PharmaSUG China 2025 - Paper SA-121

Propensity Score Weighting in a two-arm observational study with binary outcome

Jun Li, Sanofi

ABSTRACT

Randomized Controlled Trials (RCTs) are considered as the gold standard in clinical trials because of their rigorous design and ability to generate causality. However, they are often limited in practice due to ethical and cost issues. With the growing availability of large, real world health care data, there is a growing interest of non-randomized observational study for assessing the real-world causal effects of treatments. Without randomization, a direct comparison of the outcome between the two groups of participants could be biased because of the imbalances in important participant baseline characteristics between the two groups. Propensity score (PS) methods, which aim to adjust for confounders and obtain unbiased causal treatment effect estimates, have become increasingly popular and are widely used in observational studies. The objective of this paper is to provide a roadmap and practical guidance for application of Inverse Probability of Treatment Weighting (IPTW) method for two-arm clinical trial participants. It mainly consists of five steps, PS estimation, PS feasibility check, bias reduction with IPTW, balance diagnosis and analysis with PS weights including standard error estimation. The proposed analytic framework is illustrated with the simulation data from the observational study for Sarclisa firstly approved in China on Jan2025 by using SAS Enterprise Guide 9.4 and RStudio 4.4.1.

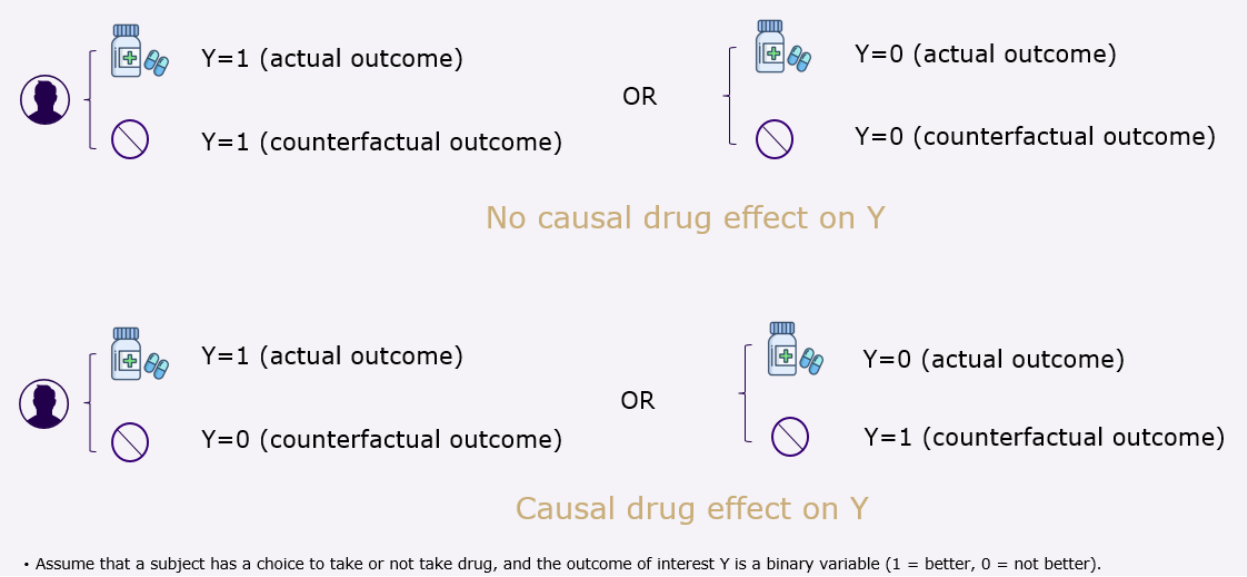
INTRODUCTION

Causal inference methods for observational data represent an alternative to randomized controlled trials when they are not feasible or when real-world evidence is sought. However, in observational research, treatment assignment is usually not random. This means that similarities between groups of participants receiving different treatments cannot be assumed. Then the estimated treatment effect could be biased due to the imbalances in important participant baseline characteristics. Inverse-probability-of-treatment weighting (IPTW), one of the most popular approaches to account for confounding, is commonly used in observational studies to balance the distributions of the baseline confounders between groups to mimic randomization. This paper introduces the basics of casual inference and propensity score, then focus on the process for how to estimate casual treatment effect with the propensity score weighting. The detailed best practices, associated SAS/R code and examples are presented using simulated data.

CASUAL INFERENCE

In health care research, it is often of interest to identify whether an intervention is “causally” related to a sequence of outcomes. But how to define and prove “causation”? Hume (1748) stated that causation is the relation that holds between two temporally simultaneous or successive events when the first event (the cause) brings about the other (the effect). Figure 1 is an example to illustrate causation. It is assumed that a subject has a choice to take or not take the drug, and the outcome of interest Y is a binary variable (1 = better, 0 = not better). If you are somehow able to know both the actual outcome of taking drug and the counterfactual outcome, then you could assess whether a causal effect exists between drug and Y. If this subject has the same outcome no matter take or not take drug, then you could claim “There is No causal drug effect on Y”; On the contrary, If this subject has the different outcome when take and not take the drug, you could claim “There is Causal drug effect on Y”. Unfortunately, in reality, it is impossible to observe both the outcome and its “counterfactual” outcome simultaneously on the same subject.

Figure 1. Causal Effect Scenarios



Such confusion is not clarified until Fisher brought the idea of randomized experiment. With a perfect randomization, the control group could provide counterfactual outcomes for the observed performance in the treatment group, so that the causal effect can be estimated. It means randomization could create balance in baseline confounders between treatment and control groups, and with this design, we assume that outcome differences among the groups are caused by differences in the efficacy of treatments. However, in observational study, there's no randomization. A direct comparison of the outcome between those two groups of patients could be biased because of the imbalances in important patient baseline characteristics between the two groups. Thus, for real word study, the most important methodological challenge is how to control confounding bias due to the lack of randomization.

Rubin and Rosenbaum (1983) proposed the use of the propensity score in observational studies to adjust for confounders and obtain unbiased causal treatment effect estimates. The propensity score (PS) is the conditional probability for a subject of **being in the treatment group** given the measured baseline covariates:

$$e(\boldsymbol{X}) = P(Z = 1|\boldsymbol{X})$$ where Z is the treatment and $\boldsymbol{X}$ is a vector of covariates.

Conditioning on the propensity score, each subject has the same chance of receiving treatment. The distributions of measured baseline confounders are similar between treatment and control groups given the propensity score.

DATA SIMULATIONS

OBS17227 is a single arm, multi-center, observational, and prospective study in Chinese relapsed and/or refractory multiple myeloma patients (RRMM) treated with isatuximab in combination with pomalidomide and dexamethasone (IPd). Real-world historic data requested by CDE as external comparison to describe the differences of effectiveness in RRMM treatment with IPd arm. Risk ratio is calculated to describe the difference of ORR between two groups after adjusting the baseline variables using inverse probability of treatment weights (IPTW) method. Table 1 lists the key variables in the original analysis dataset from which the simulated data is formed. Simulation is performed separately for treatment group and control group. The R code is shown in Program 1.

Table 1.Key Variables

Variable Name	Variable Label
USUBJID	Unique Subject Identifier
TRT01AN	Actual Treatment for Period 01 (N)
AAGE	Analysis Age
SEMMTYP	MM subtype at study entry
LINPTOT	Total Number of Prior Treatment Lines
CREATBL	Baseline Serum Creatinine
RESPONDER	Responder (1: responder; 0: non-responder)

Program 1. Create a simulated dataset

```
### Data Simulations ###
library(utils)
# Treatment ARM (N=30)
set.seed(456)
# 1. USUBJID (Unique Subject Identifier)
USUBJID <- paste0("SAR12345-100-0001-", sprintf("%05d", seq_len(30)))

# 2. AAGE (Analysis Age)
AAGE<-round(rnorm(30, mean = 61.5, sd = 8),digits = 0)

# 3. LINPTOT (Total Number of Prior Treatment Lines)
LINPTOT<-rpois(n = 30, lambda = 4)

# 4. SEMMTYP (MM subtype at study entry)
SEMMTYP<-sample(x = c("Ig G","non-Ig G"), size = 30, replace = TRUE, prob
= c(0.5, 0.5))

# 5. CREATBL (Baseline Serum Creatinine)
CREATBL<-sample(x = c("NORMAL","ABNORMAL"), size = 30, replace = TRUE,
prob = c(0.58, 0.42))

# 6. RESPONDER (Responder)
RESPONDER<-rbinom(30, 1, 0.75)

# 7. TRT01AN (Actual Treatment for Period 01 (N))
TRT01AN<-1
```

```r
treatment_group<-data.frame(USUBJID, TRT01AN, AAGE,
LINPTOT,SEMMTYP,CREATBL,RESPONDER)

# External Control ARM
set.seed(456)

# 1. USUBJID (Unique Subject Identifier)
USUBJID <- paste0("SAR12345-100-0002-", sprintf("%05d", seq_len(200)))

# 2. AAGE (Analysis Age)
AAGE<-round(rnorm(200, mean = 59, sd = 10),digits = 0)

# 3. LINPTOT (Total Number of Prior Treatment Lines)
LINPTOT<-rpois(n = 200, lambda = 3.5)

# 4. SEMMTYP (MM subtype at study entry)
SEMMTYP<-sample(x = c("Ig G","non-Ig G",""), size = 200, replace = TRUE,
prob = c(0.47, 0.49, 0.04))

# 5. CREATBL (Baseline Serum Creatinine)
CREATBL<-sample(x = c("NORMAL","ABNORMAL",""), size = 200, replace = TRUE,
prob = c(0.54, 0.24, 0.22))

# 6. RESPONDER (Responder)
RESPONDER<-rbinom(200, 1, 0.3)

# 7. TRT01AN (Actual Treatment for Period 01 (N))
TRT01AN<-2

control_group<-data.frame(USUBJID, TRT01AN, AAGE,
LINPTOT,SEMMTYP,CREATBL,RESPONDER)

#### Combine two datasets
adsl<- rbind(treatment_group, control_group)
utils::write.csv(adsl, file = "xxx/adsl_ps.csv")
```

BEST PRACTICE TO PERFORM PROPENSITY SCORE ANALYSIS

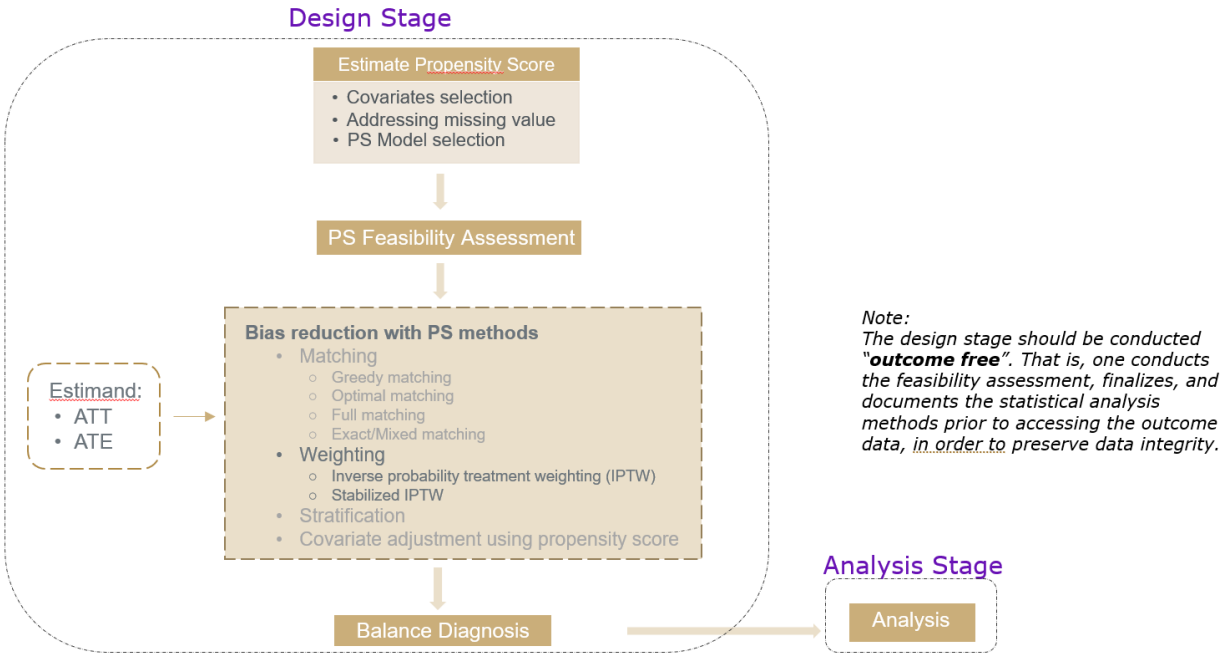


Figure 2. Roadmap

Design Stage: A Process to Mimic Randomization

STEP 1 Estimate Propensity Score

1.1 Covariates selection in the estimation model

The selection of covariates for the estimation model is an important step in propensity score analysis. It requires all baseline confounders should be identified and included appropriately.

Generally, there are two kinds of covariates that could be considered to include in the estimation model:

a. Covariates that are associated with the outcome variable

b. Covariates that are predictive of both treatment assignment and the outcome

Variables in category b are true confounders, and it is assumed that these variables should be selected. But a lot of research proved that including covariates in categories a and b has more advantages. The imbalance result from variable in category a could be addressed by including the variable in the propensity score model. However, in real data analysis, it can be difficult to identify which covariate belongs to category a or b. Some useful methods are proposed to guide the selection of covariates, Directed acyclic graphs (DAGs), expert's opinion and evidence from prior study.

1.2 Addressing missing covariates value

In real world data, the missing values are quite common and unavoidable. As the propensity score is the conditional probability for a subject to be assigned to the treatment group given the measured baseline characteristics, missing data may reduce the statistical power of a study and make the propensity score estimation more challenging.

Two methods are commonly used for handling missing data: complete case analysis (omitting) and multiple imputation.

a. Complete case (CC) analysis: To use only the patients without missing covariates values

   This method could be considered if:

   - The missing values only account for a small proportion (e.g., less than 30%).
   - Sample size is large enough.
   - The missing data are missing completely at random (MCAR).

b. Impute the missing covariates values: multiple imputation

1.3 Construction of Propensity Score Estimation Model

Logistic regression is used as the most common approach to model the binary intervention (treated or control) selection as a function of the measured baseline covariates:

$$log\left(\frac{e_i}{1-e_i}\right) = \beta_0 + \beta_1 * X_{1i} + \cdots + \beta_n * X_{ni} + \varepsilon_i$$

X: covariates

e: the propensity score (value: 0-1)

Notice, it is a simplified model that only contains main effect of the covariates, but interaction terms could be added.

STEP 2 PS Feasibility Assessment

Before propensity score analysis, sufficient PS overlap (Common Support) between treatment groups need to be examined. Imbens and Rubin (2015) stat that differences in the covariate distributions between treatment groups can manifest in some difference of the corresponding propensity score distributions, so comparisons of the propensity score distributions can provide a simple summary of the similarities of patient characteristics between treatments, and such comparisons have become a common part of feasibility assessments. Graphical display of PS Overlap between treatment groups, supplemented with statistical indicators (see below) are recommended.

a. Graphic methods

Histograms are commonly used to display the propensity score distributions of different treatment groups.

b. Standardized mean differences & Variance ratios

The standardized difference in mean (SDM) and the variance ratio (VR) have been used as indicators to quantify the difference in the PS distributions between treatment groups and assess the feasibility of PS analysis (Austin 2009, Stuart et al. 2010):

$$sdm = \frac{\overline{ps}_{treatment} - \overline{ps}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

$$\text{vr} = \frac{Var(PS1)}{Var(PS2)}$$

Austin suggested that standardized differences > 0.1 indicates significant imbalance while Stuart proposed a more conservative value of 0.25. In addition, the acceptable ranges for the ratio of variances of 0.5 to 2.0 have been cited (Austin 2009).

PS trimming to obtain a common support

Trimming patients with extreme PS is necessary, especially for PS weighting method. When PS is close to 0 or 1, the weight can be extremely large and result in inflation of the variance and reliance of the result.

a. Crump et al. *(2009)* state that for many scenarios the simple rule of trimming to an analysis data set including all estimated propensity scores between 0.1 and 0.9 is near optimal to minimizes the variance of the estimated treatment effects with a reduced sample size.

b. In some scenarios, the sample size is large and efficiency in the analysis is of less concern than excluding patients from the analysis. A commonly used approach is to trim only: 1) the Treatment A (treated) patients with propensity scores above the maximum propensity score in the Treatment B (control) group; and 2) the Treatment B patients with propensity scores below the minimum propensity score in the Treatment A group.

STEP 3 Bias reduction with Inverse Probability of Treatment Weighting

Firstly, the estimand of the study should be specified. An estimand defines what is to be estimated, it depends on the medical question of interest you want to address. Below are two most popular estimands, ATE and ATT:

- **Average treatment effect (ATE):** ATE is a commonly used estimand in comparative observational research and is defined as the average difference in the pairs of potential outcomes, averaged over the entire population. ATE can be interpreted as the difference in the outcome of interest had every subject taken treatment A versus had every subject taken treatment B.
- **Average treatment effect of treated (ATT):** Sometimes you are interested in the causal effect only among those who received one intervention of interest ("treated"). In this case the estimand is the average treatment effect of treated (ATT), which is the average difference of the pairs of potential outcomes, averaged over the "treated" population. ATT can be interpreted as the difference in the outcome had every treated subject been "treated," versus the counterfactual outcomes had every "treated" subject taken the other intervention. Notice, in a randomized experiment, ATT is equivalent to ATE.

When the estimand of interest focuses on the **full population** (average treatment effect in the full population, ATE), the weight for patient i is defined as:

$$w_{i,ATE} = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}$$

where Z_i is a flag variable denoting the treatment group (1 = Treated, 0 = Control) and e_i is the propensity score for patient i.

When the estimand of interest focuses on the **treated population** (average treatment effect among the treated, ATT), the weight for patient i is defined as:

$$w_{i,ATT} = Z_i + \frac{(1 - Z_i)e_i}{1 - e_i}$$

where Z_i is a flag variable denoting the treatment group (1 = Treated, 0 = Control) and e_i is the propensity score for patient i.

A concern with IPTW is that patients with a very low probability of their actual treatment have very high weights. Such patients then become very influential in the analysis and greatly increase the variance of any weighted analysis. To offset the impact of extreme weight, the following methods could be used:

a) Trimming

It could be considered if the extreme weights only account for a small percentage in the data (e.g., <10%).

i. Trimming extreme values by a predefined threshold (e.g. trimming weight>10)

ii. Trimming using percentiles: (e.g. trimming weight values less than the 1st percentile and greater than the 99th percentile)

b) Winsorization

If you prefer not to reduce the sample size by trimming, then winsorization can be an alternative.

i. Replace any weight above a pre-specified value R by R (e.g: Replace weight>10 by 10)

ii. Replacing any value less than the 1st percentile by the 1st percentile value and any value above the 99th percentile by the 99th percentile value

iii. Replacing any value less than the 5th percentile by the 5th percentile value and any value above the 95th percentile by the 95th percentile value

c) Stabilized weights

$$w_{i,StATE} = \frac{Z_i P(Z=1)}{e_i} + \frac{(1-Z_i)P(Z=0)}{1-e_i}$$

P(z=1) indicate the proportion of treated patients and P(z=0) indicate the proportion of control patients

Z_i is a flag variable denoting the treatment group (1 = Treated, 0 = Control) and e_i is the propensity score for patient i.

STEP 4 Balance Diagnosis

Austin (2009) states that comparing the balance between treatment groups for each potential confounder after the propensity score adjustment is the best approach to judge the success of the PS adjustment for measured baseline confounders.

Standardized mean differences is considered as the good statistics to assess the balance produced by the propensity score, because it is independent of sample size and allows for the comparison of the relative balance of variables that are measured in different units. As two very different distributions can still produce a standardized difference in means of zero (Tipton 2014), it is advisable to supplement the SDM with the variance ratio. Austin (2011) defines the standardized mean difference for continuous and binary covariates:

Continuous covariates:

$$sdm = \frac{\bar{x}_{treatment} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

Weighted mean $\bar{x}_w = \sum w_i x_i \Big/ \sum w_i$

Weighted standard deviation $s_w = \sqrt{\frac{w_i (x_i - \bar{x}_w)2}{(n-1)\sum w_i / n}}$

Binary covariates:

$$sdm = \frac{\hat{p}_{treatment} - \hat{p}_{control}}{\sqrt{\frac{\hat{p}_{treatment}(1-\hat{p}_{treatment}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}}$$

Variance Ratio

$$\text{vr} = \frac{s^2 treatment}{s^2 control}$$

Graphical methods (e.g. bar charts, box plots, cloud plots and cumulative distribution plots, etc) are also commonly used.

In practice, some imbalance on select covariates could be also observed after the above steps and the balance assessment might become an iterative process. To address the residual imbalance, you could re-construct the propensity model, trimming the population with other rules, using matching or stratification, and incorporate the covariates with imbalance into the analysis phase.

ANALYSIS STAGE: PROPENSITY SCORE WEIGHTING FOR ESTIMATING CAUSAL TREATMENT EFFECTS

STEP 5 Causal Treatment Effects Estimation

Risk Difference:

$$\widehat{\delta_{ATE}} = \left(\sum_i \frac{Z_i}{\widehat{e_i}}\right)^{-1} \sum_i \frac{Z_i Y_i}{\widehat{e_i}} - \left(\sum_i \frac{1-Z_i}{1-\widehat{e_i}}\right)^{-1} \sum_i \frac{(1-Z_i)Y_i}{1-\widehat{e_i}}$$

$$\widehat{\delta_{ATT}} = \left(\sum_i Z_i\right)^{-1} \sum_i Z_i Y_i - \left(\sum_i \frac{(1-Z_i)\widehat{e_i}}{1-\widehat{e_i}}\right)^{-1} \sum_i \frac{(1-Z_i)\widehat{e_i} Y_i}{1-\widehat{e_i}}$$

Risk Ratio:

$$\widehat{\delta_{ATE}} = \left(\sum_i \frac{Z_i}{\widehat{e_i}}\right)^{-1} \sum_i \frac{Z_i Y_i}{\widehat{e_i}} \Bigg/ \left(\sum_i \frac{1-Z_i}{1-\widehat{e_i}}\right)^{-1} \sum_i \frac{(1-Z_i)Y_i}{1-\widehat{e_i}}$$

$$\widehat{\delta_{ATT}} = (\sum_i Z_i)^{-1} \sum_i Z_i Y_i \Bigg/ \left(\sum_i \frac{(1-Z_i)\widehat{e_i}}{1-\widehat{e_i}}\right)^{-1} \sum_i \frac{(1-Z_i)\widehat{e_i} Y_i}{1-\widehat{e_i}}$$

Odds Ratio:

$$\widehat{\delta_{ATE}} = \frac{\left(\sum_i \frac{Z_i}{\widehat{e_i}}\right)^{-1} \sum_i \frac{Z_i Y_i}{\widehat{e_i}} \Big/ \left(1 - \left(\sum_i \frac{Z_i}{\widehat{e_i}}\right)^{-1} \sum_i \frac{Z_i Y_i}{\widehat{e_i}}\right)}{\left(\sum_i \frac{1-Z_i}{1-\widehat{e_i}}\right)^{-1} \sum_i \frac{(1-Z_i)Y_i}{1-\widehat{e_i}} \Big/ \left(1 - \left(\sum_i \frac{1-Z_i}{1-\widehat{e_i}}\right)^{-1} \sum_i \frac{(1-Z_i)Y_i}{1-\widehat{e_i}}\right)}$$

$$\widehat{\delta_{ATT}} = \frac{(\sum_i Z_i)^{-1} \sum_i Z_i Y_i / (1 - (\sum_i Z_i)^{-1} \sum_i Z_i Y_i)}{\left(\sum_i \frac{(1-Z_i)\hat{e}_i}{1-\hat{e}_i}\right)^{-1} \sum_i \frac{(1-Z_i)\hat{e}_i Y_i}{1-\hat{e}_i} \Big/ (1 - \left(\sum_i \frac{(1-Z_i)\hat{e}_i}{1-\hat{e}_i}\right)^{-1} \sum_i \frac{(1-Z_i)\hat{e}_i Y_i}{1-\hat{e}_i})}$$

Standard error estimation

Use of weights induces within-subjects correlation, which invalidates the naïve SE estimation. The risk is to underestimate of the standard error, which would lead to too narrow confidence interval and type I error inflation.

As the estimator is RR and not sure if normal approximation performs well, I use the 97.5% and 2.5% quantiles of the simulated samples directly to obtain the 95% CI.

EXAMPLE OF IPTW ANALYSIS USING SIMULATED DATA

Based on the prior evidence in Phase 3 study, AAGE, SEMMTYP, LINPTOT and CREATBL are selected as the covariates in the propensity score model. In the simulated data, 24% participants are with missing CREATBL and 4.5% participants are with missing SEMMTYP, all are from external control arm. To address the missing covariates value, complete case method is used. Program 2 provide the SAS code to conduct the IPTW analysis.

Program 2. SAS code to perform propensity score weighting analysis

```
/* Read the data */
proc import datafile="xxx/adsl_ps.csv"
    out=sim
    dbms=csv
    replace;
    getnames=yes;
run;

/******DESIGN STAGE******/
****STEP 1 ESTIMATE PROPENSITY SCORE;
******---Covariates selection in the estimation model:
Based on the prior evidence in Phase 3 study, AAGE, SEMMTYP, LINPTOT and
CREATBL are selected as the covariates in the propensity score model;
******---Addressing missing covariates value: Complete case method is
used;
data ps;
    set sim;

    if aage=. or semmtyp='' or  linptot=. or  creatbl='' then delete;

    drop var1;
run;

*****---Construction of Propensity Score Estimation Model;
proc logistic data=ps;
   class trt01an semmtyp creatbl ;
```

```
   model trt01an(event="1")=aage semmtyp linptot creatbl;
   output out=pscore pred=PS;
run;

*****STEP 2&3&4 Bias reduction with IPTW and balance diagnosis;
*** Method 1 IPTW - ATE weights;
title1 'IPTW without trimming';
ods graphics on;
proc psmatch data=pscore region=allobs;
 class trt01an semmtyp creatbl;
 psdata treatvar = trt01an(treated='1') ps = ps;
 psweight weight = atewgt(stabilize=No) nlargestwgt=10;
 assess ps var = (aage semmtyp linptot creatbl)/plots=all stddev =
pooled(allobs=no) ;
 output out(obs=all)=iptw1 weight=atewgt;
run;
ods graphics off;

*** Method 2 Stabilized IPTW - ATE weights;
title1 'Stabilized IPTW';
ods graphics on;
proc psmatch data=pscore region=allobs;
 class trt01an semmtyp creatbl;
 psdata treatvar = trt01an(treated='1') ps = ps;
 psweight weight = atewgt(stabilize=Yes) nlargestwgt=10;
 assess ps var = (aage semmtyp linptot creatbl)/plots=all stddev =
pooled(allobs=no) ;
 output out(obs=all)=iptw2 weight=atewgt;
run;
ods graphics off;

*** Method 3 IPTW with winsorization - ATE weights (Replace weight>10 by
10);
data pscore_winsorization;
    set pscore;
    if trt01an=1 and ps<0.1 then ps=0.1;
    if trt01an=2 and ps>0.9 then ps=0.9;
run;

title1 'IPTW with winsorization';
ods graphics on;
proc psmatch data=pscore_winsorization region=allobs;
 class trt01an semmtyp creatbl;
 psdata treatvar = trt01an(treated='1') ps = ps;
 psweight weight = atewgt(stabilize=No) nlargestwgt=10;
```

```
 assess ps var = (aage semmtyp linptot creatbl)/plots=all stddev = pooled
(allobs=no);
 output out(obs=all)=iptw3 weight=atewgt;
run;
ods graphics off;

*Conclusion: Balance achieved with Method 3;

/******ANALYSIS STAGE******/
*Compute RR;
proc freq data=iptw3;
    tables trt01an*responder/nocum norow relrisk;
    weight atewgt;
    output out=rr_estimate(keep=_rrc2_ rename=(_rrc2_=rr)) relrisk2;
run;

data rr_ci;
    attrib RR  length=8 format=best12. label="Relative Risk";
    stop;
run;

*95% CI for RR with bootstrap;
%macro iptw(data=,n=,seed=);
proc surveyselect data=&data. out=sample_all noprint
     seed=&seed. outhits
     method=urs /*requests unrestricted random sampling, which is
selection with equal probability and with replacement. For more
information, see the section*/
     samprate=1 reps=&n.;
     strata trt01an;
run;

%do i=1 %to &n.;
data sample;
  set sample_all;
  if replicate = &i.;
run;

title1 'IPTW with winsorization';
ods graphics on;
proc psmatch data=sample region=allobs;
 class trt01an semmtyp creatbl;
 psmodel trt01AN(Treated="1") = aage linptot semmtyp creatbl;
 psweight weight = atewgt(stabilize=No) nlargestwgt=10;
```

```
 assess ps var = (aage semmtyp linptot creatbl)/plots=all stddev = pooled
(allobs=no);
 output out(obs=all)=iptw3_rep weight=atewgt ps=ps;
run;
ods graphics off;

data iptw3_rep2;
    set iptw3_rep;
    atewgt_ana=atewgt;

    if atewgt > 10 then atewgt_ana = 10;
run;

proc sort data=iptw3_rep2;
   by trt01an responder;
run;

*95% CI for RR;
proc freq data=iptw3_rep2;
    tables trt01an*responder/nocum norow relrisk;
    weight atewgt_ana;
    output out=rr_rep(keep=_rrc2_ rename=(_rrc2_=rr)) relrisk2;
run;
data rr_ci;
  set rr_ci rr_rep;
run;

proc sql;
  drop table rr_rep;
quit;

%end;
%mend;

%iptw(data=ps,n=5000,seed=17227);

proc univariate data=rr_ci;
    var RR;
    output out=rr_ci2
    pctlpts = 2.5 97.5
    pctlpre = P_;
run;

data rr_ci3;
   length __col_0 $2000 __col_1 $200;
```

```
   set rr_estimate(in=a) rr_ci2(in=b);
   if a then do;
   __col_0="Risk ratio";
   __col_1=strip(put(rr,8.2));
   end;
   if b then do;
   __col_0="  95% CI";
   __col_1= strip(put(P_2_5,8.2))||' to '||strip(put(P_97_5,8.2));
  end;

  keep __col:;
run;
```

The following tables and figures contain the output from **Method 3 IPTW with winsorization** pertaining to the assessment of the propensity score and the associated inverse probability of treatment weights. From Figure 3, Observations=All shows the unweighted sample and Observations= Weighted shows the weighted sample. After weighting, the standardized differences and variance ratios demonstrate improved and reasonable balance: standardized differences all reduced from the unweighted sample and are within 0.1, and variance ratios are between 0.5 and 2.0. Figure 4 shows the standardized mean differences by plot and allows for a graphical comparison of the balance of each covariate before and after implementing IPTW analysis. The cloud plot for the distribution of the propensity scores is provided in Figure 5. This demonstrates the substantial overlap in propensity score for the two groups despite the clearly different distributions.

The estimated treatment effect are provided in Table 2. Risk ratio is 1.99 (95% CI: 1.40 to 2.66), suggesting the ORR in the treatment arm is ~1.99 folds higher than external historical arm.

Figure 3. IPTW analysis covariate balance

IPTW with winsorization

The PSMATCH Procedure

Standardized Mean Differences (Treated - Control)						
Variable	Observations	Mean Difference	Standard Deviation	Standardized Difference	Percent Reduction	Variance Ratio
Prop Score	All	0.077655	0.102214	0.759732		1.4410
	Region	0.077655	0.102214	0.759732	0.00	1.4410
	Weighted	0.020414	0.095447	0.213880	71.85	0.9359
AAGE	All	4.965517	9.422984	0.526958		0.9254
	Region	4.965517	9.422984	0.526958	0.00	0.9254
	Weighted	0.959876	9.324081	0.102946	80.46	0.8663
LINPTOT	All	0.386207	1.970897	0.195955		0.8885
	Region	0.386207	1.970897	0.195955	0.00	0.8885
	Weighted	0.187152	1.869831	0.100090	48.92	0.7274
SEMMTYP	All	0.141379	0.477788	0.295904		0.8517
	Region	0.141379	0.477788	0.295904	0.00	0.8517
	Weighted	0.030173	0.490508	0.061514	79.21	0.9763
CREATBL	All	0.062069	0.481526	0.128901		1.0727
	Region	0.062069	0.481526	0.128901	0.00	1.0727
	Weighted	0.005839	0.478005	0.012216	90.52	1.0075

Figure 4. Balance Assessment: Standardized mean differences plot for all variables

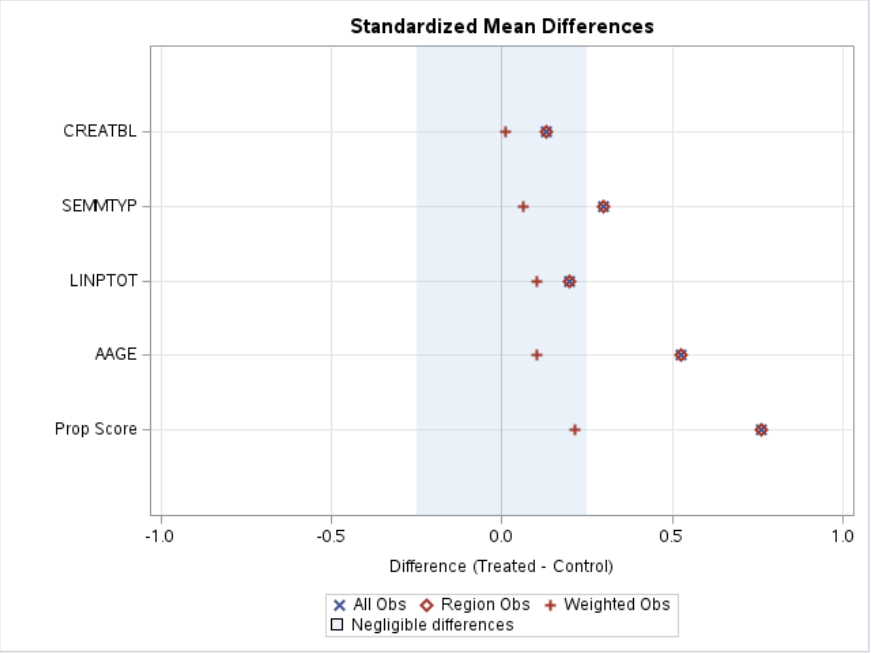


Figure 5. Cloud Plot of Propensity Score Distributions

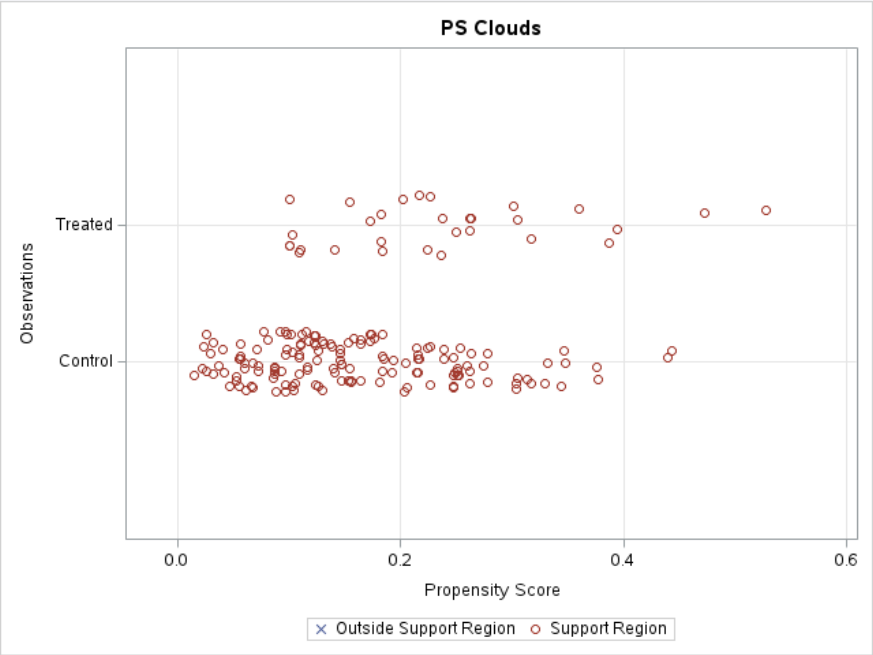


Table 2. IPTW Analysis: Estimated Treatment Effects

	Treatment	External control
Risk ratio	1.99	
95% CI	1.40 to 2.66	

Program 3 provide the R code to conduct the IPTW analysis. The estimated risk ratio is 1.99 (95% CI: 1.38 to 2.66), very similar to the result computed by SAS.

Program 3. R code to perform propensity score analysis

```
#Load the packages
library(haven)
library(dplyr)
library(survey)
library(tableone)
library(WeightIt)
library(cobalt)
library(boot)


######---DESIGN STAGE---######
#Recode TRT01AN
adsl$TRT01AN = ifelse(adsl$TRT01AN == 2, 0 ,1)

# Convert to factor;
adsl$SEMMTYP <- factor(adsl$SEMMTYP,levels=c("non-Ig G","Ig G"))
adsl$CREATBL <- factor(adsl$CREATBL,levels=c("ABNORMAL","NORMAL"))

adsl_ana <- adsl[complete.cases(adsl[, c("AAGE", "LINPTOT", "SEMMTYP",
"CREATBL")]), ]

# Balance diagnosis before implementing IPTW
covars <- c("AAGE","LINPTOT","SEMMTYP","CREATBL")
Before <- CreateTableOne(vars = covars,strata="TRT01AN", data=adsl_ana)
print(Before, smd=TRUE)

#  Method 1 IPTW - ATE weights
m1 <- weightit(TRT01AN ~ AAGE+LINPTOT+SEMMTYP+CREATBL, data=adsl_ana,
method ="glm", estimand = "ATE")
love.plot(m1, abs=TRUE,stats = c("mean.diff", "variance.ratio"),thresholds
= c(m = .1, v = 2), star = "raw")

Svy.design <- svydesign(ids = ~ 1, data = adsl_ana, weights = m1$weights)
After <- svyCreateTableOne(vars = covars, strata = "TRT01AN", data =
Svy.design, test = FALSE)
print(After, smd=TRUE)

#  Method 2 Stabilized IPTW - ATE weights;
m2 <- weightit(TRT01AN ~ AAGE+LINPTOT+SEMMTYP+CREATBL, data=adsl_ana,
method ="glm", estimand = "ATE",stabilize = TRUE)
```

```
love.plot(m2, abs=TRUE,stats = c("mean.diff", "variance.ratio"),thresholds
= c(m = .1, v = 2), star = "raw")

Svy.design_s <- svydesign(ids = ~ 1, data = adsl_ana, weights =
m2$weights)
After_s <- svyCreateTableOne(vars = covars, strata = "TRT01AN", data =
Svy.design_s, test = FALSE)
print(After_s, smd=TRUE)

#  Method 3 IPTW with winsorization - ATE weights (Replace weight>10 by
10);
m3 <- m1
m3$weights <- pmin(m3$weights, 10)
love.plot(m3, abs=TRUE,stats = c("mean.diff", "variance.ratio"),thresholds
= c(m = .1, v = 2), star = "raw")

Svy.design_w <- svydesign(ids = ~ 1, data = adsl_ana, weights =
m3$weights)
After_w <- svyCreateTableOne(vars = covars, strata = "TRT01AN", data =
Svy.design_w, test = FALSE)
print(After_w, smd=TRUE)

# Conclusion: Balance achieved with Method 3 #

######---ANALYSIS STAGE---######
# Estimate RR and 95% CI with Bootstrap

# Define bootstrap function
boot_rr <- function(data, indices) {
  # Use resampling data
  d <- data[indices, ]

  # Calculate weight
  m_boot <- weightit(TRT01AN ~ AAGE + LINPTOT + SEMMTYP + CREATBL, data=d,
                     method ="glm", estimand = "ATE")

  # Fit the model
  m_boot2 <- m_boot
  m_boot2$weights <- pmin(m_boot2$weights, 10)

  boot_data <- d %>%
    mutate(weights = m_boot2$weights)

  boot_data$weights_resp <- boot_data$weights * boot_data$RESPONDER
```

```
  boot_data.t <- boot_data[which(boot_data$TRT01AN==1),]
  boot_data.c <- boot_data[which(boot_data$TRT01AN==0),]

  #Calculate Risk ratio
  RR <- (sum(boot_data.t$weights_resp)/sum(boot_data.t$weights))/
        (sum(boot_data.c$weights_resp)/sum(boot_data.c$weights))
}
set.seed(17227)
# Excute bootstrap
boot_results <- boot(adsl_ana, boot_rr, R = 5000)

# Return RR
orig <- boot_results$t0

# Compute bootstrap CI
ci <- boot.ci(boot_results, type = "perc")
perc_lower <- ci$percent[4]
perc_upper <- ci$percent[5]

# Print result
print(paste("Risk Ratio:", round(orig, 2),
            "(95% CI:", round(perc_lower, 2), "-", round(perc_upper, 2),
")"))

#[1] "Risk Ratio: 1.99 (95% CI: 1.38 - 2.66 )"
```

CONCLUSION

This paper has presented three different approaches to adjust for confounders in observational studies: IPTW, Stabilized IPTW and IPTW with winsorization. However, in practice, some imbalance on select covariates may still exist. Then you could revise the propensity score model, trimming the population, using exact matching or stratification on a critical covariate, and incorporate the covariates with imbalance into the analysis phase to address the residual imbalance. This is an iterative process aimed to find a good way to produces balance in measured confounders.

ACKNOWLEDGMENTS

The authors would like to thank Qiuyan Wang, Monica Li and Liang Zhao for their technical assistance with the statistical analysis. We are also grateful to all colleagues who contributed to this project to deliver this drug to Chinese patients in a short time via real world data.

REFERENCES

Austin P (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. Communications in Statistics: Simulation and Computation 38(6):1228–1234

Austin PC (2009). Balance Diagnostics for Comparing the Distribution of Baseline Covariates Between Groups Propensity-score Matched Samples. Stat Med 28(25): 3083-3107.

Austin PC and Stuart EA (2015). Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies. Stat Med 34:3661-3679.

Baser O (2007). Propensity Score Matching with Limited Overlap. Economics Bulletin, 9(8):1-8.

Brookhart MA, et al. (2006). Variable selection for propensity score models. American journal of epidemiology 163(12): 1149-1156.

Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2009). Dealing with Limited Overlap in the Estimation of Average Treatment Effects. Biometrika 96(1): 187-199.

Imbens GW and Rubin DB (2015). Causal Inference for Statistics, Social, and Biomedical Sciences. New York: Cambridge University Press. Kish L (1965). Survey Sampling. New York: Wiley.

Shrier I, Platt RW, Steele RJ (2007). Re: Variable selection for propensity score models. American journal of epidemiology 166(2): 238-239

Stuart E (2010). Matching Methods for Causal Inference. A Review and a Look Forward. Stat Sci. 25(1): 1–21.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jun Li
Jun1.Li@sanofi.com