristics and the composition effect measures the component attributable to the differences in the observable characteristics. Both Oaxaca (1973) and Blinder (1973) pointed out the decomposition in equation (20) is not unique as an equally convincing alternative exists:

$$
\begin{aligned}
\widehat{\Delta_O} &= \overline{X}_B\widehat{\beta}_B - \overline{X}_A\widehat{\beta}_B + \overline{X}_A\widehat{\beta}_B - \overline{X}_A\widehat{\beta}_A \\
&= \overline{X}_A(\widehat{\beta}_B - \widehat{\beta}_A) + (\overline{X}_B - \overline{X}_A)\widehat{\beta}_B \\
&= \widehat{\Delta_S} + \widehat{\Delta_X}
\end{aligned} \tag{15}
$$

There is no reason to believe which particular way of decomposition is better, and we will use both alternatives. We will, besides the two traditional alternatives, also adopt a

decomposition form which is a combination of the two alternatives and gives the result of the structural and composition effects as the average of two forms:

$$\begin{aligned}\widehat{\Delta_O} &= (\overline{X}_B\widehat{\beta}_B - \frac{1}{2}(\overline{X}_B\widehat{\beta}_A + \overline{X}_A\widehat{\beta}_B)) \\ &\quad + (\frac{1}{2}(\overline{X}_B\widehat{\beta}_A + \overline{X}_A\widehat{\beta}_B) - \overline{X}_A\widehat{\beta}_A) \\ &= \widehat{\Delta_S} + \widehat{\Delta_X}\end{aligned}$$

3. Estimation strategy and identification issue

The RIF regression method used in this paper is a direct application of the general decomposition method derived in Heckley et al. (2016) but we also extend it to the health achievement index. Heckley et al. (2016) covered the decomposition method of a family of bivariate rank dependent indices in their paper and derived all related empirical estimation formulas. The proofs can be seen in their paper (Heckley et al., 2016), and I will cover only the necessary formulas and assumptions needed for this paper's application here.

3.1. RIF regression

Continuing with the end of section 2.4, recentering the IF implies this:

$$v^I(F_{H,F_Y}) = \int_{-\infty}^{\infty} RIF(h, F_Y(y); v^I) \cdot dF_{H,F_Y}(h, F_Y(y)) = E[RIF(H, F_Y; v^I)]$$
$$= \int_{-\infty}^{\infty} E[RIF(H, F_Y; v^I)|X = x] \cdot dF_X(x) \quad (16)$$

Following Firpo et al. (2009)'s notations, the RIF regression identifies two parameters of interest: the marginal effect of covariates on a functional, an individual effect, and the unconditional partial effect, a population effect measure. The unconditional partial effect measures the impact of a marginal location shift in a continuous covariate or the impact of a marginal change in the conditional distribution of a binary covariate holding everything else constant (Heckley et al., 2016). Relating this to the objective of this paper, the unconditional partial effect measures how an equal marginal increase in a socioeconomic independent variable for everyone would change the CI and the AI.

The first parameter of interest, the marginal effect of the covariates, is given by this formula:

$$\frac{dE[RIF(H, F_Y; v^I)|X = x]}{dx} \quad (17)$$

which is the partial derivative of the regression estimates.

The second parameter of interest is the unconditional partial effect, the response of the index, *I*, to a marginal change in the covariate. For binary covariate, this unconditional partial effect measures the response of *I* to the marginal changes in the conditional distribution of the binary covariate (Heckley et al., 2016), which is expressed as:

$$\int_{-\infty}^{\infty} \frac{dE[RIF(H, F_Y; v^I)|X = x]}{dx} \cdot dF_X(x) \tag{18}$$

3.2. RIF-OLS regression

The calculated RIF values are regressed with OLS regression on various socioeconomic variables across the sample population to find the unconditional partial effects which measure how an equal marginal increase in any socioeconomic independent variable for everyone would change the CI and the AI. For this method to produce its intended estimates, we need the following assumptions:

Additive linearity. Assuming the RIF regression model has a linear-parameter functional form with an additive error term, we have the expectation of the RIF as:

$$E[RIF(H, F_Y; v^I)|X = x] = X\beta + \mu \tag{19}$$

Zero conditional mean. $E[\mu|X]=0$.

Linearity and conditional mean independence imply a constant marginal effect of the covariates along the distribution of *X* and now the two parameters of interest are the same:

$$\frac{dE[RIF(H, F_Y; v^I)|X = x]}{dx} = \int_{-\infty}^{\infty} \frac{dE[RIF(H, F_Y; v^I)|X = x]}{dx} \cdot dF_X(x) = \beta \tag{20}$$

From Heckley et al. (2016), one can derive the following estimation formulas for the IFs and RIFs of the CI and AI:

$$\widehat{IF}(h_i, y_i; CI) = \frac{-h_i}{\widehat{\mu}_H}\widehat{CI} + 1 - \frac{h_i}{\widehat{\mu}_H} + 2\frac{h_i}{\widehat{\mu}_H}\widehat{F}_Y(y_i) - 2\widehat{C}(\widehat{F}_Y(y_i)) \quad (21)$$

$$\widehat{RIF}(h_i, y_i; CI) = \frac{\widehat{\mu}_H - h_i}{\widehat{\mu}_H}\widehat{CI} + 1 - \frac{h_i}{\widehat{\mu}_H} + 2\frac{h_i}{\widehat{\mu}_H}\widehat{F}_Y(y_i) - 2\widehat{C}(\widehat{F}_Y(y_i)) \quad (22)$$

$$\widehat{IF}(h_i, y_i; AI) = 2h_i - 2\widehat{\mu}_H + 2\widehat{\mu}_H\widehat{CI} - 2h_i\widehat{F}_Y(y_i) + 2\widehat{\mu}_H\widehat{C}(\widehat{F}_Y(y_i)) \quad (23)$$

$$\widehat{RIF}(h_i, y_i; AI) = 2h_i - \widehat{\mu}_H + \widehat{\mu}_H\widehat{CI} - 2h_i\widehat{F}_Y(y_i) + 2\widehat{\mu}_H\widehat{C}(\widehat{F}_Y(y_i)) \quad (24)$$

where

$$\widehat{F}_Y(y_i) = \frac{\sum_{j=i}^{N} 1(y_j \leq y_i)}{N} \quad (25)$$

$$\widehat{C}(\widehat{F}_Y(y_i)) = \frac{\sum_{j=1}^{N} h_j \cdot 1(y_j \leq y_i)}{N} \quad (26)$$

The $1(y_j \leq y_i)$ term in the numerators in equations (25) and (26) is equal to 1 if the argument in the bracket is true and equal to 0 if the argument is wrong. The empirical estimation of RIF values is carried out for each individual in the sample population. Then the empirical RIF values are used as the dependent variable in an OLS regression against various socioeconomic variables to yield the unconditional partial effects. What has to be noted is that the unconditional partial effects given by the RIF-OLS regression are a local effect estimate of a small change in the covariates, and should only be considered for relatively small changes (Heckley et al., 2016).

3.3. Oaxaca-Blinder decomposition of the RIF -OLS results

Now we apply the Oaxaca-Blinder decomposition method to give a detailed decomposition of the RIF-I-OLS measured unconditional partial effects. Following sections 2.5 and 3.2 and using the CI as an example, we can now decompose the empirical RIF regression results:

$$
\begin{aligned}
\widehat{\Delta_{CI}} &= E[CI_B] - E[CI_A] \\
&= \widehat{RIF}(h_{B,i}, y_{B,i}; CI) - \widehat{RIF}(h_{A,i}, y_{A,i}; CI) \\
&= \overline{X}_B\widehat{\beta}_B - \overline{X}_B\widehat{\beta}_A + \overline{X}_B\widehat{\beta}_A - \overline{X}_A\widehat{\beta}_A \\
&= (\overline{X}_B(\widehat{\beta}_B - \widehat{\beta}_A)) + ((\overline{X}_B - \overline{X}_A)\widehat{\beta}_A) \qquad (28) \\
&= \overline{X}_B\widehat{\beta}_B - \overline{X}_A\widehat{\beta}_B + \overline{X}_A\widehat{\beta}_B - \overline{X}_A\widehat{\beta}_A \\
&= (\overline{X}_A(\widehat{\beta}_B - \widehat{\beta}_A)) + ((\overline{X}_B - \overline{X}_A)\widehat{\beta}_B) \qquad (29) \\
&= (\overline{X}_B\widehat{\beta}_B - \frac{1}{2}(\overline{X}_B\widehat{\beta}_A + \overline{X}_A\widehat{\beta}_B)) + (\frac{1}{2}(\overline{X}_B\widehat{\beta}_A + \overline{X}_A\widehat{\beta}_B)) - \overline{X}_A\widehat{\beta}_A) \qquad (30) \\
&= \widehat{\Delta_S} + \widehat{\Delta_X}
\end{aligned}
$$

The form in equation (30) of the Oaxaca-Blinder decomposition simply gives the structural and composition effects as the average of the effects produced by the two traditional alternatives. The same decomposition forms are also applied to the achievement index. Moreover, the decompositions in equation (28), (29), and (30) can be carried out for any particular covariate in the RIF-OLS regression and the form remains the same except changing the vectors to scalars.

4. Data

This paper uses the National Health Interview Survey (NHIS), a major program of the National Center for Health Statistics in the United States. The NHIS covers the civilian noninstitutionalized population in the United States and is carried out by cross-sectional household interviews and follows a sampling plan of a multistage area probability design which permits a representative sampling.

The core data of the NHIS contain four major components: Household, Family, Sample Adult, and Sample Child. The household component collects basic demographic information on individuals in a particular house. The family component collects additional demographic and health information on members from each family in the house. For each surveyed family, one sample adult and one sample child, if there is any, are randomly selected and information is collected with the Sample Adult and Sample Child questionnaires. The NHIS also includes supplement of information which provide various additional details on a particular subject.

The particular data we use in this study are from the 2005 and 2015 data releases of the NHIS. The Sample Adult files under the core data files are used to derive the information on physical activity, and the Imputed Income files are used to examine income. The demographic variables used in the RIF regressions are extracted from the Person File under the core data files. We focus on the adult population as currently only Sample Adult questionnaire covers questions on physical activity.

5. Data analysis and results

5.1. Summary statistics

I first give the descriptions and summaries of the variables:

Table 1: Summary Statistics

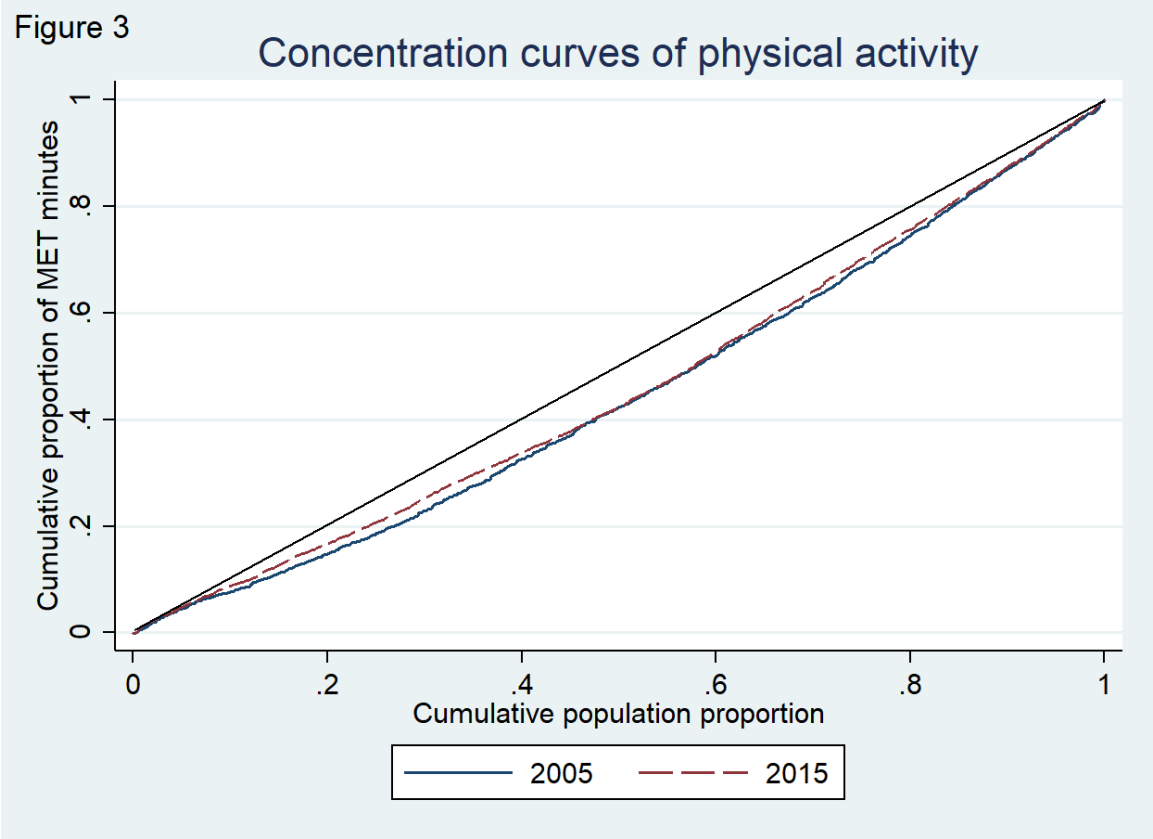
Variable	Description	2005 mean	2015 mean
mm	Weekly moderate activity minitues	108.94	122.16
vm	Weekly vigorous activity minitues	79.492	102.10
income	Family income to poverty line ratio	3.958	4.022
age	Age	45.06	46.90
age1825	Dummy for age 18-25	0.151	0.142
age2535	Dummy for age 25-35	0.187	0.178
age3545	Dummy for age 35-45	0.201	0.168
age4555	Dummy for age 45-55	0.190	0.177
age5565	Dummy for age 55-65	0.130	0.161
age65+	Dummy for age above 65	0.140	0.174
edu1	Dummy for education lower than high school	0.114	0.085
edu2	Dummy for education high school diploma	0.220	0.165
edu3	Dummy for education lower than Bachelor degree	0.309	0.321
edu4	Dummy for education Bachelor degree	0.215	0.242
edu5	Dummy for education graduate school	0.141	0.187
male	Dummy for sex male	0.483	0.481
marital1	Dummy for marital status married	0.563	0.521
marital2	Dummy for marital status divorced, separated, widowed	0.164	0.172
marital3	Dummy for marital status never married	0.200	0.220
marital4	Dummy for marital status others	0.072	0.086
race1	Dummy for race white	0.790	0.746
race2	Dummy for race African American	0.115	0.123
race3	Dummy for race others	0.095	0.131
fsize	Size of family	2.83	2.81
fchild	Number of children present in the family	0.729	0.661
felder	Number of elders present in the family	0.257	0.352
MET	Weekly MET minitues	803.8	877.0

The Metabolic Equivalent of Task (MET) is a measure of the energy cost of physical activities. One MET is roughly equivalent to the energy cost of sitting quietly. According to the website of the Office of Disease Prevention and Health Promotion (ODPHP), a division of the US Department of Health and Human Services, moderate-intensity activities are defined as activities having between 3.0 to 5.9 METs and vigorous-intensity activities are defined as having 6.0 METs or more. Walking at 3.0 miles per hour is regarded as a moderate-intensity activity and running at 10 minutes per mile is a vigorous-intensity activity. Checking the glossary section under Special Topics of the NHIS website, the NHIS agrees with the standard. Using the lower limit, moderate activity is assigned a MET value of 3 and vigorous activity is assigned a MET value of 6. Weekly MET minutes for each individual are then calculated by adding up the MET minutes from moderate activity and vigorous activity. It is worth mentioning that a large portion of the population reports zero leisure time physical activity minute while a few individuals report nearly impossible values. The maximum threshold is set at 7000 total physical activity minutes per week which is equivalent to over 16 hours every day. Out of the 29,214 observations in 2005, 3 exceed this value and are excluded. Out of the 32,256 observations in 2015, 10 observations are excluded for the same reason. In contrast, 12,334 (42.2%) out of the remaining 29,211 observations in 2005 report 0 minute in physical activity and 10,416 (32.3%) out of the remaining 32,246 observations in 2015 report 0 minute. We might think anyone will exert physical effort in their daily life. But the NHIS treats light physical activity as relatively valueless for health and does not include it in the questionnaire. Moreover, this high proportion of 0 MET minute in the

observations is not a problem as all observations will have a different socioeconomic rank and will produce a different empirical RIF value.

The mean values of the variables are calculated with sampling weights. The dummy variables are created with the information from raw data. The ratio of family income to poverty line is used as the socioeconomic variable to rank individuals. With the ranks and the weekly MET minutes of the individuals we can now plot the concentration curves:

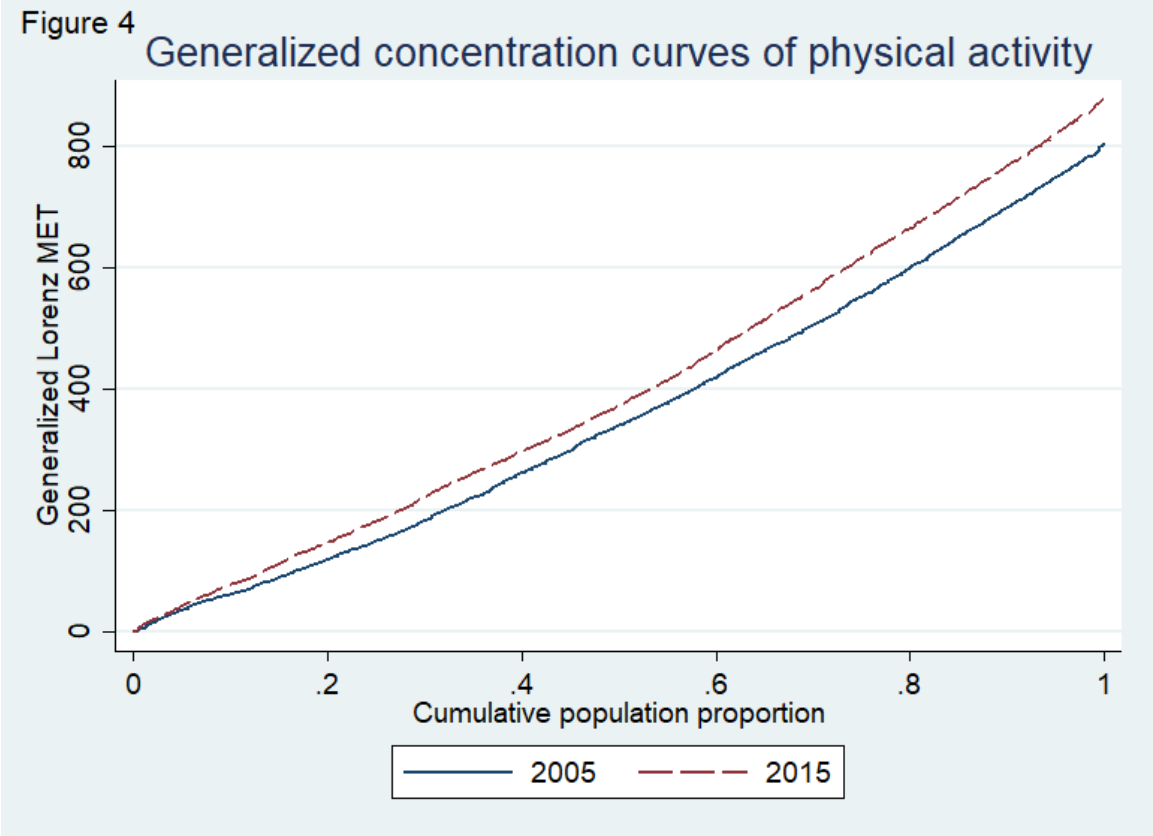
Figure 3



We can see the 2015 concentration curve of physical activity is a bit closer to the 45-degree line, which implies a smaller socioeconomic inequality. Moreover, both curves lie below the 45-degree line which implies a disproportional distribution of more physical activity to the well-off. As the two concentration curves are close to each other, a generalized concentration curve, obtained by multiplying the y-axis Lorenz ordinates of

the concentration curve with the mean of the outcome variable MET, can better show the differences:

Figure 4



A generalized concentration curve takes both relative inequality and the population mean into account. The Y-coordinate of a point on the curve is equal to the average MET minutes among the respective X cumulative proportion of ranked population.

With the concentration curves and the mean of weekly MET minutes we can now calculate the CI and AI:

Table 2: Indexes

Index	2005	2015
CI	0.1055	0.0864
AI	719.0	801.3

2015 has a smaller CI, which is consistent with the graph. Referring to table 1, the mean MET minutes is 877 in 2015, compared to 803.8 in 2005. A smaller CI and a larger mean MET in 2015 produces a larger AI.

5.2. RIF regression results

Using the empirical estimation formulas, listed in section 3.2, we can calculate the IF and RIF values for each individual. Regressing the empirical RIF values on the covariates, we can carry out the measurement of the unconditional partial effects of the covariates on the indexes:

Table 3: CI RIF regression results

	2005 (1)	2015 (1)	2005 (2)	2015 (2)	2005 (3)	2015 (3)
income	0.016***	0.013***	0.017***	0.013***	0.018***	0.014***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
male	-0.018	-0.01	-0.015	-0.008	-0.013	-0.007
	(0.015)	(0.017)	(0.015)	(0.016)	(0.015)	(0.016)
edu2	-0.084***	-0.1***	-0.081**	-0.102***	-0.083**	-0.109***
	(0.032)	(0.035)	(0.032)	(0.035)	(0.033)	(0.035)
edu3	-0.178***	-0.137***	-0.174***	-0.139***	-0.176***	-0.147***
	(0.032)	(0.033)	(0.032)	(0.033)	(0.033)	(0.033)
edu4	-0.143***	-0.111***	-0.136***	-0.111***	-0.138***	-0.12***
	(0.034)	(0.035)	(0.034)	(0.035)	(0.035)	(0.035)
edu5	-0.111***	-0.077	-0.103***	-0.075	-0.107***	-0.083**
	(0.04)	(0.04)	(0.04)	(0.039)	(0.04)	(0.039)
age2535	0.118***	0.034***	0.109***	0.035	0.123***	0.064
	(0.033)	(0.035)	(0.034)	(0.034)	(0.034)	(0.033)
age3545	0.135***	0.09***	0.124***	0.092***	0.135***	0.118***
	(0.033)	(0.034)	(0.033)	(0.034)	(0.033)	(0.033)
age4555	0.131***	0.075**	0.118***	0.073**	0.135***	0.094***
	(0.032)	(0.035)	(0.033)	(0.035)	(0.032)	(0.034)
age5565	0.115***	0.109***	0.101***	0.103***	0.129***	0.13***
	(0.034)	(0.035)	(0.035)	(0.038)	(0.036)	(0.038)
age65+	0.152***	0.111***	0.136***	0.1***	0.175***	0.111**
	(0.033)	(0.034)	(0.034)	(0.036)	(0.05)	(0.046)
marital2			0.034**	0.035	0.062***	0.074***
			(0.017)	(0.021)	(0.021)	(0.024)
marital3			-0.015	0.008	0.01	0.031
			(0.023)	(0.025)	(0.025)	(0.025)
marital4			0.041	0.028	0.056	0.044
			(0.029)	(0.024)	(0.029)	(0.024)
race2			0.028	-0.046	0.025	-0.048
			(0.027)	(0.028)	(0.027)	(0.028)
race3			0.012	-0.051***	-0.001	-0.068***
			(0.023)	(0.019)	(0.024)	(0.02)
fsize					0.029**	0.04***
					(0.013)	(0.01)

Table 3: CI RIF regression results

	2005 (1)	2015 (1)	2005 (2)	2015 (2)	2005 (3)	2015 (3)
fchild					-0.012	-0.033***
					(0.016)	(0.014)
felder					-0.005	0.016
					(0.021)	(0.023)
Constant	0.274***	0.171***	0.266***	0.177***	0.165***	0.054
	(0.043)	(0.041)	(0.045)	(0.041)	(0.05)	(0.049)

**: Coefficient significant at the 5% significance level.

***: Coefficient significant at the 1% significance level.

The bracketed values are standard errors.

(1) is regression with income, gender, education, and age. (2) is regression with an addition of covariates indicating marital status and race. (3) is regression with further addition of family size, number of children in family, and number of elders in family. Discussion is focused on regression (3) results.

Table 4: AI RIF regression results

	2005 (1)	2015 (1)	2005 (2)	2015 (2)	2005 (3)	2015 (3)
income	17.9***	18.1***	16.3***	19***	13.6***	17.4***
	(3.8)	(4.2)	(3.9)	(4.4)	(4.2)	(4.5)
male	262***	222.5***	256.8***	218.9***	254.4***	219.6***
	(23.7)	(25)	(23.6)	(25.1)	(23.5)	(24.9)
edu2	76.3	98.2	65.4	94.8	75.8	108.8
	(54.3)	(62.7)	(54.6)	(62.7)	(56.1)	(62.8)
edu3	356.2***	233.8***	342.9***	228.7***	352.5***	244.3***
	(55.1)	(57.9)	(55.5)	(57.9)	(56.8)	(57.9)
edu4	365.9***	260.6***	352.7***	258.2***	361.2***	276.5***
	(55.5)	(58.7)	(55.9)	(58.9)	(56.9)	(58.5)
edu5	383.1***	329***	372.1***	328.6***	384.5***	344.6***
	(60.7)	(63.6)	(60.8)	(63.3)	(61.2)	(63.4)
age2535	-362.2***	-194.8***	-301***	-145.5***	-350.1***	-210.6***
	(55)	(54.7)	(56.7)	(54.9)	(56.2)	(55.3)
age3545	-400.4***	-335.1***	-327.9***	-270.3***	-367.9***	-332***
	(54.7)	(52.1)	(56.4)	(55.2)	(57.1)	(56.8)
age4555	-490.2***	-424.4***	-415.6***	-360***	-458***	-413***
	(51.6)	(50.4)	(53.9)	(54.6)	(54.1)	(54.6)
age5565	-491.3***	-454.5***	-416.9***	-391.3***	-483.1***	-467.7***
	(55.2)	(49)	(59.1)	(56.7)	(60.7)	(56.4)
age65+	-573.3	-511	-506.2	-446.3	-576.2	-598.4
	(55)	(51.4)	(58.5)	(56.8)	(85.8)	(75.2)
marital2			8	12.9	-79.3**	-47.2
			(28.5)	(36.5)	(34.8)	(40.9)
marital3			130.1***	107.3***	68.6	59.1
			(40.2)	(40.2)	(43.1)	(40.3)
marital4			35.3	-12.8	-3.6	-40
			(49.7)	(44.2)	(50.9)	(44.2)
race2			-123.2***	25.5	-116.3**	31
			(47.6)	(47.7)	(47.7)	(47.5)
race3			-151.8***	-73.4**	-114.2***	-39
			(38)	(32.4)	(39.5)	(32.9)
fsize					-90.9***	-87.7***
					(21.1)	(14.9)

Table 4: AI RIF regression results

	2005 (1)	2015 (1)	2005 (2)	2015 (2)	2005 (3)	2015 (3)
fchild					61.8**	77.3***
					(28)	(21.8)
felder					-5.8	48.7
					(37.4)	(33.3)
Constant	556.6***	719.8***	515.8***	650***	801.8***	907.4***
	(69.5)	(69.4)	(73.7)	(72.4)	(86)	(83.5)

**: Coefficient significant at the 5% significance level.

***: Coefficient significant at the 1% significance level.

The bracketed values are standard errors.

(1) is regression with income, gender, education, and age. (2) is regression with an addition of covariates indicating marital status and race. (3) is regression with further addition of family size, number of children in family, and number of elders in family. Discussion is focused on regression (3) results.

From table 3, we can see that income, age, education, and family size have significant coefficients in both years. One dummy variable in each group has to to be omitted in the regression, and the regression coefficients will have to be interpreted with respect to the the omitted group. In contrast, the interpretation of the coefficients is rather straighforward for contiuous variables.

Take family size (fsize) for example. It has a significant coefficient of 0.029 in 2005 and 0.04 in 2015, which implies a universal increase of family member in every family would increase the concentration index by 0.029 from 0.1055 in 2005 and by 0.04 from 0.0864

in 2015. This means that if there are more large families, the distribution of physical activity will be more socioeconomically unequal. To be more clear, this measured positive effect, combined with the positive CI, not only suggests that the less well-off people in a larger family participate less in physical activity than the better-off people, but also that the socioeconomic wellness has heavier influences on physical acitivity for people in a larger family.

Every dummy variable for education has, in regression (3), a significant coefficient. The omitted is the group of people who have an education level lower than complete high school education. We can see each step up in education level has significant effects on inequality. The coefficents are all negative and large, which implies highly educated people are under smaller influences of socioeconomic wellness on their physical acitivity participation. In other words, the inequality in physical activity between the rich and poor is less severe for people with higher education levels. Another significant covariate is age. The base group for comparison is the group of people of age between 18 and 25. Compared to young adults, all older groups have a large and significant effect. The significant effects suggest that older people's participation in physical acitivities is more socioeconomiclly unequal. From another angle, young people exerise more uniformly regardless of their socioeconomic wellness.

Compared to the married people, the seperated, divorced, or widowed (marital2) is shown to have a significant positive effect on CI. This suggests that, compared to the people still

in marriage, people who exited their marriage face more influence from socioeconomic wellness.

The significant variables in the CI regression should largely stay significant in the regression of AI as the AI is closely related to CI. From table 4, we can see income, age, education, and family size stay significant. The signs for the coefficients are opposite to the signs of the respective coefficients in table 3 with the exception of income. This is expected as a mitigating effect on CI contributes positively to AI.

However, income's effect is positive in both tables. The measured effect is significant at the 1% significance level for both indexes and years. This seems puzzling at the first look but a good explanation exists: people with more income exercise more, thus the positive effect of income on AI, despite a marginal increase in family income for everyone will enlarge the current inequality, thus the positive coefficient in the CI regression. As long as the effect of income on the mean of physical activity dominates its effect on CI, the coefficients can be positive in both CI and AI RIF regressions.

Race, sex, and number of children in family start to become significant in the regression of AI. The difference of the AI is that it also takes the mean of the health variable into consideration. Both an effect on the mean and an effect on the CI can contribute to the overall effect on the AI. This possibly explains why some variables start to have significant coefficients in the regression of AI.

From table 4, being male has a positive partial effect of 254.4 in 2005 and 219.6 in 2015. We can see the unchanged direction and the large magnitude of the effect for the two years, which suggests a higher achivement of physical activity in males compared to females in the US. However, compared to the coefficient of the male variable in table 3, we can see the coefficient of male in the CI regression is close to 0 and insignificant. This suggests the effect of the male covariate on AI should largely come from a higher mean value in physical acitivity among male population.

Moreover, race and family children number also become significant in the regression of AI. Race1 is the dummy for white and is ommited. We can see peole of black and other races generally have a lower socioeconomic achivement, compared to white people in 2005. However, this effect becomes obsolete in 2015, suggesting a possible reduced inequality in achivement between races. The number of children in family (fchild) has negative coefficient in the CI regression and positive coefficient in the AI regression, which suggests people in families with more children tend to experience less inequality in physical activity.

5.3. Oaxaca-Blinder decomposition of the RIF measured effects

Now that we have the coefficients for each covaraite in the RIF regressions we carry out the Oaxaca-Blinder decomposition:

Table 5: CI decomposition

Predicted CI 2015	0.189					
Predicted CI 2005	0.319					
Difference	-0.13					
Decomposition form	1	2	3	1	2	3
	Composition	Com	Com	Structural	Str	Str
income	0.001	0.001	0.001	-0.016	-0.016	-0.016
male	0	0	0	0.003	0.003	0.003
age2535	-0.001	-0.001	-0.001	-0.011	-0.01	-0.011
age3545	-0.004	-0.004	-0.004	-0.003	-0.003	-0.003
age4555	-0.001	-0.002	-0.002	-0.008	-0.007	-0.008
age5565	0.004	0.004	0.004	0	0	0
age65+	0.004	0.006	0.005	-0.009	-0.011	-0.01
edu2	0.006	0.005	0.005	-0.006	-0.004	-0.005
edu3	-0.002	-0.002	-0.002	0.009	0.009	0.009
edu4	-0.003	-0.004	-0.003	0.004	0.004	0.004
edu5	-0.004	-0.005	-0.004	0.003	0.004	0.004
race2	0	0	0	-0.008	-0.009	-0.009
race3	-0.002	0	-0.001	-0.006	-0.009	-0.008
marital2	0.001	0	0	0.002	0.002	0.002
marital3	0.001	0	0	0.004	0.004	0.004
marital4	0.001	0.001	0.001	-0.001	-0.001	-0.001
fsize	-0.001	-0.001	-0.001	0.033	0.032	0.032
fchild	0.002	0.001	0.002	-0.015	-0.014	-0.014
felder	0.002	0	0.001	0.005	0.008	0.007
Constant				-0.111	-0.111	-0.111
Total	0.002	-0.001	0	-0.132	-0.129	-0.13

Table 6: AI decomposition

Predicted AI 2015	787.48					
Predicted AI 2005	628.31					
Difference	159.17					
Decomposition form	1	2	3	1	2	3
	Composition	Com	Com	Structural	Str	Str
income	1.11	0.87	0.99	14.87	15.11	14.99
male	-0.26	-0.3	-0.28	-16.83	-16.78	-16.8
age2535	2	3.32	2.66	26.13	24.81	25.47
age3545	10.93	12.11	11.52	7.22	6.04	6.63
age4555	5.62	6.23	5.93	8.56	7.95	8.26
age5565	-14.45	-14.93	-14.69	2	2.48	2.24
age65+	-20.34	-19.59	-19.97	-3.12	-3.87	-3.5
edu2	-6.04	-4.21	-5.13	7.27	5.43	6.35
edu3	3.05	4.41	3.73	-33.41	-34.76	-34.08
edu4	7.26	9.48	8.37	-18.28	-20.5	-19.39
edu5	15.7	17.52	16.61	-5.64	-7.46	-6.55
race2	0.23	-0.87	-0.32	16.97	18.07	17.52
race3	-1.42	-4.16	-2.79	7.15	9.89	8.52
marital2	-0.34	-0.57	-0.46	5.28	5.51	5.39
marital3	1.21	1.41	1.31	-1.9	-2.1	-2
marital4	-0.57	-0.05	-0.31	-2.63	-3.15	-2.89
fsize	2.3	2.38	2.34	9.18	9.1	9.14
fchild	-5.3	-4.23	-4.76	11.36	10.29	10.83
felder	4.62	-0.55	2.04	14.03	19.2	16.61
Constant				105.64	105.64	105.64
Total	5.31	8.28	6.79	153.86	150.89	152.38

The gap between the two years is decomposed into the contributions from the two effects for each covariate. The decomposition forms 1, 2, and 3 are corresponding to forms in equations (28), (29), and (30), respectively. After having a rough look at tables 5 and 6, we can see that a majority of the gap in the indexes is due to the difference of the constant terms between the two years. This very large constant term, an unexplained part of the decomposition, could be explained if there was a parallel time trend of increasing

physical activity in the US population. This wide increase in physical activity, due to a time trend instead of changes of the covariates, could cause the constant term in CI gap decomposition to be significantly negative and the constant term in AI gap decomposition to be significantly positive. The structural and composition effects, according to this decomposition, contribute moderately to this gap. Moreover, the different decomposition forms generally produce similar results except for the composition effects of the variables marital4 and felder in AI decomposition in table 6. This exception does not pose worries to our analysis as both variables have been shown to be insignificant in the previous RIF regressions. However, our proposed form, form 3, of decomposition, by its mathematical nature, combines two values from the traditional composition forms 1 and 2 and mitigates the difference. This mitigation effect shows the potential value of this decomposition form.

From Table 5 we can see income, age, education, and family size still contribute to the change in CI with the largest effect found in the structural effect of family size. Compared to the structural effect, the contribution of the composition effect of family size to the overal effect of the variable is small and negligible. We can see a significant positive overall and structural effect of family size on the change in CI from 2005 to 2015. This suggests there is an enlarged contribution to inequality from larger families, and this enlarged inequality mainly comes from these larger families experiencing more severe inequality. Combining the minimal composition effect and the large structural effect, we can infer that the effect from deeper inequality in larger families dominates the effect

from having smaller families, thus the variable fsize overall contributed negatively to the reduction in inequality during this period.

Age generally contributed to the reduction in CI with the exception of the 55 to 65 age group and above 65 age group. From the RIF regressions we saw older people experince deeper inequality than younger people. The positive composition effects for variables age5565 and age65+ suggest that there are more people of age above 55 in the 2015 population than in the 2005 population, which is true refering to the information in table 1. The negative structural effects, contributing to inequality reduction, in age2535, age3545, age4555, and age65+ suggest a reduction in inequality in these population groups in 2015. However, the inequality among these older groups is still larger than the base group of people from 18 to 25 years old, thus the positive coefficients in the CI RIF regression in 2015 in table 3.

Again from the summary statistics in table 1, we can see a decreased proportion of people in edu1 and edu2 categories and an increased proportion of people in edu3, edu4, and edu5. The proportion of people with complete high school education, edu2, decreased from 2005 to 2015 while the proportions of people with higher education levels, edu3-edu5, increased. In table 5, education dummies have negative composition effects and positive structural effects except edu2. A decreased population proportion in edu2 gives positive composition effect and negative structural effect in table 5, while an increased population proportion in edu3, edu4, and edu5 gives negative composition effects and positive structural effects. This association can be explained by a decreased equality

among highly educated people in 2015 compared to 2005, but the equality among highly educated people is still better than people with less than high school education, thus the negative composition effects and negative CI RIF coefficients in 2015 in table 3. For people in edu2 category, the opposite should be true: an increased equality in 2015 makes the negative structural effect dominate the positive composition effect and produces an overall contribution to the reduction in CI, thus the negative coefficients in table 3.

Table 6 reaffirms the moderate effects of income, age, and education with the respective signs opposite to the signs in table 5. However, male is shown to have negative structural effects and family size is shown to have positive structural effects. Again, AI takes into account both the mean and the relative inequality in the population. The negative structural effect for male can be explained if males experienced a small decrease in the mean physical activity from 2005 to 2015. Furthermore, the positive structural effect of family size can be explained by an increase in the mean physical activity depite an increase in relative inequality, as shown in previous tables.

6. Conclusion

This paper applies the RIF regression method and the Oaxaca-Blinder decomposition to measure and decompose the recent change of socioeconomic physical activity inequality in the US. We measure a positive CI, suggesting an existing socioeconomic inequality in physcial activity. Moreover, we find income, age, education, and family size have

significant effects on inequality. We also extend the empirical RIF regression to the achivement index, and in its RIF regression we find an addition of gender and the number of children in family having significant effects.

We measure a moderate reduction in inequality from 2005 to 2015 in the US, and in the Oaxaca-Blinder decomposition of the gap of the indexes between 2015 and 2005, we utilize 3 forms of the decomposition and find the results similar between the forms. A relatively large structural effect is found for family size, suggesting an enlarged inequality for people in large families in 2015 compared to 2005. Furthermore we find a weak evidence suggesting older people experience lightened inequality in 2015 than 2005 despite still contributing significantly to the existing inequality.

This empirical study has many limitations. First of all, all of the results from both the RIF regressions and the Oaxaca-Blinder decomposition are best seen as a statistical experiment aiming to find possible directions to investigate socioeconomic inequality in physical acitvity. The results do not imply a causal relationship between the covariates and the indexes. Additionally the models that this paper use to carry out the empirical estimations require strong assumptions. The RIF-OLS regression and Oaxaca-Blinder decomposition both require a linear functional form with an additive error term of zero conditional mean. Most importantly, the estimated effects of the covariates on the indexes from RIF regressions are by their nature local approximations and should not be used to calculate their percentage-wise contributions to the existing inequality (Heckley et al., 2016).

This paper gives interesting findings regarding the possible relationships between some socioeconomic characteristics and inequality in physical activity. The measured associations provide incentives for more investigations as they suggest a possibility for a causal relationship. More research of the important socioeconomic characteristics regarding physical activity inequality found in this paper can possibly uncover the underlying mechanism and causal relationship. This line of research can provide additional information on societal benefits or costs of some socioeconomic characteristics. Take education for example. If indeed people with higher education experience less socioeconomic inequality in physical activity and other health behaviours, there is more value and incentive for policy interventions aiming to promote education, especially to the less well-off. Policy interventions may also target some population groups with a deeper measured inequality to see the resulting effects on the inequality.

References

Beenackers, M. A., Kamphuis, C. B., Giskes, K., Brug, J., Kunst, A. E., Burdorf, A., & van Lenthe, F. J. (2012). Socioeconomic inequalities in occupational, leisure-time, and transport related physical activity among European adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 116.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human Resources*, 436-455.

Cerin, E., & Leslie, E. (2008). How socio-economic status contributes to participation in leisure-time physical activity. *Social Science & Medicine*, *66*(12), 2596-2609.

Elder, T. E., Goddeeris, J. H., & Haider, S. J. (2010). Unexplained gaps and Oaxaca–Blinder decompositions. *Labour Economics*, *17*(1), 284-290.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, *77*(3), 953-973.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition methods in economics. *Handbook of Labor Economics*, *4*, 1-102.

Haapanen, N., Miilunpalo, S., Vuori, I., Oja, P., & Pasanen, M. (1997). Association of leisure time physical activity with the risk of coronary heart disease, hypertension and diabetes in middle-aged men and women. *International Journal of Epidemiology*, *26*(4), 739-747.

Humphreys, B. R., McLeod, L., & Ruseski, J. E. (2014). Physical activity and health outcomes: evidence from Canada. *Health Economics*, *23*(1), 33-54.

Jahns, L., Adair, L., Mroz, T., & Popkin, B. M. (2012). The declining prevalence of overweight among Russian children: Income, diet, and physical activity behavior changes. *Economics & Human Biology*, *10*(2), 139-146.

Kakwani, N. C. (1977). Measurement of tax progressivity: an international comparison. *The Economic Journal*, *87*(345), 71-80.

Kakwani, N., Wagstaff, A., & Van Doorslaer, E. (1997). Socioeconomic inequalities in health: measurement, computation, and statistical inference. *Journal of Econometrics*, *77*(1), 87-103.

Koenig, W., Sund, M., Do, A., & Ernst, E. (1997). Leisure-time physical activity but not work-related physical activity is associated with decreased plasma viscosity. *Circulation*, *95*(2), 335-341.

Koolman, X., & Van Doorslaer, E. (2004). On the interpretation of a concentration index of inequality. *Health Economics*, *13*(7), 649-656.

Lechner, M. (2009). Long-run labour market and health effects of individual sports activities. *Journal of Health Economics*, *28*(4), 839-854.

Lee, I. M., & Skerrett, P. J. (2001). Physical activity and all-cause mortality: what is the dose-response relation?. *Medicine & Science in Sports & Exercise*, *33*(6), S459-S471.

MacAuley, D., McCrum, E.E., Stott, G., Evans, A.E., McRoberts, B., Boreham, C.A., Sweeney, K. & Trinick, T.R. (1996). Physical activity, physical fitness, blood pressure, and fibrinogen in the Northern Ireland health and activity survey. *Journal of Epidemiology & Community Health*, *50*(3), 258-263.

Monti, A. C. (1991). The study of the Gini concentration ratio by means of the influence function. *Statistica*, *51*(4), 561-580.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 693-709.

O'Donnell, O., Van Doorslaer, E., Rannan-Eliya, R.P., Somanathan, A., Adhikari, S.R., Harbianto, D., Garg, C.C., Hanvoravongchai, P., Huq, M.N., Karan, A. & Leung, G.M. (2007). The incidence of public spending on healthcare: comparative evidence from Asia. *The World Bank Economic Review*, *21*(1), 93-123.

O'Donnell, O., Van Doorslaer, E., Wagstaff, A., & Lindelow, M. (2008). *Analyzing health equity using household survey data.* Washington, DC: World Bank.

Paffenbarger Jr, R. S., Hyde, R., Wing, A. L., & Hsieh, C. C. (1986). Physical activity, all-cause mortality, and longevity of college alumni. *New England Journal of Medicine*, *314*(10), 605-613.

Paffenbarger Jr, R. S., Hyde, R. T., Wing, A. L., Lee, I. M., Jung, D. L., & Kampert, J. B. (1993). The association of changes in physical-activity level and other lifestyle characteristics with mortality among men. *New England Journal of Medicine*, *328*(8), 538-545.

Rosenthal, R.J., Morton, J., Brethauer, S., Mattar, S., De Maria, E., Benz, J.K., Titus, J. & Sterrett, D. (2017). Obesity in America. *Surgery for Obesity and Related Diseases*.

Rovio, S., Kåreholt, I., Helkala, E.L., Viitanen, M., Winblad, B., Tuomilehto, J., Soininen, H., Nissinen, A. & Kivipelto, M. (2005). Leisure-time physical activity at midlife and the risk of dementia and Alzheimer's disease. *The Lancet Neurology*, *4*(11), 705-711.

Sacco, R. L., Gan, R., Boden-Albala, B., Lin, I. F., Kargman, D. E., Hauser, W. A., ... & Paik, M. C. (1998). Leisure-time physical activity and ischemic stroke risk. *Stroke*, *29*(2), 380-387.

Sarma, S., Devlin, R. A., Gilliland, J., Campbell, M. K., & Zaric, G. S. (2015). The Effect of Leisure-Time Physical Activity on Obesity, Diabetes, High BP and Heart Disease Among Canadians: Evidence from 2000/2001 to 2005/2006. *Health Economics*, *24*(12), 1531-1547.

Sarma, S., Zaric, G. S., Campbell, M. K., & Gilliland, J. (2014). The effect of physical activity on adult obesity: Evidence from the Canadian NPHS panel. *Economics & Human Biology*, *14*, 1-21.

Van Doorslaer, E., Wagstaff, A., Bleichrodt, H., Calonge, S., Gerdtham, U.G., Gerfin, M., Geurts, J., Gross, L., Häkkinen, U., Leu, R.E. & O'Donell, O. (1997). Income-related inequalities in health: some international comparisons. *Journal of Health Economics*, *16*(1), 93-112.

Wagstaff, A. (2000). Socioeconomic inequalities in child mortality: comparisons across nine developing countries. *Bulletin of the World Health Organization*, *78*(1), 19-29.

Wagstaff, A. (2002). Inequality aversion, health inequalities and health achievement. *Journal of Health Economics*, *21*(4), 627-641.

Wagstaff, A., Van Doorslaer, E., & Paci, P. (1989). Equity in the finance and delivery of health care: some tentative cross-country comparisons. *Oxford Review of Economic Policy*, *5*(1), 89-112.

Wagstaff, A., Van Doorslaer, E., & Watanabe, N. (2003). On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam. *Journal of Econometrics*, *112*(1), 207-223.

Warburton, D. E., Nicol, C. W., & Bredin, S. S. (2006). Health benefits of physical activity: the evidence. *Canadian Medical Association Journal*, *174*(6), 801-809.

Wilcox, R.R., 2005. *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed. Elsevier, Amsterdam.