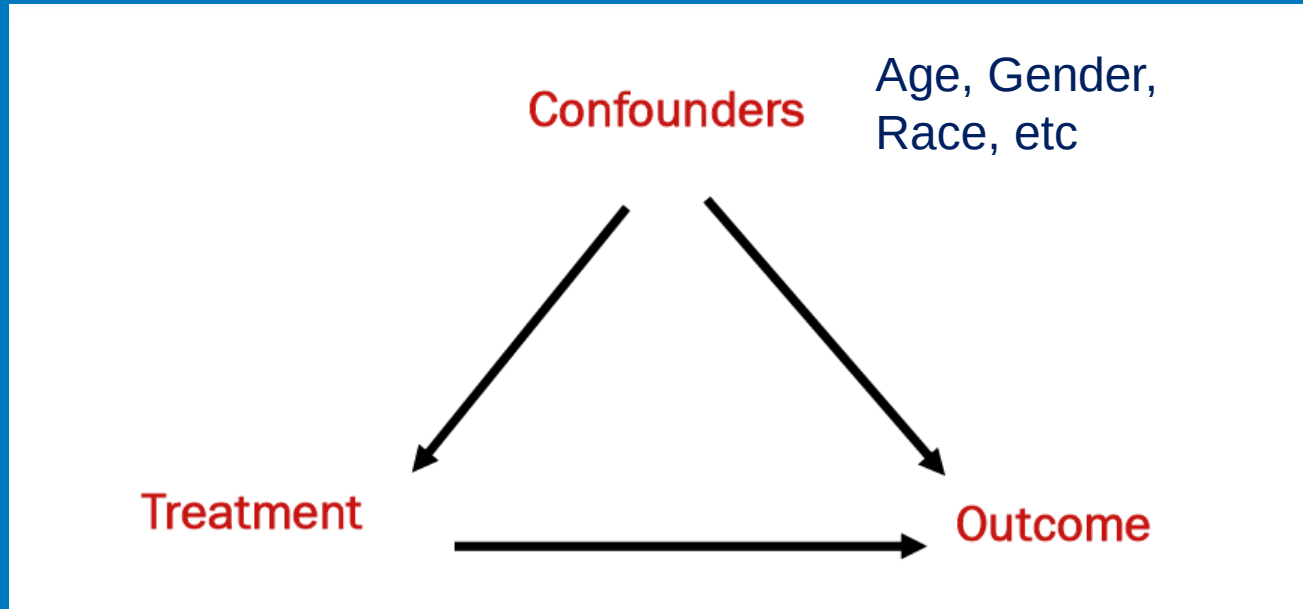


Propensity Score Matching Methods in SAS®

Nianping Hu
Health Quality Council



Confounding Bias



Confounding Bias is Responsible for Treatment and Outcome.
Confounding can Decrease, Increase, or even Change the Direction of Relation.

Strategies to Control the Confounding

- **Design Stage**

- Randomization
- Restriction
- Matching

- **Analysis Stage**

- Standardisation
- Stratification
- Matched analysis
- Adjustments (regression analysis)

Propensity Score (Why)

- If there are multiple confounders in the model, control the confounders becomes complicated and impossible.
- Propensity score is generated to convert multiple confounders in a single dimension (score) to reduce the confounding bias.

Propensity Score (When)

Strategies to control the confounding

Measured confounders

Design

- Restriction
- Matching

Analysis

- Standardization
- Stratification
- Multivariate Regression

Amenable to
Propensity
technique

Unmeasured confounders

Unmeasured but measurable in a validation study

- Two-stage sampling
- External adjustment

Unmeasurable

Design

- Crossover designs
- Active comparison group (restriction)

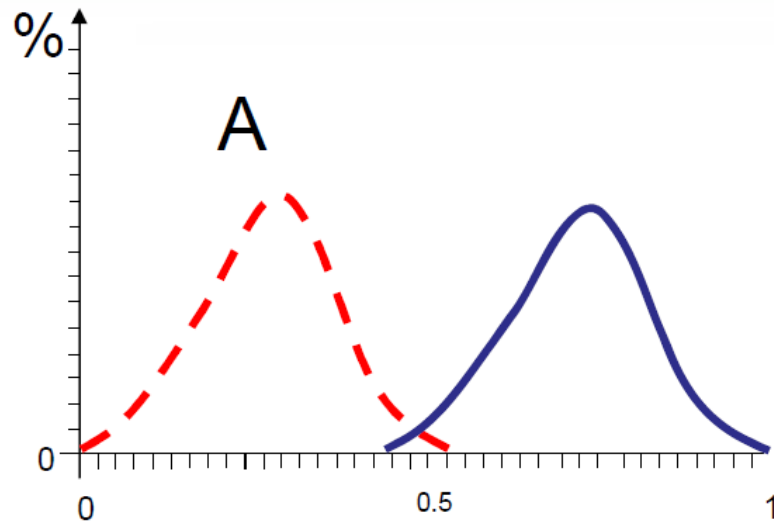
Analysis

- Instrumental Variables
- Sensitivity analysis

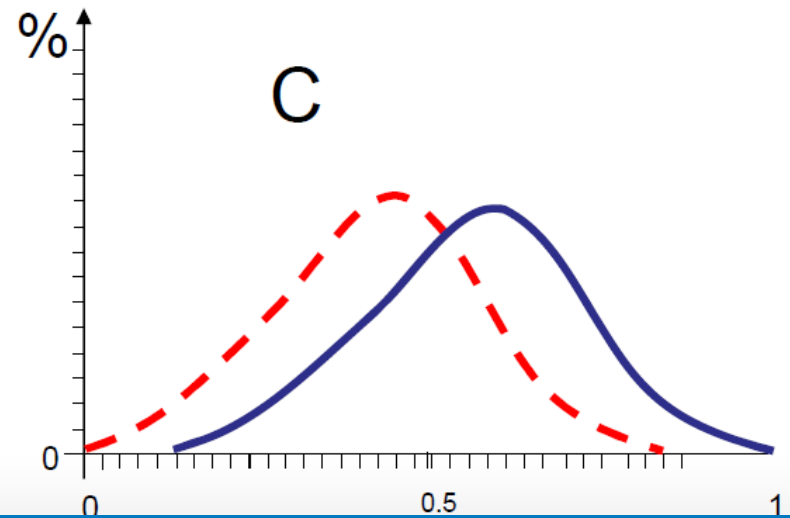
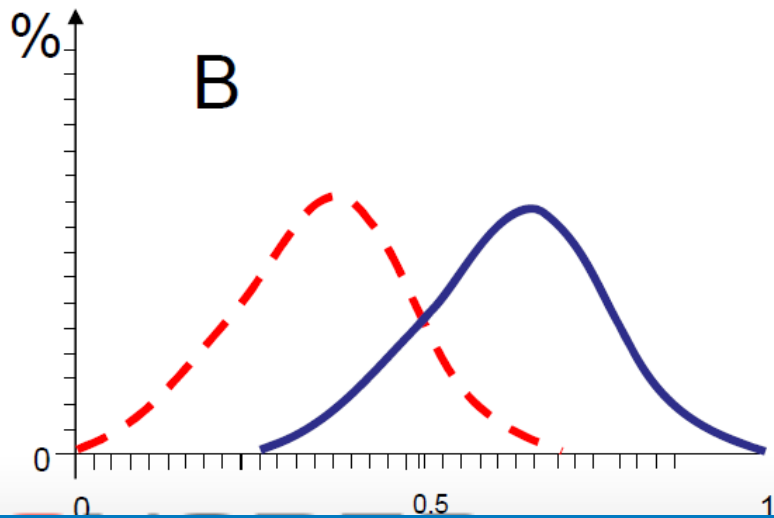
Introduction

- Propensity score is predicted probability of treatment.
- The range for propensity score is from 0 to 1.
- Propensity score is a balancing score for observed covariates.

Introduction



- A Extensive selection of treatment
- B Moderate selection of treatment
- C Little preference for treatment



Propensity Score (How)

STEP 1

- **Demographic covariates**
 - Include basic demographic information (age, sex, race, etc)
- **Additional covariates**
 - can then be identified from the various data dimensions (Such as inpatients and outpatients files).

STEP 2

- **Identify empirical candidate covariates**
- Identify the n most prevalent codes within each data dimension.
- The top n most prevalent codes were identified as candidate empirical covariates. If fewer than 100 patients were identified with a covariate, the covariate was dropped.

Example

```
*-----;
* within a data source, calculate how frequently a code occurs ; * then rank by frequency and add the variables to the output dataset. ;
*-----;
%macro hdps_get_var_frequencies(type, source_table, source_field, source_type_field);
*-----;
proc sort data = &source_table out = t_unique_by_pat(keep = &var_patient_id &source_field)    nodupkey;
    by &var_patient_id &source_field;    where &source_field is not null;
run;
proc freq data = t_unique_by_pat noprint;    table &source_field / out = t_frequency;
run;
*****;
* require at least 100 instances ;
*****;
data t_newvar;
    set t_frequency(rename = (Count = frequency)); where frequency >= 100;
    length item $50; %if &source_type_field = C %then Item = &Source_Field;;
    %if &source_type_field = N %then Item = compress(put(&Source_Field,32.));;
    Temp_id = RANUNI(-1); Type = "&type";
    Source_table = "&source_table"; Source_field = "&source_field";
    consider_for_ps = 0;    drop Percent &Source_Field;
run;
*****;
* create the new variables!    ;
*****;
data &result_diagnostic;    set &result_diagnostic    t_newvar;
run;
%mend;
*-----;
* remove any variables that aren't in the top N most frequent ;
*-----;
%macro hdps_filter_top_n;
*-----;
```

STEP 3

- **Assess code recurrence and create**
- indicator variables for each patient, for each identified code, address frequency by creating 3 variables:
- Once = 1 if that code appeared at least once within 180 days.
- Sporadic = 1 if code appeared at least more than the median.
- Frequent = 1 if code appeared at least more than the 75th percentile.

Example

```
• proc means data = &&Name&I.._Codes noprint;
•   var %do J = 1 %to &NCodes; I_&&VarId&J %end;;
•   output out = Median Median = %do J = 1 %to &NCodes; M_&&VarId&J %end;
•   Q3 = %do J = 1 %to &NCodes; Q3_&&VarId&J %end;;
• run;
• data _null_;   set Median;
•   %do J = 1 %to &NCodes;
•     call symput("Median" || compress(put(&J,8.)), compress(put(M_&&VarId&J,8.)));
•     call symput("Q3_"   || compress(put(&J,8.)), compress(put(Q3_&&VarId&J,8.)));
•   %end;
• run;
• data &&Name&I.._Codes;   set &&Name&I.._Codes;
•   %do J = 1 %to &NCodes; select;
•     when (I_&&VarId&J = .)           do; I_&&VarId&J = 0; I_%eval(&&VarId&J+1) = 0; I_%eval(&&VarId&J+2) = 0; I_%eval(&&VarId&J+3) = 0;
end;
•     when (I_&&VarId&J >= &&Q3_&J)   do; I_&&VarId&J = 1; I_%eval(&&VarId&J+1) = 1; I_%eval(&&VarId&J+2) = 1; I_%eval(&&VarId&J+3) = 1;
end;
•     when (I_&&VarId&J >= &&Median&J) do; I_&&VarId&J = 1; I_%eval(&&VarId&J+1) = 1; I_%eval(&&VarId&J+2) = 1; I_%eval(&&VarId&J+3) = 0;
end;
•     when (I_&&VarId&J >= 1)         do; I_&&VarId&J = 1; I_%eval(&&VarId&J+1) = 1; I_%eval(&&VarId&J+2) = 0; I_%eval(&&VarId&J+3) = 0;
end;
•   end;
•   if I_&&VarId&J ^= 0 then do;
•     if &&Median&J = &&Q3_&J then I_%eval(&&VarId&J+3) = .; if &&Median&J = 1   then I_%eval(&&VarId&J+2) = .; end;
•   %end;
•   keep &var_patient_id %do J = 1 %to &NCodes; I_&&VarId&J I_%eval(&&VarId&J+1) I_%eval(&&VarId&J+2) I_%eval(&&VarId&J+3) %end;;
• run;
• %end;
• data &result_diagnostic; set &result_diagnostic; array FType{3} $ 20;
•   FType1 = "once"; FType2 = "sporadic"; FType3 = "frequent"; output;
•   do I = 1 to 3; id = id + 1; frequency_type = FType(I); consider_for_ps = 1; output; end; drop FType1 - FType3 I;
• run;
```

STEP 4

- **Prioritize covariates**
- Patient characteristics related to the exposure but not outcome might cause variance and doesn't improve confounding control, sometimes maybe even introduce confounding.
- We chose prioritize covariates depending on multiplicative bias assessment within data dimensions base on their potential for controlling confounding, not on exposure and other covariates.

Example

```
%macro hdps_CalcBias;
data Bias;  set &output_detailed end = eof;  retain %do I = 1 %to &NItem;
%do J = 0 %to 3;
    Count%eval(&I*10+&J)_c0e0 Count%eval(&I*10+&J)_c0e1 Count%eval(&I*10+&J)_c1e0 Count%eval(&I*10+&J)_c1e1
    Count%eval(&I*10+&J)_c0d0 Count%eval(&I*10+&J)_c0d1 Count%eval(&I*10+&J)_c1d0 Count%eval(&I*10+&J)_c1d1
%end; %end;;

.....
*****;
* capture the prevalence of the confounder in the exp and unexp ;
*****;
if Exposed = 0 then do;  %do I = 1 %to &NItem; %do J = 0 %to 3;
    if I_%eval(&I*10+&J) = 0 then Count%eval(&I*10+&J)_c0e0 + 1;  if I_%eval(&I*10+&J) = 1 then Count%eval(&I*10+&J)_c1e0 + 1;
%end; %end;  end;

.....
*-----;
* calculate parameters about each varialbe, including its bias ;
*-----;
if eof then do;  %do I = 1 %to &NItem; %do J = 0 %to 3;  do I = 1 to 16;  Temp(I) = .;  end;

.....
*-----;
* calculate the Schelesselman bias ;  * Biasmult = [PC1(RRCD-1)+1]/ [PC0(RRCD-1)+1] ;
*-----;
bias_num = (pc_exp * (rr_cd - 1)) + 1; bias_denom = (pc_unexp * (rr_cd - 1)) + 1; bias_mult = bias_num / bias_denom;
bias_add = rd_ce * rd_cd;  abs_log_bias_mult = ABS(LOG(bias_mult));  abs_bias_add = ABS(bias_add);

.....
run;
%mend;
```

STEP 5

- **Select covariates**
- We included covariates from step 1, the top k covariates from step 4, which could be as large as $p*n*3$ when including all candidate covariates.
- $p = 8$ and $n = 200$ resulting in 4,800 candidate covariates. We selected the top $k = 500$ binary empirical covariates.

Example

```
*****;  
* only consider variables with the most bias on the mult. scale ;  
*****;  
  
proc rank data = t_ps_calc descending out = t_ps_calc;  
  var abs_log_bias_mult;  
  ranks ps_rank;  
run;  
  
*****;  
* take out the vars below the ranking threshold *;  
*****;  
  
data t_ps_calc;  
  set t_ps_calc;  
  where ps_rank <= &k;  
run;  
  
data _null_;  
  set t_ps_calc end = eof;  
  call symput ("ps_vars" || compress(put(_n_,8.)),"i_" || compress(put(id,8.)));  
  if eof then call symput ("NVars",compress(put(_n_,8.)));  
run;
```

STEP 6

- **Estimate propensity score (the hd-PS)**
- Estimate propensity score using multivariate logistic regression, including all investigator-defined covariates and the k hd-PS-selected covariates.

Example

```
*****;  
* calculate the propensity score *;  
*****;  
ods listing close;  
  
ods output Association = C ParameterEstimates =Estimate_PS;  
proc logistic data = &output_detailed(keep = exposed %do I = 1 %to &NClass; &&ClassVar&i %end;  
                                     %do I = 1 %to &NVars; &&ps_vars&I %end;  
                                     %do I = 1 %to &NFix; &&FixVar&I %end; outcome &var_patient_id)  
    descending;  
    class %do I = 1 %to &NClass; &&ClassVar&i %end;;  
    model exposed = %do I = 1 %to &NFix; &&FixVar&I %end;  
                  %do I = 1 %to &NVars; &&ps_vars&I %end;  
    output out = &output_scored_cohort(drop = %do I = 1 %to &NVars; &&ps_vars&I %end;)   
    pred=ps xbeta=logit_ps;  
run;  
  
ods listing;
```

STEP 7

- **Use the hd-PS**
- Use the hd-PS in an outcome model as any PS would be used (matching, deciling, trimming, etc.)

Example

```
%RunHighDimPropScore (  
  
    var_patient_id      = key_hsn,  
    var_exposure        = exposed,  
    var_outcome         = outcome,  
    vars_demographic    = age sex race,  
    vars_force_categorical =,  
    vars_ignore         =,  
    vars_predefined     =,  
    top_n               = 200,  
    k                   = 500,  
    input_cohort        = specific_cohort  
    input_dim1          = drug_claims      generic_name,  
    input_dim2          = outpatient_diagnoses icd9_dx,  
    input_dim3          = inpatient_diagnoses icd9_dx,  
    input_dim4          = inpatient_procedures icd9_proc,  
    input_dim5          = outpatient_procedures cpt,  
    output_scored_cohort = scored_cohort,  
    output_detailed     = detailed_cohort,  
    results_diagnostic  = variable_info  
);
```

Figure 1

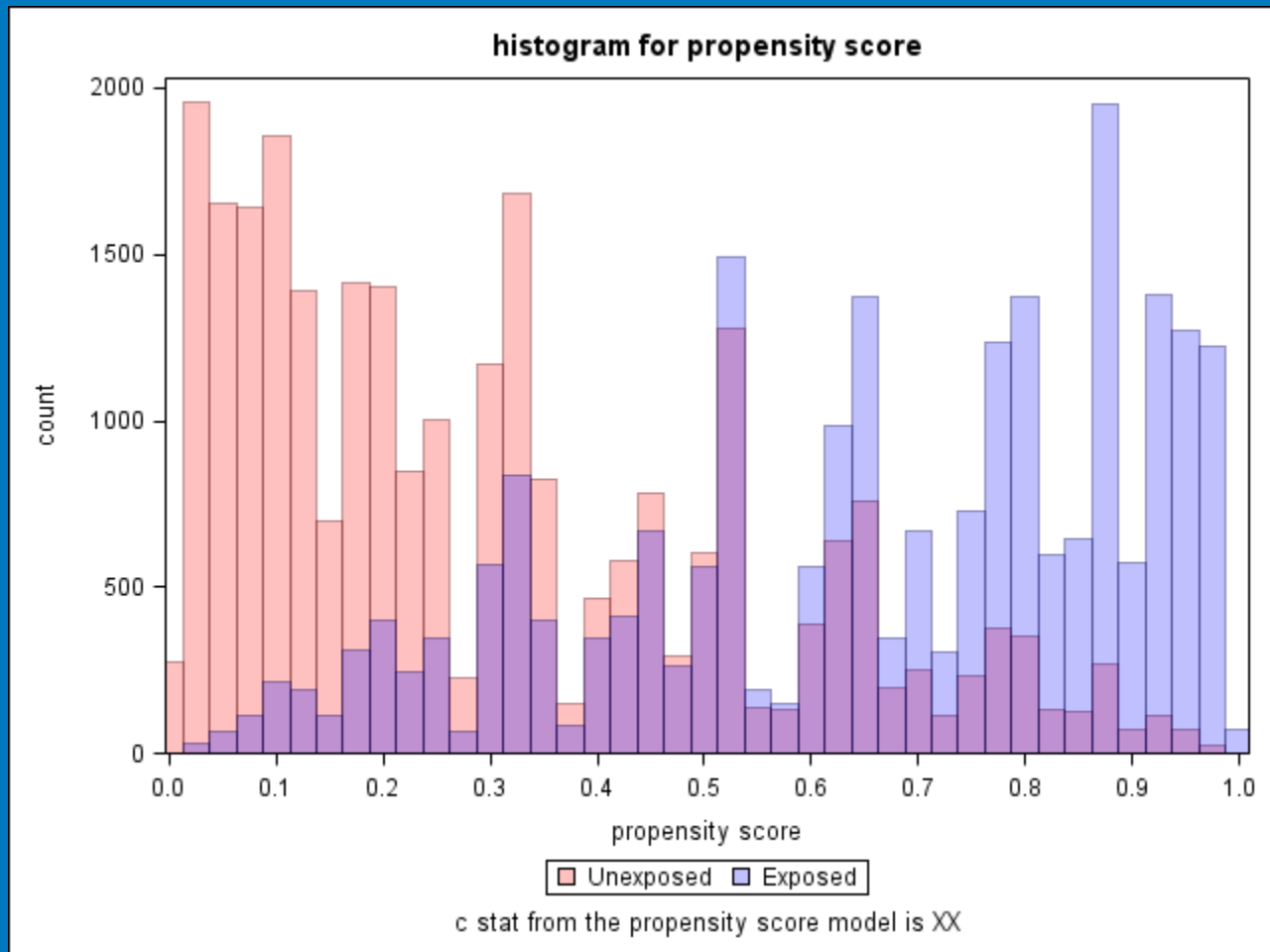
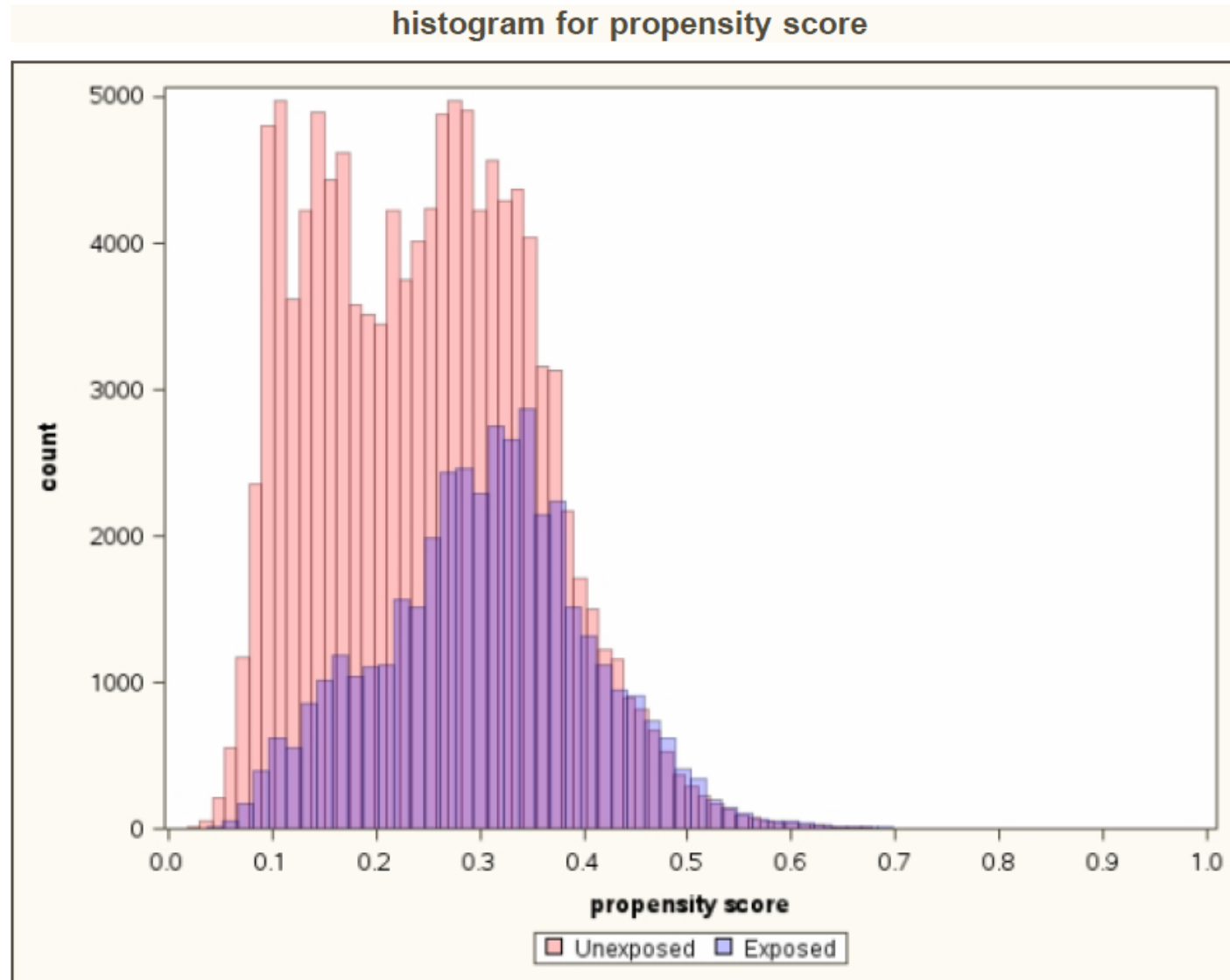


Figure 2



Reference

- High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA.* Epidemiology. 2009 Jul;20(4):512-22.
- Adjustments for Unmeasured Confounders in Pharmacoepidemiologic Database Studies Using External Information. *Stürmer T, Glynn R, Rothman K, Avorn J, Schneeweiss S.* Med Care. 2007 Oct;45(10 Supl 2):S158-65.
- A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Austin PC.* Stat Med. 2008 May 30;27(12):2037-49.
- High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S.* Eur J Clin Pharmacol. 2013 Mar;69(3):549-57.

Thanks