Open Access Medical Statistics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

Review: propensity score methods with application to the HELP clinic clinical study

This article was published in the following Dove Press journal:
Open Access Medical Statistics

Shesh N Rai[1,2,*]
Xiaoyong Wu[1,*]
Deo K Srivastava[3]
John A Craycroft[2]
Jayesh P Rai[4]
Sanjay Srivastava[4]
Robert F James[5]
Maxwell Boakye[5]
Aruni Bhatnagar[4]
Richard Baumgartner[6]

[1]Biostatistics Shared Facility, James Graham Brown Cancer Center, University of Louisville, Louisville, KY, USA; [2]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA; [3]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA; [4]Division of Cardiology, University of Louisville, Louisville, KY, USA; [5]Department of Neurosurgery, University of Louisville, Louisville, KY, USA; [6]Department of Epidemiology and Population Health, University of Louisville, KY, USA

*These authors contributed equally to this work

Correspondence: Shesh N Rai
Clinical and Translational Research Building, Room 211, 505 S. Hancock Street, Louisville, KY 40202, USA
Tel +1 502 852 4030
Email Shesh.Rai@Louisville.Edu

Abstract: Observational studies, common in clinical trials, often suffer from a lack of random assignment of the treatment. This can lead to large differences in covariates between the treated and untreated groups, which should be accounted for prior to inference, hypothesis tests, etc. Propensity score methods are frequently used to control for potentially confounding covariates when assessing causal effects of treatment on outcome. In this review, we introduce four adjustment methods based on propensity scores including matching, stratification, inverse probability of treatment weighting and covariate adjustment. Also, we give a general description of these four methods and provide some visual tools to assess covariate balance between the treated and untreated groups. We confirm the feasibility of propensity score methods by analyzing the Health Evaluation and Linkage to Primary care clinic clinical data.

Keywords: propensity score, covariate balance, observational studies, association analysis, HELP Clinic, proc glm, proc logistic, cat.psa, box.psa

Introduction

Observational studies have been widely used in statistics and medical research,[3,30] where investigations have no control over the assignment of the treatment. It is well known that in randomized studies, the randomization of subjects to the treated and untreated groups ensures that the covariates between the groups are similar. However, in observational studies, the selection of the treated and untreated groups may be typically not random. Therefore, comparing outcomes between the two groups has a challenge.[31,38] In these studies, large differences in observed covariates between the groups may lead to biased estimates of treatment effects.[26] To account for the differences of covariates, the adjustments are needed to control the confounding covariates prior to comparing outcomes between groups.

Historically, stratification as a potential approach was proposed to control the confounding from covariates in observational studies.[7] In this technique, subjects are stratified according to covariates directly and the treated and untreated subjects within the same stratum are then compared. However, such a traditional method is often limited since it can only remove the confounding from a small number of covariates. When there are multiple confounding variables involved, a simple stratification is not feasible. The number of strata increases and the sample sizes within stratum become sparse as the number of covariates increases. Stratification is often hard to adjust for many covariates.[12] More importantly, stratification cannot be used to deal with continuous covariates without a suitable discretization. Other traditional methods such

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/OAMS.S156704

as regression adjustments for covariates have also certain limitations. For example, when there is a relatively fewer number of subjects for model fitting with many covariates, the estimated parameters are subject to vast variability and less reliability. Covariate-adjusted regression modeling is not appropriate and the estimates of effect of treatments are sometimes biased.[11]

Propensity scores were proposed to adjust for difference of covariates between the treated and untreated group,[32] which has received an increasing interest in observational studies.[4,14,24,25] The propensity score of a subject is defined as the probability of being treated conditional on the subject's observed covariates. Adjustments using propensity score can reduce the bias due to covariates and lead to balanced distribution of covariates between the treated and untreated groups.[9] There are four commonly used propensity score methods: matching,[1,15,17,21,34] stratification,[22,33] inverse probability of treatment weighting [8,18,27,28] and covariate adjustment.[2,10,37] An introductory overview of propensity score methods is available in these selected publications.[6,23,31]

This review has three main aims. First, we summarize the concept of propensity scores. Second, we give some useful tools to assess covariate balance between the treated and untreated groups. Third, we show the ability of propensity score techniques to assess the association of treatment with outcomes in a real clinical study. In this case study, we identify a spurious association between the treatment and outcomes due to confounders. The paper is organized as follows. In section 2, we give a motivational case study to show that the covariates vary markedly between the treated and untreated groups. In section 3, we summarize the concept of the propensity score and describe four adjustment methods. In section 4, we provide flexible graphical tools to assess balance of covariates between the treated and untreated groups. In section 5, we provide results of the case study and compare the results from different methods for covariate adjustments. Some remarks are discussed in the last section.

Patients and methods: a case study

All subjects were recruited from a single freestanding residential detoxification unit in Boston, MA, USA, between 1 June 1997 and 1 April 1999.[35] The subjects undergoing detoxification from alcohol, heroin or cocaine who had no primary care physician were studied in a randomized control trial. The intervention consisted of a clinical evaluation at the detoxification unit in the Health Evaluation and Linkage to Primary care (HELP) clinic and control group. The data are provided in the supplement file "HELPmiss.csv". The primary outcome of interest was attendance at a primary care appointment within 12 months. Secondary outcomes assessed over 24 months were addiction severity, health-related quality of life, utilization of medical and addiction services, and HIV risk behaviors. Of 2,062 screened clients, 1,420 subjects did not meet study eligibility criteria. Among the remaining 642 eligible subjects, 470 provided written informed consent and agreed to participate in this study. After enrollment, 2 subjects died prior to the interview. Only 468 subjects received an interviewer administered baseline assessment for the demographics, short-form health survey, addiction severity index, medical and addiction care utilization in the past 6 months, depressive symptoms, alcohol and drug quantity, inventory of drug use consequences and HIV risk behaviors. An additional summary of the recruitment plan is displayed in Figure 1.

A set of 23 variables is included in this study. We have introduced abbreviations for ease of notation (Tables 1 and 2) where SAS and R are provided in the Boxes 1 and 2 along with SAS macro in the supplement file "yamgast.txt". The outcome variable of interest is physical quality of life component score (PCS). The treatment variable is the homeless status (Homeless). The list of the covariates includes sex, race, age, primary substance of abuse (Substance), any substance abuse treatment (SAT), serious thoughts of suicide in last 30 days (Suicide), randomized to HELP clinic (Randomization), use of any substance post-detox (US), post-detox linkage to primary care (LP), average number of drinks consumed per day in the past 30 days (AD), maximum number of drinks consumed per day in the past 30 days (MD), lifetime number of hospitalizations for medical problems (Hospitalization), time to first use of any substance post-detox (TS), time (in days) to linkage to primary care (TLP), number of times in past 6 months entered a detox program (Times), center for epidemiologic studies depression measure (CESD), risk assessment battery drug risk scale (RABDR), inventory of drug use consequences total score (IDUC), risk assessment battery sex risk score (RABSR), perceived social support by friends (Support), and mental quality of life component score (MCS).

Propensity score

In randomized studies, subjects are randomly assigned to either a treated group or an untreated group, which ensures that the distribution of the covariates between both groups is the same and then the effect of the treatment on the outcomes can be directly compared. However, observational studies often suffer from a lack of covariate balance between the two treatment groups, i.e., subjects are not randomly assigned to

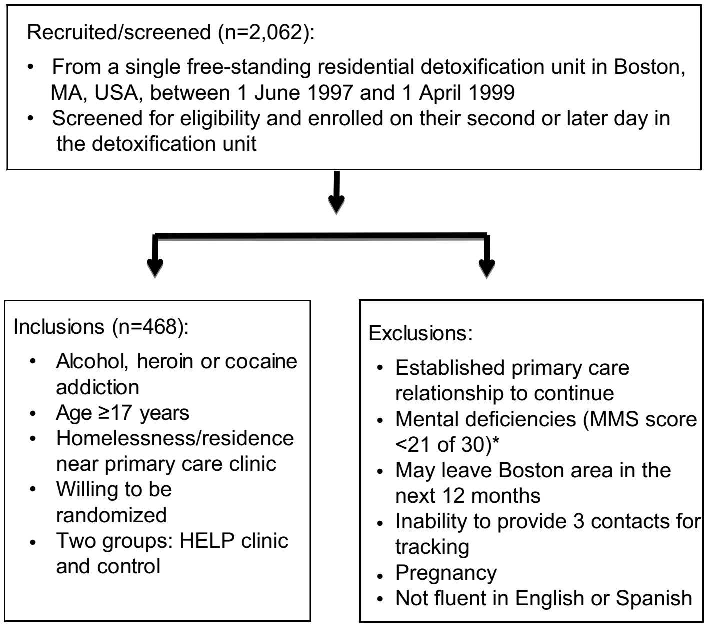
Recruited/screened (n=2,062):
• From a single free-standing residential detoxification unit in Boston, MA, USA, between 1 June 1997 and 1 April 1999
• Screened for eligibility and enrolled on their second or later day in the detoxification unit
Inclusions (n=468):
• Alcohol, heroin or cocaine addiction
• Age ≥17 years
• Homelessness/residence near primary care clinic
• Willing to be randomized
• Two groups: HELP clinic and control
Exclusions:
• Established primary care relationship to continue
• Mental deficiencies (MMS score <21 of 30)*
• May leave Boston area in the next 12 months
• Inability to provide 3 contacts for tracking
• Pregnancy
• Not fluent in English or Spanish

Figure 1 Flowchart of subjects' selection with addiction in the HELP clinic clinical study.
Note: *MMS score, max 30.
Abbreviations: HELP, Health Evaluation and Linkage to Primary care; MMS, Mini-Mental State.

Table 1 Association of Homeless with demographics characteristics and baseline measures for the HELP clinic clinical data

Variables	Total (N = 468)	Homeless (N = 217)	Housed (N = 251)	*P*-value
Sex				0.039
Female	111 (23.7)	42 (19.4)	69 (27.5)	
Male	357 (76.3)	175 (80.6)	182 (72.5)	
Race				0.027
Black	216 (46.2)	86 (39.6)	130 (51.8)	
Hispanic	51 (10.9)	22 (10.1)	29 (11.6)	
White	173 (37.0)	95 (43.8)	78 (31.1)	
Other	28 (6.0)	14 (6.5)	14 (5.6)	
Substance				<0.001[a]
Alcohol	185 (39.5)	109 (50.2)	76 (30.3)	
Cocaine	155 (33.1)	59 (27.2)	96 (38.2)	
Heroin	127 (27.1)	48 (22.1)	79 (31.5)	
Missing	1 (0.2)	1 (0.5)	0 (0.0)	
SAT				0.056
No	337 (72.0)	147 (67.7)	190 (75.7)	
Yes	131 (28.0)	70 (32.3)	61 (24.3)	
Suicide				0.003
No	335 (71.6)	141 (65.0)	194 (77.3)	
Yes	133 (28.4)	76 (35.0)	57 (22.7)	
Randomization				0.308
No	234 (50.0)	114 (52.5)	120 (47.8)	
Yes	234 (50.0)	103 (47.5)	131 (52.2)	
US				0.757
No	58 (12.4)	29 (13.4)	29 (11.6)	
Yes	195 (41.7)	93 (42.9)	102 (40.6)	
Missing	215 (45.9)	95 (43.8)	120 (47.8)	
LP				0.446
No	280 (59.8)	127 (58.5)	153 (61.0)	
Yes	165 (35.3)	81 (37.3)	84 (33.5)	
Missing	23 (4.9)	9 (4.1)	14 (5.6)	

(Continued)

Table 1 (*Continued*)

Variables	Total (N = 468)	Homeless (N = 217)	Housed (N = 251)	*P*-value
Age				0.068
Frequency	468	217	251	
Mean (95% CI)	35.74 (35.03–36.44)	36.44 (35.34–37.55)	35.12 (34.22–36.03)	
Median (min – max)	35 (18–60)	36 (18–60)	34 (21–58)	
AD				<0.001
Frequency	468	217	251	
Mean (95% CI)	18.28 (16.46–20.11)	23.48 (20.36–26.60)	13.79 (11.87–15.71)	
Median (min – max)	13 (0–142)	19 (0–142)	10 (0–76)	
MD				<0.001
Frequency	468	217	251	
Mean (95% CI)	25.06 (22.48–27.64)	31.27 (26.92–35.62)	19.68 (16.84–22.53)	
Median (min – max)	18.5 (0–184)	25 (0–179)	13 (0–184)	
Hospitalizations				0.132
Frequency	468	217	251	
Mean (95% CI)	3.09 (2.53–3.66)	3.56 (2.84–4.28)	2.69 (1.84–3.54)	
Median (min – max)	2 (0–100)	2 (0–40)	1 (0–100)	
TS				0.732
Frequency	251	122	129	
Mean (95% CI)	75.4 (68.2–82.6)	73.6 (63.1–84.2)	77.1 (67.2–87.0)	
Median (min – max)	33.0 (0.0–268.0)	32.0 (0.0–252.0)	35.0 (0.0–268.0)	
Missing	217	95	122	
TLP				0.849
Frequency	445	208	237	
Mean (95% CI)	257.6 (243.9–271.2)	256.1 (236.1–276.1)	258.8 (240.2–277.5)	
Median (min – max)	364.0 (2.0–456.0)	358.0 (2.0–449.0)	365.0 (4.0–456.0)	
Missing	23	9	14	
Times				0.032
Frequency	220	140	80	
Mean (95% CI)	2.5 (2.3–2.7)	2.8 (2.4–3.1)	2.0 (1.8–2.2)	
Median (min – max)	2.0 (1.0–21.0)	2.0 (1.0–21.0)	1.0 (1.0–11.0)	
Missing	248	77	171	
CESD				0.048
Frequency	468	217	251	
Mean (95% CI)	32.87 (31.74–34.00)	34.10 (32.46–35.73)	31.81 (30.26–33.37)	
Median (min – max)	34 (1–60)	36 (1–60)	32 (3–58)	
RABDR				0.380
Frequency	466	216	250	
Mean (95% CI)	1.9 (1.5–2.3)	2.1 (1.4–2.7)	1.7 (1.2–2.2)	
Median (min – max)	0.0 (0.0–21.0)	0.0 (0.0–21.0)	0.0 (0.0–21.0)	
Missing	2	1	1	
IDUC				<0.001
Frequency	454	210	244	
Mean (95% CI)	35.7 (35.1–36.4)	37.4 (36.5–38.2)	34.3 (33.4–35.2)	
Median (min – max)	37.5 (4.0–45.0)	39.5 (9.0–45.0)	36.0 (4.0–45.0)	
Missing	14	7	7	
RABSR				0.037
Frequency	467	216	251	
Mean (95% CI)	4.6 (4.4–4.9)	4.9 (4.5–5.3)	4.4 (4.1–4.7)	
Median (min – max)	4.0 (0.0–14.0)	5.0 (0.0–14.0)	4.0 (0.0–14.0)	
Missing	1	1	0	
Support				<0.001
Frequency	468	217	251	
Mean (95% CI)	6.67 (6.31–7.04)	5.99 (5.48–6.50)	7.26 (6.76–7.764)	
Median (min – max)	7 (0–14)	5 (0–14)	7 (0–14)	
MCS				0.138
Frequency	468	217	251	
Mean (95% CI)	31.5 (30.4–32.7)	30.6 (29.0–32.2)	32.4 (30.7–34.0)	
Median (min – max)	28.6 (6.8–62.2)	27.8 (9.2–60.5)	30.7 (6.8–62.2)	

(*Continued*)

Table 1 (*Continued*)

Variables	Total (N = 468)	Homeless (N = 217)	Housed (N = 251)	*P*-value
PCS				0.054
Frequency	468	217	251	
Mean (95% CI)	48.1 (47.1–49.0)	47.0 (45.6–48.5)	49.0 (47.6–50.3)	
Median (min – max)	48.9 (14.1–74.8)	47.0 (21.9–71.6)	51.0 (14.1–74.8)	

Note: [a]Fisher's exact test.

Abbreviations: HELP, Health Evaluation and Linkage to Primary care; SAT, any substance abuse treatment; Suicide, serious thoughts of suicide in last 30 days; US, use of any substance post-detox; LP, post-detox linkage to primary care; AD, average number of drinks consumed per day in the past 30 days; MD, maximum number of drinks consumed per day in the past 30 days; TS, time to first use of any substance post-detox; TLP, time (in days) to linkage to primary care; CESD, center for epidemiologic studies depression measure; RABDR, risk assessment battery drug risk scale; IDUC, inventory of drug use consequences total score; RABSR, risk assessment battery sex risk score; MCS, mental quality of life component score; PCS, physical quality of life component score.

Table 2 Effects of Homeless on PCS using the propensity score adjusted for covariates

Method	Estimate	Standard error	95% CI	*P*-value
Two group comparison	−1.9	1.0	(−3.87, 0.03)	0.054
Standard covariate adjustment	−1.1	1.0	(−3.04, 0.92)	0.295
Matching	−1.3	1.1	(−3.52, 0.84)	0.227
Stratification	−1.4	1.1	(−3.52, 0.69)	0.188
Inverse probability of treatment weight	−2.06	1.0	(−4.06, −0.07)	0.042
Covariate adjustment	−1.2	1.1	(−3.26, 0.97)	0.288

Abbreviation: PCS, physical quality of life component score.

Box 1 SAS codes for descriptive statistics in Table 1

```
/* Box 1 */
/* PS analysis using the data in the health evaluation HELP study */
/* The following SAS program provides the abovementioned Table 1
for manuscript */
proc import datafile = "C:\Users\x0wu0008\Documents\Projects\
PropensityScore\Analysis\HELPmiss"
out = help dbms = xls replace;
run;

/* Convert format and remove subjects with missing PCS */
data help_2;
set help;
PCS_2 = PCS + 0;
MCS_2 = MCS + 0;
TS_2 = TS + 0;
TLP_2 = TLP + 0;
RABDR_2 = RABDR + 0;
IDUC_2 = IDUC + 0;
RABSR_2 = RABSR + 0;
Times_2 = Times + 0;
if (PCS_2 = .) then delete;
run;

/* Perform association of Homeless with other factors */
%include "C:\Users\x0wu0008\Documents\Projects\PropensityScore\
Analysis\yamgast.sas";
%yamgast(dat = help_2, grp = Homeless,
vlist =
Sex \freq\
Race \freq\
Substance \freq\
SAT \freq\
Suicide \freq\
Randomization \freq\
Age \mean2 med1 \
```

(*Continued*)

Box 1 (*Continued*)

```
AD \mean2 med1 \
MD \mean2 med1 \
Hospitalizations \mean2 med1 \
TS_2 \mean2 med1 \
TLP_2 \mean2 med1 \
Times_2 \mean2 med1 \
CESD \mean2 med1 \
RABDR_2 \mean2 med1 \
IDUC_2 \mean2 med1 \
RABSR_2 \mean2 med1 \
Support \mean2 med1 \
MCS_2 \mean2 med1 \
PCS_2 \mean2 med1 \,
ncont = yes, missing = yes,
style = custom,
title = Characteristics,
footnote =,
file = C:\Users\x0wu0008\Documents\Projects\PropensityScore\
Analysis\res.rtf);

/* Perform association of homeless with other factors with missing
data */
data help_US;
set help_2;
if (US = "Missing") then delete;
run;
%include "C:\Users\x0wu0008\Documents\Projects\PropensityScore\
Analysis\yamgast.sas";
%yamgast(dat = help_US, grp = Homeless,
vlist =
US \freq\,
ncont = yes, missing = yes,
style = custom,
title = Characteristics,
footnote =,
```

(*Continued*)

Box 1 (*Continued*)

```
file = C:\Users\x0wu0008\Documents\Projects\PropensityScore\
Analysis\res.rtf);

data help_LP;
set help_2;
if (LP = “Missing”) then delete;
run;
%include “C:\Users\x0wu0008\Documents\Projects\PropensityScore\
Analysis\yamgast.sas”;
%yamgast(dat = help_LP, grp = homeless,
vlist =
LP \freq\,
ncont = yes, missing = yes,
style = custom,
title = Characteristics,
footnote =,
file = C:\Users\x0wu0008\Documents\Projects\PropensityScore\
Analysis\res.rtf);
```

Box 2 SAS codes for propensity score analyses in Table 2

```
/* Box 2*/ * the following SAS program provides the Table 2 for
manuscript */
/* Perform two group comparison */
proc glm data=help_2;
class Homeless (ref=”housed”) ;
Model PCS_2 = Homeless  / solution CLPARM;
run;

/* Keep significant factors and remove subjects with missing values for
adjustment */
data help_miss;
set help_2;
if (Homeless = “homeless”) then Homeless_2 = 1;
else Homeless_2 = 0;
if (Substance = “Missing” | IDUC_2 = . | RABSR_2 = .) then delete;
keep ID Homeless_2 Homeless Sex Race Substance Suicide AD
CESD IDUC_2 RABSR_2 Support PCS_2;
run;

/* Perform standard multiple adjustment */
proc glm data = help_miss;
class Homeless (ref = “housed”) Sex (ref = “male”) Race (ref =
“other”) Substance (ref = “heroin”) Suicide (ref = “no”);
Model PCS_2 = Homeless Sex Race Substance Suicide AD CESD
IDUC_2 RABSR_2 Support / solution CLPARM;
run;

/* Perform PS matching */
proc logistic data = help_miss desc;
class Sex (ref = “male”) Race (ref = “other”) Substance (ref =
“heroin”) Suicide (ref = “no”);
model Homeless_2 = Sex Race Substance Suicide AD CESD IDUC_2
RABSR_2 Support;
output out = help_ps pred = ps;
run;
proc sort data = help_ps out = one_match;
by Homeless_2;
run;
```

(*Continued*)

Box 2 (*Continued*)

```
proc transpose data = one_match out = data1;
by Homeless_2;
run;
data id_t (rename = (COL1-COL209 = tid1-tid209));
set data1;
if Homeless_2 = 1 and _NAME_ = ‘ID’;
run;
data ps_t (rename = (COL1-COL209 = tps1-tps209));
set data1;
if Homeless_2 = 1 and _NAME_ = ‘ps’;
run;
data id_c (rename = (COL1-COL244 = cid1-cid244));
set data1;
if Homeless_2 = 0 and _NAME_ = ‘ID’;
run;
data ps_c (rename = (COL1-COL244 = cps1-cps244));
set data1;
if Homeless_2 = 0 and _NAME_ = ‘ps’;
run;
data all;
merge id_t ps_t id_c ps_c;
caliper = 0.1;
array treat_id {*} tid1-tid209;
array ctl_id {*} cid1-cid244;
array treat_p {*} tps1-tps209;
array ctl_p {*} cps1-cps244;
array used_i {*} used1 - used244;
array matched_t {*} m_tid1-m_tid209;
array matched_c {*} m_cid1-m_cid209;
match_N = 0;
do i = 1 to 209;
min_diff = 1;
best_match = 0;
do j = 1 to 244;
if used_i[j] = . then do;
if ABS(treat_p[i] - ctl_p[j]) < caliper then do;
if ABS(treat_p[i] - ctl_p[j]) < min_diff then do;
min_diff = ABS(treat_p[i] - ctl_p[j]);
best_match = j;
end;
end;
end;
end;
if best_match > 0 then do;
match_N = match_N + 1;
used_i[best_match] = 1;
matched_t[match_N] = treat_id[i];
matched_c[match_N] = ctl_id[best_match];
end;
end;
run;
data matches;
set all;
array matched_t {*} m_tid1-m_tid209;
array matched_c {*} m_cid1-m_cid209;
do match = 1 to match_N;
Treatment_IDN = matched_t[match];
Control_IDN = matched_c[match];
output;
```

(*Continued*)

Box 2 (*Continued*)

```
end;
keep match treatment_idn control_idn;
run;
data matches_2;
set matches;
ID = Treatment_IDN;
run;
proc sort data = matches_2 out= matches_3;
by ID;
run;
data match_data;
merge help_ps matches_3;
by ID;
run;
data matches_4;
set matches;
ID = Control_IDN;
run;
proc sort data = matches_4 out= matches_5;
by ID;
run;
data match_data_6;
merge match_data matches_5;
by ID ;
run;
data final_data;
set match_data_6;
if (match = .) then delete;
run;
proc glm data=final_data;
class Homeless_2(ref = "0") Sex (ref = "male") Race (ref = "other")
Substance (ref = "heroin") Suicide (ref = "no");
model PCS_2 = Homeless_2 Sex Race Substance Suicide AD CESD
IDUC_2 RABSR_2 Support / solution CLPARM;
run;

/* Calculate estimated PS for stratification, inverse probability */
/* of treatment weight, Covariate adjustment */
proc logistic data = help_miss desc;
class Sex (ref = "male") Race (ref = "other") Substance (ref =
"heroin") Suicide (ref = "no");
model Homeless = Sex Race Substance Suicide AD CESD IDUC_2
RABSR_2 Support;
output out = help_ps_2 pred = ps_2;
run;

/* Perform PS stratification */
proc rank data = help_ps_2 groups=5 out = rank_ds;
ranks rank;
var ps_2;
run;
data quintile;
set rank_ds;
quintile = rank + 1;
run;
proc glm data = quintile;
class Homeless (ref = "housed");
model PCS_2 = Homeless quintile / solution CLPARM;
```

(*Continued*)

Box 2 (*Continued*)

```
run;

/* Perform PS inverse probability of treatment weight */
data help_iptw ;
set help_ps_2;
if Homeless = "homeless" then ps_weight = 1/ps_2;
else ps_weight = 1/(1-ps_2);
Run;
proc glm data = help_iptw;
class Homeless (ref = "housed");
Model PCS_2 = Homeless / solution  CLPARM;
run;

/* Perform PS covariate adjustment */
proc glm data=help_ps_2;
class Homeless (ref = "housed");
Model PCS_2 = ps_2 Homeless / solution CLPARM;
run;
```

either a treated group or an untreated group and thus there are often significant differences of characteristics between the two groups. Propensity score methods aim to mimic randomized studies within the context of observational studies. The differences of characteristics between the two groups must be adjusted for to reduce treatment selection bias in order to estimate treatment effect. Propensity score analysis is a statistical approach to reduce treatment selection bias.

We briefly describe notations for the propensity score-adjusted analyses. Let the triplet (Y_i, X_i, Z_i) denote response, group indicator and covariates for the i^{th} subject, respectively. For simplicity, group indicator can be for comparing two groups of interventions or treatments (such as $X_i = 1$ for the treatment group and $X_i = 0$ for the untreated group). Let $Z_i = (Z_{i1}, \ldots, Z_{in})'$ denote the vector of observed covariates for the i^{th} subject. Then the propensity score for the i^{th} subject is defined as the conditional probability that a subject will be assigned to a treatment, given a vector of observed covariates, that is,

$$P_i = P\left(X_i = 1 | Z_i\right), i = 1, .., n.$$

Since each P_i is unknown, in order to obtain its estimates, we consider a logistic regression model:

$$\text{logit}\left(P_i\right) = a_0 + a_1' Z_i ,$$

where a_0 and a_1 are regression coefficients. Then the estimate of P_i is given by:

$$\hat{P}_i = \left(1 + \exp\left(-\hat{a}_0 - \hat{a}_1' Z_i\right)\right)^{-1}, \tag{1}$$

where $\hat{a}_0$ and $\hat{a}_1$ are the maximum likelihood estimates of a_0 and a_1, respectively.

It is well recognized that adjusting for the estimated propensity score can help adjust for differences of covariates between the groups. By balancing covariates between the treated and untreated subjects, the association between the treatment and covariates is weakened or even made null, which possibly eliminates chance of confounding by covariates. Once the propensity score is estimated, there are four commonly used methods based on propensity scores, which are propensity score matching, stratification, inverse probability of treatment weighting and covariate adjustment.

Propensity score matching

Propensity score matching attempts to mimic randomization to reduce selection bias by matching the untreated group to the treated group based on the estimated propensity score such that the matched group is similar to the treated group in all the characteristics. Propensity score matching often involves studies where there are a smaller number of treated subjects and a larger number of untreated subjects. Having obtained propensity scores, these scores are used to match the untreated subjects with a treated subject with closest propensity score value. This continues until the entire treated subjects are matched. The one-to-one matching is the most commonly used propensity score matching where a treated subject is matched with the untreated subjects with similar propensity score. It has been shown[5] that the theoretical efficiency of a 1:M case-control ratio for estimating a relative risk of about one, relative to having complete information on the control population, is M/(M + 1). Therefore, increasing the number of matched controls for each case will improve the efficiency. Although in case-control studies, one-to-many matching increases efficiency, when the number of covariates is too many, another popular approach is the optimal matching that creates a series of matched sets in which each set contains at least one treated subject and at least one untreated subject. To estimate the average treatment effect, the full matching is optimal in terms of minimizing the average of the distances between each treated subject and each untreated subject within each matched set.[29]

Propensity score stratification

Propensity score stratification involves grouping subjects into strata who have similar propensity scores and balancing covariates between the treated and untreated groups. Once the propensity scores are calculated, subjects are placed into strata based on the estimated propensity scores or the quintiles of the estimated propensity scores for determining the cutoffs for the different strata. Once these strata are defined, the treated and untreated subjects within the same stratum are compared directly. A commonly used way is to stratify subjects into five approximately equal-size strata based on the quintiles of the estimated propensity scores. Specifically, we chose $\hat{Q}_j\left(j=1,..,5\right)$ such that the proportion of $\hat{P}_i$ less than or equal to $\hat{Q}_j$ is approximately equal to $\frac{j}{5}$, where $\hat{Q}_j$ is the j^{th} quantiles of $\hat{P}_i\left(i=1,..,n\right)$. The adjusted treatment effect is given by:

$$\sum_{j=1}^{5}\frac{n_j}{n}\left(\frac{1}{m_j}\sum_{i=1}^{n}X_iY_iI_{\hat{P}_i\in\Omega_j}-\frac{1}{n_j-m_j}\sum_{i=1}^{n}(1-X_i)Y_iI_{\hat{P}_i\in\Omega_j}\right),$$

where Ω_j is the interval by $\hat{Q}_j$, i.e., $\Omega_j=\left(\hat{Q}_{j-1},\hat{Q}_j\right)$, n_j is the number of subjects in the j^{th} stratum, i.e., $n_j=\sum_{i=1}^{n}I_{\hat{P}_i\in\Omega_j}$ and m_j is the number of the treated subjects in the j^{th} stratum, i.e., $m_j=\sum_{i=1}^{n}X_iI_{\hat{P}_i\in\Omega_j}$. The standard generalized linear model for the outcome is stratified for $\hat{Q}_j$. Then we obtain the estimate of the average treatment effect given by the parameter of the treatment, the standard error and the *P*-value, which tests the null hypothesis that the treatment in the model is not a significant predictor. It is shown that the approach can remove ~90 percent of the bias due to the confounders from covariates when estimating a causal effect of treatment.[33]

Inverse probability of treatment weighting using propensity score

Inverse probability of treatment weighting using propensity score is used as weights for all the subjects. If the i^{th} subject receives a treatment, then the subject's weight is defined as the inverse of propensity score. Otherwise, if the subject does not receive a treatment, then the subject's weight is defined as the inverse of 1 minus propensity score. The weight for the i^{th} subject can be written as:

$$w_i=\frac{X_i}{\hat{P}_i}+\frac{1-X_i}{1-\hat{P}_i}.$$

The average treatment effect *ATE* unadjusted by propensity score is estimated by the following equation:

$$\widehat{ATE}=\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\frac{1}{n}\sum_{i=1}^{n}(1-X_i)Y_i.$$

Thus, the average treatment effect *ATE* adjusted by propensity score is estimated by the following equation:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i Y_i}{\hat{P}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-X_i)Y_i}{1-\hat{P}_i}.$$

Note that a very large weight could produce biased estimates of the treatment effect. To decrease the variance of the estimate of the treatment effect, a potential solution is used by multiplying the weight given by:

$$w_i = \frac{X_i \sum_{j\in\Omega_T}\hat{P}_j}{n_T \hat{P}_i} + \frac{\left(1-X_i\right)\sum_{j\in\Omega_U}(1-\hat{P}_j)}{n_U(1-\hat{P}_i)},$$

where Ω_T and Ω_U are the set of treated and untreated subjects, respectively; n_T and n_U are the total number of treated and untreated subjects, respectively. This method might not be applicable when propensity scores are very large or very small, which can be seen from Table 2.

Covariate adjustment in combination with propensity score

In covariate adjustment using propensity score, subjects' propensity score is first estimated and then the outcome is regressed on the treatment and estimated propensity score. Here, a regression choice depends on the nature of the outcome and a regression model relating the outcome to treatment and propensity score should be correctly specified. For a continuous outcome Y_i, we consider the following linear model:

$$Y_i = b_0 + b_1 X_i + b_2 \hat{P}_i + \varepsilon_i,$$

where Y_i and ε_i denote the outcome and random error for the i^{th} subject, respectively, and $b_{0,}$ b_1 and b_2 are regression coefficients. For a dichotomous outcome Y_i, the following logistic regression model

$$\text{logit}\left(P\left(Y_i = 1 \mid X_i, P_i\right)\right) = b_0 + b_1 X_i + b_2 P_i,$$

is frequently used. Here $\hat{P}_i$ is a single estimate obtained by a set of covariates.

Propensity score methods can be conducted using a variety of statistical packages, for example, SAS[13] and R packages: PSAgraphics,[16] MatchIt[19] and Matching.[36]

Evaluation of covariate balance

Observational studies often suffer from a lack of covariate balance when comparing outcome between the treated and untreated groups. By using the propensity score methods, one expects that the treated and untreated subjects will produce similar distribution of the covariates. For the purpose of determining whether a propensity score method is adequately specified, we need to assess the covariate distributions between the treated and untreated groups. In the sample matched on the estimated propensity score, or stratified on the estimated propensity score or their quantiles, or weighted by the inverse probability of treatment, we use the numerical and graphical methods to assess balance of covariates between treated and untreated subjects. The measures of the differences in covariates between the treated and untreated groups have been introduced.[2]

For a continuous covariate, we can use the standardized difference,

$$D_1 = \frac{\bar{Z}_T - \bar{Z}_C}{\sqrt{\left(s_T^2 + s_C^2\right)/2}},$$

to measure differences in covariates between the treated and untreated groups, where pairs $\bar{Z}_T$, $\bar{Z}_C$ and s_T^2, s_C^2 are sample means and variances of the covariate between the treated and untreated subjects, respectively. For stratified samples, the side-by-side boxplots can be used to graphically depict the balance of covariates between the treated and untreated groups across all strata, as displayed in Figure 2, where R is provided in Box 3. It can easily be seen that for each covariate, each stratum is well balanced since the means from the treated and untreated groups for each stratum are relatively similar.

For a categorical covariate, we can use the standardized difference,

$$D_2 = \frac{\hat{P}_T - \hat{P}_C}{\sqrt{\left(\hat{P}_T\left(1-\hat{P}_T\right) + \hat{P}_C\left(1-\hat{P}_C\right)\right)/2}},$$

to measure the differences of covariates between the treated and untreated groups, where $\hat{P}_T$ and $\hat{P}_C$ are sample proportions of covariates between the treated and untreated subjects, respectively. For stratified samples, the side-by-side barplots can be used to graphically exhibit the balance of covariates between the homeless and housed groups across all strata, as displayed in Figure 3, where R is provided in Box 4. It can

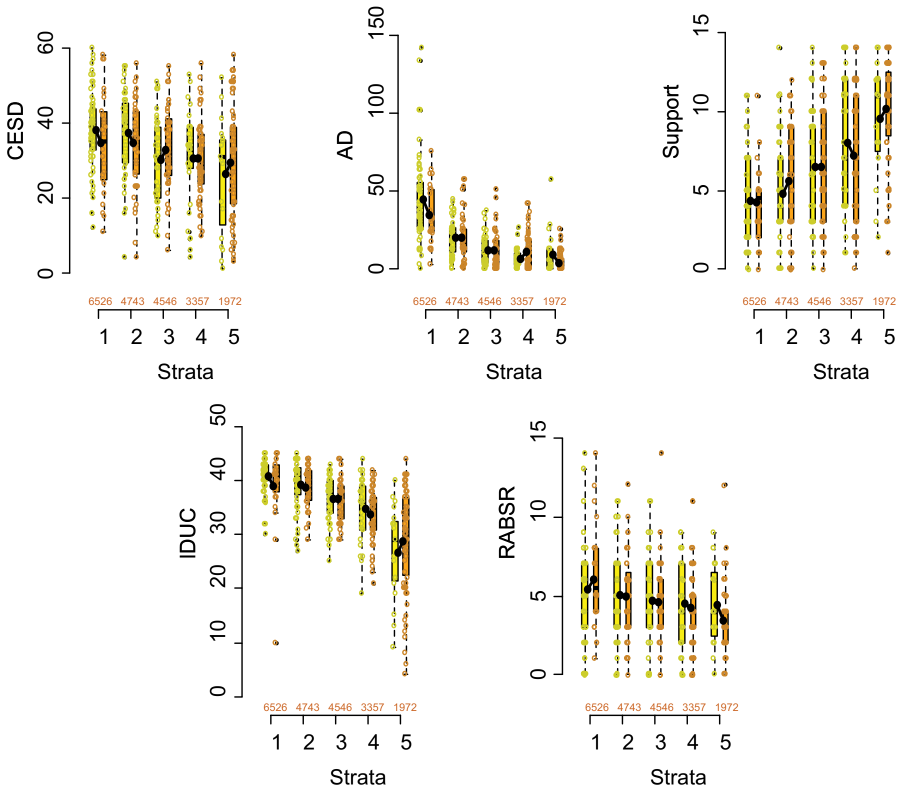

Figure 2 Side-by-side boxplot for the significant continuous covariates (CESD, AD, Support, IDUC, RABSR), which are included in Table 1.
Notes: Here yellow colors and brown colors denote the treated (homeless) and untreated (housed) subjects, respectively, the slope of the black lines denotes the expected differences in covariates between the two groups, and the numbers are the sample sizes of the subjects in each of the two groups.
Abbreviations: CESD, center for epidemiologic studies depression measure; AD, average number of drinks consumed per day in the past 30 days; Support, perceived social support by friends; IDUC, inventory of drug use consequences total score; RABSR, risk assessment battery sex risk score.

Box 3 R code for creating 5 strata for continuous variable

```
# Generate the data file strata.5 for continuous and discrete
covariates #
# The following R code involves PS stratification for balance of
covariate #
library(PSAgraphics)
path <- "C:\\Users\\x0wu0008\\Documents\\Projects\\
PropensityScore\\Analysis"
help_data <- read.csv(paste(path, "Help_Pmiss_Clean.csv", sep =
"\\"),header = T)

# fit a logistic regression model to estimate PS #
my.fit  <- glm(Homeless ~  CESD + Sex + Suicide + AD + Support
+ Race + Substance + IDUC + RABSR, data = help_data, family =
binomial)
my.ps <- my.fit$fitted
# stratify the observations based on PS #
strata.5 <- cut(my.ps, quantile(my.ps, seq(0, 1, 1/5)), include.lowest =
TRUE, labels = FALSE)

# Create side-by-side boxplot for the continuous covariates #
 attach(help_data)
par(mfrow=c(1,5))
box.psa(CESD, Homeless, strata.5, xlab = "Strata", ylab = "CESD",
legend.xy = c(2, 600), legend.labels = NULL, pts = TRUE, balance =
FALSE)
box.psa(AD, Homeless, strata.5, xlab = "Strata", ylab = "AD", legend.
```

(*Continued*)

Box 3 (*Continued*)

```
xy = c(2, 600), legend.labels = NULL, pts = TRUE, balance = FALSE)
box.psa(Support, Homeless, strata.5, xlab = "Strata", ylab =
"Support", legend.xy = c(2, 600), legend.labels = NULL, pts = TRUE,
balance = FALSE)
box.psa(IDUC, Homeless, strata.5, xlab = "Strata", ylab = "IDUC",
legend.xy = c(2, 600), legend.labels = NULL, pts = TRUE, balance =
FALSE)
box.psa(RABSR, Homeless, strata.5, xlab = "Strata", ylab = "RABSR",
legend.xy = c(2, 600), legend.labels = NULL, pts = TRUE, balance =
FALSE)
```

easily be seen that for each covariate, each stratum is well balanced since the proportions from the treated and untreated groups for each stratum are relatively similar.

Case study results

Our study was conducted to investigate whether the homeless subjects tend to be in poorer physical health than the housed subjects in our study. We first examined the association of the Homeless with other variables and the results are summarized in Table 1, where the continuous variables between the homeless and housed groups were compared using a two sample *t*-test and categorical variables between the homeless and housed groups were compared using the chi-square test.

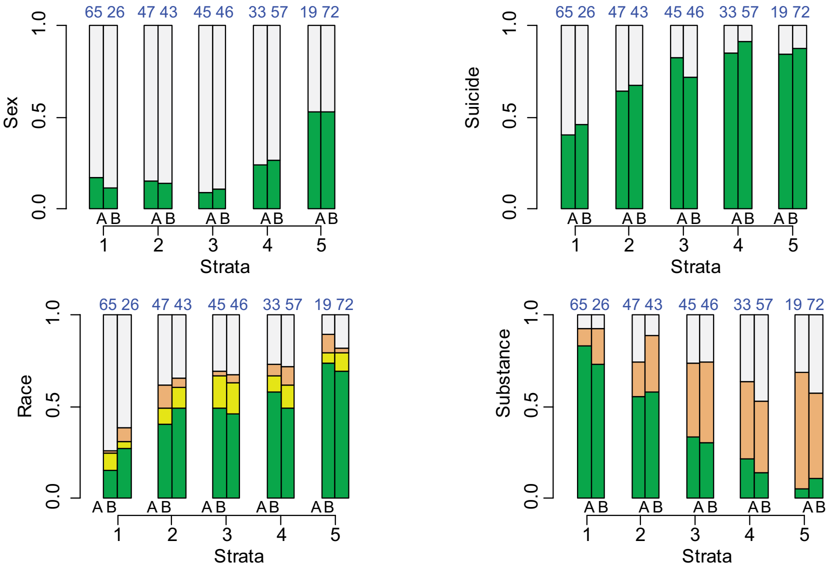

Figure 3 Side-by-side barplots for the significant categorical covariates (sex, Suicide, race, Substance), which are included in Table 1.
Note: Here the same colors represent the same level of the variables; A and B represent treated (homeless) and untreated (housed), respectively.
Abbreviations: Suicide, serious thoughts of suicide in last 30 days; Substance, primary substance of abuse.

Box 4 R code for creating 5 strata for discrete variable

```
# Obtain data file strata.5 from executing the program in box 3 #
# Create side by side barplots for categorical covariates #
par(mfrow = c(1,4))
cat.psa(Sex, Homeless, strata.5, xlab = "Strata",  ylab = "Sex", rtmar
= 0.2)
cat.psa(Suicide, Homeless, strata.5, xlab = "Strata",  ylab = "Suicide",
rtmar = 0.2)
cat.psa(Race, Homeless, strata.5, xlab = "Strata",  ylab = "Race",
rtmar = 0.2)
cat.psa(Substance, Homeless, strata.5, xlab = "Strata",  ylab =
"Substance", rtmar = 0.2)
```

We can see that the PCS for the homeless subjects appears to be marginally significantly different from the one for the housed subjects (P-value = 0.054). This means that Homeless significantly contributes to the subjects' PCS. In this comparison, no adjustments are made.

However, in this observational study, nonrandom assignment was given to homeless or housed subjects. This can lead to large differences of covariates between the homeless and housed groups. Thus, when directly comparing PCS between the two groups, the results might be misleading, that is, the association between PCS and Homeless might be spurious due to the effect of confounders from the observed covariates. Propensity score can be used to remove the effects of confounding when assessing the effects of Homeless on subjects' PCS.

In order to calculate the estimated propensity score, we first identified the covariates that were significantly associated with Homeless. From Table 1, the P-value was calculated using a two-sample t-test. We found that the significant covariates included sex, race, Substance, Suicide, AD (or MD), CESD, IDUC, RABSR, Support and Times. Although Times was significantly associated with Homeless (P-value = 0.032), it was not used as covariate for estimating propensity score since it had too many missing data (246 out of 468 subjects had missing Times). Also, 15 subjects with missing significant covariates were removed and therefore the remaining 453 subjects were used to estimate the propensity score by using logistic regression model. Once the propensity scores were obtained, we stratified subjects into five approximately equal-size strata $\hat{Q}_j\left(j=1,..,5\right)$ and used the plots to visualize the balance of each covariate between the homeless and housed groups. The side-by-side boxplots in Figure 2 assess the balance of each continuous covariate (e.g., CESD, AD, Support, IDUC and RABSR) between the homeless and housed subjects within each stratum as well as examine their distributions across 5 strata, where the dots represent the covariates, the black lines for the j^{th} stratum is visually used to compare the means of covariates C_i (i=1,..,n) between the homeless and housed subjects within j^{th} stratum, which is equal to $\frac{1}{m_j}\sum_{i=1}^{n}X_iC_iI_{\hat{P}_i\in\Omega_j}$ for the homeless group and $\frac{1}{n_j-m_j}\sum_{i=1}^{n}(1-X_i)C_iI_{\hat{P}_i\in\Omega_j}$ for the housed group, where n_j and m_j are the number of all the subjects and the number of the treated subjects in the j^{th} stratum, respectively, X_i and C_i are the treatment and a covariate for the i^{th} subject, respectively. The side-by-side barplots in Figure 3 assess the balance of each categorical covariate (e.g., sex, Suicide, race and

Substance) between the homeless and housed subjects within each stratum as well as examine their covariate distributions across 5 strata, where the green bars mean the proportions of covariates for the homeless subjects within each stratum. The plots visually have shown that the covariates between the two groups within each stratum are similar.

Thus, we can compare the effects of the Homeless on the PCS among the following methods. In Table 2, the first row uses the original data to directly compare the average PCS between the homeless and housed subjects where the *P*-value is calculated using a two-sample *t*-test. The second row uses the standard multiple regression model to determine if the Homeless is a significant predictor of the average PCS (the model also includes other significant covariates listed in Table 1). Here, the results in the first and second rows are not adjusted using any propensity scores. The Homeless is significantly associated with the PCS through the two group comparison (*P*-value = 0.054), while the Homeless is not significantly associated with the PCS through the standard covariate adjustment (*P*-value = 0.295). The remaining four rows are used to compare the average PCS between the homeless and housed subjects after adjusting for the propensity score methods indicated in section 3. We considered the multiple regression model that regresses the PCS on the Homeless and estimated propensity scores after adjustment. From Table 2, the Homeless is not significantly associated with the PCS through the propensity score matching (*P*-value = 0.227), propensity score stratification (*P*-value = 0.188) or covariate adjustment in combination with the propensity score (*P*-value = 0.288), while the Homeless is significantly associated with the PCS through the inverse probability of treatment weighting using the propensity score (*P*-value = 0.042). However, as we mentioned in section 3, the latter might be misleading since the propensity scores for some subjects are very large or very small. The evidence that the Homeless is not significantly associated with the PCS is further confirmed by the abovementioned standard covariate adjustment.

Discussion

Inference in a randomized study for comparing different groups other than the randomized groups may inherit some bias/confounding. To reduce the effects of confounding, as an alternative to multivariable regression models, the propensity score methods are sometimes used. In this paper, we introduced some visualized plots to evaluate covariate balance between the comparing groups. Any imbalance requires adjusting using propensity score method(s) prior to the inference for comparing groups.

Results from real data reveal that the propensity score methods can determine the association between treatment and outcome with propensity score adjustment being not statistically significant, which is different from the one without propensity score adjustment where the association between treatment and outcome is statistically significant. This implies that the propensity score methods remove a spurious association between treatment and outcome due to the effects of the confounding from covariates.

The multivariable model and most of the propensity score methods resulted in the same conclusion for the data set that we have considered here. If a study involves a large number of covariates, missing data, and correlated covariates, it may be difficult to get a simple set of covariates for the final multivariable model. On the other hand, if probability weights are unstable due to small number of observations within some subsets, the propensity score method based on weighting can provide conflicting results. Thus, we recommend using all these six methods (as listed in Table 2); the conclusion should be drawn on a simple majority of findings rather a simple two-group comparison approach.

Sample size calculation and stratification in many clinical studies are based on primary outcome variables, but the secondary outcome variables are often compared. As shown in this example, one can easily make a wrong conclusion by not having the sample size adjusted for covariates. Irrespective of primary or secondary outcome variable to be inferred, the sample size adjustment should be made for accommodating the effect of multiple covariates. One simple approach is adjust the sample size using R^2 contribution (explained variability) due to other covariates; a detailed approach for linear and logistic regression is given by Hsieh et al.[20]

There are some other issues in the analyses of secondary outcomes. If multiple secondary outcomes are compared using the propensity score analysis, these results must be adjusted for the multiple comparisons. Also, imputation of covariates in clinical studies is suggested before any propensity score analysis.

We provide the data set and the SAS program used in this manuscript for the ease of use in any similar data analyses.

Acknowledgments

This work was supported by the National Institute of Health grant P42 ES023716 to principal investigator: Dr S Srivastava and the National Institute of Health grant 1P20 GM113226 to principal investigator: Dr C McClain.

Disclosure

Dr SN Rai received additional support from the Wendell Cherry Chair in Clinical Trial Research. The authors report no other conflicts of interest in this work.

References

1. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083–3107.
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399–424.
3. Benson K, Hartz AJ. A comparison of observational studies and randomized controlled trials. *New Engl J Med*. 2000;342(25):1878–1886.
4. Bloch DA, Segal MR. Empirical comparison of approaches to forming strata: using classification trees to adjust for covariates. *J Am Stat Assoc.* 1989;84(408):897–905.
5. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol 1. The Analysis of Case-Control Studies*. Lyon, France: IARC Scientific Publication 32; 1980.
6. Campbell MJ. What is propensity score modelling? *Emerg Med J.* 2017;34(3):129–131.
7. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics.* 1968;24(2):295–313.
8. Curtis LH, Hammill BG, Eisenstein EL, Kramer JM, Anstrom KJ. Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Med Care*. 2007;45(10 Suppl 2):S103–S107.
9. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med.* 1998;17(19):2265–2281.
10. D'Agostino RB Jr. Propensity scores in cardiovascular research. *Circulation*. 2007;115(17):2340–2343.
11. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J Am Stat Assoc.* 1999;94(448):1053–1062.
12. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*. 1993;49(4):1231–1236.
13. Faries DE, Leon AC, Haro JM, Obenchain RL. *Analysis of Observational Health Care Data using SAS®*. Cary, NC: SAS Institute Inc.; 2010.
14. Fiebach NH, Cook EF, Lee TH, et al. Outcomes in patients with myocardial infarction who are initially admitted to stepdown units: data from the Multicenter Chest Pain Study. *Am J Med.* 1990;89(1):15–20.
15. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99(467):609–618.
16. Helmreich JE, Pruzek RM. PSAgraphics: an R package to support propensity score analysis. *J Stat Softw*. 2009;29(6):1–23.
17. Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Stat Med*. 2006;25(13):2230–2256.
18. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica.* 2003;71(4):1161–1189.
19. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw.* 2011;42(8):1–28.
20. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med.* 1998;17(14):1623–1634.
21. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86(1):4–29.
22. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937–2960.
23. Morgan SL, Winship C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press; 2007.
24. Muller JE, Turi ZG, Stone PH, et al; MILIS Study Group. Digoxin therapy and mortality after myocardial infarction. Experience in the MILIS Study. *N Engl J Med.* 1986;314(5):265–271.
25. Myers WO, Gersh BJ, Fisher LD, et al. Time to first new myocardial infarction in patients with mild angina and three-vessel disease comparing medicine and early surgery: a CASS registry study of survival. Coronary Artery Surgery Study. *Ann Thorac Surg*. 1987;43(6):599–612.
26. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. *Am J Epidemiol*. 1986;123(3):392–402.
27. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.
28. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82(398):387–394.
29. Rosenbaum PR. A characterization of optimal designs for observational studies. *J R Stat Soc Series B (Methodol)*. 1991;53(3):597–610.
30. Rosenbaum PR. *Observational Studies*. 2nd ed. New York: Springer-Verlag; 2002.
31. Rosenbaum PR. *Design of Observational Studies*. New York: Springer; 2009.
32. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41–55.
33. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984;79(387):516–524.
34. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc*. 2000;95(450):573–585.
35. Samet JH, Larson MJ, Horton NJ, Doyle K, Winter M, Saitz R. Linking alcohol and drug-dependent adults to primary medical care: a randomized controlled trial of a multi-disciplinary health intervention in a detoxification unit. *Addiction.* 2003;98(4):509–516.
36. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J Stat Softw*. 2011;42(7):1–52.
37. Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. *Stat Med*. 2014;33(23):4053–4072.
38. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med.* 2007;147(8):573–577.