AN EFFICIENT PROPENSITY SCORE METHOD FOR CAUSAL ANALYSIS WITH APPLICATION TO CASE-CONTROL STUDY IN BREAST CANCER RESEARCH

By

Azam Najafkouchak

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biostatistics – Doctor of Philosophy

2023

ABSTRACT

Propensity Score (PS) become a popular method to adjust for measured confounding factor in the absence of randomization. In real applications, a practice is to discretize these scores and use the stratification approach to estimate the causal parameter of interest. In this dissertation we introduce a novel and flexible stratification approach (continuous threshold) that uses all available information in the propensity score to improve the power for assessing the average treatment effect (ATE). This new approach requires continuous dichotomizations of the PS. Empirical processes resulting from these dichotomizations are then used to construct an integrated estimator of the causal effect, with limiting null distributions shown to be functionals of tight random processes. We illustrate our newly proposed method using simulation studies and an application to a real dataset in breast cancer (BC) research, Polish Women's Health Study (PWHS).

Based on evidence of Monte Carlo simulation study, we showed that the newly continuous threshold increases the power of test compared to PS stratification method (quintiles and median). It is also provided, closer estimation of the causal effect to the true value. Because the true value of ATE is usually unknown to the researchers, continuous threshold can be applied to improve estimation of ATE as sample size increases.

In our extensive analysis using traditional analysis of case control studies, we observed a significant reduction (approximately 50%) in breast cancer risk for high levels of total daily physical activity (PA) relative to low levels both in adolescence and adulthood. Similar reduction in risk for PA was observed for the causal effect estimated as OR's when the three PS methods: Inverse Probability Weighting (IPW), Covariate Adjustment and Stratification were applied to analyze PWHS.

When the scanning method was applied for the case study (PWHS), we showed that it was robust to the misclassification of the PS model, while other evaluated methods provided estimates of causal effect that varied under covariate misclassification.

Using Case-Control Weighted Target Maximum Likelihood Estimation (CCW-TMLE) introduced by *Rose and van der Laan, et al 2014*, we estimated ATE for total daily PA during adolescence and adulthood for our case- control study (PWHS). Our estimate of ATE was negative and significant, indicating a reduction in risk of BC for high level vs low level of PA.

In conclusion, our results contribute to the methodology of estimating causal effect by newly introduced continuous thresholding method as well as to the literature on the effect of high total daily PA in adolescence and adulthood on reduction of BC risk.

This analysis suggests that there should be more emphasis on increasing the level of PA in girls under the age of 18. In addition, to encouraging high level of adolescent PA, maintenance of higher levels of PA in adulthood should be of equal importance to gain the largest benefit from PA throughout lifetime on BC risk reduction.

To My Mom and My Children: Saman and Sepehr
Thank you for supporting me to accomplish my goals!

ACKNOWLEDGEMENTS

I would like to thank my advisor Professor David Todem for his support and guidance. I would also like to express my appreciation to my committee members, Professors Dorothy Pathak, Pramod Pathak and Joseph Gardiner. Without their help, this work could not have been completed. I appreciate all your continuous guidance and support for helping me to accomplish this research. It has been great fun working with all of you and I thank you for your patience with me.

Furthermore, I would like to thank Professors Dawn Misra (Department Chair), and David Barondess (Graduate Program Director and CMH Assistant Dean for Graduate Programs), who both supported me through this process and assisted me financially.

My biggest thanks are extended to my family, and especially to my children, for the support you have all given me throughout my graduate studies.

I would like to mention Kim Steed-Page (Director) and Laraine D. Walton (Administrative Assistance) in the MSU Family Resource **C**enter (SPOM) for supporting my family and I emotionally and in many other ways.

Finally, I want to thank the support staff in the Department of Epidemiology and Biostatistics, and especially Linda Walters and Stacy Hollon, the Graduate School, and the College of Human Medicine.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1. Introduction

Observational studies have been used recently to estimate the causal effect of treatment on outcomes. However, in observational studies, evaluation of the causal effect in common regression approaches can be difficult because of the lack of randomization, especially while involving many confounders and limited sample size. As we know, the current standard approach of regression analysis allows us to estimate *only association* between exposure and outcome, but we are usually interested to estimate *the causal effect* of exposure on the outcome.

In nonrandomized studies*, treated subjects* often *differ systematically* from *untreated subjects*. Propensity score (PS) analyses have been developed to overcome this limitation of the standard approach (logistic regression) in observational studies, and under certain conditions PS can be used to estimate the causal effect.

In this research, first we will present the definition of PS and how to estimate it. Then we will explain the major PS approaches and discuss the assumptions as well as its applications and methodologies. We will also discuss why the PS stratification method is a common approach among epidemiologists**.** Finally, we propose a new methodology as "*scanning method*" which relies on continuous thresh-holding of PS to estimate the causal effect. Our novel approach is based on PS stratification which continuously stratifies and accumulates information.

We will also illustrate all the methods discussed in this dissertation, including the new scanning approach, on both real data obtained from the breast cancer study (Polish Women's Health Study) and simulated data.

1.2. Background and Motivation for the Polish Women's Health Study (PWHS)

Based on global cancer statistics, provided by GLOBOCAN, there were an estimated 19.3 million new cancer cases and 10 million cancer deaths in 2020 (excluded non-melanoma skin cancer).[1] Breast cancer (BC) is the most diagnosed cancer worldwide, contributing 2.26 million new cases and representing 12.5% of the total number of new cases diagnosed in 2020. In addition, BC is ranked fifth for cancer deaths worldwide with 684.996 deaths (6.9% of total) after lung (18.2%) colorectum cancer (9.5%), liver cancer (8.4%) and stomach cancer (7.8%) respectfully.[1,2] Among females, in 2020, BC was the most common cancer, accounting for 25.8% of all female cancer cases, and the first common cause of deaths in women with an estimated of 684,996 female breast cancer deaths (16% of total female cancer worldwide deaths) aged 0 to 84.[2]

US has one of the highest incidence rates of BC in the world. Based on SEER (2016-2020), overall age-adjusted incidence rate (to US population in 2010) was 126.9 (per 100,000 women) and death rate was 19.6 (per 100,000 women) per year. [3] Furthermore, based on American's Cancer Society, estimation of BC new cases in the United States for 2023, will be about 353,510 (invasive and ductal carcinoma in situ) and about 43,700 women will die from BC in United States. [4] Additionally, the average risk of BC in women born today in US is 13 % (by the age of 80). [5] This means there is a 1 in 8 chance, for females now born in US to develop BC during their lifetime.[5]

Increases in BC incidence rates have been seen across the globe, in both industrialized and non-industrialized countries. In Poland, breast cancer incidence rates in Polish native women have been observed 45.4 (ASR World per 100,000) to be nearly half of those in the United States 90.3

(ASR World per 100,000) ,with a mortality rate of 14.1 (ASR World per 100,000) based on GLOBOCAN in 2012.[6,7] Lifestyle differences, such as diet or other environmental risk factors, between women living in the United States and those living in Poland have been suggested to contribute to this almost two-fold difference in incidence between these two countries. [8,9] When Polish women migrate to US and other countries with higher BC incidence, their BC risk increases and becomes almost as high as that of women in the host country in their own lifetime. [8,9,10]

While the ever-expanding field of genetics has been able to link certain diseases to specific genes, it has been estimated that the inherited causes of breast cancer (such as genetic mutations in BRCA1, BRCA2, and p53) account approximately for ten percent of all breast cancers diagnosed each year. [11,12] This leaves up to 90% of this disease to be potentially attributed to either environmental or lifestyle risk factors. [11,13] These findings suggest that primary prevention of breast cancer might be achieved through modifiable factors in a woman's life.[14,15]

Studies performed on women that have immigrated to the Unites States provide support to this hypothesis.[15,16,17] Thus, breast cancer prevention might be achieved by modifying a lifestyle factor such as diet or for example physical activity. If such proposed approach to prevention is correct, the impact on the reduction of the burden of breast cancer worldwide would be significant.

When Polish women immigrate to US or other countries, where breast cancer mortality is higher than in Poland, their breast cancer mortality, increases and becomes almost as high as those of women in their host country in their own lifetime.[8,9,10] Therefore, studying Polish immigrant women to the US offers an opportunity to evaluate effects of lifestyle changes on breast cancer risk.[16,17]

Breast cancer has multiple risk factors, some of them more consistently observed, such as family history, reproductive history, obesity, lactation, while other factors such as physical activity, diet and other environmental factors need more investigation. [16] Moreover, when identifying risk factors for breast cancer, timing when exposure occurs has been shown to play an important role in the effect size of the risk factor.[13] When a woman is exposed to risk factors, such as irradiation, certain foods, alcohol ingestion, and smoking, at a younger age, her later risk of breast cancer in adulthood is increased relative to women who had such an exposure only during adulthood.[13,18] This suggests that modifications of certain risk factors only in adulthood may not be as effective as if such risk modification occurred much earlier in life. Therefore, it has been suggested in the literature that preventative measures may be most effective in breast cancer risk reduction if they occur between late childhood and early adulthood when the breast tissue develops.[18]

1.3. The Source of Data (PWHS)

The Polish Woman's Health Study was designed to examine if changes in lifestyle factors of Polish women who immigrated to the United States, might explain the increase in breast cancer risk observed in Polish immigrant women to the US. The main aim of the study was to examine the effect of diet in women who immigrated to the United States and compare it to their counterparts who remained in Poland. Two parallel case-control studies were conducted; one of women still living in Poland and the other of Polish-born women who have immigrated to the United States before 1996. Data collected from both countries included questions about women's lifestyles during 1985-1989 as well as their lifestyles when the women were aged 12-13 years. The time- period of 1985-1989 was chosen in order to capture the traditional Polish diet before introduction of market economy in Poland after the fall of Communism in 1989.

Additionally, information was collected on other established lifestyle behaviors, and potential risk factors such as reproductive history, physical activity, family history, and occupational histories. In these analyses we will look at the relationship between physical activity during adolescent years (specifically between ages 12-13) and during adulthood on breast cancer risk in Polish immigrant women residing in the US, using data from US component from the Polish Women's Health Study. After data cleaning, the available data set, includes 411 Polish immigrant women residing in two areas, Cook County, Ill, and Metropolitan Detroit Area, Michigan (128 incident breast cancer cases from 1994-2001, and 283 controls).

Specific Aim

To assess the relationship between physical activity during adolescence (defined as ages 12-13 years of age) and adulthood and risk of breast cancer in population of Polish migrant women. Physical activity will be determined by self-reported, usual total of daily (24-hour) physical activity (summation of inactive, recreational, household, and occupational).

Hypothesis

We hypothesize that increased total daily physical activity during adolescent and adulthood will decrease women's breast cancer risk.

1.4. Organization of This Dissertation

The structure of the dissertation will be as follows:

Chapter 2 (Background and Established Methods for the Causal Treatment Effect), we will describe the potential outcomes framework; types of treatment effects and differences between randomized control trails (RCTs) and observational studies. This will then motivate us to define propensity score (PS) methods. We will also explain assumptions underlying the PS approach, discuss various considerations for estimating PS and explain several application methods that are

commonly used after the PS is estimated.

Chapter 3 We will explore our newly proposed "scanning approach" of propensity score. We will define the new method formally in mathematical notation, explore its various properties, and the properties of effect estimates stemming from their use. We will then explain the data generation steps for the simulation studies including Bootstrap Resampling method. We will present the results from the simulation study, bootstrap resampling and discuss the findings.

Chapter 4 We will illustrate the newly scanning method for the case study Polish Women Health Study (PWHS). In this chapter, we will describe the specific objectives of the study, detail the characteristics of the data set involved, explain how the propensity score methodologies were applied that provide us with estimation of causal effect in terms of ORs, and we will compare the findings from common regression methods, PS methods with scanning method (continuous threshold) and present the results.

Chapter 5 We will introduce a new approach to obtain causal effect (ATE) for the case-control study established in recent literature, using Target Maximum Likelihood Estimation (TMLE). We will present the methodology and perform the simulation study. We will then apply the new method Case-Control Weighted Target Maximum Likelihood Estimation (CCW-TMLE) developed by *Rose* and *van der Laan et al 2008, 2014* [48,56,57] for the estimation of causal effect (ATE) for our case study (PWHS), to assess the impact of total daily physical activity during adolescence and adult on breast cancer risk.

Chapter 6 The final chapter, we will provide discussion of all the approaches used in this dissertation, both established methods as well as the newly proposed method (Scanning method), and CCW-TMLE for obtaining the causal effect of treatment effect (ATE) in a case-control studies.

CHAPTER 2. BACKGROUND AND ESTABLISHED METHODS FOR CAUSAL TREATMENT EFFECT

2.1. The Potential Outcomes Framework

The potential outcomes framework, known as Rubin's Causal Model (RCM) (Holland, 1986) provides a framework of the conceptualization of causal inference.[19] Units, treatments, and potential outcomes are the three components to RCM. In the "potential outcomes framework" for any subject in the population, there are two potential outcomes: one outcome if the subject "i" is exposed to the treatment of interest ($Y_i(1)$) and one outcome if the subject "i" is not exposed to the treatment ($Y_i(0)$). Therefore, given a sample of subjects and a treatment, each subject "i" has a pair of potential outcomes. However, each subject receives only one type of treatment (control or treatment). Let Z be an indicator function denoting the treatment received (Z=1 for the exposed and Z=0 for the unexposed). We call this the "potential outcomes framework" because for any subject "*i"* we can only observe one outcome $Y_i(1)$ or $Y_i(0)$ but not both Y_i. Therefore, the counterfactual model estimating the treatment effect for a unit "i" is given by:

$$Y_i = Z_i\, Y_i(1) + (1 - Z_i)\, Y_i(0)$$ (Austin *et al* 2011)[20]

2.2. Average Treatment Effect (ATE)

Let's define for each subject the treatment effect to be $Y_i(1)$- $Y_i(0)$. Estimation of the average treatment effect (ATE) is based on the counterfactual framework, or the expected value of the differences in outcomes under two conditions (for entire population or those who received treatment). The two measurements of treatment effect are defined as:

i. *Average Treatment Effect* (ATE) is defined as the average of effect at the entire population (IMBEN *et al* 2003)[21]

$$ATE = E(Y_i(1) - Y_i(0))$$

ii. *Average treatment effect in treatment* (ATT) is defined as the average of effect on those subjects who has received the treatment (IMBEN, *et al* 2003). [21]

$$ATT = E(Y(1) - Y(0)|Z)$$

2.3. Randomized Controlled Trials (RCTs)

In most biomedical research randomized controlled trials (RCTs) are considered as the gold standard for estimating the treatment effect or interventions outcomes. In a RCT the treatment is assigned randomly to obtain the treated group and untreated group; due to randomization the treated population will not differ systematically from the overall population consequently ATE and ATT coincide (Austin et al, 2011)[20]. Given that, an unbiased estimate ATE can be obtained directly from the study data and defined as follow:

$$\mathrm{ATE} = \mathrm{E}(\mathrm{Y_i}(1)\text{-}\mathrm{Y_i}(0)) = \mathrm{E}\big(\mathrm{Y}(1)\big)\text{-}\mathrm{E}\big(\mathrm{Y}(0)\big)$$

(Lunceford & Davidian, 2004).[22] This allows the ATE to be defined, for the continuous outcome, in terms of a difference in means and for the dichotomous outcomes, a difference in proportions. For the dichotomous outcomes, alternative measures of effect include the relative risk and the odds ratio. However, in logistics regression model, marginal (average odds ratios) differ from conditional (adjusted odds ratios).[44] Marginal odds ratios are odds ratios between two variables in the marginal table while ignoring other variables, but conditional odds ratios are odds ratios between two variables for fixed levels of the other variables.[43]

Therefore "marginal" effects (the effect that we estimate) is the effect on the population but "conditional" treatment effect is the average effect on the individuals.[43]

2.4. Observational Study

In observational studies, the treated subjects often differ systematically from the untreated subjects. Therefore, in general for treated group E(Y(1) | Z=1) is not equal to overall E(Y(1)) and for untreated group E(Y(0) | Z=0)) is not equal to overall E(Y(0)). Consequently, an unbiased estimate of the ATE cannot be computed by comparing of outcomes directly between the treated subjects and untreated subjects.

2.5. Propensity Score Methodologies for Causal Inference in Observational Studies

Researchers are usually interested to estimate the causal effect of treatment on outcomes. However, in observational study, due the lack of randomization, there is a systematic bias between treated and untreated subjects. Rosenbaum and Rubin (1983) [23] developed the methodology that the probability of a subject's treated group is determined as a function of the measured covariates for that subject at the baseline. Conditioning subsequent analysis on this probability enables unbiased estimation of the average treatment effect (ATE). However, bias due to unmeasured covariates may still exist. Rosenbaum and Rubin (1983, 1985) [23,24] also show that, under the assumption of no unmeasured confounding, adjusting solely for differences in the estimated propensity score (defined below) between treated and control units *removes all systematic biases.* (IMBEN, 2003)[21]. In the next section we will define propensity score and its methodologist including assumptions, estimations, the common methods used in this study (inverse probability weighting, stratification, covariate adjusted), and finally we will explore scanning method and its methodology in the next Chapter.

2.6. Definition of Propensity Score

The propensity score (PS) is defined as the conditional probability of being exposed on observed

baseline covariates. $\mathrm{e}(x) = P_r(\mathrm{Z} = 1|X)$

2.7. Assumptions

There are several assumptions for application of PS method as follows:

2.7.1 Ignorable Treatment Assignment Assumption

The first assumption is "strongly ignorable treatment assignment" (Holmes, 2014; Rosenbaum & Rubin, 1983) [23,25] This assumption is met if two conditions hold as below:

i) Treatment assignment is independent of potential outcomes conditional on the observed baseline covariates.

$$(Y(0), Y(1) \perp Z|\, X)$$

ii) Every subject has a nonzero probability to receive either treatment or non-treatment; this aspect is referred to "Positivity "(Cole and Hernan, 2008; and Funk, 2011). [26,27]

$$(0 < (P_r(\mathrm{Z} = 1|\, \mathrm{X}) < 1)$$

2.7.2 Sufficient Common Support or Overlap

Another assumption implies that there is sufficient overlap in the distributions of the propensity scores estimated for the treatment and control groups; that is, the two groups being compared share a common support region of propensity scores in the sample data. This assumes that given similarity on background characteristics participants with the same propensity scores have an equal chance of being in either the treatment or control group. This would allow us to compare the treated and untreated subjects to estimate unbiased treatment effect.

2.7.3 Assessing Balance of Covariates

Assessing balance involves assessing whether the distributions of covariates are similar between the treated and control groups. Two common recommendations for assessing balance include the following:

i) Standardized Mean Differences

The standardized mean difference (SMD) is the one of the commons methods assessing of a single covariate balancing and it is defined as the difference in the means of each covariate between treatment groups and it is standardized by a standardization factor so that it is on the same scale for all covariates. For continuous covariates, SMD is calculated by:

$$SMD = \frac{\bar{X}_{ex} - \bar{X}_{un}}{\sqrt{(S_{ex}^2 + S_{un}^2)/2}}$$

Where $\bar{X}_{ex}\ and\ \bar{X}_{un}$ are sample mean for treated and untreated; $S_{ex}^2\ and\ S_{un}^2$ are sample variance for treated and treated group respectively.

$$\bar{X}_{ex} = \frac{1}{n_{ex}} \sum_{i=1}^{n_{ex}} X_i \ with\ S_{ex}^2 = \frac{1}{n_{ex} - 1} \sum_{i=1}^{n_{ex}} \left(X_{ex,i} - \bar{X}_{ex}\right)^2$$

$$\bar{X}_{un} = \frac{1}{n_{un}} \sum_{i=1}^{n_{un}} X_i \ with\ S_{un}^2 = \frac{1}{n_{un} - 1} \sum_{i=1}^{n_{un}} \left(X_{un,i} - \bar{X}_{un}\right)^2$$

For binary covariates SMD is obtained as:

$$SMD = \frac{\hat{P}_{ex} - \hat{P}_{un}}{\sqrt{\frac{\hat{P}_{ex}(1-\hat{P}_{ex}) + \hat{P}_{un}(1-\hat{P}_{un})}{2}}}$$

Where $\hat{P}_{ex}$ and $\hat{P}_{un}$ are denoted as proportion of successes (treated) and fails (untreated) respectively. All SMD between -0.25 and 0.25, indicates good balance for a given covariate (*Cochran and Rubin ,1973* and *Rosenbaum and Rubin, 1985*).[28,24]

ii) Variance Ratios

The variance ratio is the ratio of the sample variance of a covariate in one group to that in the other group. For instant: S_{ex}^2/S_{un}^2 where $S_{ex}^2\ and\ S_{un}^2\ denoted\ sample$ variance of treatment and sample variance of control defined above in "i". For dichotomous outcome variance ratio is

obtained by: $\frac{(\hat{P}_{ex}(1-\hat{P}_{ex}))}{(\hat{P}_{un}(1-\hat{P}_{un}))}$ ($\hat{P}_{ex}$ and $\hat{P}_{un}$ are denoted as previous part).Variance ratios close to 1 indicate the good balance because they imply the variances of the samples are similar (Austin 2009).[29]

2.8. Estimation of Propensity Score and Variables Selection

For the binary treatment, propensity score for unit i ($e(X_i)$ can be estimated from logistics regression as : $\ln(\frac{e(X_i)}{1-e(X_i)}) = \beta\, X_i$ where β is a vector of the regression coefficients and X_i is the baseline vector of covariates. [20,29,45]

For the variable selection in the above model, Rubin and Thomas (using PS matching), have suggested including all covariates that are correlated to the outcome in the PS model, regardless of being associated with exposure or not (Rubin DB, Thomas,1996). [30] However, in many applications of PS analysis, researchers often review the *ROC* statistic (the area under the receiver operating characteristic curve which is equivalent to the c-statistic in logistic regression) for determining which variable to include or exclude in the PS model. According to this approach, any variable that increases the ROC statistic should be retained in the PS model.[31,32]

2.9. Major Propensity Score Approaches

Several methods of propensity score (PS) analyses have been used such as stratification, inverse probability weighting (IPW), covariate adjustment and matching. However, we will describe three methods which have been used in this research.

2.9.1 Inverse Probability Weighting (IPW)

Let Z_i be an indicator variable denoting whether the i^{th} subject is treated or not; $where\ Z_i = 1$ and $Z_i = 0$ represent treated and untreated group respectfully. Also, let $e(X_i)$ denote the propensity score for the i^{th} subject. Then weights can be defined as W_i for i^{th} subject as follow:

$$W_i = \frac{Z_i}{e(X_i)} + \frac{1-Z_i}{1-e(X_i)}$$

Let Y_i be a binary outcome for subject i, where $Y_i = 1$ and $Y_i = 0$ denote case and control respectfully. IPTW can be then defined as an estimator of ATE (risk differences) as follow (Austin *et al*, 2011) [20]:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{z_i Y_i}{\hat{e}(x_i)}$$

$$\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-z_i)Y_i}{1-\hat{e}(x_i)}$$

$$\widehat{ATE} = \hat{\mu}_1 - \hat{\mu}_0$$

Where n is the number of subjects. From these two marginal estimates ($\hat{\mu}_1$ *and* $\hat{\mu}_0$), we could also estimate a relative risk ($\hat{\mu}_1/\hat{\mu}_0$) as well as an odds ratio ($\hat{\mu}_1/(1-\hat{\mu}_1)/(\hat{\mu}_0/(1-\hat{\mu}_0))$ for a binary outcome. [22]

2.9.2 Covariate Adjustment

Covariate adjustment is another common propensity score method where outcome is regressed on the propensity score and the treatment. For the continuous outcome, a linear model would be chosen. However, when the outcome is binary, the logistic regression would be the natural choice in which we can estimate ORs. [45] In addition, for this method we assume that there is no misclassification for the PS model (Austin *et al*, 2011). [20]

2.9.3 Stratification

One of the most common approaches among epidemiologist is to use PS stratification method because it is easily implementable and has some connection to meta-analysis**.** [20] In this method sample is divided into approximately equal number of subjects per group (strata) based on the

chosen cut-point values of the PS in the population (to estimated ATE) or in the treated group (to estimate ATT) (Rosenbaum & Rubin, 1984). [20]

Stratification on the propensity score (PS) can be considered as a meta-analysis, a set of Quasi - randomized control trails. In this method, [20] we divide subjects into strata with the similar PS. The effect of treatment on outcomes can be then estimated by comparing outcomes between treated (exposure) and untreated (control) subjects. [20,45] The overall treatment effect can be then estimated by combining stratum-specific treatment effect in each stratum, for example, using weighted average. For binary outcome we estimate the ATE as follow:

$$\widehat{ATE}_{\text{str}}=\sum_{i=1}^{k} w_i(\hat{p}_{\text{ex,i}}-\hat{p}_{\text{un,i}})/\sum_{i=1}^{k} w_i \quad (w_i=\frac{1}{SE_i^2})$$

Where weight (w_i) is $\frac{1}{SE_i^2}$ denote $\hat{p}_{\text{ex,i}}$ and $\hat{p}_{\text{un,i}}$ represent the proportion of treated and untreated observations in the i^{th} stratum in cases and controls, respectively.

The standard error of the pooled estimate is: $\sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}$ (Austin, 2011)). [20]

Cochran (1968) indicated that stratifying on the quintiles of a continuous confounding variable eliminated approximately 90% of the bias due to that variable. Rosenbaum and Rubin *et al,* 1984. [33] extended this result to stratification on the propensity score, stating that stratifying on the quintiles of the propensity score eliminates approximately 90% of the bias due to measured confounders when estimating a linear treatment effect.

CHAPTER 3: THE NEW APPROACH -CONTINUOUS THRESHOLD

Even though the standard stratification approach is easy to implement, the risk of loss of efficiency is real especially when the number of cut-points is small. We introduce a novel approach that continuously stratifies and accumulates information. This approach does not rely on arbitrary discretization but uses continuum of all available information in the propensity score to improve the power to evaluate the average treatment effect. We scan across continuous dichotomizations of the propensity score to construct an integrated estimator of the causal effect.

3.1. Methods and Definitions

3.1.1 Definition

Let denote $X_{i1}, X_{i2}, \ldots, X_{ip}$ as the p explanatory random variables (covariates) for each subject i, that are associated with an observed dichotomous outcome Y_i denotes , 1 for BC cases, 0 for controls and the exposure variable by Trt_i a binary variable, denoted 1 for exposed (treated), 0 for unexposed(no-treated) for each subject i. We assume that the relationship between the p explanatory covariates, exposure status, and the logit of the probability of a dichotomous outcome defined by the following regression equation:

$$\text{Logit}(\Pr(Y_i = 1)) = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \ldots + \alpha_p X_{ip} + \beta\, Trt_i$$

As we explained in chapter 2, for estimating of the PS $\left(e(X_i)\right)$ we just need to regress the treatment received on baseline covariates using logistic regression for binary exposure as below:

$$\text{Logit}(e(X_i)) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}$$

3.1.2 The Scanning Model (Continuous Thresholding)

In this section we explain the scanning method using arbitrary number of cut points. In this

method the PS is discretized into binary variable, yielding two strata as follows:

$$I(PS_i \leq s)) = \begin{cases} 1\ if\ \ 0 < e(x_i) \leq s \\ 0\ if\ \ s < e(x_i) \leq 1 \end{cases} \quad \text{For } 0<s<1$$

$$Where\ \ s : Cutpoints\ \ and\ \ e(x_i)\ :\ Propensity\ score$$

For each value of s, we look at the subset of subjects for whom $I(PS_i \leq s))$ =1. For this, we now define the model as below:

$$Logit(P(\ Y_i = 1|Trt, I(PS_i \leq s\ ))= \alpha_0(s) + \alpha(s)\ Trt + \beta(s)\ I(PS_i \leq s\ )$$

Continuous discretization generates a cut-point specific causal effect which is denoted by $\alpha(s)$. The coefficient $\beta(s)$ is the determinant of the scanning coefficient. For each cut point s, the effect of treatment on dichotomous outcomes is estimated within each stratum. Stratum-specific treatment effects are then pooled to obtain an overall treatment effect. We used 10 thresholds for this study.

3.2. Estimation of Causal Effect

Integrated Estimator (Inverse Variance Weighted)

To obtain an optimum estimate of the causal effect α (s) we use the inverse variance-weighted average method (IVW) which summarizes effect sizes from multiple independent cut points by calculating the weighted mean of the effect sizes using the inverse variance (inverse of covariance matrix for dimension >1) of the individual estimates as weight. The solution is given by the following equation for dimension>1:

$$\hat{\alpha} = ArgMin_{\alpha} \int_0^1 (\hat{\alpha}(s)\ - \alpha)' \left(Cov\big(\hat{\alpha}(s)\big)\right)^{-1} (\hat{\alpha}(s) - \alpha)\ ds$$

$$WE\ assume\ \hat{\alpha}(s)\ are\ independent\ for\ each\ threshold\ therefore\ we\ will\ have$$

$$Cov(\hat{\alpha}(s_i), \hat{\alpha}(s_j)) = \begin{cases} Var(\hat{\alpha}(s_i) & for\ i = j \\ 0 & for\ i \neq j \end{cases}$$

This equation will be as below for the one dimension:

$$\hat{\alpha} = ArgMin_{\alpha} \int_0^1 (\hat{\alpha}(s) - \alpha)^2 \left(Var\left(\hat{\alpha}(s)\right)\right)^{-1} ds$$

This leads to the solution of:

$$\hat{\alpha} = \int_0^1 \frac{\hat{\alpha}(s)\, ds}{Var\left(\hat{\alpha}(s)\right)} \Big/ \int_0^1 \frac{ds}{Var\left(\hat{\alpha}(s)\right)}$$

Let $\omega(s) = \frac{1}{Var(\hat{\alpha}(s)))}$ then the deterministic version is $\hat{\omega}(s) = \frac{1}{\hat{V}ar\left(\hat{\alpha}(s)\right)}$

In terms of $\hat{\omega}(s)$, $\hat{\alpha}$ is estimated by:

$$\hat{\alpha} = \int_0^1 \hat{\alpha}(s)\, \hat{\omega}(s)\, ds \,/ \int_0^1 \hat{\omega}(s)\, ds$$

For $Var(\hat{\alpha})$ and its estimation, we need to explain the Donsker's Central Limit Theorem and for continuous process as well as its application in the following section.

3.3. Application of Donsker's Central Limit Theorem for Continuous Processes

Theorem 1)

$\sqrt{n}(\hat{\alpha}(\boldsymbol{s}) - \alpha(s)) \rightarrow$ Zero mean normal process with a covariance Kernel $\sigma^2(.,.)$. For s_1, s_2 :

$\sigma^2(s_1, s_2) = \lim_{n\to\infty} n(Cov(\hat{\alpha}(s_1), \hat{\alpha}(s_2))$

$$= \lim_{n\to\infty} nE\{(\hat{\alpha}(s_1) - \alpha(s_1))(\hat{\alpha}(s_2) - \alpha(s_2))'\}$$

Under basic regularity conditions, the empirical process $\hat{\alpha}(s)$ converges to a Gaussian process with mean zero and a covariance Kernel. The integrated estimator $\hat{\alpha}$ converges to a normal distribution. Specifically, Process $\sqrt{n}(\hat{\alpha}(s) - \alpha(s))$ in $\mathbb{Z}$ converges to a normal in process with mean 0 and covariance Kernel: $\sigma_{\alpha}^2(s_1, s_2)$ for two thresholds s_1 and s_2.

Theorem 2)

$$\sqrt{n}(\hat{\alpha} - \alpha) \rightarrow N(0, \sigma_{\alpha}^2)$$

Where $\hat{\alpha}$ is the weighted average of $\hat{a}(s)\ for\ 0 < s < 1\ such\ as\ \hat{a}(s) = \frac{\int_0^1 \omega(s)\ \hat{a}(s)\ ds}{\int_0^1 \omega(s) ds}$

We assume $\hat{\omega}(s) \rightarrow \omega(s)\ (uniform\ convergence\ in\ s)$ where $\hat{\omega}(s) = \frac{1}{\hat{V}ar(\hat{a}(s))}$. We then define an estimator $\hat{\alpha}$ for α to be as $\hat{\alpha} = \frac{\int_0^1 \hat{\boldsymbol{\omega}}(\boldsymbol{s})\hat{\boldsymbol{\alpha}}(\boldsymbol{s})ds}{\int_0^1 \hat{\boldsymbol{\omega}}(\boldsymbol{s})ds}$

Donsker's Theorem

Assuming the following regularity conditions

1) $\sqrt{(n}(\hat{\alpha}_-(s) - \alpha_-(s)) = \frac{1}{\sqrt{n}} \sum h_{i,-}(s) + op(1)$

2) $\sqrt{(n}(\hat{\alpha}_+(s) - \alpha_+(s)) = \frac{1}{\sqrt{n}} \sum h_{i,+}(s) + op(1)$

3) $\sqrt{n}(\hat{\alpha} - \alpha) = \frac{1}{\sqrt{n}} \sum h_i(s) + op(1)$

Theorems 1 and 2 hold.

Proof of these regularity conditions is beyond the scope of this dissertation.

3.4. Bootstrap Resampling for Estimation of Variance

Analytical form of the asymptotic variance of α is not trivial due to dependency of α to s , across various values of s. We will use the bootstrap resampling approach to estimate the variance of $\hat{\alpha}(s)$. This involves repeated resampling with replacement of the given data, a large number of times, and using the sampling variance based on these replicates, to estimate the underlying variance of each $\hat{\alpha}(s)$.

3.5. Evaluation of Causal Treatment Effect

Let the statistical null hypothesis denotes that the average treatment effect is zero and alternative hypothesis denotes that the average treatment effect is not zero.

Statistical Hypothesis Testing

We proposed:

$$\begin{cases} H_0: Treatmenteffect = 0 \\ H_1: Treatment\ effect \neq 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} H_0: \alpha = 0 \\ H_1: \alpha \neq 0 \end{cases}$$

Asymptotic Distribution

Using Central Limit theorem, we can use $\mathbb{Z}$ standard distribution for large enough sample size such that: $\mathbb{Z} = \frac{(\hat{\alpha} - \alpha)}{\hat{\sigma}_{\hat{\alpha}}}$ and under the null hypothesis we have $\mathbb{Z} = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$ **.** We reject H_0 for large values of $\mathbb{Z}$**.**

3.6. Simulation Study

In this section we will describe the data generating process based on iterative Monte Carlo simulation. We conducted several simulation studies to evaluate the finite sample performance of the scanning method as compared to some of the PS's methods. Simulations are conducted for binary outcome. Furthermore, the evaluations are based on the relative bias, MSE and empirical Standard errors.

3.6.1 Data Generation Steps

We randomly simulated the data that were associated with dichotomous outcome Y, in two steps.[43] We also generated Monte Carlo simulation with three different sample sizes to evaluate the performance of methods based on PS methods as well as the Scanning Method. For the both steps we generated independent covariates $X = (X_1, X_2, X_3 X_4)$ as follow:

$X_i's, i.i.d\ , X_1, X_2, X_3 \sim N(0,1)\ and, X_4 \sim Bernolli(0.5)$. However, in individual applications of this data-generating process, both the number of subjects per simulated dataset and the number of simulated datasets could be modified according to the design of the specific simulations. We also set all coefficients for both steps. The choices presented here are only for illustrative

purposes.

Step1 (Exposure Model):

We generated 4 independent random variables X_1 to X_4 as explained above with iteration for each of 10,000 subjects. Then we calculated the empirical PS as follow by setting $\beta_0 = -1.5$, $\beta_1 = -0.1$ and $\beta_3 = \beta_4 = 0$ which indicates that X1, X2 are correlated to exposure but X3 and X4 are not in the following model:

$$LogitProb(Trt = 1|X_i = x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \quad \ldots\ldots\ldots\ldots (1)$$

Step2 (Outcome Model):

In the outcome model we used those 4 covariates which were simulated as described above and we set $\beta = 3$ $(coefficient\ of\ treatment)$. Our goal was to select β to generate a desired risk difference of approximately 0.29. Since we need to compute the expectation of $Prob(Y = 1)$ we are required to fix the values of α_0 to α_4 as $\alpha_0 = \alpha_2 = 1$, $\alpha_3 = -1$, $\alpha_4 = -0.5$ (coefficient of covariates). Then we calculated the probability of outcome with iteration of 10,000 to estimate the empirical risk differences from below model:

$$LogitProb(Y = 1|X_i = x_i\ ,\ \mathrm{Trt_i} = \mathrm{trt_i}) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + \beta\ \mathrm{trt_i} \ldots\ldots\ldots\ldots (2)$$

For estimating the empirical risk difference, once treatment was included in the model to compute the probability of Y=1, if each subject exposed ($P_{i,1}$) and once without having treatment in the model to compute the probability of Y=1, if each subject not exposed ($P_{i,0}$) as follow: [45]

$$P_{i,1} = \frac{1}{1+e^{\alpha_0+\beta T+\alpha_1 x_1+\alpha_2 x_2+\alpha_3 x_3+\alpha_4 x_4}} \qquad P_{i,0} = \frac{1}{1+e^{\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\alpha_3 x_3+\alpha_4 x_4}}$$

We then calculated mean of each these probabilities $\overline{P_1}$ (exposed) and $\overline{P_0}$ (unexposed) as below:[45]

$\overline{P_1} = \frac{1}{10000} \sum_{1=1}^{10000} P_{i,1}$ and $\overline{P_0} = \frac{1}{10000} \sum_{1=1}^{10000} P_{i,0}$

We then computed the empirical risk differences (RD) as follow:

RD = $\overline{P_1}$ - $\overline{P_0}$ where $\overline{P_1}$ and $\overline{P_0}$ $are\ mean\ of\ P_{i,1}, P_{i,0}$ respectfully. Now with identified the empirical risk differences of approximately 0.29 we can consider it as the true risk difference. [43]

Then we generated the Monte Carlo simulation with 3 different sample sizes, n=200, n=400 and 600 with iterative of 1000 with the same covariates X_1 to X_4 explained above with the same coefficients as well as the same β (coefficient of treatment effect) in the data generation outcome step 2 with the model (2).

3.6.2 Bootstrap Resampling Method

Since in the continuous threshold method, using 10 thresholds in our study, introduced dependent between estimate obtained from each cut points we could not calculate the Standard Deviation (SD) of our estimate by the known methods, and it introduces complexity in the variance formula. Therefore, we used 100 Bootstrapping to estimate Asymptotic Standard Deviation (ASD), Empirical Standard Deviation (ESD) for continuous threshold method in order to estimate Mean Square Error (MSE), Mean Estimated Standard Error (Asymptotic and Empirical) and Empirical Coverage Rate and other statistics used in this chapter in which definitions are available in Table 3.1 We used simple random sampling with replacement (SRS) using proc surveyselect statement (in SAS) having 10 thresholds in the model as the strata. We then calculated all statistics in the tables 3.2-3.4, comparing the results of Monte Carlo simulations for different methods of PS with Scanning method using ATE for n=200, 400 and 600 with replication=1000.

Table 3.1 Definition of Criteria Reported in the Simulation Study

Let $\hat{\theta}_k$ and $\hat{s}(\hat{\theta}_k)$ Average Treatment Effect (ATE) and its Standard Deviation of estimated in Kth simulated data set and θ be the true ATE for K (K=1, . . . ,100). The Criteria used in our simulation studies are as follow:

1) Relative Differences ($\hat{\theta}, \theta$) $= \frac{1}{100}\sum_{k=1}^{100}(abs(\hat{\theta}_k - \theta)/\theta)$

2) Mean Square Error (MSE)= $\frac{1}{100}\sum_{k=1}^{100}(\hat{\theta}_k - \theta)^2$

3) The Empirical Standard Deviation (ESD)= $\sqrt{\frac{1}{99}\sum_{k=1}^{100}(\hat{\theta}_k - \theta)^2}$

4) The Asymptotic Standard Deviation (ASD)= $\sqrt{\frac{1}{100}\sum_{k=1}^{100}(\hat{\theta}_k - \theta)^2}$

5) Empirical Coverage Rate of the Normal 95 % Confidence Interval 95 % : $\frac{1}{100}\sum_{k=1}^{100}(I(95\%\ CI_{inf,k} \leq \theta \leq 95\%\ CI_{sup,k}))$where $95\%\ CI_{inf,k}$ and $95\%\ CI_{sup,k}$ are Lower and Upper bounds of C.I. Estimated by bootstrap in the kth simulated data set respectfully.

3.7. Results

In below we denoted results of the three simulation studies for n=200, 400 and 600.

Table 3.2 Results Monte Carlo Simulations Methods Examining Different Propensity Score Methods for Average Treatment Effect (n=200, Replication=1000)

True RD	Crude	Stratification Quintiles	Stratification Median	Continuous Threshold
		Mean Estimated ATE		
0.29397	0.3128349	0.2914532	0.2992	0.2980739
		(%) Relative Difference Between the True ATE and Selected Method ATE		
	6.793	0.519	1.78	1.3
		Mean-squared Error (MSE) of Estimated ATE		
	0.0040136	0.0034422	0.0043437	0.0032097
		Mean Estimated Standard Error (Empirical)		
	0.0602221	0.0586773	0.0616916	0.0564617
		Mean Estimated Standard Error (Asymptotic)		
	0.0607531	0.05172	0.0604790	0.056809
		Empirical Coverage Rate of 95% of Confidence Intervals		
	0.905	0.943	0.886	0.933

Table 3.3 Results Monte Carlo Simulations Methods Examining Different Propensity Score Methods for Average Treatment Effect (n=400, Replication=1000)

True RD	Crude	Stratification Quintiles	Stratification Median	Continuous Threshold
		Mean Estimated ATE		
0.2938974	0.3175092	0.2794353	0.2909748	0.2904112
		(%) Relative Difference Between the True ATE and Selected Method ATE		
	8.034	4.921	0.292	1.186
		Mean-squared Error (MSE) of Estimated ATE		
	0.0024695	0.0018343	0.0015464	0.0018945
		Mean Estimated Standard Error (Empirical)		
	0.0437476	0.0403337	0.0392879	0.0434073
		Mean Estimated Standard Error (Asymptotic)		
	0.0429642	0.0453117	0.039308	0.0379732
		Empirical Coverage Rate of 95% of Confidence Intervals		
	0.875	0.960	0.952	0.913

Table 3.4 Results Monte Carlo Simulations Methods Examining Different Propensity Score Methods for Average Treatment Effect (n=600, Replication=1000)

True RD	Crude	Stratification Quintiles	Stratification Median	Continuous Threshold
Mean Estimated ATE				
0.2936191	0.3141636	0.2894840	0.2893247	0.2957849
(%) Relative Difference Between the True ATE and selected Method ATE				
	6.997	1.408	1.1462	0.737
Mean-squared Error (MSE) of estimated ATE				
	0.0015904	0.0010351	0.00137	0.000941996
Mean Estimated Standard Error (Empirical)				
	0.0341983	0.0319882	0.03686	0.0305873
Mean Estimated Standard Error (Asymptotic)				
	0.0354349	0.0288179	0.03132	0.0333567
Empirical Coverage Rate of 95% of Confidence Intervals				
	0.892	0.897	0.9	0.962

3.8. Conclusions for Simulation Studies

The power of the test can be improved when we use the continuous threshold compared to the other common PS approaches. As shown by results in Monte Carlo simulation studies, scanning threshold has the least SE and least MSE compared to the other approaches (Tables 3.2-3.4). The proposed continuous threshold method in the simulation studies, performs well, especially when sample size went up from 200 to 600. All estimations for evaluated measures such as mean ATE, SE, MSE and 95% C.I, worked very well. Coverage Rate performed better than in the other methods. Given that the true ATE is usually unknown to the analyst, continuous threshold works well especially for sample size 600. The percent relative difference between true ATE and estimate from the scanning was only 0.7% which was less than those observed for the other PS methods.

CHAPTER 4: CASE STUDY-POLISH WOMEN'S HEALTH STUDY (PWHS)

In this chapter we assess the relationship between physical activity during adolescence (defined as ages 12-13 years of age) and adulthood and risk of BC in population of Polish migrant women. We also evaluate the causal effect of physical activity during adolescence and adulthood on BC risk by using PS methods. We then compare the findings from common regression methods, PS methods with the scanning method.

4.1. Literature Review of Epidemiologic Studies on Physical Activity and Breast Cancer Risk

Physical activity during lifetime through its effect on weight reduction and hormonal levels has been shown to be protective against breast cancer. [35,36] In a systematic review of the literature on physical activity and breast cancer, it was concluded that there was an inverse association between physical activity, measured during various lifetime periods, such as adolescence or adulthood, and breast cancer risk. However, it is not clear whether the magnitude of the protective effect is similar for physical activity during adolescence vs. adulthood and whether there is an additional protective effect for those who are active throughout their lifetime. [37] The reduction in risk was observed to be stronger for postmenopausal women. [38]

4.2. Design and Participants of PWHS

4.2.1 Study Design and Data Collection

This study consists of two parallel population-based case-control studies, one conducted among Polish-immigrant women in Cook County (Chicago) and Detroit Metropolitan Area, and one in Poland in 4 centers: Katowice, Gliwice, Poznan and Bialystok.

The analyses for this dissertation are based on the US component of the Polish Women's Health Study. The study was designed to evaluate the effects of diet and other lifestyle factors on breast cancer risk in Polish immigrant women to the United States and Polish natives. The study was funded by the National Institute of Health/ National Cancer Institute (NIH/NCI), in 1997 with Dr. Dorothy R. Pathak as the Principal Investigator (PI). The data collection for the study started in 2000 concurrently in Poland and the United States.

4.2.2 Study Population in US

The main challenge in identifying both cases and controls was confirmation that the women were born in Poland and currently residing in the two study areas. Cases were identified by the Illinois State Cancer Registry (ISCR) for the Cook County area and by the SEER Registry located at Karmanos Cancer Institute/Wayne University for the Detroit Metropolitan Area.

Cases had to be histologically, or cytologically confirmed incident invasive breast cancer diagnosed between January 1st 1994 and December 31st 2001 in the age group 20 -79 years. As a first step in case identification and recruitment, a letter had been sent out to the physician asking for permission to contact the patient to evaluate their eligibility for the study. All White cases with unknown place of birth were first screened for being Polish born. All Polish-born cases were then approached about participation in the study. This involved an introductory letter describing the study and letting them know that an interviewer will contact them to answer any questions and set-up a time for an interview if they agree to participate.

Controls were frequency matched to cases on age (within 5-year age groups) and area of residence. Random Digit Dialing (RDD) approach was used to identify controls under the age of 65, and controls between the ages 65-79 were supplemented from a sample obtained for the Health Care Financing Administration (HCFA) for the female population at the two sites. All

controls were screened for place of birth. If a Polish-born female between ages 20-79 was identified, the additional exclusion criteria were previous diagnosis with any other cancer except squamous or basal cell carcinoma.

4.2.3 Reason for Choosing the Time-Period 1985-1989 for Data Collection

Since the main hypothesis for the PWHS was to determine an effect of the traditional Polish diet and breast cancer risk, diet information was collected for the time- period 1985-1989, the last 5-year period prior to introduction of market economy in Poland when Western style foods became available on the market. Although diet in US did not undergo such drastic changes as diet in Poland after 1990, we chose to ask immigrant women in US to recall their diet during the same time- period (1985-89) in order to have similar recall bias between Poland and US.

Food frequency questionnaire (FFQ) was used to capture the usual dietary intake for that time-period. Given the choice of the time-period, 1985-1989 for dietary assessment, information for other factors such as physical activity, obesity and body size index were also collected for that time-period.

4.2.4 Other Risk Factors Assessed in Questionnaire

Information on other established and potential risk factors such as age, age at first full term pregnancy, parity, age at menarche, age at menopause, use of oral contraceptive pills, hormonal replacement therapy, family history of breast cancer, height and weight were collected up to the time of interview. This allowed for the adjustment for these standard reproductive and other risk factors in our analyses. Each participant provided information on the date of immigration to the US allowing for the calculation of the duration of stay in the US.

4.3. Description of Variables

4.3.1 Dependent Variable

Our outcome variable was a dichotomous outcome; the breast cancer case/control status.

4.3.2 Exposure: Physical Activity

Assessment of Physical Activity (1985-1989-Adulthood, 12-13-Adolesence):

Physical activity was assessed using a questionnaire modeled on validated physical activity questionnaires which included daily activities like sitting, reclining and household activities such as sweeping, gardening, cooking, stair climbing and sleeping. It also included recreational activities such as recreational sports, walking, bicycling, aerobic exercise as well as job activity. Intensity of activity was expressed in terms of MET's (Metabolic Equivalent Task), which were extracted from the Compendium of Physical Activities.[29] To calculate total MET-h per day, the hours spent on each type of activity that has its own unique MET-h value are multiplied by that value and added together. Information on physical activity was missing for 2 participants, as they were unable to recall their activity in 1985-1989.

4.3.3 Main Exposure: Total Daily Physical Activity During Adolesce or Adulthood

The same questions were asked for the time -period 1985-1989 and for their adolescent time-period (12-13 years old) or adulthood. Many of these women during their adolescence were in labor camps during WWII and thus had also entries for occupational activity. The process for derivation of total hours spent in each activity was the same for both time periods as defined above.

Description below of calculating physical activity has been derived by Renee Bloome (Department of Epidemiology, MSU, 2008) in her thesis entitled: "Adolescent Physical Activity & Breast Cancer Risk: A Look into the Polish Women's Health Study ".

i) **Adolescent/Adulthood Non-Occupational Physical Activity**

Average number of hours spent doing activities such as: school athletic participation, other recreational, household, occupational, and inactive/sedentary actives including hours of sleep per day were reported. Women could report the number of hours spent participating in the given activity in terms of per day, per week, per month, and in some types of activities such as sports, per year. For the purpose of these analyses, we converted all answers into hours spent per day. To accomplish this, all hours of activities reported per week were divided by 7, those reported per month were divided by 30.4 (an average number of days/month), and activities reported per year were divided by 365.25. Hours of each activity were then summed up to calculate the total number of hours each woman reported per day. To calculate total MET-h per day, the hours spent on each type of activity that has its own unique MET value were multiplied by that value and added together. [39]

Since interviewers were instructed not to question respondent's answers, the derived number of non-occupational hours/day when combined with occupational hours of physical activity to create total hours per day, was both under and over 24 hours. Therefore, the total reported hours per day needed to be standardized to a 24-hour day for each subject. To calculate the adjustment fraction, number of hours reported were divided by 24. Subsequently, the total calculated MET-h for each subject, were divided by their same unique standardization fraction. This ensured that hours reported for each activity were uniformly adjusted by the inverse of the standardization fraction for each subject.

MET- h from stair climbing were not included in our total daily physical activity variable for the following reason. Literature states that the average person takes about only one second per step to climb stairs (Bassett 1997).[40] When this value is multiplied by the number of stairs (or flights)

climbed and converted into a per day measure, MET h calculated from hours spent climbing stairs is insignificant compared to all other activities. For example, if a woman were to climb even 6 flights of stairs a day, with an average of 20 stairs per flight, and a MET value for stair climbing set to 8.5, her total MET-h from stairs contribute less than 0.3 of a MET-h/day. For this reason, we have not included MET-h/d from stair climbing in our total daily MET-h for either the adult period or the adolescent period collected from our subjects.

Similar process was used to calculate non-occupational daily activity for adulthood.

ii) **Adolescent/Adulthood Occupational Physical Activity**

Women reported the number of months or years they were employed during adolescence and adulthood. For the time when they were 12-13 years old, the maximum time was reported either as 24 months or 2 years. For adulthood, 1985-1989, it was either 60 months or 5 years. Women had the option of reporting their work activities, either in months or in years. For the adolescent years, women who reported working during the summer, when not in school, were assigned their work duration to be 3 months/year. Women were then asked to report their average number of hours they worked in a given week.

For measuring adolescent occupational physical activity, only 42 women reported being employed when they were 12-13 years old. For the women who reported holding an occupation during their adolescence, the number of hours worked per week ranged from 3 hours per week to 84. It is important to note that for some women, the time-period during which they were 12-13 years old overlapped with the years of World War II. Several of our study participants indicated they had worked in a concentration and/or labor camp during this time-period. This often resulted in seemly high reporting of hours worked per week as well as percent of those hours spent in strenuous physical activity.

The typical work and school weeks in Poland prior to 1985 consisted of six school and workdays as opposed to the typical five-day work week found in the United States. To incorporate this knowledge, we assumed that any women reporting an adolescent or adult occupation worked this six-day work week thus the number hours reported working per week was divided by 6 instead of 7 as for non-occupational activities.

If a woman reported that she worked a certain number of months during these two time periods, the number of months reported was multiplied by 26.1 to convert the number of months into the number of workdays during that time-period. The value 26.1 was calculated by multiplying the average number of days in a month, 30.4 by (6/7) to account for the 6-day work week. The result gave us the average number of workdays that each woman worked during either the 2-year adolescent time period or during the 5 years in adulthood.

If a woman reported that she worked a certain number of years in either adolescence or adulthood, then the number of years was multiplied by 312 working days per year. The conversion of 312 was calculated by multiplying 52 weeks per year times the number of working days per week, 6. The result gave us an average number of workdays during the specific time-period.

To calculate the average number of days worked/year, the above calculated number of days was divided by 2 for adolescence and 5 for adulthood. Since respondents were reporting the average number of hours worked/week, that number was divided by 6 to account for the average number of hours worked/day. To calculate the final number of average hours worked per day during the given time -period (adolescence or adulthood), the number of working days/year was multiplied by the average hours worked/day and then divided by 365.25.

In the questionnaire, after reporting the number of hours worked per week, women were asked to indicate either the number of hours, or percent of time, they spent each week sitting, standing, walking without lifting, walking with lifting less than 25 pounds, and doing heavy physical work. Each of these occupational activities carries a different MET value. Sitting was assigned a MET value of 1.5, standing was coded at 2.3 MET value, walking without lifting 3.0 MET value, walking with some lifting (less than 11.5 kg or 25lbs) 4.0 MET value, and heavy physical work 7.0 MET value (Ainsworth 2000). Percentage of time in each type of activity was then multiplied by corresponding MET-h value. To obtain the total MET-h for the occupational daily activity, the occupational daily hours were weighted by percentage of time spent in that activity times the corresponding MET-h and summed over all types of activities involved in their workday. The standardization fraction that was required to adjust each day to 24 hours and initially calculated by summing the hours of non-occupational and occupational daily activities (described in the non-occupational physical activity section above), was then applied to the total MET-h calculated for the occupational activity.

iii) **Total Daily Physical Activity During Adolescent or Adulthood**

After both occupational and non-occupational physical activity during adolescent or adulthood converted to MET-h per day were calculated, all activity for each time-period was summed up creating a variable of total daily physical activity for both the adolescent and adult periods.

Total Adolescent MET-h/Day Tertiles for Table 4.4.2 and 4.4.3

Tertiles were created using the total daily MET-h/Day as reported by controls. The adolescent tertiles created from the data were as follows: low = 0-<45.9 MET-h/day, medium = 45.9-<55.7 MET-h/day, and high as greater than or equal to 55.7 MET-h/day.

Total Adulthood MET-h/Day Tertiles for Table 4.4.2 and 4.4.3

Adult physical activity tertiles were calculated by a process similar to those for adolescent physical activity. The adult tertiles were as follows: low = 0 - < 48.8 MET-h/day, medium = 48.8 - <59.6 MET-h/Day, and high as greater than or equal to 59.6 MET-h/Day.

Joint Levels of total daily Adolesce and Adulthood PA Activities for Table 4.4.4

To examine the association of joint adolescence PA and adult PA on BC risk, we created nine categories from the tertiles derived (low, medium, high) for both the adolescent and adult physical activity. The nine categories (3×3) are with adolescent tertile level followed by the adult in the following notation: low/low, low/med low/high, med/low, med/med, med/high, high/low, high/med, and high/high.

4.3.4 Other Risk Factors and Covariates

Age at Menarche: Age at menarche was assessed by the onset of natural menstruation. Median age at menarche was 14 years. Age at menarche was divided into 3 categories as follows: less than 13 years, 13 to less than 15, and 15 years and older.

Menopausal Status and Age at Menopause: Subjects who reported having menstrual cycles were considered as pre-menopausal. Subjects who provided age at natural menopause were considered post-menopausal. Subjects who were uncertain about their menopausal status were categorized as follows: women, who had hysterectomy without removal of ovaries, were considered pre-menopausal if their age was less than 50 and postmenopausal if they were 50 years or older and their age at menopause was assigned to be 50. Women who reported that they were post-menopausal but did not provide information about their age at menopause, were assigned age 50 for their age at menopause.

Age at First Full Term Pregnancy: Full term pregnancy was defined as any pregnancy with

gestational age more than 24 weeks or 6 months, irrespective of the outcome. Age at first full term pregnancy was divided into 3 categories: less than 22, equal to 22 to less than 30 and equal to and greater than 30 years. We also included nulliparous as a separate group.

First Degree Family History: History of breast cancer in the mother, sister or daughter was considered as positive first-degree family history. This was analyzed as a binary variable (0=no family history, 1=with family history).

Adult Alcohol Consumption: Alcohol consumption was assessed by total intake of beer, wine and hard liquor during 1985 -1989. Among those who consumed alcohol, the median consumption was 0.5 drinks per week, which is equivalent approximately to 6 grams of alcohol per week (1 drink is approximately equivalent to 12 grams of alcohol).[29] This variable was consider as the continuous to our analysis. Tertiles were formed from the total average weekly alcohol consumption: low =0 – <0.21 serving/week; medium = 0.21 - <1.09 servings/week; and high $\geq$ 1.09 servings/week.

Adult Total Caloric Intake: As part of the questionnaire women filled out a Food Frequency Questionnaire (FFQ) to capture usual intakes of certain foods during 1985-89. Calories assigned to an average serving for a specific food were then multiplied by the frequency of consumption standardized to daily consumption. For a few foods including breads, eggs, and alcohol, where the number of servings was also reported, the frequency of consumption was multiplied by number of servings, subsequently multiplied by calories per serving. Total caloric intake was then calculated as the sum of the caloric intake from each food-type in the questionnaire. Tertiles were then formed based on the distribution in controls: less than 2047 calories/day; between 2047 and less than 2660 calories/day; and over 2660 calories/day.

BMI (1985-1989): Body Mass Index (BMI) was calculated from height and weight in 1985 -

1989 using the following formula - weight (kg)/ [height]2 (m). BMI was considered as 4 categories: under-weight (<18.5), normal (18.5-<25), overweight (≥25-≤30) and obese (>30). If height was available, but weight in 1985-1989 was missing, weight at age of 18, 30, 40, 50 or 60, based on subject's age range in 1985 was used for BMI calculation. Thus, if age in 1985-1989 was less than 25 years, then weight at 18 was used for calculations, similarly if age in 1895-89 was between any one of these age categories, 25-<35, 35-<45, 45-<55, 55-70 then weight at respective closest age decades 30, 40, 50 and 60 were used respectively. If weight for the closest decade as described above was missing (13 participants), mean weight specific to case/control status, for a particular age range in 1985 was used for BMI calculation for these individuals. If height was missing, the BMI remained as missing. If the individual's age in 1985 was lower than the age at which maximum height was attained, their 1985 BMI was also assigned to a missing category. Totally, information on BMI in 1985-1989 was missing for 13 participants.

Adolescent Body Size: Height and weight for the time when participants were aged 12 to 13 years were not obtained. Instead, women were asked to identify their closest body size and shape from a series of nine pictures (*Koprowski 2001*). Women selecting one of the first two figures were labeled as 'underweight,' those selecting one of the subsequent three figures were considered 'normal weight," and women reporting a figure of six through nine were grouped into an 'overweight' category (see Appendix C).

Duration Living in the US by 1985: During interview each participant provided information on date of migration from Poland to the US. Using this information, we calculated the duration of stay in the US prior to 1985. We considered three categories for this variable as follows: recently moved (moved after 1985), less than 10 years prior to 1985, and moved more than 10 years prior to 1985.

Hormone Replacement Therapy: All subjects were asked if they ever used hormone replacement therapy in form of pills or skin patches, creams, suppositories or injectables for relief of menopausal symptoms and/or prevention of bone loss. The response was recorded as yes/no and analyzed as a binary variable.

Oral Contraceptive Use (OC): Use of hormonal preparations for birth control was asked for all subjects and recorded as ever used / never used. This was analyzed as a binary variable.

Age in 1989 Controls were matched on to the distribution of age at diagnoses of cases however, case identification ranged from 1994 through 2001. Therefore in 1989, cases were older than controls. This variable was created to adjust the models for the potential age difference between cases and controls during 1985-1989. This variable (presented in Table 4.4.1) was divided into six categories: < 18 years; 18 - < 35 years; 35 - < 45 years; 45 - < 55 years; 55 - < 65 years; ≥ 65 years. Since we used the years between 1985 through 1989 as a proxy for adulthood, women in the < 18 years of age category (n = 15 controls) were excluded from all analyses since they had not started adulthood (defined as > 18 years) between 1985 and 1989.

Joint strata of age and site: We used joint strata of age at diagnosis (cases) or interview (controls) such as : (1) age<35 years and site Cook County, (2) age<35 years and 35 - < 44 years (combined to the small sample size) and site DMA, (3) age 35- <44 years and site Cook County; (4) age 45- < 54 years and site DMA, (5) age 45–54 and site Cook County, (6) age 55-64 years and site DMA, (7) age 55–64 y and site Cook County, (8) age 65-74 years and site DMA, (9) age 65–74 years and Cook County, (10) age ≥ 75 years and site AMS, (11) age ≥ 75 and site Cook County.

4.4. Statistical Analyses

For descriptive analysis we used the cross tabulations for comparing distribution of cases and

controls for Polish-born women residing in two sites (Cook County, IL and Detroit Metropolitan Area, MI) with selected risk factors of breast cancer (Table 4.1). Group differences between case and control for categorical variables, were tested in conditional logistic regression models within the joint strata , age at diagnosis (cases) or interview (controls) (<35 y; 35–44 y; 45–54 y; 55–64 y; 65–74 y; ≥ 75 y) and site (Cook County, DMA); for DMA ages <35 and 35–44 was combined due to small sample size (Table 4.1).

For the continuous variable of Met hours/day for adolescent and adulthood physical activity of total, PROC GLM was used to obtain p-value for the differences in the least square means adjusted for the joint age and site strata (Table 4.2).

To assess the effect of selected risk factors on breast cancer risk we performed 2 models. For both models, first we used odds ratios for evaluating the association of case status on physical activity. Furthermore, we run the analysis for both models using adolescent physical activity (in tertiles) as the main exposure as well as adult physical activity (in tertiles) as the main exposure. Focus of this dissertation is on both adolescent physical activities, though will also be evaluating joint adolescent physical activity and adulthood physical activity in certain models to evaluate interaction between them.

For the first model we used conditional logistics regression within the joint strata of age and site, age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y; ≥ 75 y) and site (Cook County, DMA). For DMA (ages <35y and 35-44y) combined due to small sample size to obtain odds ratios and 95% CI for the association between physical activity (adolescence and adult) and breast cancer risk.

In the multivariate analysis (model 2) we also perform conditional logistics regression, using joint strata of age and area of the residence, and additionally adjusted for potential risk factors

included: total energy intake in 1985-1989, age at menarche, age at first term pregnancy, family history of breast cancer among first degree relatives, use of contraception, use of hormonal replacement therapy, alcohol consumption, BMI in 1985-1989 and duration of stay in US. The dependent variable was binary (case/control) in the logistic regression model. We obtained OR's and 95% CI utilizing both adolescent physical activity as the main exposure and adulthood physical activity as the main exposure (Table 4.3).

4.4.1 Results for Accessing Association Between Total Daily PA During Adolescence or Adults and BC Risk

According to Table 4.1, we didn't see any significant differences between cases and controls with respect to age (diagnosis for cases, interview for controls), age in 1985, migration status in 1985, age at menarche, oral contraceptive use, hormone replacement therapy and menopausal status (at time of diagnosis for cases and interview for controls), BMI, total energy intake, and alcohol intake during 1985–1989. However, this difference was statistically significant for first degree of family history of BC and age at full term pregnancy (P-value<0.05).

Table 4.2 indicates distribution of total physical activity (METs-h /Day) during adolescence and adulthood. In this table, we examined the association between total daily PA, categorized as low, medium, and high during adolescence and adulthood with BC risk applying the traditional case-control logistic regression analyses which provides us with estimated OR's. Based on the results in this table, cases had lower levels of total daily physical activity for each percentile of distribution and significantly lower mean total daily physical activity compared to controls both for the adolescent (P-value<0.05) and for the adult(P-value<0.01).

For evaluating the association of total physical activity (MET h/Day) and case status, we used conditional logistic regression within joint strata of age and site (Model 1-unadjusted, Table 4.3),

as well as multivariate model adjusting for all potential risk factors (Model 2, Table 4.3), independently for total daily adolescent physical activity and total daily adulthood physical activity. We observed that ORs did not change substantially between Model 1 (unadjusted) and Model 2(multivariate adjusted), for both total daily PA in adolescence and adulthood. The results show that women in the highest level of PA during adolescence (greater than 55.7 MET h/Day), have a significant 45% reduction in BC risk (OR = 0.55) and women in the highest level of PA during adulthood (greater than 59.6 METs-h/Day) have a significant 47% reduction in BC risk (OR=0.53) (Table4.3).

Table 4.4 provides ORs for the joint effect of total daily PA during adolescence and adulthood, adjusted for potential risk factors. The result in this table shows that for *all women* the high total daily adolescent PA reduces BC risk for the medium and high levels of total daily PA in adulthood, reaching statistical significance for the high adolescence and high adulthood category (OR=0.29 and 95% CI :0.11-0.77).

In Table 4.4 we also presented the results for the evaluating effect of the joint PA during adolescence and adulthood by *menopause status*, adjusted for all potential risk factors as well. For the *premenopausal women*, we observed that the high levels of adolescent PA were protective for BC irrespective of the level of PA in adulthood and OR's estimates were statistically significant. These associations were also statistically significant for moderate adolescent PA with rigorous level of PA in adulthood in *premenopausal women*. ORs with 95% CI are as follow: median adolescent/ high adult (OR=0.10 with 95% CI :0.013-0.68), high adolescent/ low adult (OR=0.003 with 95% CI: 0.002-0.38), high adolescent/ median adult (OR=0.14 with 95% CI: 0.003-0.84) and high adolescent/ high adult (OR=0.14 with 95% CI: 0.11-0.77).

For the *postmenopausal women,* the high adolescent PA also provides the reduction in BC risk for all levels of adult PA, however, these reductions were only statistically significant. The observed lack of significance in the high adolescent/high adult and medium adolescent/high adult physical activity *in postmenopausal women* were most likely due to the decreased sample size. Therefore, our findings for this section, supported the hypothesis that increased total daily adolescent PA decreases BC risk in women especially for *premenopausal women*. Although the estimates of ORs for *postmenopausal women* were in the direction of protective effect of high adolescence PA, none of the OR's reached statistically significant.

Table 4.1 Selected Characteristics of women resident in Cook County IL(CC) or Detroit Metropolitan Areas MI (DMA), interviewed Between 2000-2003, by Case-Control Status

Selected Covariates	Cases (128) %	Controls (283) %	P-Value[1]
Study site			0.6[2]
Cook County – Chicago	78.91	76	
Detroit Metropolitan Area	21.09	24	
Age (y) at diagnosis (cases) / interview (controls)			0.3[3]
<35	4.6	5.6	
35-44	19.1	14.8	
45-54	30.5	27.8	
55-64	20.6	16.2	
65-74	19.1	23.9	
≥ 75	6.1	11.6	
Migrant status in 1985			0.2
In Poland	43.7	44.5	
In US <10 y	23.4	25.1	
In US ≥10 y	32.8	30.4	
First degree family history of breast cancer	15.6	8.1	**0.02**
Age at menarche (y)			0.3
<13	23.4	20.8	
13-<15	53.9	48.1	
≥ 15	22.7	31.1	
Age At Full Term Pregnancy			**0.02**
Nulliparous	11.7	9.2	
<22	21.1	28.3	
22-<30	50.8	54.1	
≥30	16.4	8.5	
Ever used oral contraception	14.9	12.0	0.4
Ever used hormonal replacement therapy	13.3	11.7	0.6
Premenopausal at diagnosis (cases) / interview (controls)	55.50	61.5	0.6
BMI in 1985-89 (kg/m^2)			0.8
<8.5	3.9	3.9	
18.5-<25	64.8	59.4	
25-<30	25.0	26.1	
≥30	6.3	8.8	
Total energy intake in 1985-89 (kcal/d)			0.2
<1935	40.6	33.2	
1935-<2365	32.0	33.9	
2365-<2880	27.3	32.9	
≥30	18.0	25.1	
Alcohol intake in 1985-1989 (drinks/week)			0.5
None	9.4	14.8	
<0.7	46.1	42.1	
≥0.7	44.4	41.1	

[1] Comparison between cases and controls adjusting for the age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y;≥ 75 y) and site (CC, DMA). For DMA (ages <35y; 35-44 y) were combined due to small sample size.

[2] Adjusted for age at diagnosis (cases) or interview (controls)

[3] Adjusted for site.

* Bold indicates statistical significance of $p < 0.05$.

Table 4.2 Percentile Distribution and Least Square Means (LSMean)and the SE of the Mean (SEM) for PA (MET-hs /day) during Adolescence and Adulthood

	12-13 year Adolescence MET-h/day			1985-1989 Adulthood Met-h/day		
Percentile LSMean SEM	**Control (283)**	**Case (128)**	**Delta Control Case**	**Controls (283)**	**Case (128)**	**Delta Control Case**
75th	59.15	55.88	3.27	63.35	55.86	7.49
50th	49.91	47.08	2.83	53.62	50.03	3.59
25th	43.38	41.96	1.42	46.51	42.02	4.49
LSMean (SEM)	**53.05** (0.77)	**50.17** (1.08)	**2.88***	**53.68** (0.84)	**48.36** (1.18)	**5.32****

[a]LSMean and SEM adjusted for age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y;≥ 75 y) and site Cook County, Detroit Metropolitan Area-DMA). For DMA (ages <35y; 35-44 y) were combined due to small sample size. * p<0.05, ** p<0.01.

Table 4.3 Risk of Breast Cancer on PA during adolescence or adulthood among Polish-born women residing in Cook County, IL or the Detroit Metropolitan Area, MI

PA	Daily PA	Number Cases, Controls	Model 1 [a] OR (95% C.I.)	Trend [b] P-value [b]	Model 2 [c] OR (95% C.I.)	Trend [b] P-value [b]
At Age 12-13	<45.9	55, 95	1 (Ref.)		1(Ref.)	
	≤ 45.9 −<55.7	41 ,91	0.79 (0.48-1.31)	P_{Trend}=**0.024**	0.75(0.43-1.28)	P_{Trend}= **0.043**
	≥55.7	32, 97	**0.54 (0.32-0.92)**		**0.55(0.31-0.98)**	
Adult	<48.8	57, 94	1 (Ref.)		1(Ref.)	
	≤ 48.8 −<59.6	46, 95	0.82 (0.50-1.34)	P_{Trend}=**0.003**	0.93(0.55-1.58)	P_{Trend}=**0.041**
	≥59.6	25, 94	**0.43 (0.25-0.76)**		**0.53(0.29-0.97)**	

[a] OR within the combined strata of age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y;≥ 75 y) and site (Cook County, DMA). For DMA (ages <35y and 35-44y) combined due to small sample size.

[b] Linear trend (MET-h/day) on median values for categories, modeled as a continuous variable

[c] OR additionally adjusted for BMI in 1985-1989 (<18.5 kg/m^2; 18.5 – 24.99 kg/m^2; 25.0 – 29.99 kg/m^2; ≥ 30.0 kg/m^2), total energy intake in 1985-1989 (<1935; 1935-<2365; 2365-2880; ≥2880) , family history of breast cancer (yes; no), age at menarche (<13 y; 13-14 y; ≥ 15 y), reproductive history (nulliparous; first full term pregnancy < 22y; first full term pregnancy 22-29 y; first full term pregnancy ≥ 30 y), oral contraceptive use (ever; never), hormone replacement therapy use (ever; never), menopausal status at diagnosis (cases) or interview (controls) (premenopausal; postmenopausal), alcohol intake in 1985-1989 (none; < 0.7 serving/week; ≥ 0.7 serving/week) and migration status in 1985 (Poland; in US < 10y; in US ≥10y). Bold represents OR's that are statistically significant at p < 0.05.

Table 4.4 Breast cancer risk in Polish-born women residing in Cook County, IL or the Detroit Metropolitan Area, MI: Adjusted ORs for Joint Physical Activity (Met-h/Day) During Adolescence and Adulthood for all Women and Menopausal Status

	Adolescent Physical Activity Low		Medium		High	
Adult Physical Activity	**OR**	**95% C.I.**	**OR**	**95% C.I.**	**OR**	**95% C.I.**
All Women						
Low	1.00	Ref	0.97	(0.41-2.27)	0.91	(0.37-2.26)
Medium	0.99	(0.42-2.32)	0.95	(0.43-2.10)	0.30	(0.27-1.73)
High	1.21	(0.48-3.05)	0.35	(0.12-1.05)	**0.29**	**(0.11-0.77)**
Premenopausal						
Low	1.00	Ref	0.27	(0.05-1.47)	**0.03**	**(0.002-0.38)**
Medium	0.66	(0.14-3.00)	0.83	(0.23-3.00)	**0.14**	**(0.03-0.84)**
High	1.11	(0.23-5.30)	**0.12**	**(0.02-0.81)**	**0.11**	**(0.02-0.66)**
Postmenopausal						
Low	1.00	Ref	1.24	(0.41-3.70)	2.34	(0.73-7.47)
Medium	1.33	(0.41-4.31)	0.56	(0.15-1.99)	0.97	(0.28-3.36)
High	0.88	(0.23-3.70)	0.30	(0.06-1.45)	0.31	(0.08-1.22)

Within the combined strata of age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y;≥ 75 y) and site (Cook County, DMA). For DMA (ages <35y and 35-44y) combined due to small sample size, adjusted for BMI in 1985-1989 (<18.5 kg/m^2; 18.5 – 24.99 kg/m^2; 25.0 – 29.99 kg/m^2; ≥ 30.0 kg/m^2), total energy intake in 1985-1989 (<1935; 1935-<2365; 2365-2880; ≥2880) , family history of breast cancer (yes; no), age at menarche (<13 y; 13-14 y; ≥ 15 y), reproductive history (nulliparous; first full term pregnancy < 22y; first full term pregnancy 22-29 y; first full term pregnancy ≥ 30 y), oral contraceptive use (ever; never), hormone replacement therapy use (ever; never), menopausal status at diagnosis (cases) or interview (controls) (premenopausal; postmenopausal), alcohol intake in 1985-1989 (none; < 0.7 serving/week; ≥ 0.7 serving/week) and migration status in 1985 (Poland; in US < 10y; in US ≥10y). Bold represents OR's that are statistically significant at $p < 0.05$.

4.5. Comparing Effect of Total Physical Activity on BC Risk Using Common Logistic Regression, PS Methods, and Scanning Method

In previous section we used conditional logistic regression models (models 1 and 2) to evaluate the association between BC risk and total daily physical activity (MET h/day) during adolescence and adulthood. However, we want o to assess the causal effect of physical activity on BC risk by using PS methods, which had been introduced in chapter 2. Additionally, we will evaluate the causal effect of PA during adolescence and adulthood, using newly proposed scanning method. To accomplish these analyses, we will convert our main exposure (total daily adolescent PA/total daily adult PA) to binary variables. In this section to create our binary exposure for adolescents, we cut total daily adolescent physical activity (MET h/day) less than the median were coded as 0, and those above the median were assigned value of 1. Therefore, PS for adolescent, was defined as the probability of total adolescent physical activity > 47 (MET h/day) (treated) conditional on measured participants' covariates. Similarly, to create our binary exposure total PA for adults, we cut total adult physical activity (MET h/per day) at the median values for the controls (54 MET h/day). Individuals with physical activity less than the median were coded as 0, and those above the median were assigned value of 1. Thus, PS for adult, was defined as the probability of total adult physical activity >54 (MET h/day) (treated) conditional on measured participants' covariates Thus, PS for adult, was defined as the probability of total daily adult physical activity >54 (MET h/Day) (treated) conditional on measured participants covariates. PS for both exposures (total adolescent PA/total adult PA) were then estimated using conditional logistic regression (adjusted for potential covariates) where the binary exposure (total adolescent PA/total adult PA) was the dependent variable, and the covariates were the independent variables.

4.5.1 Estimation of PS and Assessing the Balance of Covariates Using Total Daily PA During Adolescence and Adulthood as a Binary Main Exposure

In order to assess the impact of potential misclassification in estimating both in traditional analyses as well as the causal inference utilizing PS estimation, we conducted the analysis with and without a covariate age1989 in the model for both exposures (total daily adolescence PA and total daily adult PA).

We estimated PS based on four PS models as follows:

i) **PS Model <u>not Includes</u> Age1989 *<u>Total Daily PA During Adolescence</u>***

$$Logit\ Prob(\ T = 1 \mid X_i = x_i\ , Tn(PA)_i = \ \mathrm{t_i}) \ = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12}(x_{i6} * x_{i6}) + \beta_{13}(x_{i9} * x_{i9})$$

Where $Tn(PA)$ denotes exposure: *Total daily adolescent Physical Activity*

X_1: age at interview; X_2: menopausal status; X_3: total daily alcohol consumption; X_4: first family history of BC; X_5: HRT (hormone replacement therapy); X_6:age at menarche; X_7:OC (Oral Contraceptive); X_8: AFFP (age at full term pregnancy); X_9; BMI; X_{10}: duration in US; X_{11}:total daily calory intake

ii) **PS Model <u>not Includes</u> Age1989 *<u>Total Daily PA During Adulthood</u>***

$$Logit\ Prob(\ T = 1 \mid X_i = x_i\ , Ad(PA)_i = \ \mathrm{t_i}) \ = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12}(x_{i6} * x_{i6}) + \beta_{13}(x_{i9} * x_{i9})$$

Where $Ad(PA)$ denotes exposure: ***<u>Total daily adult Physical Activity</u>***

X_1: age at interview; X_2: menopausal status; X_3: total daily alcohol consumption; X_4: first family history of BC; X_5: HRT (hormone replacement therapy); X_6:age at

menarche; X_7:OC (Oral Contraceptive); X_8: AFFP (age at full term pregnancy); X_9; BMI; X_{10}: duration in US; X_{11}:total daily calory intake

iii) **PS Model <u>Includes</u> Age1989 for Total Daily *<u>PA During Adolescence</u>***

$$Logit\ Prob(\ T = 1 \mid X_i = x_i\ ,\ Tn(PA)_{\mathrm{i}} = \mathrm{t_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12}(x_{i6} * x_{i6}) + \beta_{13}(x_{i9} * x_{i9}) + \beta_{14} x_{i12}$$

Where $\boldsymbol{Tn(PA)}$ **denotes exposure: *<u>Total Daily Adolescent PA</u>***

X_1: age at interview; X_2: menopausal status; X_3: total daily alcohol consumption; X_4: first family history of BC; X_5: HRT (hormone replacement therapy); X_6:age at menarche; X_7:OC (Oral Contraceptive); X_8: AFFP (age at full term pregnancy); X_9; BMI; X_{10}: duration in US; X_{i11}:total daily calory intake; X_{i12}:**Age at 1989**

iv) **PS Model <u>Includes</u> Age1989 for *<u>Total Daily PA During Adulthood</u>***

$$Logit\ Prob(\ T = 1 \mid X_i = x_i\ ,\ Ad\ (PA)_{\mathrm{i}} = \mathrm{t_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12}(x_{i6} * x_{i6}) + \beta_{13}(x_{i9} * x_{i9}) + \beta_{14} x_{i12}$$

Where $\boldsymbol{Ad\ (PA)}$ **denotes exposure: *Total daily adolescent PA.***

X_1: age at interview; X_2: menopausal status; X_3: total daily alcohol consumption; X_4: first family history of BC; X_5: HRT (hormone replacement therapy); X_6:age at menarche; X_7:OC (Oral Contraceptive); X_8: AFFP (age at full term pregnancy); X_9; BMI; X_{10}: duration in US; X_{i11}:total daily calory intake; $\boldsymbol{X_{i12}}$:**Age at 1989**

To ensure that the PS estimation provides balance in the covariates for treated and untreated, we conducted analysis needed to assess the assumption of PS for all four PS models that we introduced them in above.

As we explained in chapter two (chapter 2 section 7), there are several assumptions regarding the PS methods which are required to be held. The first assumption for PS model is "strongly ignorable treatment assignment". Therefore, we've provided, assessing the balance of covariates for all four PA models (Figures 4.1 to 4.16).

i) Assessing the Balance of Covariates for Total Daily Adolescent PA PS-Model does not Include Age1989

Figure 4.1 shows the distribution of the propensity scores for both treated (Adolescent PA=1) and untreated (Adolescent PA=0) distributions. The two distributions visually are similar (overlap assumption) and range from 0.06 to about 0.76 (positivity). Therefore, the ignitability of treatment assumption holds.

Figure 4.2 provides PS clouds for treated and untreated. According to this, all PS estimated are in support region for both treated and untreated.

Figure 4.3 shows the boxplot of PS distributions by treated (PA) and untreated. PS for the treated group is shifted to higher values than untreated group, ranged from 0.27 to 0.91 for treated and from 0.13 to 0.88 for untreated group.

Figure 4.4 presents covariates standardized mean differences. All covariates standardized mean differences are between -0.25 and 0.25, for all observed PS as well as weighted PS.

Figure 4.5 denotes covariates densities for continuous covariates. As it shown in this figure, there is a good overlapping of weighted densities of PS for BMI, menarche, and alcohol consumption.

Figure 4.1 Estimation of PS for Adolescent PA without Age1989

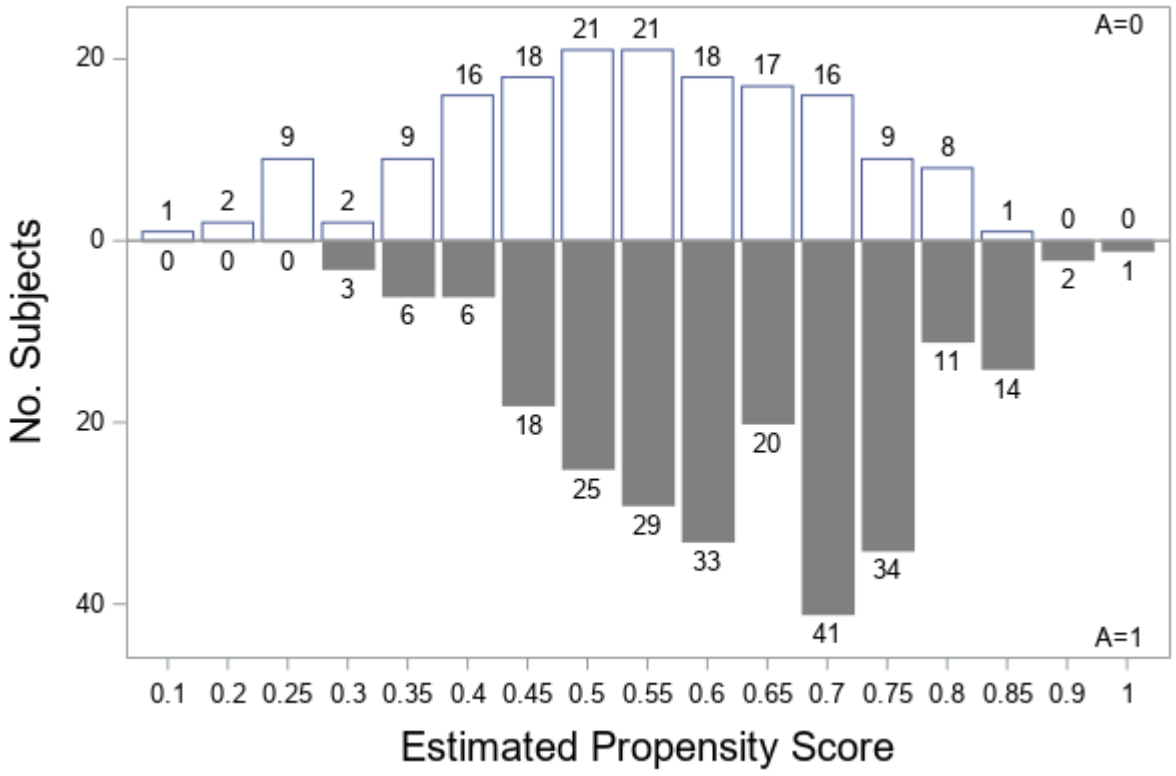


Figure 4.2 PS Clouds for Adolescent PA and Controls

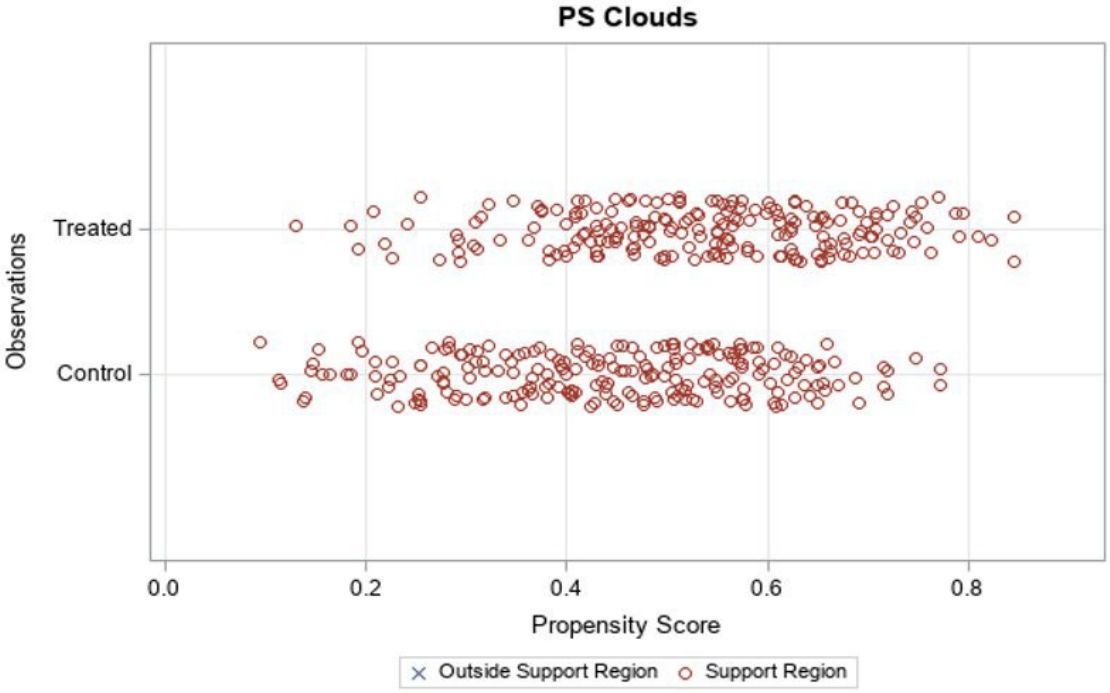


Figure 4.3 Boxplot of PS Distribution for Adolescent PA and Controls

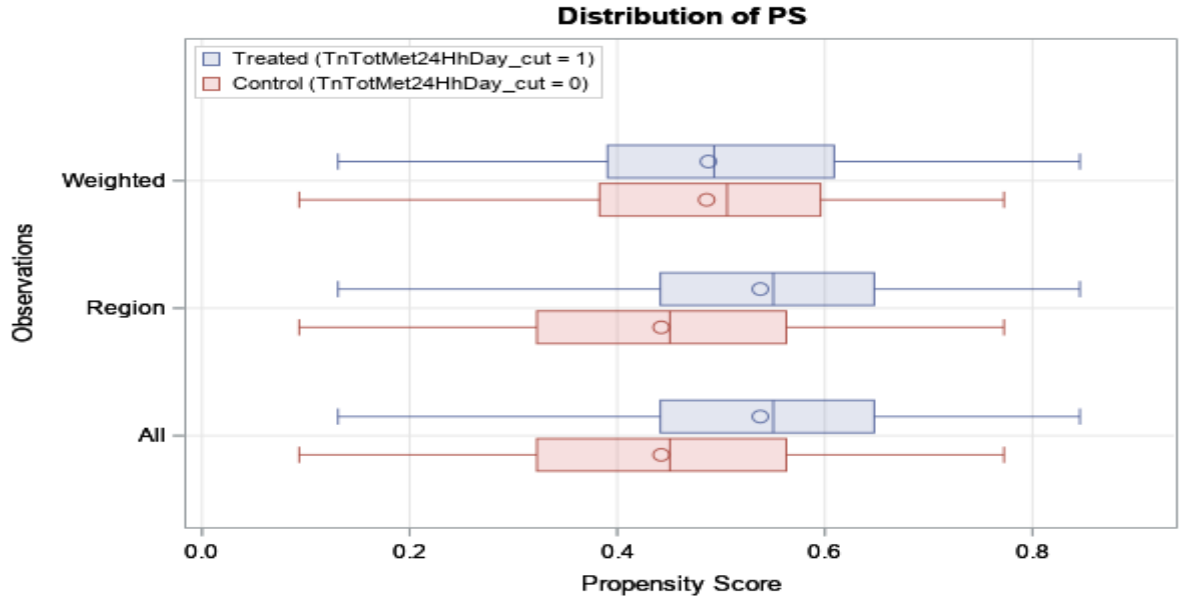


Figure 4.4 Covariates Standardized Mean Differences of PS

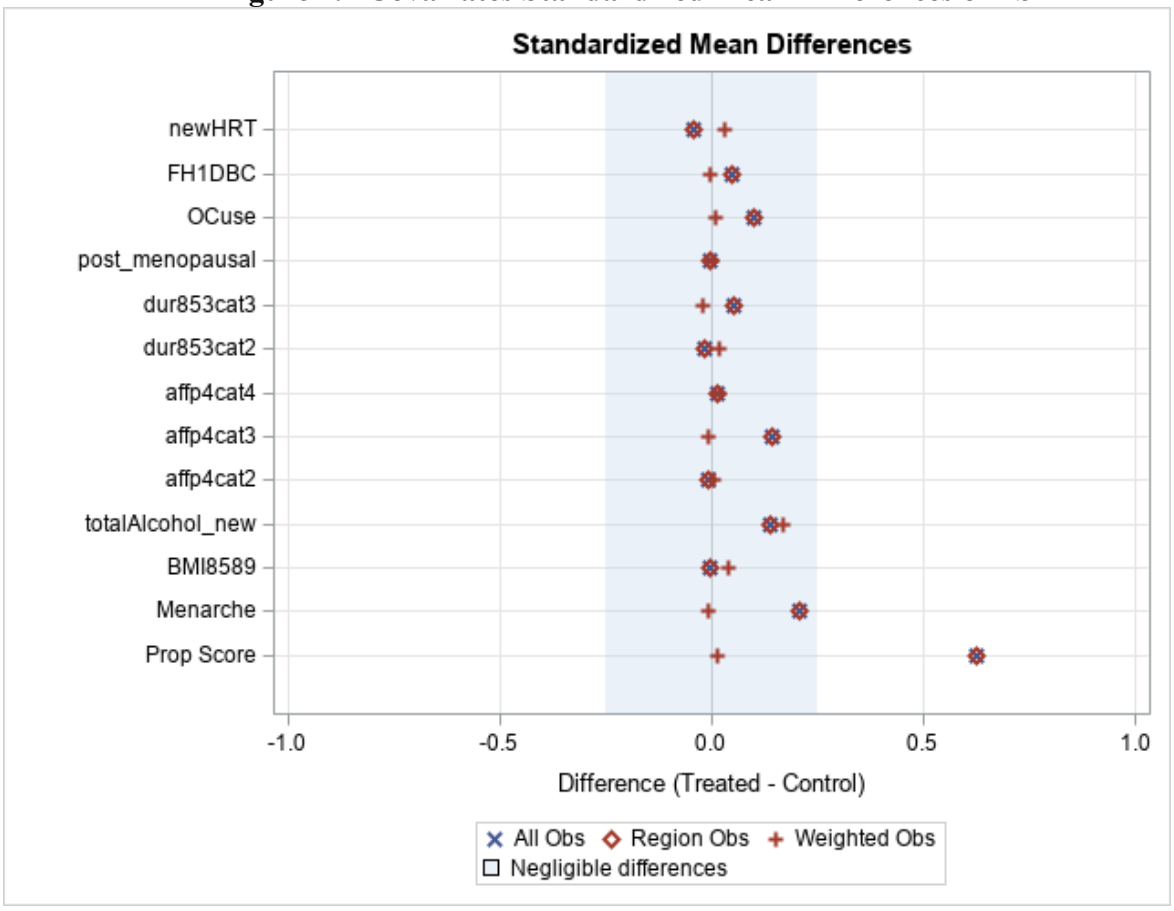
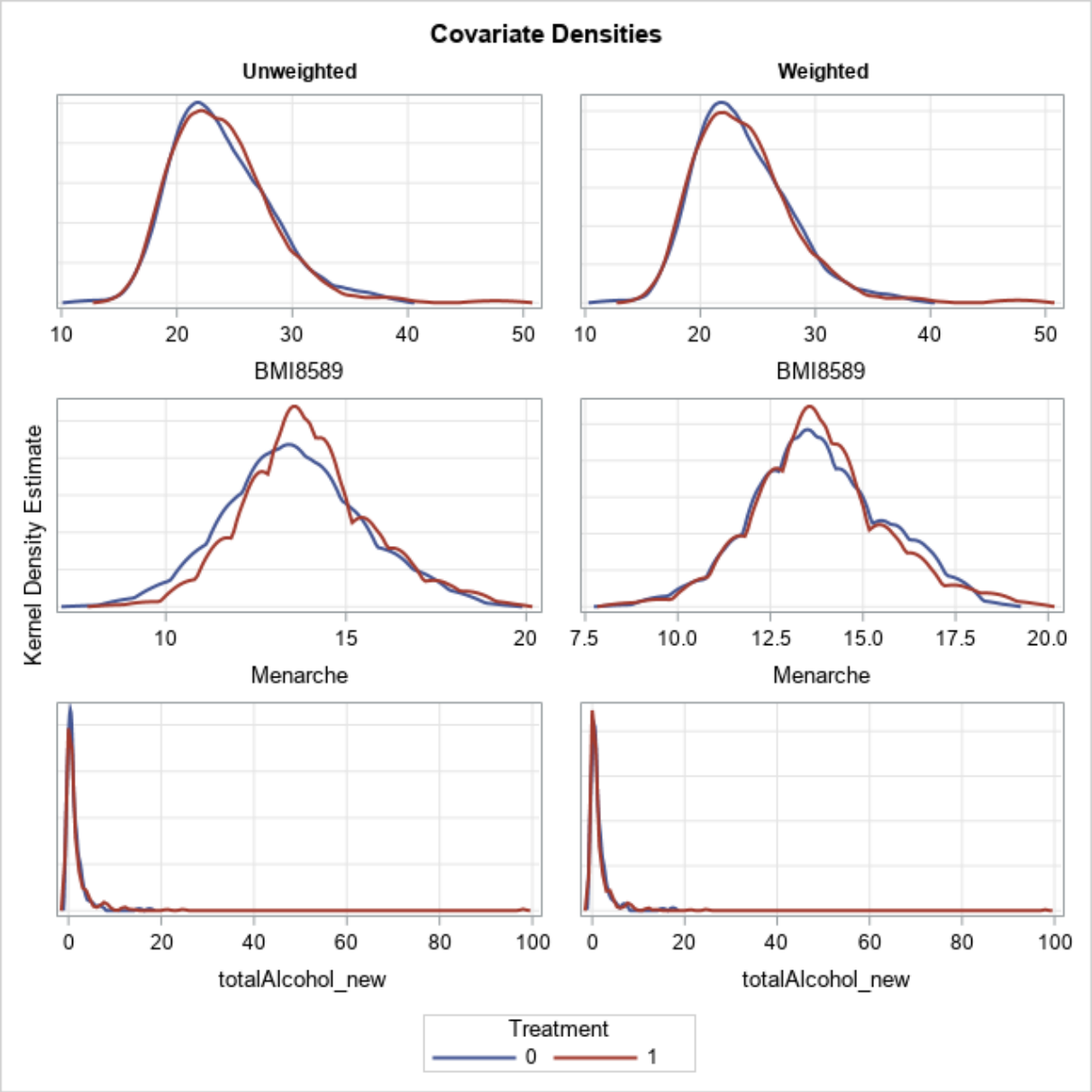


Figure 4.5 Covariates Densities for Continuous Covariate



ii) Assessing the Balance of Covariates for Total Daily Adult PA PS-Model does not Include Age1989

Figure 4.6 shows the distribution of the propensity scores for both treated (total daily Adult PA=1) and untreated (total daily Adult PA=0) distributions. The two distributions visually are similar (overlap assumption) and range from 0.06 to about 0.76 (positivity). Therefore, the ignitability of treatment assumption holds.

Figure 4.7 provides PS clouds for treated and untreated. According to this, all PS estimated are in support region for both treated and untreated.

Figure 4.8 shows the boxplot of PS distributions by treated (PA) and untreated. PS for the treated group is shifted to higher values than untreated group, ranged from 0.151 to 0.76 for treated and from 0.06 to 0.76 for untreated group.

Figure 4.9 presents covariates standardized mean differences. All covariates standardized mean differences are between -0.25 and 0.25, for all observed PS as well as weighted PS. Only one category of AFFP (age at full term pregnancy) is not at the range of -0.25 and 0.25.

Figure 4.10 denotes covariates densities for continuous covariates. As it shown in this figure, there is a good overlapping of weighted densities of PS for BMI, menarche, and alcohol consumption.

Figure 4.6 Estimation of PS for Adult PA without Age1989

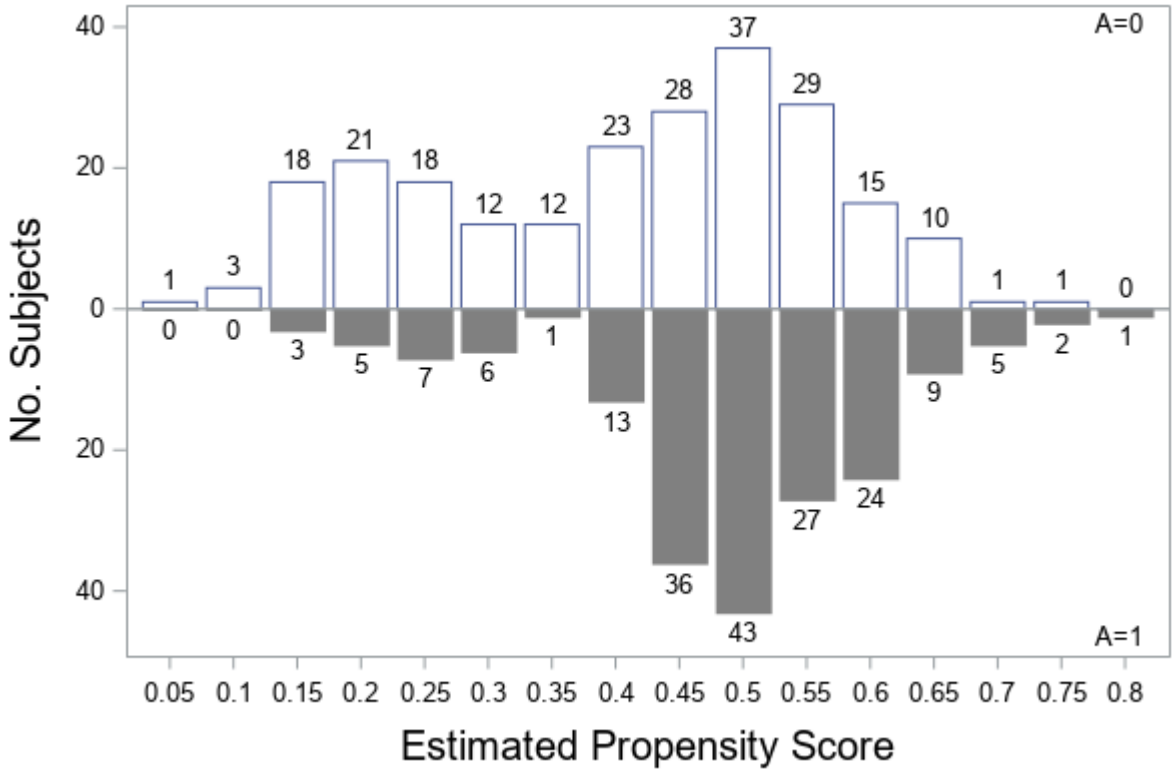


Figure 4.7 PS Clouds for Treated (Total Daily Adult PA) and Controls

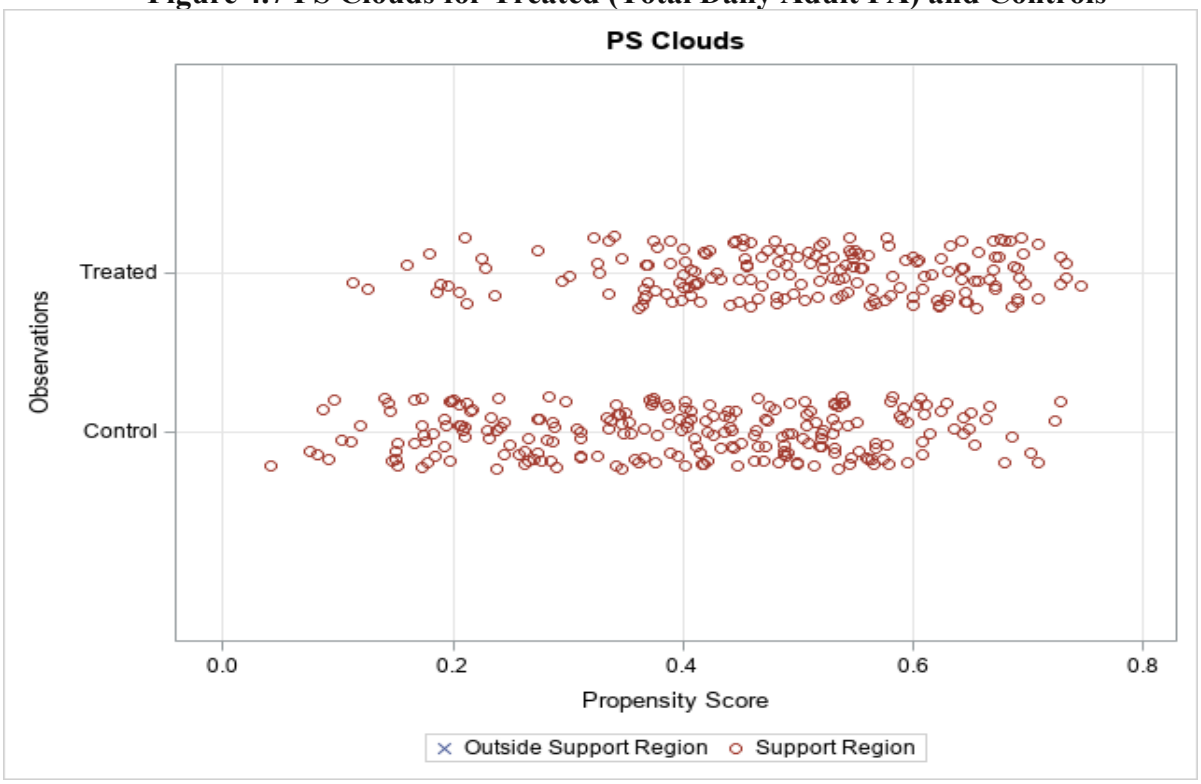


Figure 4.8 Boxplot of PS Distribution by Treated and Control

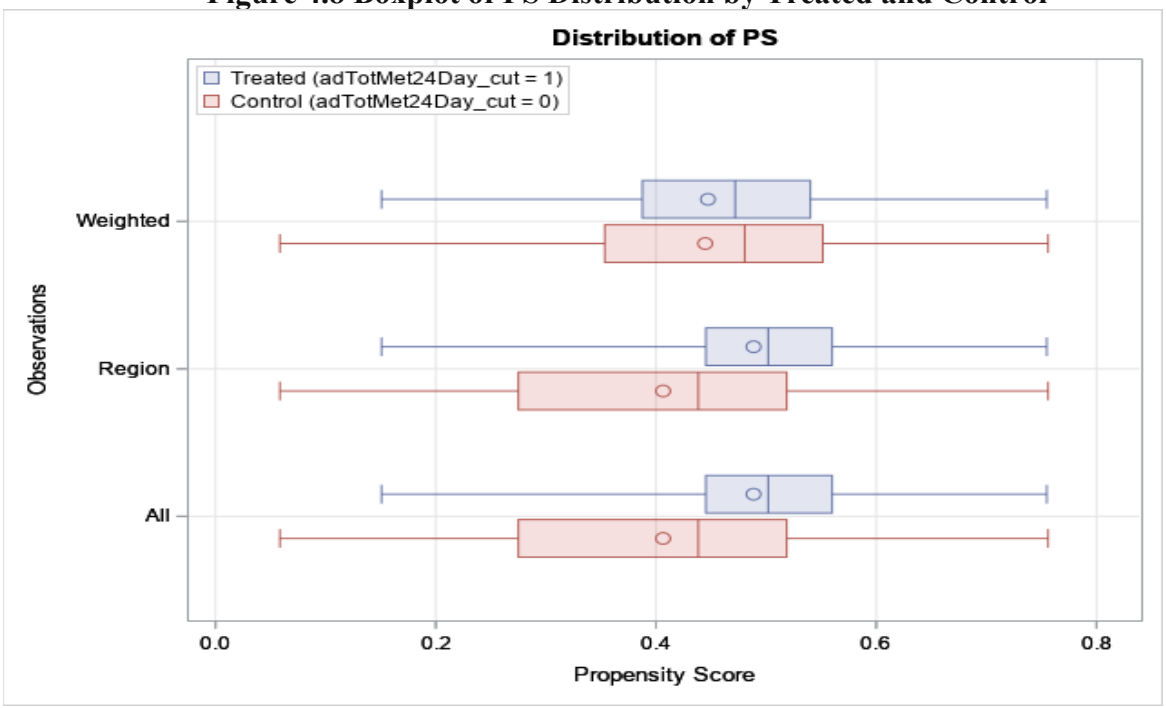


Figure 4.9 Covariates Standardized Mean Differences of PS

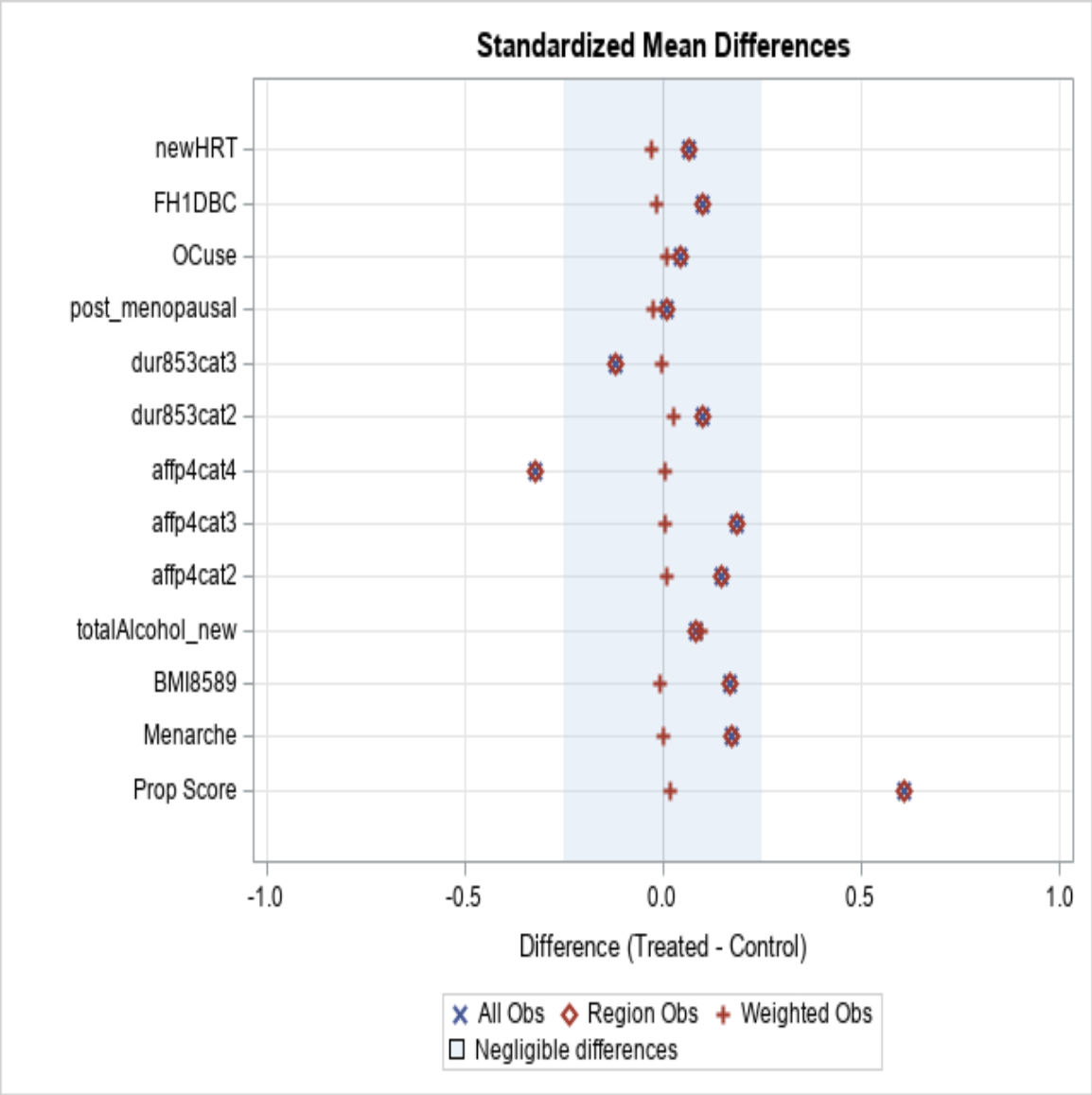
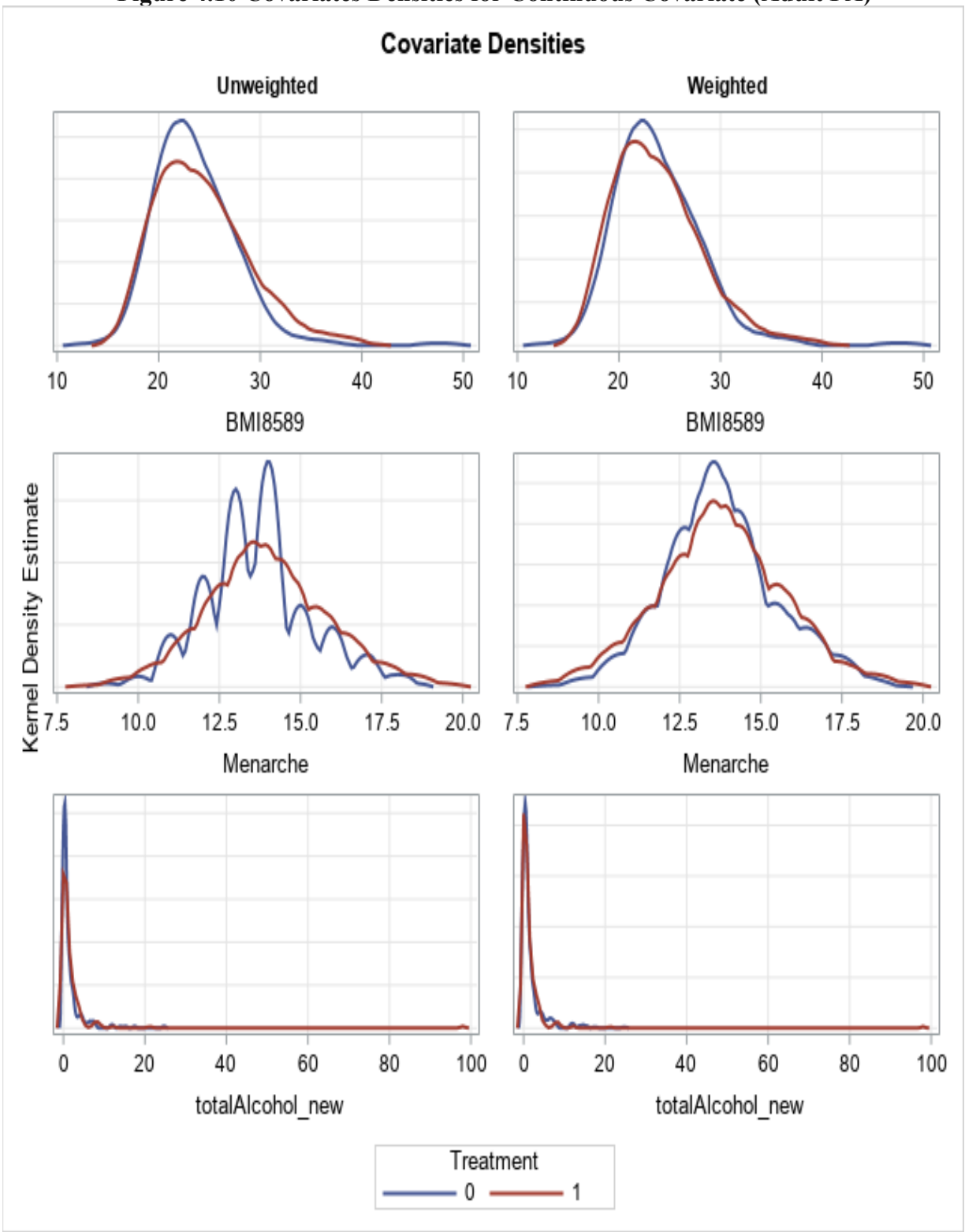


Figure 4.10 Covariates Densities for Continuous Covariate (Adult PA)



iii) Assessing the Balance of Covariates for Total Daily Adolescent PA PS -Model Includes Age1989

Figure 4.11 shows the distribution of the propensity scores for both treated (total daily adolescent PA=1) and untreated (total daily adolescent PA=0) distributions. The two distributions visually are similar (overlap assumption) and range from zero to about 0.99 (positivity). Therefore, the ignitability of treatment assumption holds.

Figure 4.12 provides PS clouds for treated and untreated. According to this, all PS estimated are in support region for both treated and untreated.

Figure 4.13 shows the boxplot of PS distributions by treated (PA) and untreated. PS for the treated group is shifted to higher values than untreated group, ranged from 0.28 to 0.91 for treated and from 0.13 to 0.88 for untreated group.

Figure 4.14 presents covariates standardized mean differences. All covariates standardized mean differences are between -0.25 and 0.25, for all observed PS as well as weighted PS.

Figure 4.15 denotes covariates densities for continuous covariates. As it shown in this figure, there is a good overlapping of weighted densities of PS for BMI, menarche, and alcohol consumption and age 1989.

Figure 4.11 Estimation of PS for Adolescent PA with Age1989

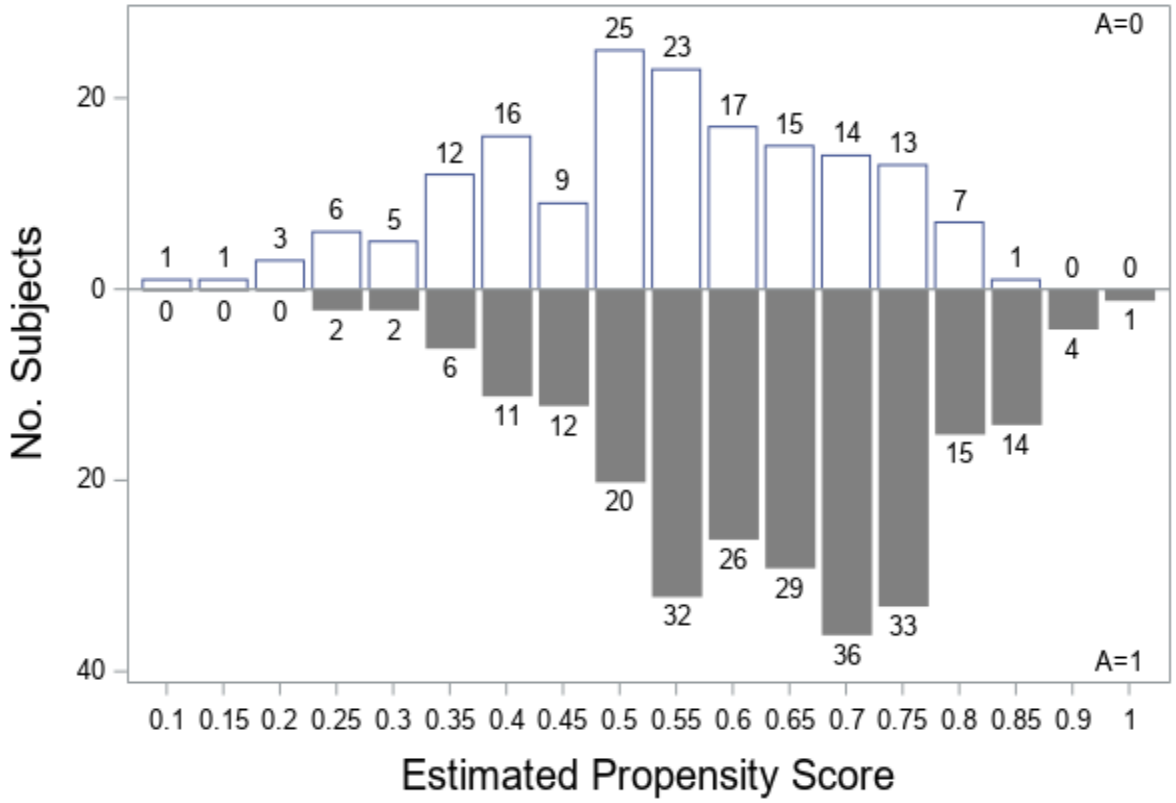


Figure 4.12 PS Clouds for Treated (Adolescent PA) and Controls

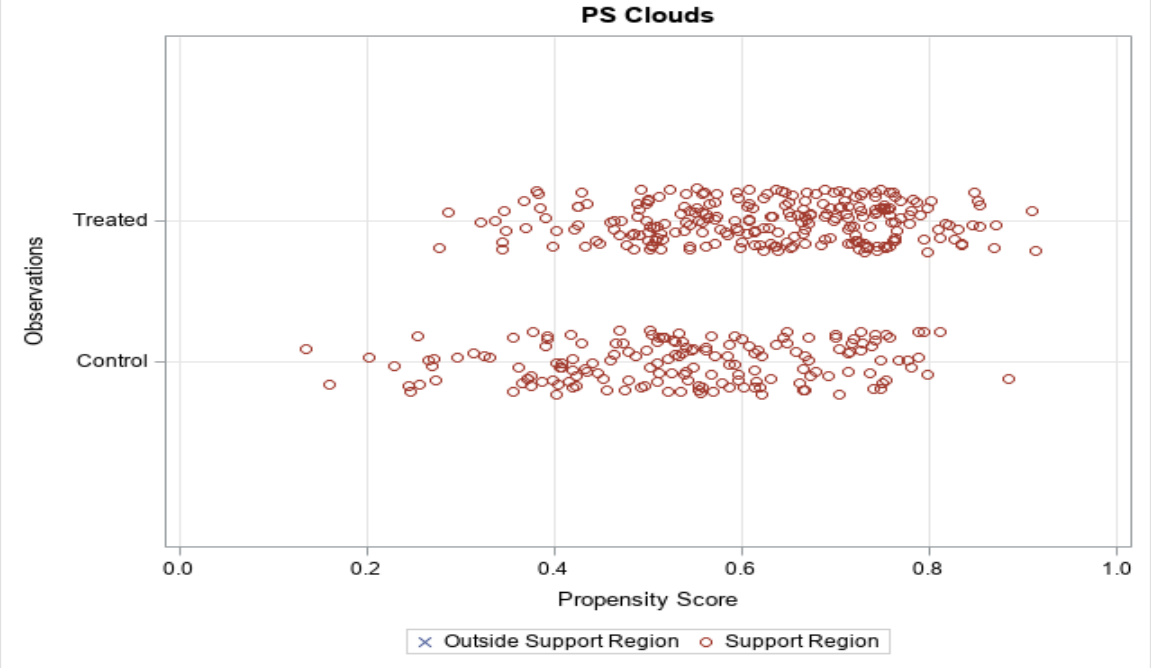


Figure 4.13 Boxplot of PS Distribution by Treated (Adolescent PA) and Control

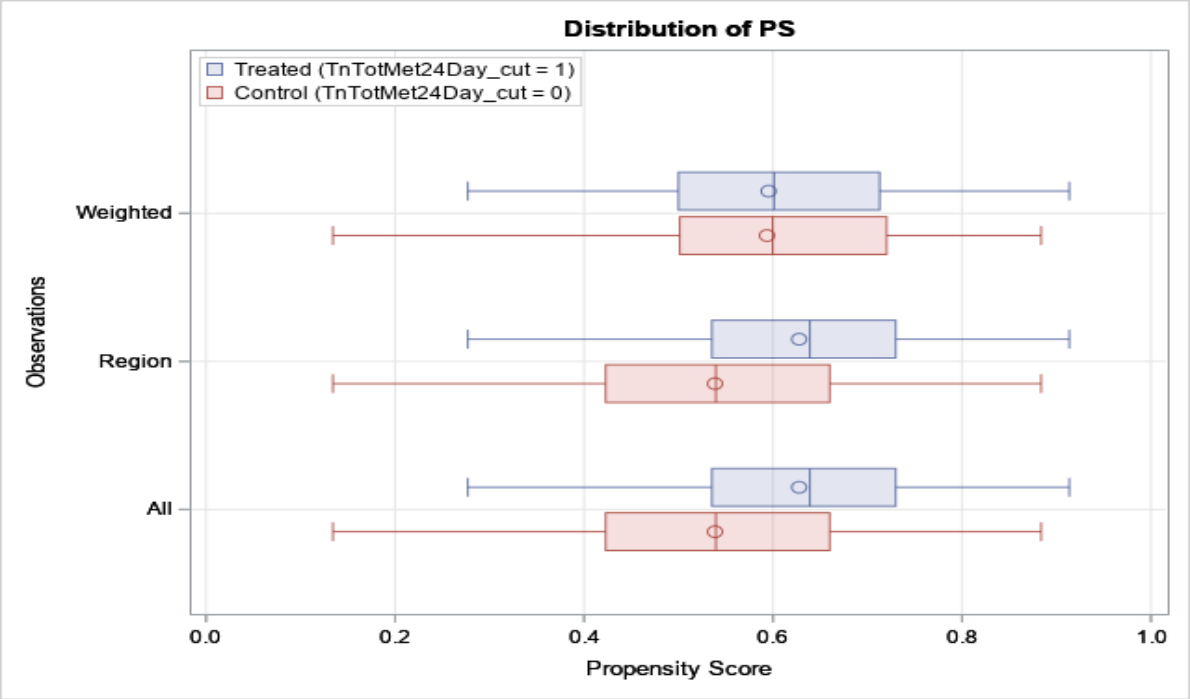


Figure 4.14 Covariates Standardized Mean Differences of PS (Adolescent PA)

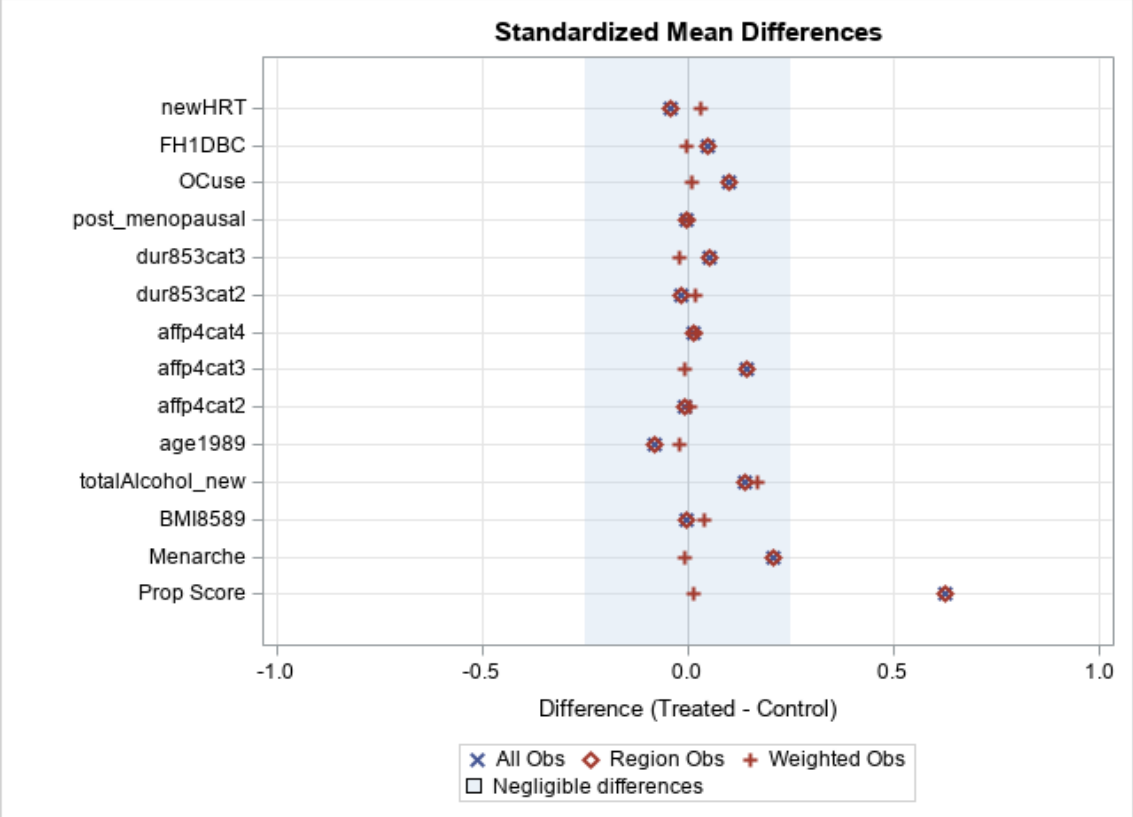
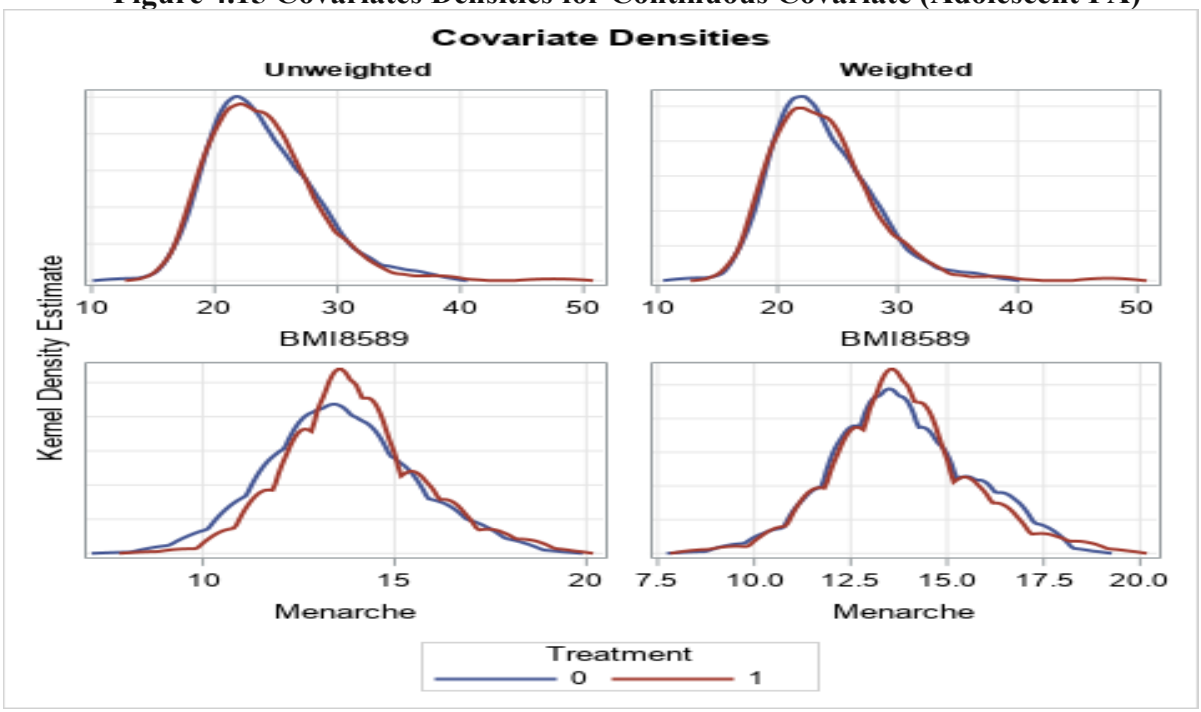
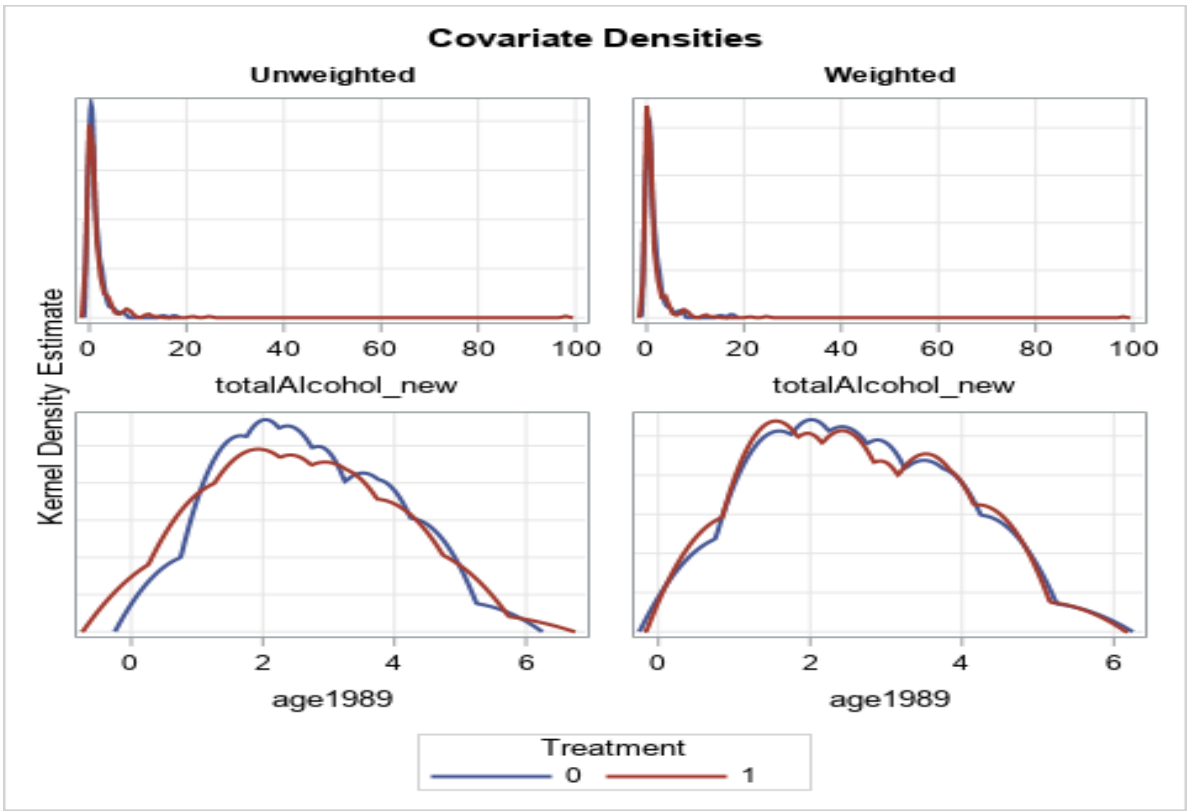


Figure 4.15 Covariates Densities for Continuous Covariate (Adolescent PA)





iv) Assessing the Balance of Covariates for Total Daily Adult PA PS-Model Includes Age1989

Figure 4.16 shows the distribution of the propensity scores for both treated (total daily adult PA=1) and untreated (total daily adult PA=0) distributions. The two distributions visually are similar (overlap assumption) and range from 0.06 to about 0.75 (positivity). Therefore, the ignitability of treatment assumption holds.

Figure 4.17 provides PS clouds for treated and untreated. According to this, all PS estimated are in support region for both treated and untreated.

Figure 4.18 shows the boxplot of PS distributions by treated (PA) and untreated. PS for the treated group is shifted to higher values than untreated group, ranged from 0.15 to 0.75 for treated and from 0.06 to 0.75 for untreated group.

Figure 4.19 presents covariates standardized mean differences. All covariates standardized mean differences are between -0.25 and 0.25, for all observed PS as well as weighted PS. Again, age at full term pregnancy(cat4) is not in the region. Never mind, all other covariates are between -0.25 and 0.25.

Figure 4.20 denotes covariates densities for continuous covariates. As it shown in this figure, there is a good overlapping of weighted densities of PS for BMI, menarche, and alcohol consumption and age 1989.

Figure 4.16 Estimation of PS for Adult PA with Age1989

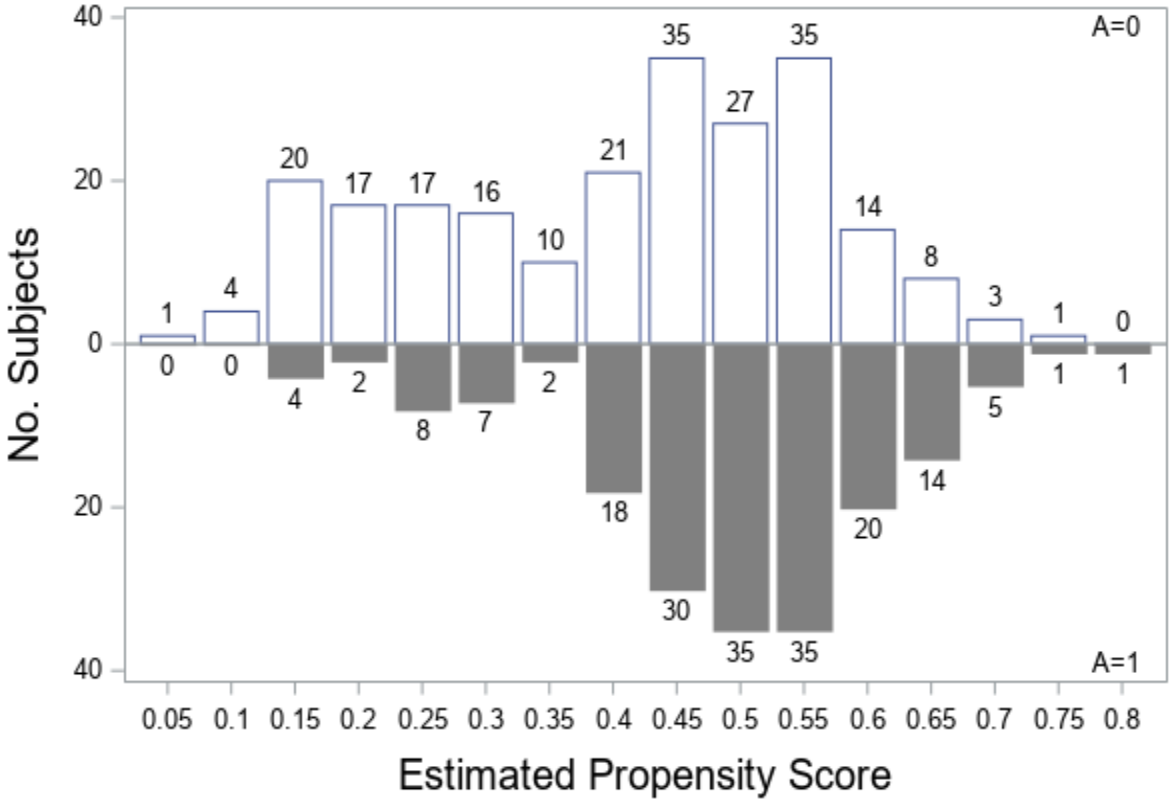


Figure 4.17 PS Clouds for Treated and Controls for Total Daily Adult PA

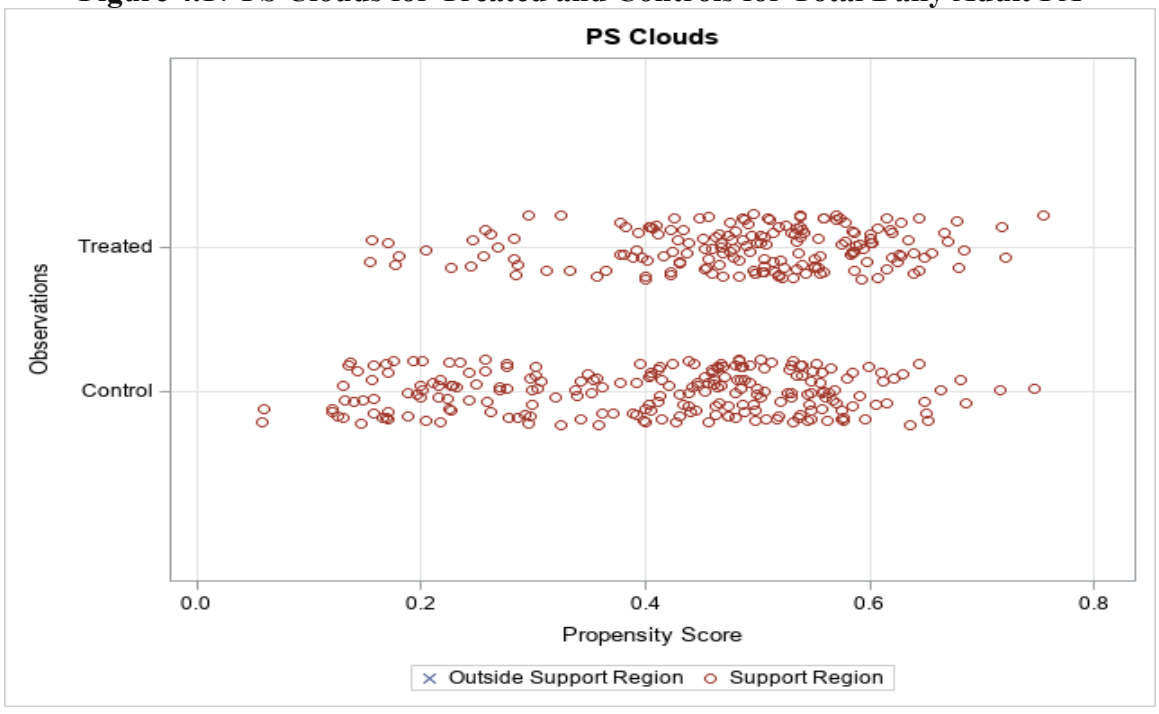


Figure 4.18 Boxplot of PS Distribution by Treated (Adult PA) and Control

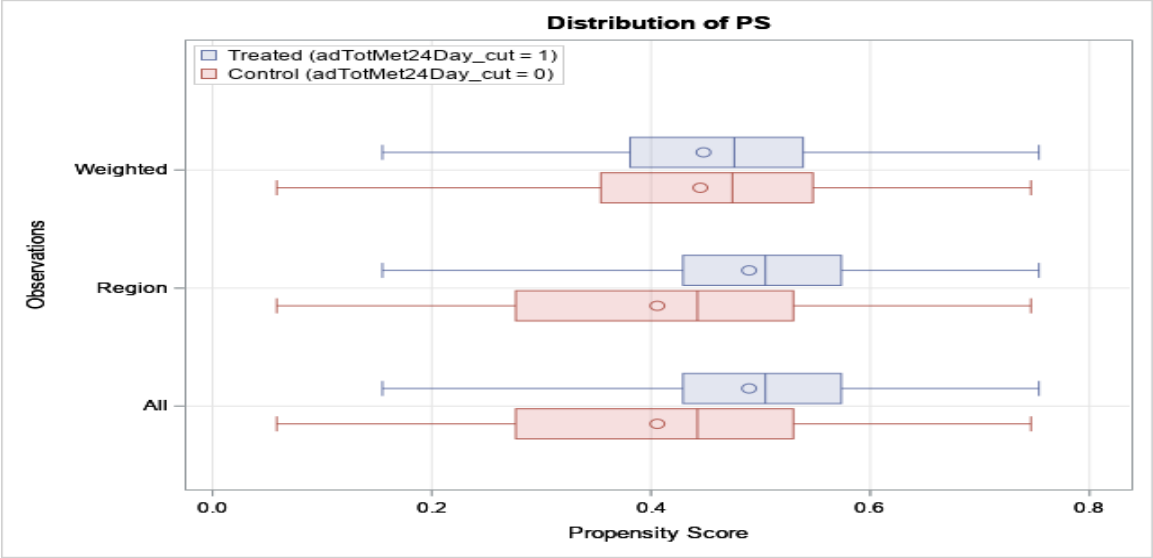


Figure 4.19 Covariates Standardized Mean Differences of PS

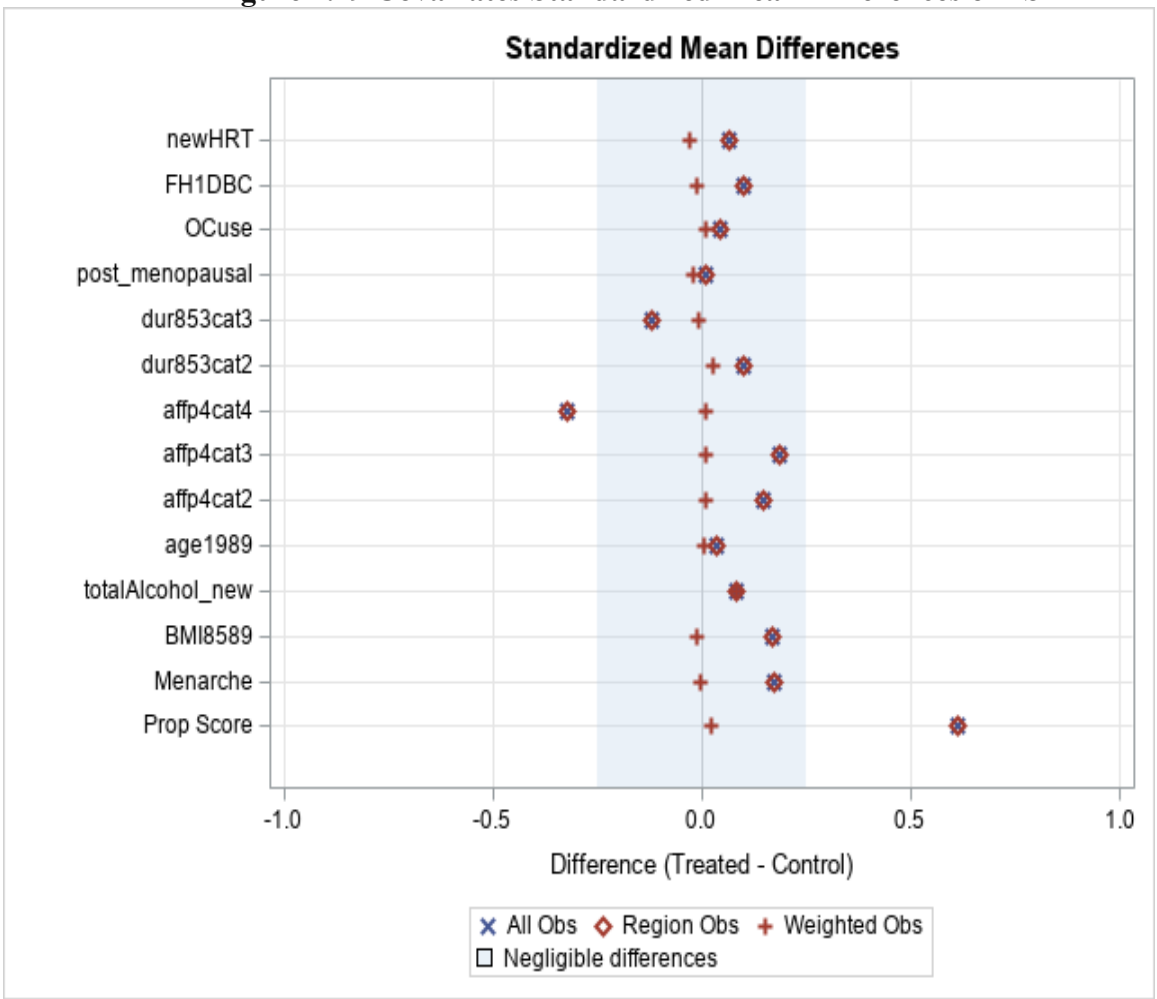
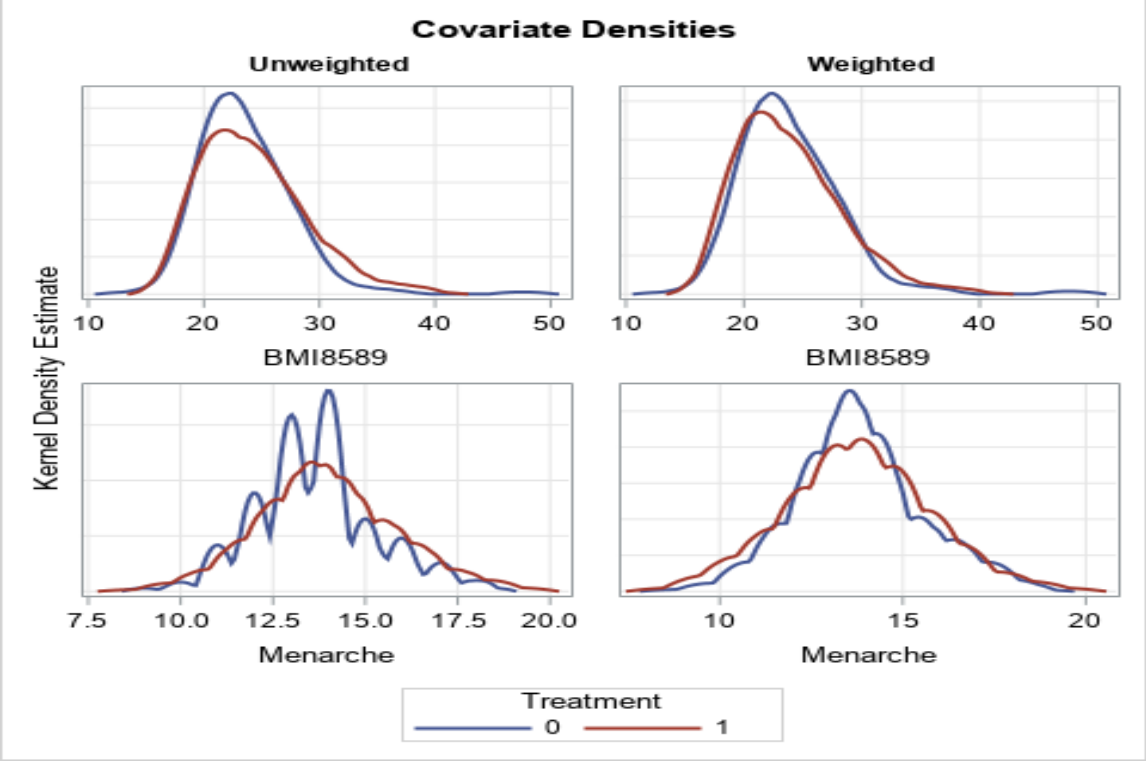
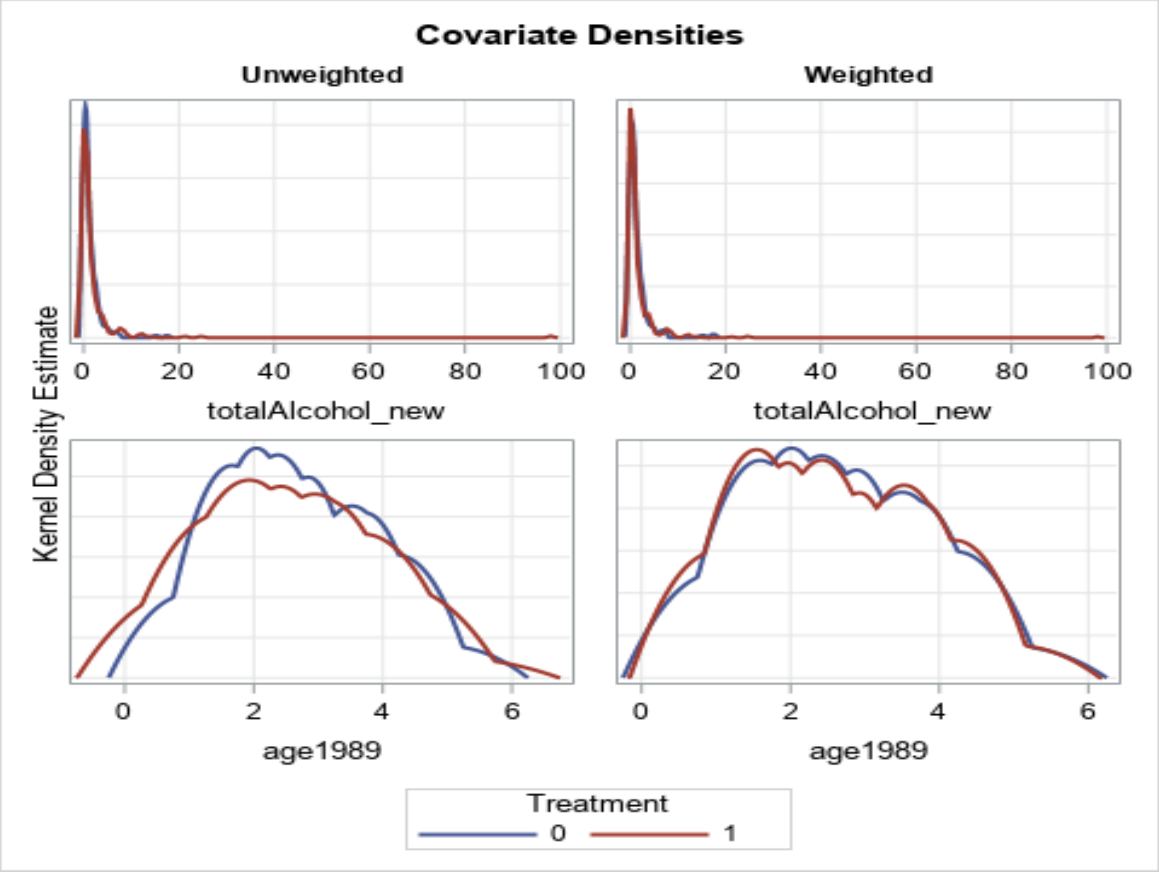


Figure 4.20 Covariates Densities for Continuous Covariate





4.5.2 Statistical Analysis

In this section we compared the common logistic regression methods (with the same binary cut for exposure), common PS methods and finally the scanning method. We've used conditional logistic regression for all analysis to obtain odds ratios and 95% confidence intervals. We used common logistic regression methods: model 1 (unadjusted), model 2 (Adjusted for joint strata of age and site) and, model 3 (adjusted for all potential risk factors). We've also used logistic regression for 2 methods of propensity score: model 4 (IPW), model 5 (using PS as a covariate adjustment). For the model 6 (PS-stratification), we used Mantel-Haenszel method for estimate ORs.

The newly developed scanning method (model 7) which is based on continuous discretization generates a cut-point (s) specific causal effect, denoted by $\alpha(s)$ as it has been described in chapter 3. For each cut point s, the effect of treatment on dichotomous outcomes is estimated within each stratum. We used 10 thresholds for this study. The stratum-specific estimates of treatment effect are then pooled across stratum to estimate an overall treatment effect, using weighted average (w_i=$\frac{1}{SE_i^2}$). For our case status outcome (BC) we estimated the causal effect as follow:

$$\hat{\alpha} = \int_0^1 \hat{\alpha}(s)\,\hat{\omega}(s)\,ds \,/\, \int_0^1 \hat{\omega}(s)\,ds \quad \text{where } \hat{\omega}(s) = \frac{1}{\hat{V}ar(\hat{a}(s))}$$

To estimate standard error of $\hat{\alpha}$ we used bootstrap resampling and then calculated 95% CI for the scanning method. In this table, we have compared the results of the different methods for assessing the effect of total daily adolescent/adulthood activity on BC risk.

We provide the results for:

1) PS Model for Total Daily PA During Adolescence without age1989 (Table 4.5)
2) PS Model for Total Daily PA During Adulthood without age1989 (Table 4.6)

3) Additionally, Age1989 were added to PS Model for Total Daily PA During Adolescence (Table 4.7)
4) Additionally, Age1989 were added to PS Model for Total Daily PA During Adulthood (Table 4.8)

4.5.3 Results/Conclusions

Based on the traditional analyses, *not including the age1989* in the model, total daily *adolescent physical activity* reduces breast cancer risk by approximately 40 % (*Table 4.5*). Based on PS analyses, the estimates of our odds ratios are similar in magnitude and of similar statistical significance (P-value<0.05). However, the newly proposed scanning method provided the shortest confidence interval with the similar point estimate as the other PS methods.

In *table 4.6*, we provide estimates of *total daily adult physical activity* on BC risk using PS methods and scanning method (*without age1989 in the models*). Using PS methods, total daily adult PA had 45% reduction on BC risk. Total daily adult PA for Adjusted logistic regression analysis (model 3) had the same reduction as PS methods. The scanning method provides an estimate of approximately 49% reduction on BC risk. In addition, the scanning method provides the shortest confidence interval with the similar point estimate as the other PS methods.

Age 1989 was added to our models to *assess the potential impact* of covariate misclassification on estimation of *causal effec*t using developed standard *PS* methodologies and our newly developed *scanning method*.

Based on the traditional and PS analyses, while including the age1989 in the model, several of the estimates of ORs for the effect of total daily adolescent PA did not reach statistical significance, although they still provided estimates indicating reduction in risk (Table 4.7). However, the estimate obtained from scanning method remained approximately the same as

when age1989 was not in the model and was statistically significant.

In table 4.8, we provide estimates of total daily adult PA on BC risk using PS methods and scanning method (with age1989 in the models). Estimates of ORs for the total daily adult PA were similar to those obtained when the model was run without age1989. The scanning method continued to provide the shortest confidence interval.

The comparison of the estimates from models with and without age1989, points to the dilemma that researchers often face which covariate should be included in the model. This comparison allows us to point out that the scanning method is least sensitive to the misclassification of covariate in estimation of causal effect using OR's.

However, interpretation of the causal effects in terms of odds ratios is more complex than interpretation of causal effect in terms of risk difference. Furthermore, the goal of the causal inference is to estimate risk difference due to treatment effect which cannot be estimated directly in case control studies. Consequently, in the next chapter we will apply a case-control weighted Target Maximum Likelihood Method (CCW-TMLE) proposed by Rose et al, 2017 which under certain assumptions allows us to estimate the causal risk difference (ATE) for a case-control study. [41]

Table 4.5 Comparison of Different Methods for Evaluating Effect of Total Daily Adolescent Physical Activity on BC Risk in Polish Migrants in US PS-Model does not Include Age1989

Method	OR 95% C.I.	P-Value
Model 1 (Common Logistic Regression Unadjusted)	0.61 (0.40, 0.93)	0.0211*
Model 2 (Common Logistic Regression Adjusted only for site and age) [a]	0.58 (0.38, 0.91)	0.0167*
Model 3 (Common Logistic Regression Adjusted for all risk factors) [b]	0.60 (0.38, 0.95)	0.0276*
Model 4 (PS _ IPW)	0.64 (0.47,0.85)	0.0027*
Model 5 (PS_ Covariate Adjusted)	0.64 (0.41,1.003)	0.0517*
Model 6 (PS _ Stratification Using Quintiles)	0.63 (0.40,0.98)	0.0391*
Model 7 (Scanning Method) §	0.66 (0.64,0.69)	<0.0001**

[a] OR adjusted for joint age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y; ≥ 75 y) and site (Cook County, DMA). For DMA (ages <35y and 35-44y) combined due to small sample size. Adolescence physical activity median cut point was used.

[b] Additionally adjusted for BMI, total energy intake in 1985-1989 (quartiles) , family history of breast cancer (yes; no), age at menarche, reproductive history (nulliparous; first full term pregnancy < 22y; first full term pregnancy 22-29 y; first full term pregnancy ≥ 30 y), oral contraceptive use (ever; never), hormone replacement therapy use (ever; never), menopausal status at diagnosis (cases) or interview (controls) (premenopausal; postmenopausal), alcohol intake in 1985-1989 (none; < 0.7 serving/week; ≥ 0.7 serving/week) and migration status in 1985 (Poland; in US < 10y; in US ≥10y). Adolescence physical activity median cut point was used.

***** Bold indicates statistical significance of $p < 0.05$.

§ SE and C.I. in scanning method were estimated through Bootstrapping

Table 4.6 Comparison of Different Methods for Evaluating Effect of Total Daily Adult Physical Activity on BC Risk in Polish Immigrant Women to US PS-Model does not Include Age1989

Method	OR 95% C.I.	P-Value
Model 1 (Common Logistics Regression Unadjusted)	0.48 (0.31, 0.74)	0.0009**
Model 2 (Common Logistics Regression Adjusted for join site and age) [a]	0.48 (0.31, 0.74)	0.0011**
Model 3 (Common Logistics Regression Adjusted for all risk factors) [b]	0. 53 (0.32, 0.85)	0.0088**
Model 4 (PS_IPW)	0.55 (0.41,0.74)	0.0001**
Model 5 (PS_ Covariate Adjusted as a Continuous Variable)	0.55 (0.35,0.87)	0.0113*
Model 6 (PS_ Stratification Using Quintiles)	0.55 (0.34,0.78)	0.0118*
Model 7 (PS_ Scanning Method) §	0.51 (0.48,0.53)	<0.0001**

[a] OR adjusted for joint age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y; ≥ 75 y) and site (Cook County, DMA). For DMA (ages <35y and 35-44y) combined due to small sample size. Adolescence physical activity median cut point was used.

[b] Additionally adjusted for BMI, total energy intake in 1985-1989 (quartiles) , family history of breast cancer (yes; no), age at menarche, reproductive history (nulliparous; first full term pregnancy < 22y; first full term pregnancy 22-29 y; first full term pregnancy ≥ 30 y), oral contraceptive use (ever; never), hormone replacement therapy use (ever; never), menopausal status at diagnosis (cases) or interview (controls) (premenopausal; postmenopausal), alcohol intake in 1985-1989 (none; < 0.7 serving/week; ≥ 0.7 serving/week) and migration status in 1985 (Poland; in US < 10y; in US ≥10y). Adult physical activity median cut point was used.

***** Bold indicates statistical significance of $p < 0.05$.

§ SE and C.I. in scanning method were estimated through Bootstrapping

Table 4.7 Comparison of Different Methods for Evaluating Effect of Total Daily Adolescent Physical Activity on BC Risk in Polish Immigrant Women to US PS-Model Includes Age1989

Method	OR 95% C.I.	P-Value
Model 1 (Common Logistic Regression Adjusted only for age1989)	0.61 (0.40, 0.93)	0.0215*
Model 2 (Common Logistic Regression Adjusted for age1989, site and age at diagnosis/interview) [a]	0.64 (0.42, 1.003)	0.052
Model 3 (Common Logistic Regression Adjusted for age1989, site, age at diagnosis and for all potential risk factors) [b]	0. 69 (0.42, 1.12)	0.1322
Model 4 (PS Method_ IPW)	0.71 (0.53,0.96)	0.025*
Model 5 (PS_ Covariate Adjusted PS Model include age1989)	0.71 (0.45,1.115)	0.1368
Model 6 (PS_ Stratification Using Quintiles)	0.70 (0.44,1.05)	0.1264
Model 7 (Scanning Method) §	0.66 (0.64,0.69)	<0.0001**

[a] OR adjusted for age at 1989 within the combined strata of age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y;≥ 75 y) and site (Cook County, DMA). For DMA (ages <35y and 35-44y) combined due to small sample size. Adolescence physical activity median cut point was used.

[b] OR additionally adjusted for Age at 198, BMI, total energy intake in 1985-1989 (quartiles) , family history of breast cancer (yes; no), age at menarche, reproductive history (nulliparous; first full term pregnancy < 22y; first full term pregnancy 22-29 y; first full term pregnancy ≥ 30 y), oral contraceptive use (ever; never), hormone replacement therapy use (ever; never), menopausal status at diagnosis (cases) or interview (controls) (premenopausal; postmenopausal), alcohol intake in 1985-1989 (none; < 0.7 serving/week; ≥ 0.7 serving/week) and migration status in 1985 (Poland; in US < 10y; in US ≥10y). Adolescence physical activity median cut point was used.

.* Bold indicates statistical significance of $p < 0.05$. ** Bold indicates statistical significance of $p < 0.01$

§ SE and C.I. in scanning method were estimated through Bootstrapping

Table 4.8 Comparison of Different Methods for Evaluating Effect of Total Daily Adult Physical Activity on BC Risk in Polish Immigrant Women to US PS-Model Includes Age1989

Method	OR 95% C.I.	P-Value
Model 1 (Common Logistics Regression adjusted only for age1989)	0.47 (0.31, 0.74)	0.0008**
Model 2 (Common Logistic Regression Adjusted for age1989, site and age at diagnosis/interview) [a]	0.49 (0.31, 0.78)	0.0028**
Model 3 (Common Logistic Regression Adjusted for age1989, site, age at diagnosis and for all potential risk factors) [b]	0. 55 (0.34, 0.90)	0.0174*
Model 4 (PS_IPW)	0.58 (0.43,0.79)	0.0004**
Model 5 (PS_ Covariate Adjusted as a continuous variable)	0.58 (0.36,0.91)	0.0190*
Model 6 (PS_ Stratification Using Quintiles)	0.55 (0.35,0.88)	0.0129*
Model 7 (Scanning Method) §	0.52 (0.49,0.54)	<0.0001**

[a] OR adjusted for age at 1989 within the combined strata of age at diagnosis(cases) or interview (controls) (<35y; 35-44 y; 45-54 y; 55-64 y; 65-74 y;≥ 75 y) and site (Cook County, DMA). For DMA (ages <35y and 35-44y) combined due to small sample size. Adolescence physical activity median cut point was used.

[b] OR additionally adjusted for age at 1989, BMI, total energy intake in 1985-1989 (quartiles) , family history of breast cancer (yes; no), age at menarche, reproductive history (nulliparous; first full term pregnancy < 22y; first full term pregnancy 22-29 y; first full term pregnancy ≥ 30 y), oral contraceptive use (ever; never), hormone replacement therapy use (ever; never), menopausal status at diagnosis (cases) or interview (controls) (premenopausal; postmenopausal), alcohol intake in 1985-1989 (none; < 0.7 serving/week; ≥ 0.7 serving/week) and migration status in 1985 (Poland; in US < 10y; in US ≥10y). Adult physical activity median cut point was used.

* Bold indicates statistical significance of $p < 0.05$. ** Bold indicates statistical significance of $p < 0.01$.

§ SE and C.I. in scanning method were estimated through Bootstrapping

CHAPTER 5. AVERAGE TREATMENT EFFECT FOR CASE-CONTROL STUDIES UTILIZING TARGET MAXIMUM LIKELIHOOD EATIMATION (TMLE)

5.1. Introduction

Although in case-control studies the sampling is based on the disease status, we are interested in estimating the risk of disease given exposure ($P[D=1 \mid E=e]$) as well as the risk difference $RD = P[D=1 \mid E=1] - P[D=1 \mid E=0]$. When confounding variables of **x** are present the risk difference depends on **x**, $RD(\mathbf{x}) = P[D=1 \mid E=1, \mathbf{x}] - P[D=1 \mid E=0, \mathbf{x}]$. The average risk difference is $E\big(RD(\mathbf{x})\big)$ where the expectation is with respect to the distribution of **x**.

Because *D* is a binary, logistic regression for $\pi(\mathbf{x}, e) = P[D=1 \mid E=e, \mathbf{x}]$ would be the natural choice. However, due the case-control sampling on disease status, the likelihood function constructed from the distributions $f(\mathbf{x}, e \mid D=1, s=1)$ and $f(\mathbf{x}, e \mid D=0, s=1)$ modifies the intercept in the logistic regression model. The indicator $s=1$ is retained to show that selection is made in the in the case-control populations. The intercept term becomes $\beta_0^* + \beta_0 = \log(\tau_1 / \tau_0) + \beta_0$ where $\tau_1 / \tau_0 = P[s=1 \mid D=1] / P[s=1 \mid D=0]$ is the sampling fraction of cases to controls. Unless we have specific information on how the case-control samples were obtained we cannot estimate parameters that depend on β_0^*. The odds ratio can be estimated, but not the relative risk and risk difference.

Recent developments have produced methods for causal inference in case-control studies building upon the vast literature on estimation of marginal causal (treatment) effects in cohort studies. We use some of these techniques in our case-control study (PWHS) to estimate average

risk differences $E\big(RD(\mathbf{x})\big)$ for total daily physical activity during adolescence and adulthood.

We are guided by several key publications, van der Laan , 2008, [56,57] Rose and van der Laan, 2014,[59] and applications by Abdollahpour *et al*, 2021,[61] Almasi-Hashiani *et al*, 2021.[62]

We will use either *Y* or *D* to denote outcome or disease status. These are binary variables. Also, either *E* or *T* will denote exposure status—again as binary variables. Covariates are denoted by **x.**

5.1.1. Brief literature review

van der Laan *et al*, 2008 [48] explore the *Marginal Causal Estimation Theory* for case-control studies. This methodology, which is a non-parametric double robust estimation for marginal causal effects, rely on knowledge of prevalence $P_0(Y = 1) = q_0$, to mimic the bias of the case-control sampling design. Many researchers (Anderson, 1972; Greenland, 1981; Prentice and Breslow, 1978; Morise *et al*, 1996; Wacholder, 1996; Greenland, 2004) [49-53] have discussed using $\log\left(q_0/(1-q_0)\right)$ to update the intercept of the traditional logistic regression model for the case-control study.[49-53] However, others (Robins,1999) and Mansson *et al*, 2007) [54,55] suggested a causal inference method for the case-control study applying propensity score (PS) methods which relies on the mechanism of exposure among control subjects as a weight to update a logistic regression of disease status on exposure. Mansson and colleagues[54] also illustrated that the weighted case-control method creates unbiased estimation and close to those of the cohort method utilizing methods of the PS. Additionally, the inverse probability of treatment weighting (IPTW) by PS has recently been used by many researchers as the double robust causal inference with assumption of no misspecification in the PS model.[20-27,45-48, 54,55] Robins and Mansson [54,55] also discussed the IPTW by PS which is based on only the marginal structure logistic regression model for the case-control study. They indicated that using this method does not require the knowledge of prevalence probability, but prevalence of disease

should be close to zero. They presented the procedure for the estimated PS (exposure mechanism) among control subjects to update a logistic regression of *Y* (binary outcome) on *T* (exposure). They also noted this IPTW estimator is a non-identifiable parameter and has a nonparametric distribution which highly demands the correct specification of the exposure model (PS model). Wooldridge also described three approaches (for cohort study) to estimate treatment effect based on Rubin causal model effect under the assumption of ignorability and overlap (Wooldridge, 2010 Chapter 21) [60]. Author in this book, mentioned these approaches as follow: regression-based methods, propensity score methods and combinations of regression-based methods and propensity score methods.

In this chapter we will discuss the new weighting method which applies the new probability weighting for the case-control study design utilizing the prevalence of disease. This double robust method for the causal effect has been explored by Rose and van der Laan *et al*, 2008 and 2014 [41,48] and illustrated in two recent applications by Abdollahpour *et al*, 2021, [61] Almasi-Hashiani *et al*, 2021. [62] The last two articles describe the data gathering process, in particular how case and controls were sampled from their respective populations.

5.2. Estimations for Case- Control Studies

Rose and van der Laan presented the procedure for the case-control weighted targeted maximum likelihood estimation (WTMLE).[48,56,57,59,63,64] They used "targets" of the parameter of interest instead of the distribution of interest. They defined $G^* = (Y, T, X) \sim P_0^*$ as the unobserved full data with true distribution of interest P_0^* which includes a dichotomous outcome *Y* (case/control status), binary exposure/treatment *T*, and vector of baseline covariates *X*. [48,56] Therefore P_0^* indicated the population from which all cases and controls have been sampled. They also defined several marginal causal effect parameters such as the causal risk difference (RD), the

causal risk ratio (RR) and the causal odds ratio (OR). They also denoted the causal effect parameter $\varphi_0^* = \boldsymbol{\varphi}^*(P_0^*) \in \mathbb{R}^d$ of distribution $P_0^* \in M^*$.[56,57] These marginal causal effects for a binary exposure $T \in \{0,1\}$ are as follows:

$$\varphi_{0,\mathrm{RD}}^* = \mathbb{E}_0^*\{\mathbb{E}_0^*(Y|T=1,X) - \mathbb{E}_0^*(Y|T=0,X)\} \quad (1)$$

$$= \mathbb{E}_0^*\{(Y_1) - \mathbb{E}_0^*(Y_0)\}$$

$$= P_0^*(Y_1=1) - P_0^*(Y_0=1)$$

$$\varphi_{0,\mathrm{RR}}^* = \frac{\mathbb{E}_0^*\{\mathbb{E}_0^*\{(Y|T=1,X)\}\}}{\mathbb{E}_0^*\{\mathbb{E}_0^*\{(Y|T=0,X)\}\}} = \frac{\mathbb{E}_0^*\{(Y_1)}{\mathbb{E}_0^*\{(Y_0)} = \frac{P_0^*(Y_1=1)}{P_0^*(Y_0=1)} \quad (2)$$

$$\varphi_{0,\mathrm{OR}}^* = \frac{P_0^*(Y_1=1)P_0^*(Y_0=0)}{P_0^*(Y_1=0)P_0^*(Y_0=1)} \quad (3)$$

Where Y_0 *and* Y_1 are the counterfactual outcomes for binary exposure T and $(X, T, Y = Y_T)$ as the time-ordered missing data structure on (X, Y_0, Y_1), the full data structure.[56,57] In this research we focus on case-control-weighted TMLE for the RD.

Case-Control Sampling and its Probability Distribution

van der Laan and Rubin (2006) [63] or Moore and van der Laan (2007)[64] suggested that if we assume n observations of $G_1, G_2, \ldots, G_n$ *are IID such that* $G_n \sim P_0^*$, we could apply the locally efficient target MLE of φ_0^*. Moreover, van der Laan and Robins (2002) used the double robust estimation function methodology. Furthermore, van der Laan [57] indicated two types of case-control sampling: independent (un-matched) case-control sampling and matched case-control sampling. Our application to the PHWS is based on unmatched case-control sampling.

Independent Case-Control Sampling

In this sampling, we first sample a case by sampling (X_1, T_1) from the conditional distribution (X, T) given Y=1. Then, subsequently, one samples J controls (X_0^j, T_0^j) from conditional distribution of (X, T) given Y=0 *and* j=1, 2,..,J. This results in the data:

$$G = ((X_1, T_1), (X_0^j, T_0^j : j = 1, \dots, J) \sim P_0 \text{ with}$$

$$(X_1, T_1) \sim (X, T | Y = 1)$$

$$(X_0^j, T_0^j) \sim (X, T | Y = 0)$$

where the sampling distribution of the data structure will be as above P_0. Thus, a case-control data set includes *n* observations $G_1, G_2, \dots, G_n$ with sampling distribution P_0. This cluster includes one case and J controls with marginal distribution of cases and controls as P_0.[57]

The Estimation Problem

van der Laan (2008) also explained the statistical problem for estimating the parameter $\varphi_0 = \boldsymbol{\varphi}^*(P_0^*)$ of the population distribution $P_0^* \in M^*$ of (X, T, Y), known to be an element of some specified model M^* based on the case-control data set $G_1, G_2, \dots, G_n \sim q_0$. Model M^* regardless of knowledge of prevalence (q_0) generated models for marginal distribution of case (X_1, T_1) and controls (X_0^j, T_0^j) $j = 1, \dots, J$.[57]

Known or Sensitivity Analysis Parameters/Weights

Additionally, van der Laan *et al* 2008,[57] defined:

$q_0 \equiv P_0^*(Y = 1)$ and $q_0(\delta|\mathrm{M}) \equiv P_0^*(Y = \delta|\mathrm{M})$

as the marginal probability of being a case, and the conditional probability of being a case/control, conditional on the matching variable. Furthermore, he defined:

$$\bar{q}_0(M) \equiv q_0 \frac{P_0^*(Y_1 = 0|\mathrm{M})}{P_0^*(Y_1 = 1|\mathrm{M})} = q_0 \frac{q_0(0|M)}{q_0(1|M)}$$

He also denotes that $\bar{q}_0(M)$ will be determined by q_0 and $q_0(1|M) = P_0^*(Y_1 = 1|\mathrm{M})$; and $\mathbb{E}_0(\bar{q}_0(M)) = 1 - q_0$. Hence, q_0 for the unmatched case-control study and $\bar{q}_0(M)$ for matched case-control study, will be used to weight cases and controls to obtain a valid estimation procedure.[57] It is assumed that q_0 and $\bar{q}_0(M)$ are known respectively for unmatched and

matched case-control study. [57]

In our study, we focus on independent case-control sampling (unmatched).

Definition (Case-control weighted function)

Given a $D^*(G^*)$=$D^*(X,T,Y)$ we define the case-control weighted version D^* as

$$D_{q_0}(G) \equiv q_0 D^*(M_1, X_1, T_1, 1) - \frac{1}{J}\sum_{j=1}^{J} \bar{q}_0(M_1)\, D^*(M_1, X_2^j, T_2^j, 0)$$

Where in the special case of unmatched case-control Design, we have $\bar{q}_0(M_1)$=1 − q_0

Theorem: Unbiased estimating function mapping

Let $D^*(G^*)$=$D^*(X,T,Y)$ be a function so that $P_0^* D^* = \mathbb{E}_{P_0^*}^* D^*(G^*) = 0$. Then in an unmatched case-control study $D_{q_0} \equiv q_0 D^*(X_1, T_1, 1) + (1-q_0)\frac{1}{J}\sum_{j=1}^{J} D^*(X_2^j, T_2^j)$ satisfies $P_0 D_{q_0} = 0$.

In more generality, for any function D^* and corresponding case control weighted function D_{q_0} we have $P_0 D_{q_0} = P_0^* D^*$.

(The proof is available in the article Estimation Based on Case-Control Designs with Known Incidence Probability by van der Laan). [57]

5.3. Case-Control weighting of estimation procedures developed for Prospective Sampling

Sherri Rose and Mark van der Laan *et al,* 2014,[48] explored TMLE method for estimation of causal effects in 6 steps. They presented an example of case-control-weighted TMLE for the marginal risk difference. In this chapter we will follow their methods in 6 steps using Polish Women Health Study (PWHS) which is an unmatched case control study.

Let define G= (Y, T, X) ~ P_0 as unobserved full data experimental unit G with binary outcome Y, binary exposure T, and vector of covariate X, and the true underlying distribution of interest

P_0. Define RD as:

$$\text{RD}= E_{\boldsymbol{x},0}(E_{G,0}(Y|T=1,\mathbf{x}) - E_{G,0}(Y|T=0,\mathbf{x}))$$

Suppose the observed data are from an unmatched case-control study with N_1 cases from the conditional distribution of (X, T) given $Y = 1$ and N_0 controls from the conditional distribution of (X, T) given $Y = 0$. Denote $J = N_0/N_1$ which will be used in the case-control weights as the average number of controls per case. For this data structure, the procedure for calculating the case-control-weighted TMLE will be as follow:

STEP1: Assign the weights such that q_0 (prevalence) to the cases and $(1 - q_0)/J$ to the controls. We use these weights in the case-control-weighted TMLE procedure for all steps.

STEP2: Estimate the conditional outcome Y=1 given exposure and covariates $P_{G,n}(Y|T,X)$. We may use a case-control-weighted parametric logistic regression or any procedure that allows for weighted observations. We use the logistic procedure in SAS which incorporates the weights from STEP1.

STEP3: Estimate the probability of exposure T given covariates X, $P_{G,n}(T|X)$, using case-control weighted logistic regression. We use the logistic procedure in SAS similar to STEP2. This step gives us propensity scores (PS).

STEP4: Calculate subject-specific weights $\boldsymbol{H_i}$ denoted by

$$H_i = \frac{I_i(T=1)}{P_{G,n,i}(T=1|X)} - \frac{I_i(T=0)}{1-P_{G,n,i}(T=0|X)}$$

where $\boldsymbol{H_i}$ is regarded as a covariate. The form of this covariate depends on the type of parameter being estimated. We focus on the RD parameter.

STEP5: Update the initial fit achieved in STEP2 by holding the coefficients of $P_{G,n}(Y|T,X)$ fixed, while estimating a coefficient ε for H (T, X) using case-control-weighted maximum

likelihood estimation in the following sub-model:

$$P_{G,n}^{update}[Y = 1|T, X] = \text{expit}\left(\log\left(\frac{P_{G,n}(Y|T,X)}{1-P_{G,n}(Y|T,X)}\right) + \varepsilon H(T,X)\right)$$

STEP6: Estimate the parameter given in equation 1 by plugging the updated estimate of $E_{G,n}(Y|T=1,x)$ and $E_{G,n}(Y|T=0,x)$

Let $\hat{Y}_{i1}$ and $\hat{Y}_{i0}$ denote the updated probability outcome from STEP6 assuming *all subjects are exposed,* or *all subjects are not exposed* respectively. This would be expressed as follows:

$$\hat{Y}_{i1} = \frac{\exp(\mathbf{x}_i'\hat{\beta} + \hat{\gamma} + \hat{\varepsilon}H_i)}{1+\exp(\mathbf{x}_i'\hat{\beta} + \hat{\gamma} + \hat{\varepsilon}H_i)} \quad \text{and} \quad \hat{Y}_{i0} = \frac{\exp(\mathbf{x}_i'\hat{\beta} + \hat{\varepsilon}H_i)}{1+\exp(\mathbf{x}_i'\hat{\beta} + \hat{\varepsilon}H_i)}$$

Then the estimated averages for exposed and unexposed are:

$$\hat{\mu}_1 = n^{-1}\sum_{i=1}^{n}\hat{Y}_{i1} \quad \text{and} \quad \hat{\mu}_0 = n^{-1}\sum_{i=1}^{n}\hat{Y}_{i0}.$$

The *RD estimate* (*ATE*) *will be then calculated as* $\hat{\mu}_1 - \hat{\mu}_0$.

5.4. Estimating ATE in a Cohort Study

Wooldridge (2010, Chapter 21)[60] provides extensive details on three methods to estimate ATE based on Rubin causal model (RCM) in a cohort study. We outline only the regression adjustment method and regression adjustment with propensity score weighting.

i) Regression Adjustment

The observed data are a random sample $\{(Y_i, T_i, \mathbf{x}_i): 1 \le i \le n\}$ on independent units for outcome, treatment indicator and covariate vector. The estimation strategy for ATE is based on the regression functions $m_0(x)$= 𝔼 (y| x, t=0) and $m_1(x)$= 𝔼 (y| x, t=1) and subsequently estimate the parameter τ_{ATE} =𝔼 $[m_1(x) - m_0(x)]$. To estimate τ_{ATE} we proceed as follows:

(1) Use the subsample $\{(Y_i, T_i = 1, \mathbf{x}_i): 1 \le i \le n_1\}$ of treated to estimate the parameter δ_1 in the

logistic regression model $P[Y_i = 1 \mid T_i = 1, \mathbf{x}_i] = G(\mathbf{x}_i'\boldsymbol{\delta}_1)$. Score the entire data set *assuming all subjects are treated* to get the predicted $G(\mathbf{x}_i'\hat{\boldsymbol{\delta}}_1)$.

(2) Use the subsample $\{(Y_i, T_i = 0, \mathbf{x}_i) : 1 \le i \le n_0\}$ of untreated to estimate the parameter $\boldsymbol{\delta}_0$ in the logistic regression model $P[Y_i = 1 \mid T_i = 0, \mathbf{x}_i] = G(\mathbf{x}_i'\boldsymbol{\delta}_0)$. Score the entire data set *assuming all subjects are untreated* to get the predicted $G(\mathbf{x}_i'\hat{\boldsymbol{\delta}}_0)$.

Then τ_{ATE} is estimated by $\hat{\tau}_{ATE,reg} = n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\hat{\boldsymbol{\delta}}_1) - G(\mathbf{x}_i'\hat{\boldsymbol{\delta}}_0)\right)$.

Applying the Uniform Weak Law of Large Numbers shows that $\hat{\tau}_{ATE,reg}$ is a consistent estimator of τ_{ATE}. The asymptotic normality of $\sqrt{n}(\hat{\tau}_{ATE,reg} - \tau_{ATE})$ is established by applying the Central Limit Theorem, and a consistent estimator of the asymptotic variance can be derived. The derivations are provided in Wooldridge, 2010, Chapter 21 using the large sample properties of the MLE $\hat{\boldsymbol{\delta}}_1$, $\hat{\boldsymbol{\delta}}_0$.

ii) Regression Adjustment with Propensity Score Weighting

Wooldridge describes the steps to obtain the double-robust estimator of τ_{ATE}. The steps are as follows:

(1) Estimate the propensity score model $e(\mathbf{x}) = P[T = 1 \mid \mathbf{x}] = G(\mathbf{x}'\hat{\gamma})$ using the data $\{(T_i, \mathbf{x}_i) : 1 \le i \le n\}$ and obtain the propensity scores $\{\hat{e}(\mathbf{x}_i) = G(\mathbf{x}_i'\hat{\gamma}) : 1 \le i \le n\}$.

(2) Use the subsample $\{(Y_i, T_i = 1, \mathbf{x}_i) : 1 \le i \le n_1\}$ of treated to estimate the parameter $\boldsymbol{\delta}_1$ in the *weighted* logistic regression model $P[Y_i = 1 \mid T_i = 1, \mathbf{x}_i] = G(\mathbf{x}_i'\boldsymbol{\delta}_1)$. The MLE $\hat{\boldsymbol{\delta}}_1$ from this weighted regression optimizes the log-likelihood

$$\sum_{i=1}^{n}\left\{\frac{[T_i = 1]}{\hat{e}(\mathbf{x}_i)}\left(Y_i \log G(\mathbf{x}_i'\delta_1) + (1 - Y_i)\log(1 - G(\mathbf{x}_i'\delta_1))\right)\right\}.$$

Score the entire data set *assuming all subjects are treated* to get the predicted $G(\mathbf{x}_i'\hat{\boldsymbol{\delta}}_1)$.

(3) Use the subsample $\{(Y_i, T_i = 0, \mathbf{x}_i) : 1 \le i \le n_0\}$ of untreated to estimate the parameter $\boldsymbol{\delta}_0$ in the *weighted* logistic regression model $P[Y_i = 1 \mid T_i = 0, \mathbf{x}_i] = G(\mathbf{x}_i'\boldsymbol{\delta}_0)$. The MLE $\hat{\boldsymbol{\delta}}_0$ from this weighted regression optimizes the log-likelihood

$$\sum_{i=1}^{n} \left\{ \frac{[T_i = 0]}{1 - \hat{e}(\mathbf{x}_i)} \left(Y_i \log G(\mathbf{x}_i'\delta_0) + (1 - Y_i) \log(1 - G(\mathbf{x}_i'\delta_0)) \right) \right\}.$$

Score the entire data set *assuming all subjects are untreated* to get the predicted $G(\mathbf{x}_i'\hat{\boldsymbol{\delta}}_0)$.

Then τ_{ATE} is estimated by $\hat{\tau}_{ATE,pswreg} = n^{-1} \sum_{i=1}^{n} \left(G(\mathbf{x}_i'\hat{\boldsymbol{\delta}}_1) - G(\mathbf{x}_i'\hat{\boldsymbol{\delta}}_0) \right)$.

The large sample properties of $\hat{\tau}_{ATE,pswreg}$ are derived from Uniform Weak Law of Large Numbers and Central Limit Theorem. We note that $\hat{\tau}_{ATE,pswreg}$ depends on the MLE ($\hat{\boldsymbol{\delta}}_1$, $\hat{\boldsymbol{\delta}}_0$, $\hat{\gamma}$). It has the double robustness property in the sense that either logit model for the PS or the logit models for outcome needs correct specification for get consistency of $\hat{\tau}_{ATE,pswreg}$.

5.5. Properties of the ATE Estimator in Case-Control Studies

In Section 5.3 we described the 6 steps to obtain TMLE method for estimation of ATE by the method of Rose and van der Laan *et al,* 2014. [59] At the end of Step 6, we have the estimate of ATE $\hat{\mu}_1 - \hat{\mu}_0 = n^{-1} \sum_{i=1}^{n} \left(G(\mathbf{x}_i'\hat{\beta} + \hat{\gamma} + \hat{\varepsilon} H_i) - G(\mathbf{x}_i'\hat{\beta} + \hat{\varepsilon} H_i) \right)$.

Consistency follows from applying the Uniform Weak Law of Large Numbers (WLLN). We see that $\hat{\mu}_1 - \hat{\mu}_0$ converges in probability to $\mu_1 - \mu_0$ where:

ATE= $E\left(G(\mathbf{x}_i'\beta + \gamma + \varepsilon H_i)\right) - E\left(G(\mathbf{x}_i'\beta + \varepsilon H_i)\right) \equiv \mu_1 - \mu_0$. The expectation is with respect to the distribution of $(\mathbf{x}_i, H_i)$.

Next, we want the asymptotic distribution of $\hat{\mu}_1 - \hat{\mu}_0$, and in particular the asymptotic variance. This could be derived formally from a similar derivation of the ATE estimator that combines both regression adjustment and propensity score weighting (Wooldridge, 2010 Chapter 21)[60] as in section 5.4 Our estimator $\hat{\mu}_1 - \hat{\mu}_0$ has exactly the same functional form as the ATE estimator of Wooldridge but the latter was established for cohort studies.

After STEP6, we may use the bootstrap to get the variance (and standard error) of $\hat{\mu}_1 - \hat{\mu}_0$.

Rose and van der Laan (2008)[56] provide another approach to inference based on the case-control weighted $\hat{\mu}_1 - \hat{\mu}_0$ which is using the influence function (IC):

$IC(Y_i, T_i, \mathbf{x}_i) \equiv H_i\left(Y_i - \pi_i^{(update)}\right) + \left(\hat{Y}_{i1} - \hat{Y}_{i0}\right) - \widehat{ATE}$ where $\widehat{ATE} = \hat{\mu}_1 - \hat{\mu}_0$.

Form the cluster of one case with *J* randomly selected controls. The weighted IC is:

$$IC_w(i) \equiv q_0 IC(1, T_i, \mathbf{x}_i) + \frac{(1-q_0)}{J}\sum_{j=1}^{J} IC(0, T_{ij}, \mathbf{x}_{ij}) \qquad (1)$$

The subscript (*i*) now indexes the case with data $(Y_i = 1, T_i, \mathbf{x}_i)$ and $(Y_{ij} = 0, T_{ij}, \mathbf{x}_{ij}) : 1 \le j \le J$ are the data for the *J* controls of that case. The sample variance of $\{IC_w(i) : 1 \le i \le n_1\}$ is

$S_{IC}{}^2 = n_1^{-1}\sum_{i=1}^{n_1}\left(IC_w(i) - \overline{IC_w}\right)^2$ where $\overline{IC_w}$ is the sample mean of $\{IC_w(i) : 1 \le i \le n_1\}$. Note that we must use n_1 --the sample size of the number of clusters. The recommend estimator of standard error of $\hat{\mu}_1 - \hat{\mu}_0$ is $S / \sqrt{n_1}$.

A classical approach to deriving the distribution of $\hat{\mu}_1 - \hat{\mu}_0$ would start from

$\sqrt{n}\left(\hat{\mu}_1 - \hat{\mu}_0 - (\mu_1 - \mu_0)\right) = \sqrt{n}\left((\hat{\mu}_1 - \mu_1) - (\hat{\mu}_0 - \mu_0)\right)$ and the two terms

$\sqrt{n}\left(\hat{\mu}_1 - \mu_1\right) = \sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\hat{\beta} + \hat{\gamma} + \hat{\varepsilon}H_i) - G(\mathbf{x}_i'\beta + \gamma + \varepsilon H_i)\right)\right) + \sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\beta + \gamma + \varepsilon H_i) - \mu_1\right)\right)$

and

$$\sqrt{n}\left(\hat{\mu}_0-\mu_0\right)=\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\hat{\beta}+\hat{\varepsilon}H_i)-G(\mathbf{x}_i'\beta+\varepsilon H_i)\right)\right)+\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\beta+\varepsilon H_i)-\mu_0\right)\right)$$

If we ignore the variation in $(\hat{\beta},\hat{\varepsilon})$ the difference between the two terms is

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\beta+\gamma+\varepsilon H_i)-\mu_1\right)\right)-\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\beta+\varepsilon H_i)-\mu_0\right)\right)$$

$$=\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\beta+\gamma+\varepsilon H_i)-G(\mathbf{x}_i'\beta+\varepsilon H_i)\right)-\left(\mu_1-\mu_0\right)\right).$$

Applying the Central Limit Theorem, we get asymptotic normality mean zero and asymptotic variance $\sigma^2\equiv E\left\{\left(G(\mathbf{x}_i'\beta+\gamma+\varepsilon H_i)-G(\mathbf{x}_i'\beta+\varepsilon H_i)\right)-\left(\mu_1-\mu_0\right)\right\}^2$. We then estimate σ^2 by $\hat{\sigma}^2=\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\hat{\beta}+\hat{\gamma}+\hat{\varepsilon}H_i)-G(\mathbf{x}_i'\hat{\beta}+\hat{\varepsilon}H_i)\right)-\left(\hat{\mu}_1-\hat{\mu}_0\right)\right)^2$ by simply plugging in the estimators ($\hat{\beta},\hat{\varepsilon}$, $\hat{\mu}_1$ $\hat{\mu}_0$). This gives the standard error of $\hat{\mu}_1-\hat{\mu}_0$ as $\sqrt{\hat{\sigma}^2/n}$.

However, a complete analysis cannot ignore the variation in $(\hat{\beta},\hat{\varepsilon})$. We should examine the distribution of

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\hat{\beta}+\hat{\gamma}+\hat{\varepsilon}H_i)-G(\mathbf{x}_i'\beta+\gamma+\varepsilon H_i)\right)\right)-\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\left(G(\mathbf{x}_i'\hat{\beta}+\hat{\varepsilon}H_i)-G(\mathbf{x}_i'\beta+\varepsilon H_i)\right)\right)$$

incorporating the asymptotic variance (matrix) **V** of $\hat{\beta},\hat{\gamma},\hat{\varepsilon}$. Doing the complete analysis is a daunting exercise. An easier approach is to use the bootstrap resampling to get the estimate the standard error of ATE estimator ($\hat{\mu}_1-\hat{\mu}_0$).

5.6. The Causal Effect and Impact of Physical Activity on Breast Cancer Risk Using TMSE for a Case-Control Study in US (PWHS)

In this section we used TMLE method to estimate RD for our case-control study (PWHS) using 6-steps, planning to estimate the double robust causal effect of physical activity on BC risk. The framework of TMLE is sufficient for both observational and RCT (Rose *et al, 2014 and*

2017).[40,41] We used weighted maximum likelihood estimation for our case-control study since we cannot estimate the RD from common regression analysis because the intercept term will have the selection probability and disease prevalence. Therefore, if the prevalence of disease is known we could calculate the weights and apply them to our study to estimate the intercept and mimic the bias. However, Abdollahpour *et al*, 2021 [61] and Almasi-Hashiani *et al*, 2021 [62] presented a new approach for the STEP1. They obtained weights for cases and controls such as to simulate a cohort study. Steps are as follow:

STEP1: We first estimated the population of Polish immigrants to US for both sites (Detroit and Chicago). Approximately 70,000 Polish immigrants to US were residing in Chicago in late 1990's, of which 45% were women (31,000). From 31,000 about 90% were women aged between 20 to 79, therefore about 28000 immigrant women aged 20-79 were residing in Chicago. Number of controls for Chicago Metropolis was 215 in our study therefore weight for controls will be approximately 130 (28000/215).

In Detroit, about 10,000 Polish immigrants to USA were residing of whom 4500 (45%) were women, and approximately 4000 were women aged between 20 to 79 therefore weight for controls in Detroit will be 58.8 (4000/68).

For each site, cases were identified by respective cancer registry. For Chicago, at Illinois State Cancer Registry 3,341 white BC cases were screened for place of birth; 266 cases were identified as being Polish born (6.5%). Additionally for 1,008 cancer cases the registry was unable to determine their eligibility. Therefore, we assumed that potentially 6.5% of those for whom eligibility was undetermined were potentially Polish born. Therefore, we estimated the total potentially Polish born cases for Chicago area to be

266+(1008*0.065) =266+65.52=331.52 or approximately 332 cases. Interviews were completed

with 116 out of 266 cases. The dataset for these analyses consists of 101 cases since 15 cases had to be eliminated for incomplete data. Therefore, our weight for the Chicago cases was calculated as follows: 332 /101=3.28 needed to be adjusted by a fraction of 1.15 calculated as 116/101=1.15. Therefore, the final weight for the Chicago cases would be 3.77.

For Detroit Metropolitan area, the screening process was more involved since place of birth was not available for almost 90% of white cases. Therefore 20,721 white BC cases were screened for place of birth; 62 cases were identified as being Polish born (0.03%). Additionally for 3,065 cancer cases the registry was unable to determine their eligibility. Therefore, we assumed that potentially 0.03% of those for whom eligibility was undetermined were potentially Polish born. Therefore, we estimated the total potentially Polish born cases for the Detroit area to be 62+(3,065*0.003) =62+9=71 or approximately 71 cases. Interviews were completed with 36 out of 62 cases. The dataset for these analyses consists of 27 cases since 9 cases had to be eliminated for incomplete data. Therefore, our weight for the Detroit cases was calculated as follows: 71 /27=2.63 which needed to be adjusted by a fraction of 1.33 calculated as 36/27=1.33. Therefore, the final weight for the Detroit cases would be 3.5.

Finally, the case and control weights described above were assigned to the cases and controls for each site, in order to simulate a cohort study and using case-control MLE.

STEP2: We estimated the conditional outcome distribution $P_{G,n}(Y|T,X)$ by a logistic regression model (in SAS) that incorporates the weights in STEP1, where Y is dichotomous denoted Y=0 no BC disease (control) and Y=1 BC disease (case), exposure (T) is define as the total adolescent physical activity(PA) (Table 5.7.1) and total adulthood physical activity (PA) (Table 5.7.2).

STEP3: Estimate the conditional exposure (PA) distribution $P_{G,n}(T|X)$ using again a logistic regression model that incorporates the weights from STEP1. This is similar to getting the

propensity scores $\{\hat{e}_i : 1 \le i \le N\}$.

STEP4: Subject-specific weights are defined by:

$H_i = \frac{I_i(T=0)}{\hat{e}_i} - \frac{I_i(T=0)}{1-\hat{e}_i}$ where $I_i(T=0)$ is 0 for untreated (low level of total PA), and 1 for treated (high level of total PA). This cut-points are as follow: for total adolescent PA > 47 (MET h/day) for treated, total adolescent PA ≤ 47 (MET h/day for untread and for total adult PA > 54 (MET h/day) for treated, total adult PA ≤ 54 (MET h/day) for untreated as we explained them in chapter 4 section5.

Here H_i is used for the RD parameter (ATE).

STEP5: Update the estimated conditional outcome distribution $P[Y=1 \mid T, \mathbf{x}]$ in STEP2 as follows: from the logistic regression in STEP2 we have $\pi(\mathbf{x},t) = \frac{\exp(\mathbf{x}'\beta + \gamma t)}{1+\exp(\mathbf{x}'\beta + \gamma t)}$ we get the estimated $\mathbf{x}'\hat{\beta} + \hat{\gamma}t$. The updated logistic regression will be defined as:

logit($\pi(\mathbf{x},t)$ /(1- $\pi(\mathbf{x},t)$))= $(\mathbf{x}_i'\hat{\beta} + \hat{\gamma}t_i) + \varepsilon H_i$ where ε is a parameter to be estimated.

This updated model has no intercept, $\hat{\beta}, \hat{\gamma}$ are held fixed and the weights in STEP1 are used.

Therefore, the updated probability is:

$\pi^{(update)}(\mathbf{x}_i, t_i) = \frac{\exp(\mathbf{x}_i'\hat{\beta} + \hat{\gamma}t_i + \hat{\varepsilon}H_i)}{1+\exp(\mathbf{x}_i'\hat{\beta} + \hat{\gamma}t_i + \hat{\varepsilon}H_i)}$ where $H_i = \frac{t_i}{\hat{e}_i} - \frac{(1-t_i)}{(1-\hat{e}_i)}$

STEP6: Let $\hat{Y}_{i1}$ denote the updated probability outcome from STEP5 assuming *all subjects are exposed*. This would be $\hat{Y}_{i1} = \frac{\exp(\mathbf{x}_i'\hat{\beta} + \hat{\gamma} + \hat{\varepsilon}H_i)}{1+\exp(\mathbf{x}_i'\hat{\beta} + \hat{\gamma} + \hat{\varepsilon}H_i)}$. Get the average as

$\hat{\mu}_1 = n^{-1}\sum_{i=1}^{N} \hat{Y}_{i1}$. Similarly define $\hat{Y}_{i0}$ as the updated probability outcome from STEP5 assuming

all subjects are unexposed which will be defined as: $\hat{Y}_{i0} = \frac{\exp(\mathbf{x}_i'\hat{\beta} + \hat{\varepsilon}H_i)}{1+\exp(\mathbf{x}_i'\hat{\beta} + \hat{\varepsilon}H_i)}$. Get the average as: $\hat{\mu}_0 = n^{-1}\sum_{i=1}^{N}\hat{Y}_{i0}$.

The ***RD Estimate (ATE)*** *using TMLE will be then calculated by* $\hat{\mu}_1 - \hat{\mu}_0$.

5.7. Estimation of Standard Error for ATE

For estimation of SE_{ATE} we used two methodologies as follows:

1) We obtained asymptotic standard error from:

$$S_{ATE}{}^2 = (n-1)^{-1}\sum_{i=1}^{n}\left(ATE_{Rose,weight}(i) - \overline{ATE_{Rose,weight}}\right)^2$$

2) We estimate $SE_{\mathbf{ATE}}$ by Bootstrap Resampling Approach for N=1000 bootstrap samples.

5.8. Results for Double Robust Causal Effect of Physical Activity on BC risk Using CCE-TMLE for Case Study (PWHS)

We examine the causal effect of total daily PA during adolescence or adulthood, using CCW-TMLE, on BC risk (Table 5.1 and 5.2). The observed estimate of ATE on BC risk, for high total daily adolescent PA compared with low adolescent PA using CCW-TMLE was -0.0090010. Using two different techniques we obtained the values for estimated $SE_{\mathbf{ATE}}$ as follows: 0.000280649, 0.000029465 respectively for the first approach and bootstrap resampling approach. Using bootstrap resampling method (N=1000), we obtained lower estimated $SE_{\mathbf{ATE}}$, however, ATE was statistically significant using either estimated values of $SE_{\mathbf{ATE}}$ (P-value<0.0001). Our observed ATE indicated a reduction in BC risk for higher- level total daily adolescent PA compared to lower-level adolescent PA (Table 5.1).

Using CCW-TMLE, we assessed the causal effect of total adult daily PA on BC risk. We again found the reduction of BC risk for the higher-level of total adult daily PA compared to lower-

level adult PA. The estimated ATE was -0.0090644 with estimated $SE_{\mathbf{ATE}}$ 0.000232381 and 0.000016792 for the two approaches respectively (Table 5.2).

Table 5.1 Results for Causal Effect of Total Daily Adolescent PA on BC Risk in Polish Immigrant Women to US Using Case-Control Weighted Target Maximum Likelihood CCW-TMLE

Method	ATE	SE	95% C.I.	P-Value
$\hat{\tau}_{ATE,pswreg}$ [a]	-0.0090010	0.000280649	(-0.0095526, -0.0084493)	<0.0001
$Bootstrap$ [b]	-0.0090010	0.000029465	(-0.009058751, -0.00894239)	<0.0001

[a] Asymptotic Approach
[b] Bootstrap Resampling Approach (N=1000)

Table 5.2 Results for Causal Effect of Total Daily Adult PA on BC Risk in Polish Immigrant Women to US Using Case-Control Weighted Target Maximum Likelihood CCW-TMLE

Method	ATE	SE	95% C.I.	P-Value
$\hat{\tau}_{ATE,pswreg}$ [a]	-0.0090644	0.000232381	(-0.0095212, -0.0086076)	<0.0001
$Bootstrap$ [b]	-0.0090644	0.000016792	(-0.009097312, -0.009031488)	<0.0001

[a] Asymptotic Approach
[b] Bootstrap Resampling Approach (N=1000)

5.9. Simulation Study

For evaluating of physical activity (PA) on BC risk using a case-control-weighted TMLE methodology, we generated a single dataset which represent BC disease in population with estimated prevalence of $q_0 = 0.00125$. Hence, we applied the age-adjusted incidence rate for the

white women in US based on SEER registry for estimation of q_0 (125 per 100,000)[3]. In this simulation we generated 1,000,000 target population with prevalence 0.00125 with exposure and covariates as described below:

Age is one of the risk factors for developing BC. Some other risk factors are family history, menstrual and reproductive history, inherited genetic mutation and menopause. We also assumed 15% of women with breast cancer have a first family member with BC. In addition, studies show that the alcohol daily drinker, have about 10% increase of developing BC compared to female non-drinkers. Researchers also reported that considering overall PA, women who get regular exercise have a 10%-20% lower risk of BC relative to women who do not exercise regularly. Furthermore, an estimated of 2.1 million BC cases were diagnosed in women worldwide in 2018 of which 11.4% were premenopausal cases and 88.6% were postmenopausal BC cases per 100,000, therefore we generated menopausal status (yes/no) with probability of 0.9. We simulated the truncated normal for age distribution (year) with mean of 55.74 and standard error of 13.35, truncated on the interval [29, 81]. Average age at Menarche (year) was simulated with a normal distribution with 13.8 and standard error of 1.75. All binary covariates were simulated by a Bernoulli (p=0.5).

We then selected the sample size of 750 with 250 cases and 500 controls (1:2) from 1,000,000 (target population) to be close to our case study (PWHS). We conducted this simulation study to evaluate the causal RD of total daily adult PA on BC risk, utilizing case-control-weighted TMLE. [59] We then followed all the steps which has been described in 5.3 with using the prevalence of 0.00125 in all steps in Rose *et al, 2014*.[40,]

The model that we used contains age (x1), age at menarche(x2) , family history of BC(x3), alcohol drinker(x4), menopausal status (x5) and treatment(PA) with coefficients

α_0=log ($\frac{0.00125}{1-0.00125}$) , $\alpha_1 = 0.09$, $\alpha_2 = -0.03$, $\alpha_3 = 0.15$, $\alpha_4 = 0.1$ and β =-0.45. Our goal was to generate a case-control (1:2) study with prevalence of breast cancer 0.00125 and some of established risk factors which were in our study that were associated with BC risk. We defined coefficient of exposure (β), approximately as the PA's coefficient with defined above risk factors with BC, to generate a desired risk difference of approximately -0.1. This gives us the below model:

$$LogitProb(Y = 1|X_i = x_i\ ,\ \mathrm{PA_i} = \mathrm{t_i}) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + \alpha_5 x_{i5} + \beta\ \mathrm{trt_i}$$

Then we calculated the estimated of ATE based on one selected sample as well as its SE, using TMLE.

We then applied the methodology on bootstrap resampling with iteration of 1000 using all STEPS which incorporates the weights of prevalence (0.00125) to estimate the empirical standard error, 95% of empirical CI and P-value from above model.

Results from Simulation Study

The fitted model of simulation our target population study (N=1,000,000) with prevalence of disease =0.00125 was as follows:

$$\begin{aligned} LogitProb(Y = 1|X_i = x_i\ ,\ \mathrm{PA_i} = \mathrm{t_i}) \\ = -6.6568 + 0.1096 \times age - 0.0298 \times \mathrm{Menarche} + 0.1601 \times \mathrm{FH} \\ + 0.1021 \times \mathrm{Alcohol} - 0.9049 \times \mathrm{Menopause} - 0.4524 \times \mathrm{Physical\ Activity} \end{aligned}$$

The estimated prevalence of BC disease in our target population study was 1.283×10^{-3} which is close to the population prevalence (1.125×10^{-3}). Based on this target population, we were interested to estimate the causal effect of total daily adult PA on BC risk using CCW- TMLE.

We then selected one sample (n=750, $n_{ca} = 250$, n_{co}=500) from the target population with fitted model:

$$LogitProb(Y = 1|X_i = x_i\ ,\ \mathrm{PA_i} = \mathrm{t_i})$$

$$= -6.597 + 0.115 \times age - 0.05 \times \text{ Menarche} + 0.276 \times \text{ FH} + 0.437 \times \text{ Alcohol}$$

$$- 0.763 \times \text{Menopause} - 0.41 \times \text{Physical Activity}$$

We then applied weighted case-control sampling for cases and controls as follows: 0.00125 and (1-0.0015)/j where j=2, respectively for cases and controls for all steps in section 5.3. Using CCW- TMLE we observed the causal ATE of total daily PA on BC risk -0.000665 with SE=0.000051299. We also used bootstrap resampling (N=1000) from our selected sample in which n=750, $n_{ca} = 250$, n_{co}=500 (using surveyselect in SAS) to estimate double robust SE for ATE using 6 steps. SE was 0.000001617 using CCW- TMLE. The empirical fitted model is then estimated as follows:

$$LogitProb(Y = 1|X_i = x_i\ ,\ \mathrm{PA_i} = \mathrm{t_i})$$

$$= -6.27310 + 0.11 \times age - 0.03 \times \text{ Menarche} + 0.17 \times \text{ FH} + 0.11 \times \text{ Alcohol}$$

$$- 0.92 \times \text{Menopause} - 0.45533 \times \text{Physical Activity}$$

Table 5.3 Simulation Study's Results for the Causal Effect of Total Daily Adult PA on BC Risk in a Simulated Polish immigrant Women in US Using Case-Control Weighted Maximum Likelihood Estimation CCW-TMLE

Method	ATE	SE	95% C.I.	P-Value
$\hat{\tau}_{ATE,\,pswreg}$ [a]	-0.000665	0.000051299	(-0.0007659, -0.0005645)	<0.0001
$Bootstrap$ [b]	-0.000665	0.000001617	(-0.000668168, -0.000661832)	<0.0001

[a] Asymptotic Approach
[b] Bootstrap Resampling (N=1000)

5.10. Conclusions/Discussions

In this chapter we have focused on ATE (RD) estimated by CCW-TMLE, for its simplicity of interpretation of the treatment/exposure effect of disease risk which was developed by Rose (2014). Using this method of estimation of ATE, we observed a significant reduction in BC risk for individuals with high total daily PA, a relative to individuals with low PA. We also conducted a simulation study for a population of size 1,000,000 where we assumed the prevalence of BC to be 0.00125. In the simulation study we used estimates of the effect of PA on BC and effect of our included covariates in the model to be similar to those observed in PWHS. The applied method to estimate ATE in a case-control study, as described by Rose, worked very well, and provided reliable results for the effect of PA on BC risk.

CHAPTER 6: DISCUSSIONS/CONCLUSIONS

We introduced a motivation example and research goals as well as our case study in chapter 1. In chapter 2 we described the Robin counterfactual for observational study. We also defined PS, methods, and its application in chapter 2.

In chapter 3, based on simulation studies, we compared ATE estimation using the stratification PS methods with the newly proposed, scanning method with continuous thresholding. Because the ATE is usually unknown to the analyst, continuous threshold works well especially as sample size increased from 200 to 600. The percent relative difference between true ATE and estimate from the scanning method was only 0.7% with sample size of 600. This observed difference between true ATE and an estimated one was lower than the other observed differences based on other stratification PS methods. The coverage rate for 95% CI, also performed best as sample increases to 600. Therefore, in chapter 3, we illustrated that the newly proposed scanning method improved the power of the test to detect difference between treated and untreated groups, compared to stratification PS methods.

In chapter 4, in the first part, we evaluate association between total daily PA, categorized as low, medium, and high, during adolescence and adulthood and BC risk applying the traditional case-control logistic regression analyses which provides us with estimated OR's. Based on the results in this section, cases had lower levels of total daily physical activity for each percentile of distribution and significantly lower mean total daily physical activity compared to controls both for the adolescent (P-value<0.05) and for the adult(P-value<0.01).

For evaluating the association of total daily physical activity (MET h/Day) and case status, we used conditional logistic regression within joint strata of age and site (Model 1-unadjusted, Table 4.3), as well as multivariate model adjusting for all potential risk factors (Model 2, Table 4.3),

independently for total daily adolescent physical activity and total daily adulthood physical activity. We observed that ORs did not change substantially between Model 1 (unadjusted) and Model 2(multivariate adjusted), for both total daily PA in adolescence and adulthood. The results show that women in the highest level of PA during adolescence (greater than 55.7 METs-h/Day), have a significant 45% reduction in BC risk (OR = 0.55) and women in the highest level of PA during adulthood (greater than 59.6 METs-h/Day) have a significant 47% reduction in BC risk (OR=0.53) (Table4.3).

Table 4.4 provides ORs for the joint effect of total daily PA during adolescence and adulthood, adjusted for potential risk factors. The result in this table shows that for *all women* the high total daily adolescent PA reduces BC risk jointly with medium and high levels of total daily PA in adulthood, however, reaches statistical significance only for the high adolescence and high adulthood category (OR=0.29 and 95% CI :0.11-0.77).

Also, in Table 4.4 we presented the results for the evaluating effect of the joint PA during adolescence and adulthood by *menopause status*, adjusted for all potential risk factors as well. For the *premenopausal women*, we observed that high levels of adolescent PA were protective for BC irrespective of the level of PA in adulthood and OR's estimates were statistically significant. For medium adolescent total daily PA, the estimated OR was statistically significant only for high level of PA in adulthood (OR=0.10 with 95% CI :0.013-0.68).

For the *postmenopausal women*, the high adolescent PA also provides reduction in BC risk for all levels of adult PA, however, these reductions were not statistically significant. The observed lack of significance in the high adolescent/high adult and medium adolescent/high adult physical activity in *postmenopausal women* were most likely due to the decreased sample size.

Therefore, our findings in chapter 4, first part, assessing the association between PA and BC risk,

supported the hypothesis that increased total daily PA during adolescence and adulthood, decreases BC risk in *women* especially for *premenopausal women*. Although the estimates of ORs for *postmenopausal women* were in the direction of protective effect of high adolescence PA for BC, none of the OR's reached statistical significance. In the second part of chapter 4, our goal was to evaluate the causal effect, measured in terms of OR's, of total daily PA during adolescence and adulthood on BC risk by using the PS methods which have been introduced in chapter 2. Additionally, we were interested in estimating these causal effects using the newly proposed scanning method also in terms of estimated OR, and compared it with the estimated form common regression methods and PS methods. To estimate PS, which was used in these analyses, we chose to divide daily total PA at the median rather than three levels. Additionally, we were interested in evaluating the impact of the choice of covariates to be included in the models on the estimates of OR's using the traditional PS and the scanning method. For the comparison purposes we estimated PS when the variable age1989 was/was not included in the PS model. We then estimated ORs for 7 models, models 1-3 denoted 3 common regression analyses (unadjusted, adjusted for combined strata of site and age, multivariate adjusted for potential risk factors), Models 4-6 denoted common PS methods analyses (IPW, covariates adjustment, stratification-quintiles) and finally model 7 denoted the newly scanning method (Table 4.5-4.8).

When age1989 was not included as a covariate in the models, based on the traditional analyses, *total daily adolescent physical activity* reduces breast cancer risk by approximately 40 % (Table 4.5). Based on PS analyses, the estimates of our odds ratios are similar in magnitude and of similar statistical significance (P-value<0.05) to the traditional analyses. However, the newly proposed scanning method provided the shortest confidence interval with the similar point

estimate as the other PS methods.

Table 4.6 provides estimates of *total daily adult physical activity* on BC risk using PS methods and scanning method again, without age1989 included in the 7 models. Using multivariate analysis (model 3) and PS methods (models 4-6), total daily adult PA had approximately 45% reduction on BC risk. The scanning method (model 7) provided similar estimate of approximately 49% reduction in risk, with the shortest 95 % CI.

Subsequently, Age1989 was added to our models to assess the potential impact of covariate misclassification on estimation of causal effect using developed standard PS methodologies and our newly developed scanning method.

Based on the traditional and PS analyses, while *including the age1989* in the model, several of the estimates of ORs for the effect of *total daily adolescent PA* did not reach statistical significance, although they still provided estimates indicating reduction in risk (Table 4.7). However, the estimate obtained from scanning method remained approximately the same as when age1989 was not in the model and was statistically significant.

In table 4.8, we provide estimates of *total daily adult physical activity* on BC risk using PS methods and scanning method (*with age 1989* in the 7 models). Estimates of ORs for the total daily adult PA were similar to those obtained when the model was run without age1989. The scanning method continued to provide the shortest confidence interval.

The *comparison* of the estimates from models *with and without age1989*, points to the dilemma that researchers often face which covariate should be included in the model. This comparison allows us to point out that *the scanning method* is *least sensitive* to the *misclassification* of covariate in estimation of causal effect using OR's.

However, interpretation of the causal effects in terms of odds ratios is more complex than

interpretation of causal effect in terms of risk difference (ATE). This motivated us to search if any methodologies have been developed to estimate ATE (RD) in the context of case-control study. Although the processes to estimate ATE have often been discussed in the literatures in the context of cohort study, the first process describing estimation of ATE, in a case-control study (TMLE) was proposed by *Rose et al,2014and 2017* [41] which under certain assumptions allows us to estimate the causal risk difference (ATE) for a case-control study. [41] In their 2014 publication(*Rose and van der Lann*) [41], they proposed a case-control weighted Maximum Likelihood Estimation (CCW-TMLE) which is a double robust approach for assessing the causal effect of ATE in case-control studies. Subsequently this method was used by *Almasi-Hashiani et al, 2021and Abdollahpour et al, 2021* where they modified the weights to be used in STEP1 of Rose's proposed CCW- TMLE. We followed *Abdollahpour et al, 2021* processes for weights' estimation.

In chapter 5, we examined the causal effect of *total daily PA during adolescence* and *adulthood*, using *CCW-TMLE*, on BC risk (Table 5.1 and 5.2). The observed estimate of ATE on BC, for high total daily adolescent PA compared with low adolescent PA using CCW-TMLE was -0.0090010 and was statistically significant (P-value<0.0001). Our observed ATE indicated a reduction in BC risk for higher- level total daily adolescent PA compared to lower-level adolescent PA (Table 5.1).Similarly, we observed reduction of BC risk for the higher-level of total adult daily PA compared to lower-level adult PA. The estimated ATE was -0.0090644 and was statistically significant (P-value<0.000) (Table 5.2).

We also conducted a simulation study for a population of size 1,000,000 where we assumed the prevalence of BC to be 0.00125. Our goal was to select β (exposure's coefficient) to generate a desired risk of approximately -0.1 to be similar to our case study (PWHS). Similarly, we chose

coefficients for the effects of specific covariates on BC risk to be similar to those observed in the PWHS or other sources for BC risk. We then applied the method described by Rose, to estimate ATE in a simulated case-control study. We then selected one sample from the simulated target population of 1,000.000 subjects, to estimate the ATE as proposed by Rose and van der Laan, in their 6 steps CCW-TMLE approach. To estimate SE of ATE, we chose bootstrap resampling (N=1000). We again applied 6 steps Rose (CCW-TMLE) for these 1000 bootstrap resampling to estimate empirical standard error for ATE (Table 5.3). Using CCW-TMLE, ATE was observed -0.000665 with double robust SE: 0.000001617 (95% CI -0.000668168, -0.000661832). Based on simulation study we observed a significant reduction of PA on BC risk (P-value<0.0001).

Strengths

We developed a new method "continuous *threshold* "for estimation the causal effect in observational studies. We proposed the novel and flexible stratification approach which uses all available information in the propensity score which improves the power of test for evaluating the average treatment effect. Existing methods potentially may not provide us with the true estimate of causal effect, especially when the sample size is small, and we are dealing with many confounders. Therefore, in chapter 3, we demonstrated by Monte Carlo simulation study, that the newly proposed scanning method improved the power of the test to detect difference between treated and untreated groups, compared to stratification PS methods, especially when the sample size was increasing.

We also obtained the results using the PWHS, a case-control study, for the scanning method and comparing it with three common logistic regression (Models 1 to 3, tables 4.1-4.4) and three PS methods (modes 4 to 6, tables 4.1-4.4). The newly proposed scanning method provided the shortest confidence interval with the similar point estimate as the other PS methods in all 4 tables

(tables 4.1-4.4). The comparison of the estimates from models with and without age1989, illustrates the new scanning methos is robust to misclassification of covariate that should be included in the model. This comparison allows us to point out that *the scanning method is least sensitive to the misclassification* of covariate in *estimation of causal effect* using OR's.

To our knowledge our case study (PWHS) is the first research study assessing total daily PA by creating summation of inactive (such as sleeping and sitting), as well as active times (such as recreational and household physical activity), and occupational, in contrast to many publications on the effect of recreational PA on BC risk. Assessment of PA in the PWHS questionnaire captured data from two separate time periods: during adolescence (ages 12-13) and adulthood (years between 1985-1989). Therefore, we were able to assess the effect of total daily PA during two different time periods (adolescence and adulthood) as well as the joint effect between PA in the two separate life periods. The questionnaire also captured a wide range of non-physical activity information including most, if not all, other BC risk factors. By having access to information on all these BC risk factors allowed us to account and adjust for them, thus giving us a more accurate estimation of total daily adolescent and total daily adult physical activity's true effect on breast cancer risk.

Wooldridge (2010, Chapter 21)[60] addressed that combining common regression adjustment method with PS methods allow us to attain some robustness because of misspecification in the regression model or PS model. The author also mentioned that the estimator is a doubly robust if only one of the models correctly be specified not both. Since CCW-TMLE is a special case of combination of common regression analysis with PS method (IPW), therefore if only one of the models (regression model or PS model) be correctly specified we would gain some robustness of misspecification in the regression model or PS model.

Limitations

Analytical form of the asymptotic variance of α (ATE) using continuous threshold, is not trivial due to dependency of α to s, across various values of s. However, existing formulas for standard error, require that each cut, be independent of each other. Thus, we did not have any parametric estimation for standard error, so we used the bootstrap resampling approach to estimate the variance of $\hat{\alpha}(s)$ which involves repeated resampling with replacement of the given data, a large number of times, and using the sampling variance based on these replicates, to estimate the underlying variance of each $\hat{\alpha}(s)$. Although, not having a closed form formula for estimation of standard error is a limitation for this method, we were able to overcome this limitation by using of bootstrap resampling to estimate standard error of ATE for each cut-point ($\hat{S}_{\hat{\alpha}(s)}$).

The other limitation for this research was that our case study was a case-control study. There exists a large body of literature on estimating the causal effect of treatment in term of ATE for cohort studies, the literature for estimating ATE in a case-control study is very limited. There exists only one process proposed *Sherri Rose and Mark van der Laan et al 2014,* which we did apply to our study in chapter 5. As a future direction we propose to compare the scanning method with the process by Rose as well. However, we used the Monte Carlo simulation for comparing the scanning method with other PS-Stratification models for estimating causal effect (OR) in the context of a cohort study.

Application of CCW-TMLE approach as described by Rose & van der Laan, requires the knowledge of BC prevalence in population. However, two publication *Ibrahim Abdollahpour, et al 2021 & Amir Almasi-Hashiani*, *et al 2021*, modified *Rose and van der Laan's* approach by recreating the population that gave rise to the cases and controls in their studies. Through this process they were able to recreate weights for cases and controls (Rose and van der *Laan*

STEP1) that were subsequently used for all additional 5 steps proposed by them.

We estimated the resident Polish born population in two study areas from late 1990's to estimate the weights for controls in our study. To estimate the weights for the cases we used information provided to us by the two Cancer Registries (Illinois State Cancer Registry and, SEER Registry at Karmanos Cancer Center), who identified all Polish born Cancer cases that were part of our study. With that information we were able to reconstruct the total population of Polish born BC cases during the time period of our study for two study areas. We then applied our calculated weights for cases and controls in STEP1 in our estimation of ATE for causal effect of PA on BC risk. Although we didn't have the prevalence of BC in Polish born women for these two studies areas, we were able to overcome that limitation by using our calculated weights.

For simulation study in chapter 5, we also needed to simulate the target population based on prevalence of BC for Polish immigrant women in US. Since that information was not available to us, we used the incidence of BC for white women in US, provided by SEER.

Conclusions

We introduced an approach which does not rely on arbitrary discretization but uses continuum of all available information in the propensity score to improve the power to evaluate the average treatment effect. Based on evidence of Monte Carlo simulation study, we showed that the newly continuous threshold increases the power of test compared to PS stratification method (quintiles and median). It is also providing closer estimation of the causal effect to the true value (tables 3.1-3.3). Because the true value of ATE is usually unknown to the analyst, continuous threshold works well especially when sample size increases from 200 to 600. Consequently, using the scanning method, we get the lower MSE and better coverage rate.

When the scanning method was applied for the case study, PWHS, it was robust to the

misclassification of the PS model when one covariate was not included in the model. Furthermore, using CCW-TMLE, we could estimate ATE in terms of RD instead of OR's which is easy to interpret.

In conclusion, our results enhance the current findings in the literature about the effect of total daily PA during adolescent and adulthood on BC risk.

This analysis suggests that there should be more emphases on increasing the level of PA in girls under the age of 18. In addition, to encouraging high level of adolescent PA, maintenance of higher levels of PA in adulthood should be of equal importance to gain the largest benefit from PA through lifetime on BC risk reduction.

BIBLIOGRAPHY

1. GLOBONCAN Cancer Facts Sheets: All Cancers. http://globocan.iarc.fr/factsheets/cancers/all.asp

2. World Cancer Funds International https://www.wcrf.org/cancer-trends/worldwide-cancer-data/

3. National Cancer Institution. **Surveillance, Epidemiology, and End Results (SEER)** https://seer.cancer.gov/statfacts/html/breast.html

4. American Cancer Society https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html

5. American Cancer Society https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

6. GLOBONCAN Cancer Facts: Incidence, Breast, Female https://gco.iarc.fr/overtime/en/dataviz/bars?sexes=2&sort_by=value0&cancers=14&years=2012

7. GLOBONCAN Cancer Facts: Mortality, Breast, Female https://gco.iarc.fr/overtime/en/dataviz/bars?sexes=2&sort_by=value0&cancers=14&years=2012&age_end=17&types=1

8. Jerzy Staszewski, William Haenszel.*"* Cancer Mortality Among the Polish-Born in the United States". JNCI: Journal of the National Cancer Institute, Volume 35, Issue 2, August 1965, Pages 291–297, https://doi.org/10.1093/jnci/35.2.291

9. Jerzy Staszewski, M G Mccall & N S Stenhouse. "Cancer Mortality in 1962-66 Among Polish Migrants to Australia". British Journal of Cancer volume 25, 1971,pages 599–610.

10. Adelstein, A., Staszewski, J. & Muir, C. "Cancer mortality in 1970-1972 among Polish-born migrants to England and Wales". Br J Cancer 40, 464–475 (1979). https://doi.org/10.1038/bjc.1979.202

11. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P. "Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families". The Breast Cancer Linkage Consortium. Am J Hum Genet. 1998 Mar; 62(3):676-89.

12. K McPherson, C M Steel, J M Dixon. "ABC of Breast. Breast cancer-epidemiology, risk factors, and genetics". BMJ VLUME 321 9 Sep 2000.

13. Colditz GA, Frazier AL.” Models of breast cancer show that risk is set by events of early life: prevention efforts must shift focus”. Cancer Epidemiol Biomarkers Prev. 1995 Jul-Aug; 4(5):567-71.

14. Trygve Lofterød, Hanne Frydenberg, Vidar Flote, Anne Elise Eggen, Anne McTiernan, Elin S. Mortensen, Lars A. Akslen, Jon B. Reitan[1] Tom Wilsgaard, and Inger Thune *“*Exploring the effect of lifestyle on breast cancer risk, age at diagnosis, and survival: the EBBA-Life study”, 2020; 182(1):215-227

15. J R C Sainsbury, T J Anderson, D A L Morgan. "ABC of breast diseases Breast cancer". BMJ VOLUME 321 23 SEPTEMBER 2000 bmj.com 745.

16. Thomas DB, Karagas MR. (Cancer in first- and second-generation Americans”. Cancer Res. 1987 Nov 1; 47(21):5771-6.

17. Kurpinski, Jerzi . ” Migration and mental health - a comparative study”, 2007:49.57

18. Cho E, Spiegelman D, Hunter DJ, Chen WY, Zhang SM, Colditz *GA, Willett WC.* “Premenopausal intakes of vitamins A, C, and E, folate, and carotenoids, and risk of breast cancer”. Cancer Epidemiol Biomarkers Prev. 2003 Aug; 12(8):713-20.

19. Shadish, W. R. Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. Psychological Methods, 2010; 15(1), 3–17. https://doi.org/10.1037/a0015916.

20. Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Pages 399-424 | Received 02 Mar 2011, Published online: 09 Jun 2011.https://doi.org/10.1080/00273171.2011.568786.

21. KiEisuKE HIRANO, GUIDO W. IMBENS, AND GEERT RIDDER. Efficient of average treatment effects using the estimated propensity score. 2003, July; Econometrica, No 4, Vol. 71, 1161-1189.

22. Lunceford JK , Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine, 01 Oct 2004, 23(19):2937-2960.

23. Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika*;* Vol. 70, No. 1 (Apr 1983), pp. 41-55 (15 pages).

24. Paul R. Rosenbaum and Donald B. Rubin. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician: *Vol. 39, No. 1 (Feb. 1985), pp. 33-38 (6 pages).*

25. Takeshi Emura, Jingfang Wang and Hitomi Katsuyama. Assessing the Assumption of Strongly Ignorable Treatment Assignment Under Assumed Causal Models. April 23, 2008.

26. Stephen R. Cole1 and Miguel A. Herna´n. Constructing Inverse Probability Weights for Marginal Structural Models. American Journal of Epidemiology, Volume 168, Issue 6, 15 September 2008, Pages 656–664, https://doi.org/10.1093/aje/kwn164

27. Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stu¨rmer, M. Alan Brookhart, and Marie Davidian. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology, Volume 173, Issue 7, 1 April 2011, Pages 761–767, https://doi.org/10.1093/aje/kwq439

28. William G. Cochran and Donald B. Rubin. Controlling Bias in Observational Studies: A Review. Sankhyā: The Indian Journal of Statistics, 1973 Series A, Vol. 35, No. 4, 417-446 (30 pages).

29. Peter C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. 2009 Nov 10; 28(25): 3083-107. doi: 10.1002/sim.3697

30. D B Rubin, N Thomas. Matching using estimated propensity scores: relating theory to practice. Biometrics, 1996 Mar;52(1):249-64.

31. Sherry Weitzen, Kate L Lapane, Alicia Y Toledano, Anne L Hume, Vincent Mor. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf, 2004 Dec;13(12):841-53. doi: 10.1002/pds.969.

32. Til Stürmer, MD, MPH, Manisha Joshi,, MSW Robert J. Glynn, Jerry Avorn, Kenneth J. Rothman, and Sebastian Schneeweiss. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006 May; 59(5): 437–447.

33. Paul R. Rosenbaum and Donald B. Rubin. Bias in Observational Studies Using Sub-Classification on the Propensity Score. September 1984, Journal of the American Statistical Association 79(387).

34. Cle´mence Leyrat, Shaun R Seaman, Ian R White, Ian Douglas, Liam Smeeth, Joseph Kim, Matthieu Resche-RigonJames R Carpenter, and Elizabeth J Williamson. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Statistical Methods in Medical Research 2019, Vol. 28(1) 3–19.

35. Wenji Guo, Georgina K. Fensom, Gillian K. Reeves & Timothy J. Key. Physical activity and breast cancer risk: results from the UK Biobank prospective cohort. British Journal of Cancer volume 122, pages726–732 (2020).

36. Anne McTiernan, Christine M. Friedenreich Peter T. Katzmarzyk, Kenneth E. Powell, Richard Macko, David Buchner, Linda S. Pescatello, Bonny Bloodgood, Bethany Tennant, Alison Vaux-Bjerke, Stephanie M. George, Richard P. Troiano, and Katrina L. Piercy. Physical Activity Guidelines Advisory Committee Physical Activity in Cancer Prevention and Survival: A Systematic Review. Med Sci Sports Exerc. 2019 June; 51(6): 1252–1261. doi:10.1249/MSS.1937.

37. Monninkhof EM, Elias SG, Vlems FA, van der Tweel I, Schuit AJ, Voskuil DW. 'Physical activity and breast cancer: a systematic review". Epidemiology. 2007 Jan; 18(1):137-57.

38. Regan A. Howard, Michael F. Leitzmann, Martha S. Linet, and D. Michal Freedman. Physical Activity and Breast Cancer Risk among Pre- and Postmenopausal Women in the U.S. Radiologic Technologists Cohort. Cancer Causes Control. 2009 Apr; 20(3): 323–333.

39. B E Ainsworth, W L Haskell, M C Whitt, M L Irwin, A M Swartz, S J Strath, W L O'Brien, D R Bassett Jr, K H Schmitz, P O Emplaincourt, D R Jacobs Jr, A S Leon. Compendium of physical activities: an update of activity codes and MET intensities. Med Sci Sports Exerc 2000 Sep;32(9 Suppl):S498-504. doi: 10.1097/00005768-200009001-00009.

40. BASSETT, DAVID R.; VACHON, JOHN A.; KIRKLAND, ARISTOTLE O.; HOWLEY, EDWARD T.; DUNCAN, GLEN E.; JOHNSON, KELLY R. Energy cost of stair climbing and descending on the college alumnus questionnaire. Medicine & Science in Sports & Exercise 29(9):p 1250-1254, September 1997.

41. Megan S. Schuler, Sherri Rose. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. American Journal of Epidemiology, Volume 185, Issue 1, 1 January 2017, Pages 65–73, https://doi.org/10.1093/aje/kww165

42. Sherri Rose and Mark van der Laan. A Double Robust Approach to Causal Effects in Case-Control Studies. American Journal of Epidemiology, Vol. 179, No. 6 DOI: 10.1093/aje/kwt318. Advance Access publication January 31, 2014.

43. Peter C. Austin & James Stafford. The Performance of Two Data-Generation Processes for Data with Specified Marginal Treatment Odds Ratios. Published online: 22 May 2008. https://doi.org/10.1080/03610910801942430

44. M. H. GAIL, S. WIEAND, S. PIANTADOSI. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariate. Biometrika, Volume 71, Issue 3, December 1984, Pages 431–444, https://doi.org/10.1093/biomet/71.3.431

45. Peter C. Austin. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. Published online 27 January 2010 in Wiley Interscience. DOI: 10.1002/sim.3854.

46. Peter C. Austin. Data-Generation Process for Data with Specified Risk Differences or Numbers Needed to Treat. Communications in Statistics—Simulation and Computation®, 39: 563–577, 2010.

47. Mia S. Tackney, Tim Morris, Ian White, Clemence Leyrat, Karla Diaz-Ordaz, and Elizabeth Williamson. A comparison of covariate adjustment approaches under model misspecification in individually randomized trials. Tackney *et al*. Trials (2023) 24:14. https://doi.org/10.1186/s13063-022-06967-6

48. Sherri Rose and Mark J. van der Laan. Simple Optimal Weighting of Cases and Controls in Case-Control Studies. The International Journal of Biostatistics Volume 4, Issue 1 2008 Article 19.

49. Anderson JA. Separate sample logistic discrimination. Biometrika, 1972;59:19–35.

50. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. Am J Epidemiol. 2004;160(4):301–305.

51. Prentice RL, Breslow NE. Retrospective studies and failure time models. Biometrika. 1978;65(1):153–158

52. Morise AP, Diamon GA, Detrano R, Bobbio M, Gunel Erdogan. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. Med Decis Making. 1996;16:133–142

53. Wacholder S. The case-control study as data missing by design: Estimating risk differences. Epidemiology. 1996;7(2):144–150

54. Robins JM. [choice as an alternative to control in observational studies]: Comment. Statistical Science. 1999;14(3):281–293.

55. Mansson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. Am J Epidemiol. 2007;166(3):332–339

56. Mark J. van der Laan. Estimation Based on Case-Control Designs with Known Prevalence Probability. The International Journal of Biostatistics. Volume4, Issue, 2008, Article17.

57. Mark J. van der Laan. Estimation Based on Case-Control Designs with Known Incidence Probability. University of California, Berkeley U.C. Berkeley Division of Biostatistics Working Paper Series. Year 2008. Paper 234.

58. Stephen C Newman. Causal analysis of case-control data. Epidemiol Perspect Innov. 2006; 3: 2. Published online 2006 Jan 27. doi: 10.1186/1742-5573-3-2.

59. Sherri Rose and Mark van der Laan. A Double Robust Approach to Causal Effects in Case-Control Studies. Am J Epidemiol. 2014 Mar 15; 179(6): 663–669.

60. Jeffrey M Wooldridge. Book: "Econometric Analysis of Cross Section and Panel Data" (second edition)

61. Ibrahim Abdollahpour, Saharnaz Nedjat, Amir Almasi-Hashiani, Maryam Nazemipour, Mohammad Ali Mansournia, and Miguel Angel Luque-Fernandez. Estimating the Marginal Causal Effect and Potential Impact of Waterpipe Smoking on Risk of Multiple Sclerosis Using the Targeted Maximum Likelihood Estimation Method: A Large, Population-Based Incident Case-Control Study. American Journal of Epidemiology, Volume 190, Issue 7, July 2021, Pages 1332–1340, https://doi.org/10.1093/aje/kwab036

62. Amir Almasi-Hashiani, Saharnaz Nedjat, Reza Ghiasvand, Saeid Safiri, Maryam Nazemipour Nasrin Mansournia and Mohammad Ali Mansournia. The causal effect and impact of reproductive factors on breast cancer using super learner and targeted maximum likelihood estimation: a case-control study in Fars Province, Iran. Almasi-Hashiani et al. BMC Public Health (2021).Published: 24 June 2021. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-11307-5

63. Kelly L. MooreMark J. van der Laan. Targeted Maximum Likelihood Estimation. University of California, Berkeley U.C. Berkeley Division of Biostatistics Working Paper Series: Year 2006, Paper 213.

64. Kelly L. MooreMark J. van der Laan. Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation. University of California, Berkeley U.C. Berkeley Division of Biostatistics Working Paper Series: Year 2007, Paper 215.

APPENDIX A: ETHICAL CONSIDERATIONS

The study was reviewed and approved by the Institutional Review Board (IRB) at Michigan State University. The approval was renewed in until 1/2024.

APPENDIX B: PHYSICAL ACTIVITY QUESTIONNAIRE FROM POLISH WOMEN'S HEALTH STUDY

I'll start with asking questions about gym and sports during your school years.

H1. When you were 12 and 13 years old, about how many hours per day or per week did you spend…(READ FIRST.NEXT ACTIVITY):

a. …participating in gym classes as part of your school program?

____ Hours per
a. Day b. Week c. Month d. N/A *(circle one)*

b. …participating in sports such as basketball, volleyball, soccer, or swimming as part of a competitive team in or outside of school? Please include time spent at practices.

____ Hours per
a. Day b. Week c. Month d. N/A *(circle one)*

Now, I'd like to ask you about recreational activities and transportation...

H2. When you were 12 and 13 years old, about how many hours per day or per week did you spend… (READ FIRST.NEXT ACTIVITY) Then ask: and during 1985- 1989?

a. (SHOW CARD H-1) … doing activities while you were sitting or reclining, such as eating, watching television, reading, playing cards, sewing or knitting, or just doing nothing? Please include the time you spent sitting in class.

____ Hours per
a. Day b. Week c. Month d. N/A *(circle one)*

b. (SHOW CARD H-2) … participating in recreational sports, such as softball, soccer, volleyball, skating, swimming, calisthenics, running or jogging, skiing or cross-country skiing. Please do not include time spent training for a competitive sports team or doing aerobics.

____ Hours per
a. Day b. Week c. Month d. Year e. N/A *(circle one)*

c. … walking to get to places such as school, work, or shopping, or for recreation or exercise, including walking with persons, pets?

____ Hours per
a. Day b. Week c. Month d. N/A *(circle one)*

d. … bicycling to get to places such as school, work, or shopping, or for recreation. Please include time you spent exercising on a stationary bike.

____ Hours per

a. Day b. Week c. Month d. N/A *(circle one)*

e. … dancing or doing aerobic exercise?

____ Hours per

a. Day b. Week c. Month d. N/A *(circle one)*

Next, I will ask you to estimate the number of stairs you walked up on an average day, week, or month. You can give me you answer as the number of stairs climbed in a day, week, or month, or as a number of floors climbed in a day, week, or month. One floor is about 20 steps. Please remember to include stairs in your home, at school, at work, or other places such as where you shop.

H3. When you were 12 and 13 years old, about how many hours per day or per week did you spend…(READ FIRST.NEXT ACTIVITY) Then ask: and during 1985- 1989?

a. Please estimate the number of stairs or floors that you walked up on an average day, week, or month in you home, at, work, or other places.

____ Stairs or Floors (*circle one*) per

a. Day b. Week c. Month d. N/A *(circle one)*

Next, I'd like to ask you about outdoor activities related to unpaid garden or farm work done after regular working hours or another paying job. About outdoor actives related to jobs that you were paid for and outdoor activities related to work on your own or family farm. I will ask later.

H4. When you were 12 and 13 years old, about how many hours per day or per week did you spend… (READ FIRST.NEXT ACTIVITY) Then ask: and during 1985- 1989?

a. (SHOW CARD H-3) … moderate and heavy chores such as gardening, sweeping, raking, mowing, digging, planting, weeding, shoveling, chopping wood, milking cows, animal pens, feeding and caring for large animals (cow, pigs, horses, etc.)?

____ Hours per

a. Day b. Week c. Month d. Year e. N/A *(circle one)*

Next, I'd like to ask you about household activities. Again, please do not include household activities that you may have done as part of a job for which you were paid.

H5. When you were 12 and 13 years old, about how many hours per day or per week did you spend…(READ FIRST.NEXT ACTIVITY) Then ask: and during 1985- 1989?

a. …light chores such as cooking, cleaning, washing dishes, making beds, sweeping or vacuuming, mopping, doing laundry (by machine) m and light shopping or standing in lines?

____ Hours per

a. Day b. Week c. Month d. N/A *(circle one)*

b. …moderate or heavy chores such as heavy cleaning, scrubbing, hand washing clothes with a washboard or using an impeller-type machine, making home repairs, and heavy shopping like caring and lifting groceries?

____ Hours per
a. Day b. Week c. Month d. N/A *(circle one)*

c. …moderate or heavy chores related to children, others, or small pets such as bathing, dressing , carrying children, and active play with children or pets?

____ Hours per
a. Day b. Week c. Month d. N/A *(circle one)*

H6. When you were 12 and 13 years old, about how many hours per day or per week did you spend… (READ FIRST.NEXT ACTIVITY) Then ask: and during 1985- 1989?

a. …sleeping? Please include naps taken during the day.

____ Hours per Day

Now I want to ask you just a few questions about your work relate physical activity. Work includes any part-time or full-time jobs, any self-employment or work for a family business, jobs in the military or on you own family farm, and any jobs you may have held during World War II.

H7. …how many months or years were you employed?

____ Months or Years (*circle one*)

H8. During those months (NUMBER OF MONTHS OR YEARS FROM H7), about how many hours per week did you usually work?

_____ Hours per Week

Now I'd like to know what percentage or how many hours of you time you spent in each of these five kinds of actives. The total should add up to 100% or the total number of hours you worked each week.

H9. When you were 12 and 13 years old, about how many hours per day or per week did you spend… (READ FIRST.NEXT ACTIVITY) Then ask: and during 1985- 1989?

a. …sitting?

____ Hours or Percent (*circle one*)

b. …standing?

____ Hours or Percent (*circle one*)

c. …walking with no lifting?

____ Hours or Percent (*circle one*)

d. …walking with some lifting (less than 11.5 kg or less than 25 lbs.)?

____ Hours or Percent (*circle one*)

e. …doing heavy physical work?

____ Hours or Percent (*circle one*)

H10. TOTAL: SHOULD EQUAL 100% OR HOURS PER WEEK FROM H8

____ Hours or Percent (*circle one*)

APPENDIX C: ADOLESCENT BODY SIZE FIGURES USED IN THE POLISH WOMEN'S HEALTH STUDY

Participants who are between 12 and 13 years old, please report closest body size and shape from a series of nine pictures: Figure C.1. in below, instead of reporting your height and weight.

Figure C.1. Adolescent Body Size Figures used in the Polish Women's Health Study