Chapter 24
Propensity Scores and Propensity Score Matching for Assessing Multiple Confounders

1 General Purpose

In the Chap. 23 methods for assessing confounders were reviewed. Propensity score are ideal for assessing confounding, particularly, if multiple confounders are in a study. E.g., age and cardiovascular risk factors may not be similarly distributed in two treatment groups of a parallel-group study. Propensity score matching is used to make observational data look like randomized controlled trial data. This chapter assesses propensity score and propernsity score matching.

2 Schematic Overview of Type of Data File

Outcome	Treatment modality	Propensity scores
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.

T.J. Cleophas, A.H. Zwinderman, *Clinical Data Analysis on a Pocket Calculator*,
DOI 10.1007/978-3-319-27104-0_24

3 Primary Scientific Question

Is propensity score and propensity score matching adequate for assessing studies with multiple confounders.

4 Propensity Scores

A propensity (prop) score for age can be defined as the risk ratio (or rather odds ratio) of receiving treatment 1 compared to that of treatment 2 if you are old in this study.

	Treatment-1	Treatment-2	odds treatment-1 / odds treatment-2
	$n = 100$	$n = 100$	(OR)
1. Age > 65	63	76	0.54 (63/76 / 37/24)
2. Age < 65	37	24	1.85 ($= OR_2 = 1/OR_1$)
3. Diabetes	20	33	0.51
4. Not diabetes	80	67	1.96
5. Smoker	50	80	0.25
6. Not smoker	50	20	4.00
7. Hypertension	51	65	0.65
8. Not hypertension	49	35	1.78
10. Not cholesterol	39	22	2.27

The odds ratios can be tested for statistical significance (see Chap. 2, odds ratios), and those that are statistically significant can, then, be used for calculating a combined propensity-score for all of the inequal characteristics by multiplying the significant odds ratios, and, then, calculating from this product the combined propensity-score = combined "risk ratio" (= combined OR / (1+ combined OR). y = yes, n = no, combined OR $= OR_1 \text{ x } OR_3 \text{ x } OR_5 \text{ x } OR_7 \text{ x } OR_9$.

	Old	Diab	Smoker	Hypert	Cholesterol	Combined OR	Combined propensity score
Patient 1	y	y	n	y	y	7.99	0.889
2	n	n	n	y	y	105.27	0.991
3	y	n	n	y	y	22.80	0.958
4	y	y	y	y	y	0.4999	0.333
5	n	n	y				
6	y	y	y				
7	….						
8	….						

Each patient has his / her own propensity score based on and adjusted for the significantly larger chance of receiving one treatment versus the other treatment.

Usually, propensity score adjustment for confounders is accomplished by dividing the patients into four subgroups, but for the purpose of simplicity we here use 2 subgroups, those with high and those with low propensity scores.

Confounding is assessed by the method of subclassification. In the above example an overall mean difference between the two treatment modalities is calculated.

For treatment zero

Mean effect ± standard error (SE) = 1.5 units ± 0.5 units

For treatment one

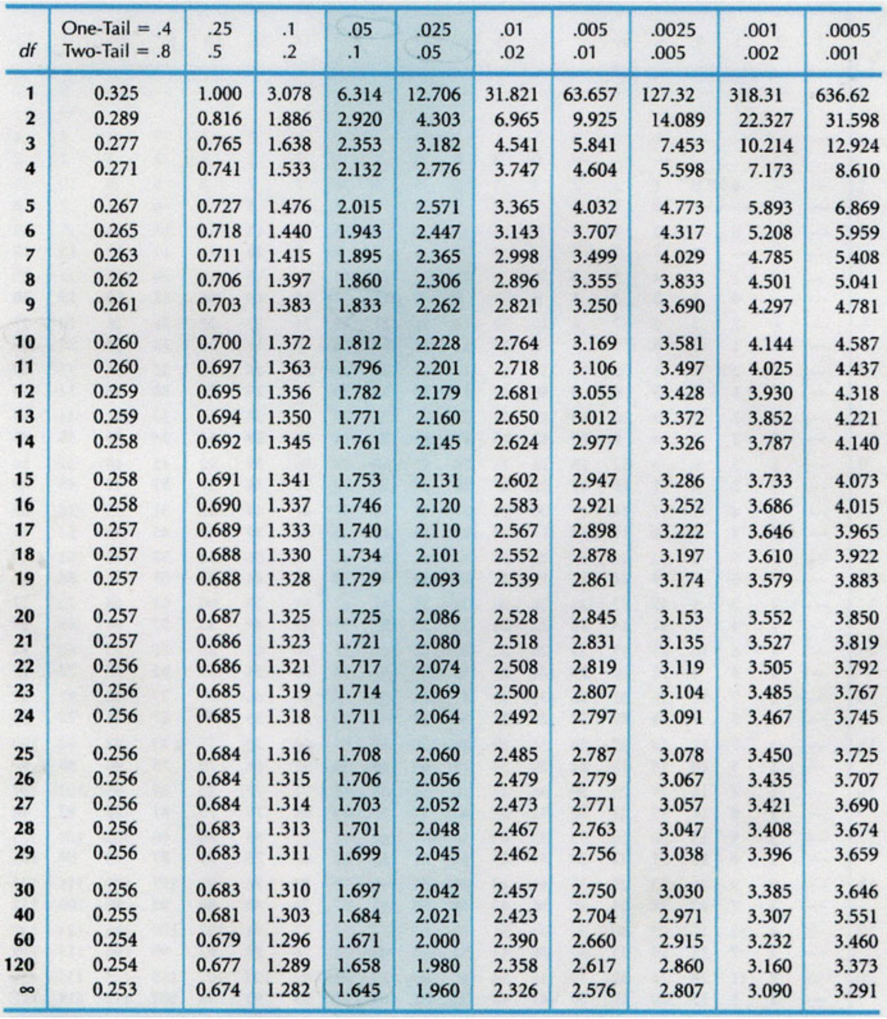
Mean effect ± SE = 2.5 units ± 0.6 units

The mean difference of the two treatments

$$
\begin{aligned}
&= 1.0 \text{ units} \pm \text{pooled standard error} \\
&= 1.0 \pm \surd\ (0.5^2 + 0.6^2) \\
&= 1.0 \pm 0.61
\end{aligned}
$$

The t-value as calculated $= 1.0/0.61 = 1.639$

The underneath t-table is helpful to determine a p-value.



df	One-Tail = .4	.25	.1	.05	.025	.01	.005	.0025	.001	.0005
	Two-Tail = .8	.5	.2	.1	.05	.02	.01	.005	.002	.001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

The t-table has a left-end column giving degrees of freedom (≈ sample sizes), and two top rows with p-values (areas under the curve = p – values), one-tail meaning that only one end of the curve, two-tail meaning that both ends are assessed simulataneously. The t-table is, furthermore, full of t-values, that, with ∞ degrees of freedom, are equal to z-values (Chap. 36). The t-values are to be understood as mean results of studies, but not expressed in mmol/l, kilograms, but in so-called SEM-units (Standard error of the mean units), that are obtained by dividing your mean result by its own standard error. With many degrees of freedom (large samples) the curve will be a little bit narrower, and more in agreement with nature.

With 200–2 (200 patients, 2 groups) = 198 degrees of freedom, a t-value > 1.96 is required to obtain a two-sided $p < 0.05$. It can be observed that our p-value > 0.05. It is even > 0.10.

In order to assess the possibility of confounding, a weighted mean has to be calculated. The underneath equation is adequate for the purpose (prop score = propensity score).

$$\text{Weighted mean} = \frac{\text{Difference}_{\text{high prop score}} \,/\, \text{its SE}^2 + \text{Difference}_{\text{low prop score}} / \text{its SE}^2}{1/\,\text{SE}^2{}_{\text{high prop score}} + 1/\,\text{SE}^2{}_{\text{low prop score}}}$$

For the high prop score we find means of 2.0 and 3.0 units, for the low prop score 1.0 and 2.0 units. The mean difference separately are 1.0 and 1.0 as expected. However, the pooled standard errors are different, for the males 0.4, and for the females 0.3 units.

According to the above equation a weighted t-value is calculated

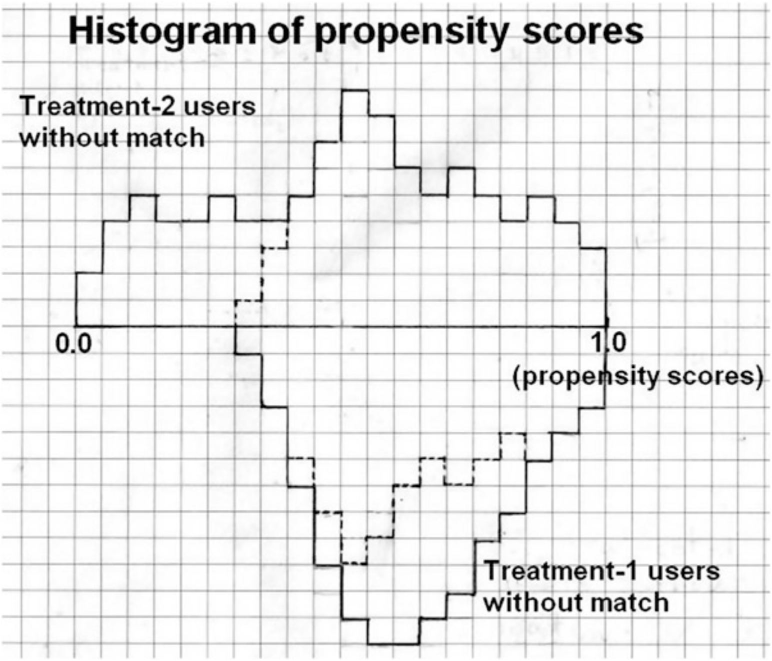
$$\begin{aligned}
\text{Weighted mean} &= \frac{\left(1.0/0.4^2 + 1.0/0.3^2\right)}{\left(1/0.4^2 + 1/0.3^2\right)} \\
&= 1.0 \\
\text{Weighted SE}^2 &= 1/(1/0.4^2 + 1/0.3^2) \\
&= 0.0576 \\
\text{Weighted SE} &= 0.24 \\
\text{t-value} &= 1.0/0.24 = 4.16
\end{aligned}$$

With 98 degrees of freedom, and a t-value of 4.16 means a two sided p-value < 0.001 is obtained.

The weighted mean is equal to the unweighted mean. However, its SE is much smaller. It means that after adjustment for the prop scores a very significant difference is observed. Instead of subclassification, also linear regression with the propensity scores as covariate is a common way to deal with propensity scores. However, this is hard on a pocket calculator.

5 Propensity Score Matching

In the study of 200 patients each patient has his/her own propensity score. We select for each patient in group 1 a patient from group 2 with the same propensity score.



The above graph is an example of the nearest neighbor watching method for matching patients with similar propensity scores. Each square represents one patient. In random order the first patient from group 1 is selected. Then, he/she is matched to the patient of group 2 with the nearest propensity score. We will continue until there are no longer similar propensity scores. Group 1 has to be summarized above the x-axis, group 2 below it. The patients with dissimilar propensity scores that cannot be matched, have to be removed from the analysis.

This procedure will end up sampling two new groups that are entirely symmetric on their subgroup variables, and can, thus, be simply analyzed as two groups in a randomized trial. In the given example two matched groups of 71 patients were left for comparison of the treatments. They can be analyzed for treatment differences using unpaired t-tests (Chap. 7) or chi-square tests (Chap. 38), without the need to further account confounding anymore.

6 Conclusion

Propensity score are for assessing studies with multiple confounding variables, e.g., age and cardiovascular risk factors, factors that are likely not to be similarly distributed in two treatment groups of a parallel-group study. Propensity score matching is used to make observational data look like randomized controlled trial data. This chapter assesses propensity score and propernsity score matching.

7 Note

More background, theoretical and mathematical information of propensity scores is given in Statistics applied to clinical studies 5th edition, Chap. 29, Springer Heidelberg Germany, 2012, from the same authors.