Evaluating the causal Effects of Hormone Replacement Therapy on Osteoporosis in Postmenopausal Women

Mei Dong, Sean Xinyang Feng, Lin Yu

Abstract

Osteoporosis (OP) is common among postmenopausal women, characterized by decreased bone mass and increased fracture risk. Hormone replacement therapy (HRT) is often prescribed to mitigate bone loss due to estrogen deficiency. This study aims to evaluate the causal effect of HRT on the occurrence of OP in postmenopausal women using data from 7,143 postmenopausal women from the CanPath Student Dataset. Outcome regression, inverse-probability treatment weighting, and targeted maximum likelihood estimation were employed. Arthritis was used as a negative control outcome to assess the robustness of the findings. The analysis showed no significant causal effect of HRT on the occurrence of OP. The finding suggests that HRT may not significantly reduce the risk of OP in postmenopausal women. Further research with comprehensive data collection and considering HRT treatment duration is necessary to validate these results and better understand the potential benefits and risks of HRT in postmenopausal women.

1 Introduction

Menopause is a natural physiological phenomenon resulting from the natural decline in ovarian function (Ji and Yu, 2015). The World Health Organization (WHO) defines natural menopause as 12 continuous months of amenorrhoea not caused by physiological or pathological factors. The average age of menopause is 51 years in industrialized countries and 48 years in poorer regions (Rogers et al., 2002).

Osteoporosis (OP) is a systemic skeletal disorder characterized by decreased bone mass and deterioration of bone tissue, leading to bone fragility and a higher risk of fractures (Sandhu and Hampson, 2011). Sex hormone is a key factor that regulates the process of bone remodelling, which is necessary for bones to maintain a healthy level of bone density. Lack of estrogen among postmenopausal women can result in bone loss (Ji and Yu, 2015). Thus, the occurrence of OP is common among postmenopausal women. Other factors, including smoking status, body mass index (BMI) and family history of OP, also contribute to the progression of bone mass recession (Bonjour et al., 2009; Steinman and Shibli-Rahhal, 2019).

Estrogen replacement, most commonly in the form of hormone replacement therapy (HRT), is recommended for postmenopausal women to prevent bone loss caused by estrogen deficiency (Wells et al., 2002; Torgerson and Bell-Syer, 2001; Cartwright et al., 2016; Crofton et al., 2010). HRT has been shown to preserve bone density at all skeletal sites and significantly reduce the risk of fracture (Gambacciani and Levancini, 2014). Cartwright et al. (2016) found that HRT significantly increased the bone density at the lumbar spine among women in a randomized clinical trial. The Global Consensus Statement on Menopausal Hormone Therapy explicitly states that HRT is an effective and appropriate approach for the prevention of osteoporosis-related fractures in at-risk women before the age of 60 years

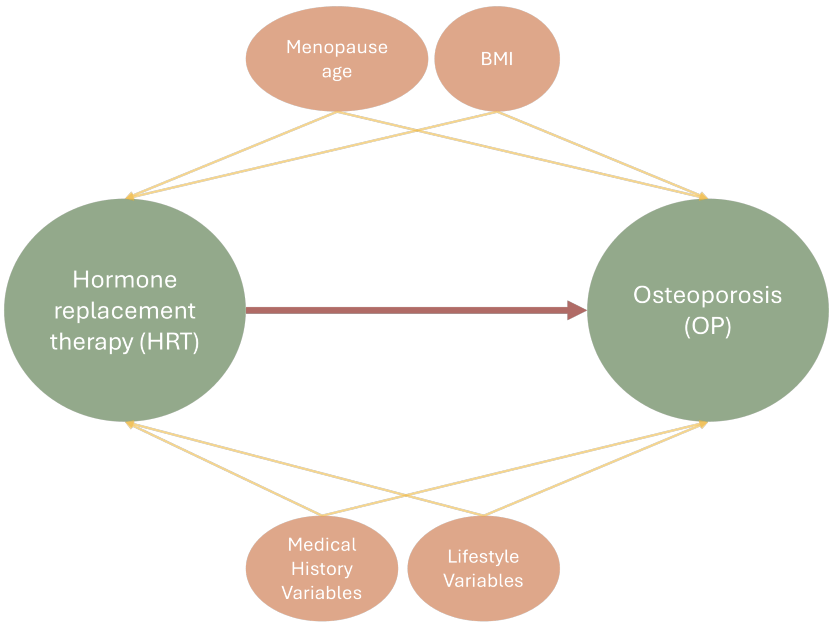

Figure 1: Directed acyclic graph showing the hypothesized causal relationship between HRT and OP. The confounding variables include Menopause age, BMI, medical history variables, and lifestyle variables. Medical history variables: hysterectomy, oophorectomy, osteoporosis family history; Lifestyle variables: smoking, alcohol, activity.

or within 10 years after menopause (De Villiers et al., 2013).

The objective of this study is to evaluate the causal effect of HRT on the occurrence of OP among postmenopausal women, aiming to demonstrate that HRT is beneficial in preventing OP. We hypothesize that women who receive HRT will have a lower risk of developing OP compared to those who do not receive HRT. Figure 1 illustrates the directed acyclic graph (DAG) showing the hypothesized causal relationship between HRT and OP. The confounding variables may have an effect on both HRT and OP, including menopause age, body mass index (BMI), medical history variables (hysterectomy, oophorectomy, osteoporosis family history) and lifestyle variables (smoking, alcohol, activity). We will use three different models, outcome regression, inverse-probability treatment weighting (IPTW), and targeted maximum likelihood estimation (TMLE) to estimate the average causal effect of HRT on OP while adjusting for the confounding variables in Figure 1.

2 Methodology

2.1 Study samples and variables

This study utilized data from the pan-Canadian data platform (CanPath; https://canpath.ca) Student Dataset. We initially selected 15,285 postmenopausal women, defined as those who have experienced menopause. The selection process is illustrated in Figure 2. To ensure the inclusion of samples where HRT occurred before the onset of OP, we excluded samples with missing data on menopause age, HRT age, or age of first OP occurrence. For women who received HRT, we excluded those whose menopause age was less than their HRT age and those whose first OP occurrence was before receiving HRT. For women without HRT, we excluded those whose first OP occurrence was before menopause occurrence. The same criteria were applied for the negative control outcome, arthritis. This process resulted in a final sample size of 7,143 women for analysis.

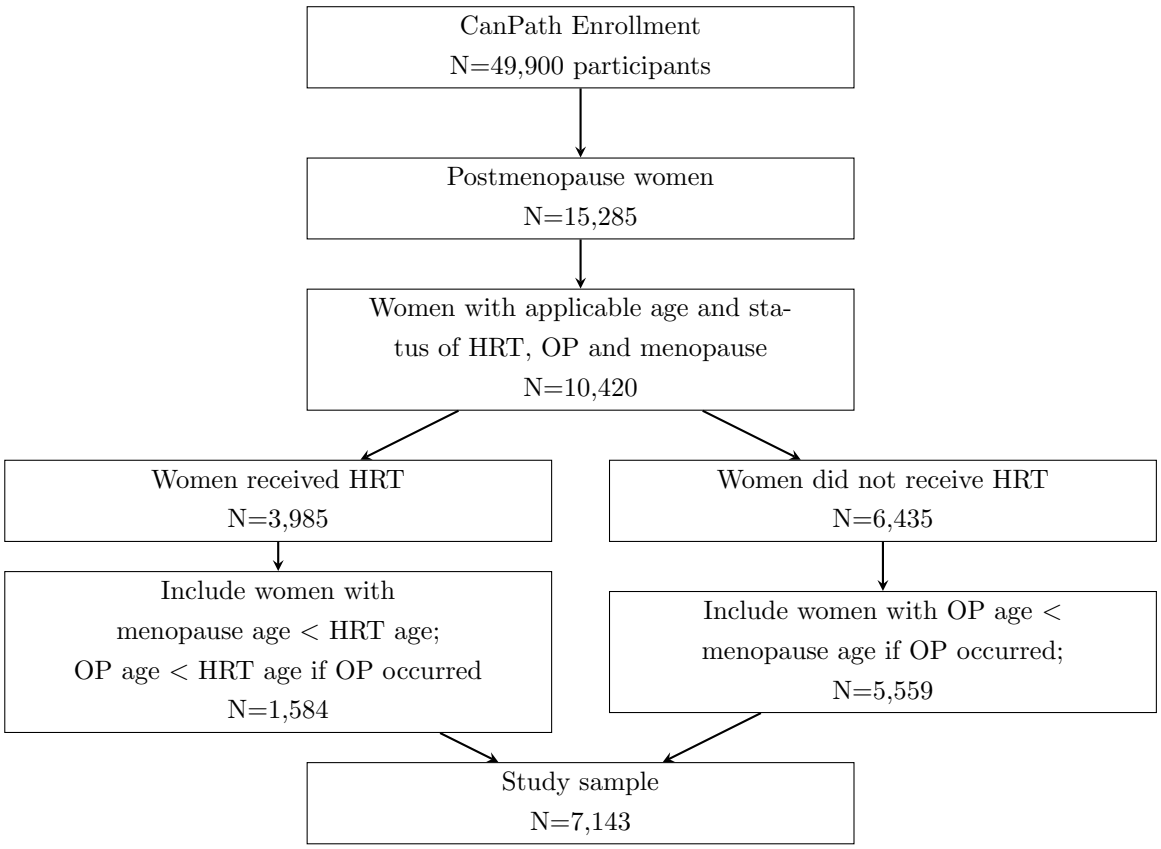

Figure 2: Study sample flowchart.

Our treatment variable is whether receiving HRT before the first occurrence of OP. The primary outcome of interest is the occurrence of OP. Potential confounders include menopause age, BMI, and medical history variables such as hysterectomy, oophorectomy, and family history of osteoporosis. Lifestyle variables considered are smoking status (categorized as never smoker, current smoker, and ever smoker), alcohol consumption (categorized as never versus ever), and physical activity. Physical activity is measured as a three-level quantitative indicator based on vigorous intensity activities in metabolic equivalent (MET)-minutes per week. The negative control outcome is the occurrence of arthritis.

2.2 Statistical analysis

We implemented three models, outcome regression, inverse-probability treatment weighting (IPTW) (Cole and Hernan, 2008), and doubly robust estimator (van der Laan, 2010) to estimate the causal effect of HRT on OP. The outcome regression is modelled using a logistic regression adjusted by all specified confounders. The treatment model was constructed using logistic regression to estimate the probability of receiving HRT treatment, incorporating all specified confounders. We assessed covariance balance before and after weighting and trimmed extreme weights at the 0.99 percentile. We used targeted maximum likelihood estimation (TMLE)(Van Der Laan and Rubin, 2006) for a doubly robust estimator, where both the outcome model and treatment are modelled through SuperLearner. The missing data were addressed using inverse probability of missingness weighting (IPMW) (Seaman and White, 2011), where the missing status was regressed on a fully observed baseline variable, menopause age. We also conducted a sensitivity analysis using the complete case data through an IPTW model.

Confidence intervals were obtained using bootstrap with 1000 replications.

To perform sensitivity analysis for the potential presence of unmeasured confounders, we utilized arthritis as a negative control outcome. Arthritis was selected because it is also a common disease in postmenopausal women. Therefore, arthritis shares common risk factors with osteoporosis, such as BMI, smoking, and physical inactivity. Recent research showed that HRT affects the progression of rheumatoid arthritis, There is no evidence showing that it is directly influenced by hormone replacement therapy (HRT). This characteristic makes arthritis an ideal candidate for assessing the robustness of our findings. All analysis was performed in R version 4.4.1.

3 Results

3.1 Participant characteristics

Table 1 summarizes the baseline characteristics of women treated and not treated with HRT in the study sample. There were 1,584 individuals received HRT and 5,559 did not. Those who received HRT were more likely to have a lower menopause age and the presence of hysterectomy and oophorectomy. Similar trends were observed for family history, BMI, smoke, alcohol, and activity level between the treated and untreated groups.

Variables	**Overall**	**Not treated**	**Treated**	**SMD**
N	7143	5559	1584	
OP = yes (%)	477 (6.7)	345 (6.2)	132 (8.3)	0.082
OP age, years (mean (SD))	56.31 (5.98)	55.79 (5.84)	57.67 (6.13)	0.315
Arthritis = yes (%)	1226 (17.2)	850 (15.3)	376 (23.7)	0.214
Arthritis age (mean (SD))	54.95 (7.02)	54.22 (6.98)	56.60 (6.85)	0.344
Menopause age, years (mean (SD))	47.09 (7.19)	47.96 (6.69)	44.05 (8.02)	0.529
HRT age, years (mean (SD))	48.97 (6.58)	–	48.97 (6.58)	–
Hysterectomy = yes (%)	1618 (25.7)	980 (20.1)	638 (44.8)	0.548
Oophorectomy = yes (%)	940 (13.2)	516 (9.3)	424 (26.9)	0.469
OP family history = yes (%)	1589 (48.8)	1204 (47.9)	385 (51.9)	0.081
BMI (mean (SD))	27.38 (5.98)	27.48 (6.08)	26.99 (5.57)	0.085
Smoke (%)				0.089
Never smoked	3337 (48.4)	2641 (49.2)	696 (45.8)	
Past smoker	2722 (39.5)	2070 (38.6)	652 (42.9)	
Current smoker	829 (12.0)	658 (12.3)	171 (11.3)	
Alcohol = yes (%)	6680 (94.9)	5195 (94.9)	1485 (94.9)	0.001
Activity level (%)				0.042
Low	1630 (26.5)	1286 (26.8)	344 (25.5)	
Moderate	2102 (34.1)	1649 (34.3)	453 (33.6)	
High	2425 (39.4)	1872 (38.9)	553 (41.0)	

Table 1: Descriptive characteristics between those treated and not treated with HRT in the study sample. Figures are n (%) or mean (SD).
OP, osteoporosis; HRT, hormone replacement therapy; BMI, body mass index; SMD, standardized mean difference; SD, standard deviation.

3.2 Casual effect estimates

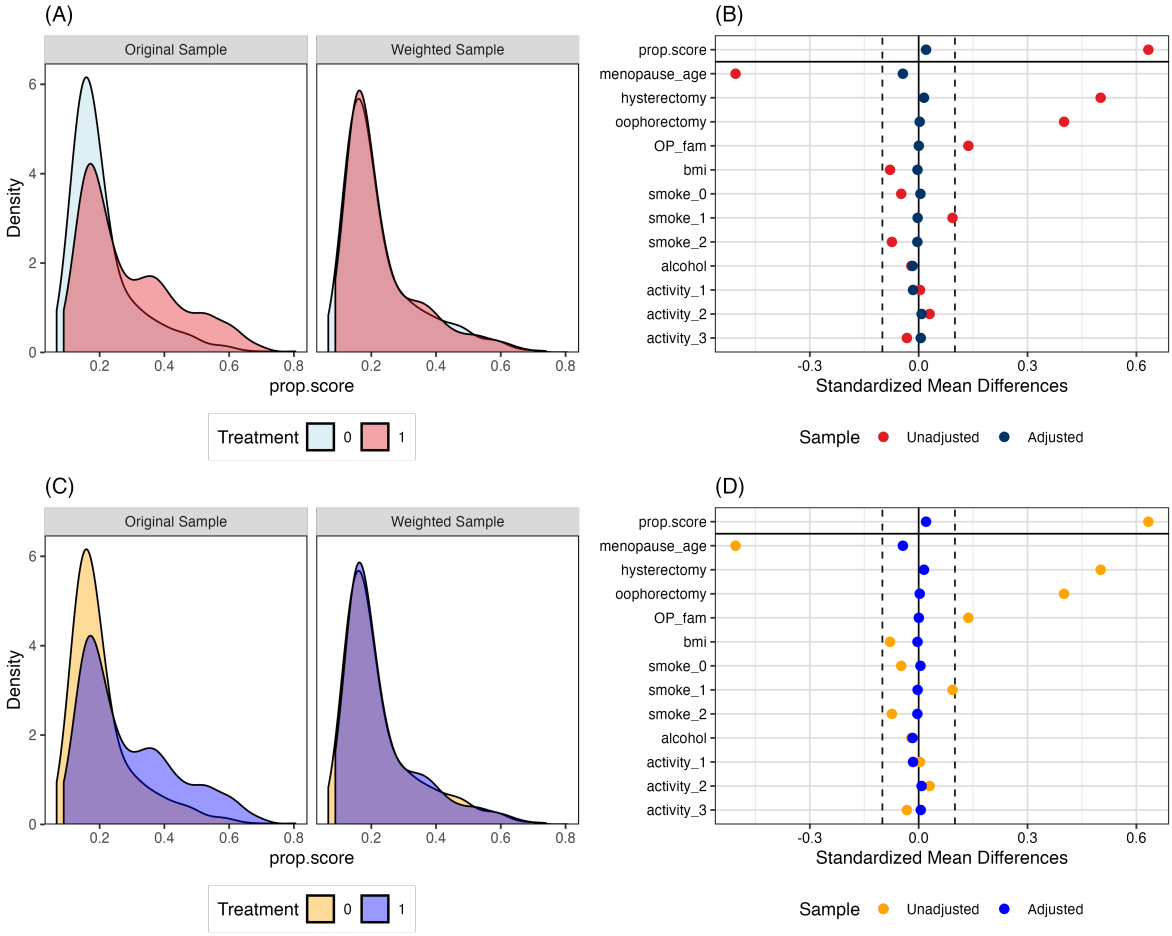

Figure 3: Propensity score distributional overlap(A & C) and standardised differences(B & D) before and after propensity score weighting; Figure (A) and (B) are for samples with inverse probability missing weights; Figure (C) and (D) are for complete data.

We assessed the distributional overlap of propensity scores among samples weith inverse probability missing weights (Figure 3 A & B) and complete data(Figure 3 C & D), respectively. We see the results from samples with missing weights and complete data are similar. Figure 3 A and C present the distributional overlap of propensity scores in the treatment and control groups. We see the distributions overlap. Figure 3 B and D show the standard mean difference for all variables approach zero, indicating the covariate distributions between two comparison groups are balanced after weighting.

Model	Point Estimate	95% bootstrapped CI
Outcome Regression	0.991	(0.657, 1.429)
IPTW	0.994	(0.682, 1.432)
TMLE	1.001	(0.707, 1.417)
CCA	1.008	(0.692, 1.396)
NCO	1.730	(1.361, 2.243)

Table 2: Point estimates and 95% Confidence Intervals of Odds Ratio. IPTW: Inverse-probability treatment weight; TMLE: targeted maximum likelihood estimation; CCA: complete cases analysis using IPTW; NCO: Negative control outcome analysis usign IPTW.

Table 2 provides a summary of the estimated odds ratios (OR) and their 95% bootstrapped confidence intervals (CI) for outcome regression, IPTW, TMLE, CCA, and NCO. The OR estimates for logistic regression, IPTW, and TMLE models are quite similar, with values of 0.991 (95% CI: 0.657, 1.429), 0.994 (95% CI: 0.657, 1.429), and 1.001 (95% CI: 0.707, 1.417), respectively. The points estimates are not different from one and the 95% CI include 1, indicating HRT has no causal effect on OP. These similar results from the three causal models suggest robustness of our causal analysis with measured confounder. CCA yielded a similar OR estimate of 1.008 (95% CI: 0.692, 1.396). However, sensitivity analysis using the negative control method produced an OR of 1.730 (95% CI: 1.361, 2.243), indicating the presence of unmeasured confounding factors.

4 Conclusion and Discussion

In this study, we utilized three models, outcome regression, IPTW, and doubly robust estimator to investigate the causal effect of HRT on the risk of OP occurrence of postmenopausal women. We found that in all models, HRT does not have a causal effect on the occurrence of OP in our data.

Recent studies have suggested that HRT can prevent women from decreasing bone mass density and reduce the risk of fracture (Gambacciani and Levancini, 2014). Since the reduction of bone mass density is a crucial factor in the development of OP, it is expected that HRT could potentially help in reducing the risk of OP occurrence. Our finding shows no causal effect of HRT on OP occurrence in postmenopausal women is counter-intuitive. There are two possible reasons behind this: 1) the CanPath student dataset has been modified to protect participant information and, therefore, may not be representative of the true data relationship; 2) the negative control outcome analysis indicates there are potential unmeasured confounders. Future analysis will be focused on applying our approach to the real CanPath data and investigating the causal DAG with expert knowledge. Given that OP is a significant risk factor for mortality in postmenopausal women and HRT has side effects such as increased risk of breast cancer (Chen et al., 2002) and ovarian cancer (Zhou et al., 2008), highlighting the need for careful consideration of its benefits and risks.

Our study has several strengths. We considered three models in our analysis, ensuring the robustness of model misspecification of our analysis. In addition, we utilized the IPMW to account for missing data. This method is more appropriate than complete case analysis by reducing potential biases related to incomplete data.

Our study also has a few limitations. For example, the time after menopause is also highly correlated with the decrease in bone mass density. During the menopausal transition period, the average reduction

in bone mass density is about 10% (Ji and Yu, 2015). Therefore, time from menopause to HRT can be an effect modifier of HRT on OP occurrence. This analysis will be limited to postmenopausal women who receive HRT. Additionally, HRT duration may also play an important role in reducing bone mass density. A time-varying design can be considered to assess the effect of the length of HRT treatment on the occurrence of OP. This will be left for future analysis.

In conclusion, our study suggests that there is no causal effect of HRT on the risk of OP in menopause women. HRT might not be an appropriate treatment for preventing OP.

References

Bonjour, J.-P., Chevalley, T., Ferrari, S., and Rizzoli, R. (2009). The importance and relevance of peak bone mass in the prevalence of osteoporosis. *Salud Pública de México*, 51.

Cartwright, B., Robinson, J., Seed, P. T., Fogelman, I., and Rymer, J. (2016). Hormone Replacement Therapy Versus the Combined Oral Contraceptive Pill in Premature Ovarian Failure: A Randomized Controlled Trial of the Effects on Bone Mineral Density. *The Journal of Clinical Endocrinology & Metabolism*, 101(9):3497–3505.

Chen, C.-L., Weiss, N. S., Newcomb, P., Barlow, W., and White, E. (2002). Hormone replacement therapy in relation to breast cancer. *Jama*, 287(6):734–741.

Cole, S. R. and Hernan, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

Crofton, P. M., Evans, N., Bath, L. E., Warner, P., Whitehead, T. J., Critchley, H. O. D., Kelnar, C. J. H., and Wallace, W. H. B. (2010). Physiological versus standard sex steroid replacement in young women with premature ovarian failure: effects on bone mass acquisition and turnover. *Clinical Endocrinology*, 73(6):707–714.

De Villiers, T. J., Gass, M. L. S., Haines, C. J., Hall, J. E., Lobo, R. A., Pierroz, D. D., and Rees, M. (2013). Global Consensus Statement on Menopausal Hormone Therapy. *Climacteric*, 16(2):203–204.

Gambacciani, M. and Levancini, M. (2014). Featured Editorial Hormone replacement therapy and the prevention of postmenopausal osteoporosis. *Menopausal Review*, 4:213–220.

Ji, M. and Yu, Q. (2015). Primary osteoporosis in postmenopausal women. *Chronic Diseases and Translational Medicine*, 1(1):9–13.

Rogers, A., Saleh, G., Hannon, R. A., Greenfield, D., and Eastell, R. (2002). Circulating Estradiol and Osteoprotegerin as Determinants of Bone Turnover and Bone Density in Postmenopausal Women. *The Journal of Clinical Endocrinology & Metabolism*, 87(10):4470–4475.

Sandhu, S. K. and Hampson, G. (2011). The pathogenesis, diagnosis, investigation and management of osteoporosis. *Journal of Clinical Pathology*, 64(12):1042–1050.

Seaman, S. R. and White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.

Steinman, J. and Shibli-Rahhal, A. (2019). Anorexia Nervosa and Osteoporosis: Pathophysiology and Treatment. *Journal of Bone Metabolism*, 26(3):133.

Torgerson, D. J. and Bell-Syer, S. E. (2001). Hormone replacement therapy and prevention of vertebral fractures: a meta-analysis of randomised trials. *BMC Musculoskeletal Disorders*, 2(1):7.

van der Laan, M. J. (2010). Targeted maximum likelihood based causal inference: Part i. *The International Journal of Biostatistics*, 6(2).

Van Der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1).

Wells, G., Tugwell, P., Shea, B., Guyatt, G., Peterson, J., Zytaruk, N., Robinson, V., Henry, D., O'Connell, D., Cranney, A., The osteoporosis methodology group, and The Osteoporosis Research Advisory Group (2002). V. Meta-Analysis of the Efficacy of Hormone Replacement Therapy in Treating and Preventing Osteoporosis in Postmenopausal Women. *Endocrine Reviews*, 23(4):529–539.

Zhou, B., Sun, Q., Cong, R., Gu, H., Tang, N., Yang, L., and Wang, B. (2008). Hormone replacement therapy and ovarian cancer risk: a meta-analysis. *Gynecologic oncology*, 108(3):641–651.

Appendix A Author Contribution

S.F. is responsible for data cleaning, writing of the introduction and part of the methodology section. L.Y. is responsible for the statistical analysis, writing of results and part of the methodology section. M.D. is responsible for statistical analysis, sensitivity analysis, writing of the conclusion and discussion, and part of the methodology section. All authors are involved in the conceptualization, analysis design, and revising the manuscript.

Appendix B R codes

Part 1: Data Cleaning

```
library(dplyr)
library(tableone)
library(WeightIt)
library(tidyverse)
library(cobalt)
library(geepack)
library(xgboost)
library(tmle)
library(DiagrammeR)
library(cobalt)
library(survey)
library(gfoRmula)
library(ltmle)
library(rsvg)

canpath_data <- read.csv("student_dataset_canue_Version2_49900par_358var.csv")

canpath_subset <- canpath_data %>%
  filter(SDC_SEX==2) %>% # include female
```

```
  filter(WH_MENOPAUSE_EVER==1) %>% # include female with meanopause==1
#  filter(DIS_OP_EVER!=2 & WH_MENOPAUSE_AGE!=-7 & WH_HRT_EVER!=-7 & DIS_OP_FAM_EVER!=2 & WH_OOPHORECTOMY_EVER !=-7) %>%
  filter(DIS_OP_EVER!=2 & WH_MENOPAUSE_AGE!=-7 & WH_HRT_EVER!=-7 & DIS_ARTHRITIS_EVER !=2) %>%
  # exclude people with non-applicable op occurrence/MENOPAUSE_AGE/HRT status
  filter(!(DIS_OP_EVER == 1 & DIS_OP_AGE == -7)) %>% ##remove patient with OP event but not OP event age
  filter(!(DIS_ARTHRITIS_EVER == 1 & DIS_ARTHRITIS_AGE==-7)) %>% ## remove patient with ARTHRITIS event but no ARTHRITIS event age
  mutate(id = row_number()) %>%
  select(id, DIS_OP_EVER, DIS_OP_AGE, WH_MENOPAUSE_AGE, WH_HRT_EVER, WH_HRT_AGE, WH_HRT_DURATION, WH_CONTRACEPTIVES_EVER, WH_HYSTERECTOMY_EVER,
         WH_OOPHORECTOMY_EVER, DIS_OP_FAM_EVER, PM_BMI_SR, SMK_CIG_STATUS, ALC_EVER, PA_LEVEL_SHORT, DIS_ARTHRITIS_EVER, DIS_ARTHRITIS_AGE)
# PA_LEVEL_SHORT: Categorical indicator of the participant's level of physical activity (IPAQ short form), value = 1,2,3
# DIS_EYE_CATARACTS_EVER, DIS_SLE_EVER are served as negative control outcome
# Occurrence of cataracts at any point during the life of the participant; Occurrence of systemic lupus erythematosus at any point during the life of the participant

# The subset with HRT==1
canpath_subset_hrt1 <- canpath_subset %>%
  filter(WH_HRT_EVER==1) %>%
  filter(WH_MENOPAUSE_AGE < WH_HRT_AGE) %>% # include menopaus age < HRTage
  filter(!(DIS_OP_EVER==1 & DIS_OP_AGE < WH_HRT_AGE)) %>% # exclude those had OP but OP age<HRT age
  filter(!(DIS_ARTHRITIS_EVER==1 & DIS_ARTHRITIS_AGE < WH_HRT_AGE)) # exclude those had ARTHRITIS but ARTHRITIS age<HRT age

# The subset with HRT==0
canpath_subset_hrt0 <- canpath_subset %>%
  filter(WH_HRT_EVER==0) %>%
  filter(!(DIS_OP_EVER==1 & DIS_OP_AGE < WH_MENOPAUSE_AGE)) %>% # exclude those had OP but OP age< MENOPAUSE age
  filter(!(DIS_ARTHRITIS_EVER==1 & DIS_ARTHRITIS_AGE < WH_MENOPAUSE_AGE)) # exclude those had ARTHRITIS but ARTHRITIS age< MENOPAUSEage

canpath_subset2 <- rbind(canpath_subset_hrt1, canpath_subset_hrt0) %>%
  arrange(id) %>%
  mutate(OP_status = as.factor(DIS_OP_EVER),
         OP_age = ifelse(DIS_OP_AGE>=0, DIS_OP_AGE, NA),
         ARTHRITIS_status = as.factor(DIS_ARTHRITIS_EVER),
         ARTHRITIS_age = ifelse(DIS_ARTHRITIS_AGE>=0, DIS_ARTHRITIS_AGE, NA),
         menopause_age = WH_MENOPAUSE_AGE,
         HRT = as.factor(WH_HRT_EVER),
         HRT_age = ifelse(WH_HRT_AGE>=0, WH_HRT_AGE, NA),
         HRT_duration = ifelse(WH_HRT_DURATION>0, WH_HRT_DURATION, NA),
```

```
          time_menopause_to_HRT = HRT_age - menopause_age,
          contraceptives = as.factor(WH_CONTRACEPTIVES_EVER),
          hysterectomy = as.factor(WH_HYSTERECTOMY_EVER),
          oophorectomy = as.factor(ifelse(WH_OOPHORECTOMY_EVER >=0, WH_
              OOPHORECTOMY_EVER, NA)),
          OP_fam = as.factor(ifelse(DIS_OP_FAM_EVER<2, DIS_OP_FAM_EVER, NA)),
          bmi = PM_BMI_SR,
          smoke = case_when(SMK_CIG_STATUS==0 ~0 ,
                            SMK_CIG_STATUS==1 ~1,
                            SMK_CIG_STATUS==2 | SMK_CIG_STATUS==3 ~2),
          smoke=as.factor(smoke),
          alcohol = as.factor(ALC_EVER),
          activity = as.factor(PA_LEVEL_SHORT)) %>%
  select(!c(DIS_OP_EVER, DIS_OP_AGE, WH_MENOPAUSE_AGE, WH_HRT_EVER, WH_HRT_AGE
      , WH_HRT_DURATION, WH_CONTRACEPTIVES_EVER, WH_HYSTERECTOMY_EVER,
                    WH_OOPHORECTOMY_EVER, DIS_OP_FAM_EVER, PM_BMI_SR, SMK_CIG_
                        STATUS, ALC_EVER, PA_LEVEL_SHORT, DIS_ARTHRITIS_EVER,
                        DIS_ARTHRITIS_AGE))


covariates <- select(canpath_subset2, -c(id, HRT))
baselines <- colnames(covariates)
baselines
tab0 <- CreateTableOne(vars = baselines,
                       data = canpath_subset2,
                       strata = "HRT",
                       test = FALSE, #mute P-value calculation;
                       smd = TRUE,
                       addOverall = TRUE)
print(tab0, smd = TRUE, showAllLevels = FALSE)

NA_count <- canpath_subset2 %>%
  group_by(HRT) %>%
  summarise(across(everything(), ~sum(is.na(.))))
NA_count <- as.data.frame(t(NA_count))[-c(1,2),]
colnames(NA_count) <- c("HRT=0", "HRT=1")
NA_count
table(canpath_subset2$OP_status)
canpath_subset2 %>%
  group_by(OP_status) %>%
  summarise(OPage_NAcount = sum(is.na(OP_age)),
            OPfam_NAcount = sum(is.na(OP_fam)))
# all 8524 people with OP_status=0 has missing data in OP_age


# DAG
grViz("
    digraph causal {
    # Nodes
```

```
    node [shape=plaintext]
    A [label = 'A']
    L [label = 'L']
    X [label = 'X']
    Y [label = 'Y']
    # Edges
    edge [color=black, arrowhead=vee]
    rankdir = LR

    A->X
    X->Y
    A->Y
    L->A
    L->Y


    # Graph
    graph [overlap=true, fontsize=14]
    }")
```

Part 2: Modelling and Sensitivity Analysis

```
library(dplyr)
library(naniar)
library(data.table)
source("data_cleaning.R")
# canpath_subset2

## variables
model_dat <- canpath_subset2 %>%
  select(-contraceptives) %>%
 # select(-OP_age, ARTHRITIS_status, -ARTHRITIS_age) %>%
  select(ARTHRITIS_status,
         OP_status,HRT,
         menopause_age,
         HRT_age, HRT_duration,
        time_menopause_to_HRT,
        # contraceptives,
        hysterectomy,oophorectomy,
         OP_fam,
         bmi,smoke,
         alcohol,activity) %>%
  mutate(HRT_duration = ifelse(is.na(HRT_duration),0,HRT_duration),
         time_menopause_to_HRT = ifelse(is.na(time_menopause_to_HRT),0,time_
             menopause_to_HRT),
         HRT_age = ifelse(is.na(HRT_age),0,HRT_age),
         )

## 1. Inverse probability missing weight-----
model_dat$missing[complete.cases(model_dat)] <- 0
```

```r
model_dat$missing[!complete.cases(model_dat)] <- 1

missed <- glm(missing ~ menopause_age,
              data = model_dat, family = "binomial")
summary(missed)
model_dat$fitted = missed$fitted.values
model_dat$IPCW = (1-mean(model_dat$missing))/(1-model_dat$fitted)
model_dat <- subset(model_dat, complete.cases(model_dat))
summary(model_dat$IPCW)


## 2. IPTW-----

baselines <- c(
  "menopause_age",
  # "HRT_age",
  # "HRT_duration",  ## effect modifier
  # "time_menopause_to_HRT"
  #"contraceptives",
  "hysterectomy",
  "oophorectomy",
  "OP_fam",
  "bmi",
  "smoke",
  "alcohol",
  "activity")


ps_formula <- as.formula(paste0("HRT ~ ",
                                paste(baselines,collapse = "+")))

IPTW <- weightit(ps_formula,
                 data = model_dat,
                 method = "glm", #using the default logistic regression;
                 stabilize = TRUE)

IPTW <- trim(IPTW, at = .99)

summary(IPTW)
### 2.1 check balance-----
library(cobalt)
library(sjPlot)
fig1 <- bal.plot(IPTW,
         which="both",
         type = "density",
         colors = c("lightblue","#E31B23"),
         grid = FALSE,
         position = 'bottom',
         alpha.weight = 0.1,
```

```
         sample.names = c("Original Sample","Weighted Sample")
         )+
  ggtitle("")

fig2 <- love.plot(IPTW,
          binary = "std",
          grid = TRUE,
          thresholds = c(m = .1),
          colors = c("#E31B23","#003366"),
          position = 'bottom')

##
library(ggpubr)


ggsave("dist_overlap.png",  # File path to save
       plot = ggarrange(fig1, fig2,
                        ncol = 2),             # Plot object to save
       width = 16,            # Width in inches
       height = 9,           # Height in inches
       dpi = 400)



ggsave(fig1,file="dist_overlap.png",dpi = 400)
ggsave(fig2,file="SMD.png",dpi = 400)
## 3. outcome model ----
model <- glm(OP_status ~ HRT,
             family = binomial(link="logit"),
             weights = (IPTW$weights)*(model_dat$IPCW),
             data = model_dat)




summary(model)

## outcome model
# model <- glm(OP_status ~ HRT*HRT_duration,
#              family = binomial(link="logit"),
#              weights = (IPTW$weights)*(model_dat$IPCW),
#              data = model_dat)
# non identifiable

## 4. complete cases--------

model_dat_cc <- canpath_subset2 %>%
  # select(-OP_age, -ARTHRITIS_status, -ARTHRITIS_age) %>%
  select(ARTHRITIS_status,
```

```
          OP_status,HRT,
          menopause_age,HRT_age,HRT_duration,time_menopause_to_HRT,
          # contraceptives,
          hysterectomy,oophorectomy,OP_fam,bmi,smoke,
          alcohol,activity) %>%
  mutate(HRT_duration = ifelse(is.na(HRT_duration),9999,HRT_duration),
          time_menopause_to_HRT = ifelse(is.na(time_menopause_to_HRT),9999,time
              _menopause_to_HRT),
          HRT_age = ifelse(is.na(HRT_age),9999,HRT_age),
  ) %>% na.omit()

model_dat_cc[model_dat_cc == 9999] <- NA

model_dat_cc1 <- model_dat_cc

model_dat_cc1[is.na(model_dat_cc1)] <- 0


baselines <- c(
  "menopause_age",
  # "HRT_age",
  # "HRT_duration",  ## effect modifier
  # "time_menopause_to_HRT"
  # "contraceptives",
  "hysterectomy",
  "oophorectomy",
  "OP_fam",
  "bmi",
  "smoke",
  "alcohol",
  "activity")


ps_formula <- as.formula(paste0("HRT ~ ",
                                paste(baselines,collapse = "+")))

IPTW_cc <- weightit(ps_formula,
                 data = model_dat_cc1,
                 method = "glm", #using the default logistic regression;
                 stabilize = TRUE)

IPTW_cc <- trim(IPTW_cc,at=0.99)

summary(IPTW_cc)

library(cobalt)
library(sjPlot)
fig1_cc <- bal.plot(IPTW_cc,
                 which="both",
```

```
                  type = "density",
                  colors = c("lightblue","#E31B23"),
                  grid = FALSE,
                  position = 'bottom',
                  alpha.weight = 0.1,
                  sample.names = c("Original Sample","Weighted Sample")
)+
  ggtitle("")

fig2_cc <- love.plot(IPTW_cc,
                   binary = "std",
                   grid = TRUE,
                   thresholds = c(m = .1),
                   colors = c("#E31B23","#003366"),
                   position = 'bottom')

##
library(ggpubr)

ggsave("dist_overlap_cc.png",  # File path to save
       plot = ggarrange(fig1_cc, fig2_cc,
                        ncol = 2),             # Plot object to save
       width = 16,            # Width in inches
       height = 9,           # Height in inches
       dpi = 400)


model_cc <- glm(OP_status ~ HRT,
              family = binomial(link="logit"),
              weights = (IPTW_cc$weights),
              data = model_dat_cc1)

summary(model_cc)

## 5. Negative control outcome-----
model_dat_no <- canpath_subset2 %>%
  select(-OP_age,  -ARTHRITIS_age) %>%
  select(ARTHRITIS_status,
         OP_status,HRT,
         menopause_age,
         HRT_age, HRT_duration,
         time_menopause_to_HRT,
        #  contraceptives,
         hysterectomy,oophorectomy,
         OP_fam,
         bmi,smoke,
         alcohol,activity) %>%
  mutate(HRT_duration = ifelse(is.na(HRT_duration),0,HRT_duration),
         time_menopause_to_HRT = ifelse(is.na(time_menopause_to_HRT),0,time_
```

```
            menopause_to_HRT),
         HRT_age = ifelse(is.na(HRT_age),0,HRT_age),
  )


model_nco <- glm( ARTHRITIS_status ~ HRT,
              family = binomial(link="logit"),
              weights = (IPTW$weights)*(model_dat$IPCW),
              data = model_dat)

model_nco_cc <- glm( ARTHRITIS_status ~ HRT,
                  family = binomial(link="logit"),
                  weights = (IPTW$weights),
                  data = model_dat_cc1)




## 6. associatioin model------

model_asso <-  glm(
  "OP_status ~ HRT + menopause_age + hysterectomy + oophorectomy + OP_fam +
      bmi + smoke + alcohol + activity",
  family = binomial(link = "logit"),
  weights = IPTW$weights,
  data = model_dat
)

summary(model_asso)$coefficients[2]

tab_model(model_asso)


## 7. doubly robust----


model <- glm(formula,
                family = gaussian(link = "identity"),
                data = model_dat
)
## predicition
Q2_pred_dat <- rbind(
  select(model_dat, -c(HRT)),
  select(model_dat, -c(HRT))
) %>%
  mutate(HRT = c(rep(1, nrow(model_dat)), rep(0, nrow(model_dat))) %>% as.
      factor())

Q2_pred <- predict(model, Q2_pred_dat, type = 'response') %>%
```

```
  data.frame()


outcome_pred_dat <- cbind(Q2_pred_dat)  %>%
  mutate(origin_HRT = c(model_dat$HRT,model_dat$HRT)) %>%
  mutate(IPTweight = c(IPTW$weights,IPTW$weights) ) %>%
  mutate(  OP_pred = Q2_pred$.)

tmp <- outcome_pred_dat %>% select(origin_HRT,HRT,OP_status,OP_pred)


DR_est <- outcome_pred_dat %>%
  mutate(
    ### indicator
    indicator = ifelse(origin_HRT == HRT,
                       IPTweight * (as.numeric(OP_status) - OP_pred),
                       0 ),
    DR = OP_pred + indicator
  ) %>%
  group_by(HRT) %>%
  summarise(DR_est = mean(DR))


# DR risk difference
DR_est$DR_est[2] - DR_est$DR_est[1]

## 8. tlme-----
require(SuperLearner)
library(tmle)

set.seed(1017)
require(SuperLearner)

model_dat$HRT <- as.integer(as.character(model_dat$HRT))
model_dat$OP_status <- as.integer(as.character(model_dat$OP_status))
covariates <- as.matrix(model_dat %>% select(all_of(baselines)))


# SL.library <- c("SL.glmnet","SL.glm")
SL.library <- c("SL.glmnet","SL.glm")
tmle_fit <- tmle(Y = model_dat$OP_status,
                 A = model_dat$HRT,
                 W = covariates,
                 Q.SL.library = SL.library,
                 g.SL.library = SL.library,
                 family = "binomial",
                 obsWeights = model_dat$IPCW
                 )

```

```
# summary(tmle_fit)
```

```
### compute CIs
rm(list=ls())
source("modelling.R")
set.seed(1017)
boot.est_model <- rep(NA, 1000)
boot.est_model_nco <- rep(NA, 1000)
boot.est_model_cc <-   rep(NA, 1000)
boot.est_model_asso <- rep(NA, 1000)
boot.est_model_tmle <- rep(NA, 1000)

for (i in 1:1000){
  boot.idx <- sample(1:dim(model_dat)[1], size = dim(model_dat)[1], replace = T)
  boot.data <- model_dat[boot.idx,]

  boot.idx1 <- sample(1:dim(model_dat_cc1)[1], size = dim(model_dat_cc1)[1], replace = T)
  boot.data1 <- model_dat[boot.idx1,]



  boot.data$missing[complete.cases(boot.data)] <- 0
  boot.data$missing[!complete.cases(boot.data)] <- 1

  missed <- glm(missing ~ menopause_age,
                data = boot.data, family = "binomial")
  summary(missed)
  boot.data$fitted = missed$fitted.values
  boot.data$IPCW = (1-mean(boot.data$missing))/(1-boot.data$fitted)
  boot.data <- subset(boot.data, complete.cases(boot.data))

  baselines <- c(
    "menopause_age",
    # "HRT_age",
    # "HRT_duration",  ## effect modifier
    # "time_menopause_to_HRT"
    #"contraceptives",
    "hysterectomy",
    "oophorectomy",
    "OP_fam",
    "bmi",
    "smoke",
    "alcohol",
    "activity")


  ps_formula <- as.formula(paste0("HRT ~ ",
```

```
                                  paste(baselines,collapse = "+")))

IPTW <- weightit(ps_formula,
                 data = boot.data,
                 method = "glm", #using the default logistic regression;
                 stabilize = TRUE)

IPTW <- trim(IPTW, at = .99)


IPTW_cc <- weightit(ps_formula,
                    data = boot.data1,
                    method = "glm", #using the default logistic regression;
                    stabilize = TRUE)

IPTW_cc <- trim(IPTW_cc, at = .99)


boot.model <- glm(OP_status ~ HRT,
             family = binomial(link="logit"),
             weights = (IPTW$weights)*(model_dat$IPCW),
             data = boot.data)



boot.model_nco <- glm( ARTHRITIS_status ~ HRT,
                  family = binomial(link="logit"),
                  weights = (IPTW$weights)*(model_dat$IPCW),
                  data = boot.data)


boot.model_cc <- glm(OP_status ~ HRT,
                family = binomial(link="logit"),
                weights = (IPTW_cc$weights),
                data = boot.data1)

boot.model_asso <-  glm(
  "OP_status ~ HRT + menopause_age + hysterectomy + oophorectomy + OP_fam +
      bmi + smoke + alcohol + activity",
  family = binomial(link = "logit"),
  weights = IPTW$weights,
  data = boot.data
)

SL.library <- c("SL.glmnet","SL.glm")
boot.tmle_fit <- tmle(Y = boot.data$OP_status,
                 A = boot.data$HRT,
                 W = covariates,
                 Q.SL.library = SL.library,
```

```
                    g.SL.library = SL.library,
                    family = "binomial",
                    obsWeights = boot.data$IPCW
  )

  ##
  boot.est_model[i] <- summary(boot.model)$coefficients[2]
  ## nco
  boot.est_model_nco[i] <- summary(boot.model_nco)$coefficients[2]
  ## complete case
  boot.est_model_cc[i] <- summary(boot.model_cc)$coefficients[2]
  ## association model
  boot.est_model_asso[i] <- summary(boot.model_asso)$coefficients[2]

  boot.est_model_tmle[i] <- boot.tmle_fit$estimates$OR$psi
}

paste0("(",round(boot.est_model %>% quantile(c(0.025,0.975)),3)[1],",",
round(boot.est_model %>% quantile(c(0.025,0.975)),3)[2],")")

## cc
paste0("(",round(boot.est_model_cc %>% quantile(c(0.025,0.975)),3)[1],",",
       round(boot.est_model_cc %>% quantile(c(0.025,0.975)),3)[2],")")

## nco
paste0("(",round(boot.est_model_nco %>% quantile(c(0.025,0.975)),3)[1],",",
       round(boot.est_model_nco %>% quantile(c(0.025,0.975)),3)[2],")")


## association
paste0("(",round(boot.est_model_asso %>% quantile(c(0.025,0.975)),3)[1],",",
       round(boot.est_model_asso %>% quantile(c(0.025,0.975)),3)[2],")")

## tmle
paste0("(",round(boot.est_model_tmle %>% quantile(c(0.025,0.975)),3)[1],",",
       round(boot.est_model_tmle %>% quantile(c(0.025,0.975)),3)[2],")")


OR_summary <- data.frame(
  model = c("Association", "MSM", 'tmle', "CC", "NCO"),
  point_est = c(
    round(exp(summary(model_asso)$coefficients[2]), 3),
    round(exp(summary(model)$coefficients[2]), 3),
    round(tmle_fit$estimates$OR$psi,3),
    round(exp(summary(model_cc)$coefficients[2]), 3),
    round(exp(summary(model_nco)$coefficients[2]), 3)
  ),
  `95% CI` = c(
    paste0(
```

```
      "(", round(exp(boot.est_model_asso %>% quantile(c(0.025, 0.975))), 3)[1], ",",
      round(exp(boot.est_model_asso %>% quantile(c(0.025, 0.975))), 3)[2], ")"
    ),
    paste0(
      "(", round(exp(boot.est_model %>% quantile(c(0.025, 0.975))), 3)[1], ","
      ,
      round(exp(boot.est_model %>% quantile(c(0.025, 0.975))), 3)[2], ")"
    ),

    paste0(
      "(", round(exp(boot.est_model_tmle %>% quantile(c(0.025, 0.975))), 3)[1], ",",
      round(exp(boot.est_model_tmle %>% quantile(c(0.025, 0.975))), 3)[2], ")"
    ),

    paste0(
      "(", round(exp(boot.est_model_cc %>% quantile(c(0.025, 0.975))), 3)[1], ",",
      round(exp(boot.est_model_cc %>% quantile(c(0.025, 0.975))), 3)[2], ")"
    ),
    paste0(
      "(", round(exp(boot.est_model_nco %>% quantile(c(0.025, 0.975))), 3)[1], ",",
      round(exp(boot.est_model_nco %>% quantile(c(0.025, 0.975))), 3)[2], ")"
    )
  )
)

library(rio)

export(OR_summary,file="OR_summary.xlsx")
```