Chapter 14
Why Match? Matched Case-Control Studies

Sherri Rose, Mark J. van der Laan

Individually matched case-control study designs are common in public health and medicine, and conditional logistic regression in a parametric statistical model is the tool most commonly used to analyze these studies. In an individually matched case-control study, the population of interest is identified, and cases are randomly sampled. Each of these cases is then matched to one or more controls based on a variable (or variables) *believed* to be a confounder. The main potential benefit of matching in case-control studies is a gain in efficiency, not the elimination of confounding. Therefore, when are these study designs truly beneficial?

Given the potential drawbacks, including extra cost, added time for enrollment, increased bias, and potential loss in efficiency, the use of matching in case-control study designs warrants careful evaluation.

In this chapter, we focus on individual matching in case-control studies where the researcher is interested in estimating a causal effect, and certain prevalence probabilities are known or estimated. In order to eliminate the bias caused by the matched case-control sampling design, this technique relies on knowledge of the true prevalence probability $q_0 \equiv P_{X,0}(Y = 1)$ and an additional value:

$$\bar{q}_0(M) \equiv q_0 \frac{P_{X,0}(Y = 0 \mid M)}{P_{X,0}(Y = 1 \mid M)},$$

where M is the matching variable. We will compare the use of CCW-TMLEs in matched and unmatched case-control study designs as we explore which design yields the most information for the causal effect of interest. We assume readers have knowledge of the information presented in the previous chapter on independent case-control study designs.

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_14,

14.1 Data, Model, and Target Parameter

We define $X = (W, M, A, Y) \sim P_{X,0}$ as the experimental unit and corresponding distribution $P_{X,0}$ of interest. Here X consists of baseline covariates W, an exposure variable A, and a binary outcome Y, which defines case or control status. We can define $\psi_0^F = \Psi^F(P_{X,0}) \in \mathbb{R}^d$ of $P_{X,0} \in \mathcal{M}^F$ as the causal effect parameter, and for binary exposure $A \in \{0, 1\}$ we define the risk difference, relative risk, and odds ratio as in the previous chapter. The observed data structure in matched case-control sampling is defined by

$$O = \left((M_1, W_1, A_1), (M_0^j = M_1, W_0^j, A_0^j : j = 1, \ldots, J)\right) \sim P_0, \text{ with}$$

$$(M_1, W_1, A_1) \sim (M, W, A \mid Y = 1) \text{ for cases and}$$

$$(M_0^j, W_0^j, A_0^j) \sim (M, W, A \mid Y = 0, M = M_1) \text{ for controls.}$$

Here $M \subset W$, and M is a categorical matching variable. The sampling distribution of data structure O is described as above with P_0. Thus, the matched case-control data set contains n independent and identically distributed observations $O_1, \ldots, O_n$ with sampling distribution P_0. The cluster containing one case and the J controls is the experimental unit, and the marginal distribution of the cluster is specified by the population distribution $P_{X,0}$. The model $\mathcal{M}^F$, which possibly includes knowledge of q_0 or $\bar{q}_0(M)$, then implies models for the probability distribution of O consisting of cases (M_1, W_1, A_1) and controls (M_1, W_2^j, A_2^j), $j = 1, \ldots, J$.

14.2 CCW-TMLE for Individual Matching

CCW-TMLEs for individually matched case-control studies incorporate knowledge of q_0 and $\bar{q}_0(M)$, where $\bar{q}_0(M)$ is defined as

$$\bar{q}_0(M) \equiv q_0 \frac{P_{X,0}(Y = 0 \mid M)}{P_{X,0}(Y = 1 \mid M)} = q_0 \frac{q_0(0 \mid M)}{q_0(1 \mid M)}.$$

Implementation of CCW-TMLE in individually matched studies echos the procedure for independent (unmatched) case-control studies, with the exception that the weights now differ. We summarize this procedure assuming the reader is already familiar with the material in the previous chapter. We focus on the risk difference $\psi_{RD,0}^F = E_{X,0}[E_{X,0}(Y \mid A = 1, W) - E_{X,0}(Y \mid A = 0, W)]$ as an illustrative example.

Implementing CCW-TMLE for Individually Matched Data

Step 0. Assign weights q_0 to cases and $\bar{q}_0(M)/J$ to the corresponding J controls.

Step 1. Estimate the conditional probability of Y given A and W using super learning and assigned weights. The estimate of $P_{X,0}(Y = 1 \mid A, W, M) \equiv \bar{Q}_0(A, W, M)$ is $\bar{Q}_n^0(A, W, M)$. Let Q_n^0 be the estimate of the conditional mean and the case-control-weighted empirical distribution for the marginal distribution of W, representing the estimator of $Q_0 = (\bar{Q}_0, Q_{W,0})$.

Step 2. Estimate the exposure mechanism using super learning and weights. The estimate of $P_{X,0}(A \mid W, M) \equiv g_0(A \mid W, M)$ is $g_n(A \mid W, M)$.

Step 3. Determine a parametric family of fluctuations $Q_n^0(\epsilon)$ of Q_n^0 with fluctuation parameter ϵ, and a case-control-weighted loss function $L_{q_0}(Q) = q_0 L^F(Q)(M_1, W_1, A_1, 1) + (\bar{q}_0(M)/J) \sum_{j=1}^{J} L^F(Q)(M_1, W_2^j, A_2^j, 0)$ such that the derivative of $L^F(Q_n^0(\epsilon))$ at $\epsilon = 0$ equals the full-data efficient influence curve at any initial estimator $Q_n^0 = (\bar{Q}_n^0, Q_{W,n}^0)$ and g_n. Since initial Q_{Wn}^0 is the empirical distribution (i.e., case-control-weighted nonparametric maximum likelihood estimation), one only needs to fluctuate $\bar{Q}_n^0$ and the fluctuation function involves a choice of clever covariate chosen such that the above derivative condition holds. Calculate the clever covariate $H_n^*(A, W, M)$ for each subject as a function of $g_n(A \mid W, M)$:

$$H_n^*(A, W, M) = \left(\frac{I(A = 1)}{g_n(1 \mid W, M)} - \frac{I(A = 0)}{g_n(0 \mid W, M)} \right).$$

Step 4. Update the initial fit $\bar{Q}_n^0(A, W, M)$ from step 1 using the covariate $H_n^*(A, W, M)$. This is achieved by holding $\bar{Q}_n^0(A, W, M)$ fixed while estimating the coefficient ϵ for $H_n^*(A, W, M)$ in the fluctuation function using case-control-weighted maximum likelihood estimation. Let ϵ_n be this case-control-weighted parametric maximum likelihood estimator. The updated regression is given by $\bar{Q}_n^1 = \bar{Q}_n^0(\epsilon_n)$. No iteration is necessary since the next ϵ_n will be equal to zero. The CCW-TMLE of Q_0 is now $Q_n^* = (\bar{Q}_n^1, Q_{Wn}^0)$, where only the conditional mean estimator $\bar{Q}_n^0$ was updated.

Step 5. Obtain the substitution estimator of the target parameter by application of the target parameter mapping to Q_n^*:

$$\psi_n^* = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(q_0 \bar{Q}_n^1(1, W_{1,i}, M_{1,i}) + \frac{\bar{q}_0(M)}{J} \sum_{j=1}^{J} \bar{Q}_n^1(1, W_{2,i}^j, M_{1,i}) \right) \right.$$
$$\left. - \left(q_0 \bar{Q}_n^1(0, W_{1,i}, M_{1,i}) + \frac{\bar{q}_0(M)}{J} \sum_{j=1}^{J} \bar{Q}_n^1(0, W_{2,i}^j, M_{1,i}) \right) \right\}.$$

Step 6. Calculate standard errors, p-values, and confidence intervals based on the influence curve of the CCW-TMLE ψ_n^*. The influence curve can be selected to be the case-control-weighted full-data efficient influence curve (just as we defined the case-control-weighted full-data loss function).

14.3 Simulations

In the following simulation studies, we compare the CCW-TMLE in independent and individually matched study designs.

Simulation 1. Our first simulation study is designed to illustrate the differences between independent case-control sampling and matched case-control sampling in "ideal" situations where control information is not discarded (e.g., data collection is expensive, and covariate information is only collected when a control is a match). The population contained $N = 35{,}000$ individuals, where we simulated a 9-dimensional covariate $W = (W_i : i = 1, \ldots, 9)$, a binary exposure (or "treatment") A, and an indicator Y. These variables were generated according to the following rules: $P_{X,0}(W_i = 1) = 0.5$, $P_{X,0}(A = 1 \mid W) = \text{expit}(W_1 + W_2 + W_3 - 2W_4 - 2W_5 + 2W_6 - 4W_7 - 4W_8 + 4W_9)$, and $P_{X,0}(Y = 1 \mid A, W) = \text{expit}(1.5A + W_1 - 2W_2 - 4W_3 - W_4 - 2W_5 - 4W_6 + W_7 - 2W_8 - 4W_9)$.

Both the exposure mechanism and the conditional mean of Y given its parents were generated with varied levels of association with A and Y in order to investigate the role of weak, medium, and strong association between a matching variable W_i and A and Y. The corresponding associations can be seen in Table 14.1. For example, W_1 was weakly associated with both A and Y. Matching is only potentially beneficial when the matching variable is a true confounder.

Another illustration of the varied association levels can be seen in Table 14.2, where we display the probability an individual in the population was a case given $W_i = w$, all the nonmatching covariates (Z), and A. For example, let's say matching variable W_2 is *age* with 1 representing < 50 years old and 0 representing ≥ 50 years old. In this population, it was not very likely (0.013) that someone who is < 50 years old will become a case, while someone who is ≥ 50 years old has a much higher chance of becoming a case (0.047), given Z and A. Therefore, W_2, W_5, and W_8 represent situations where the distribution of W_i among cases and controls is very different. The covariates W_3, W_6, and W_9 represent situations where this difference is even more extreme.

The simulated population had a prevalence probability of $q_0 = 0.030$ and exactly 1,045 cases, and the true value of the odds ratio was given by $OR = 2.302$. We sampled the population using a varying number of cases $nC = (200, 500, 1000)$ in both matched and unmatched designs, and for each sample size we ran 1000 simulations. In each sample, the same cases were used for both designs. Controls were matched to cases in our matched simulations based on one variable (W_i) for both 1:1 and 1:2 designs. The causal odds ratio was estimated using a CCW-TMLE with correctly specified case-control-weighted logistic regressions.

The matched and unmatched designs performed similarly with respect to bias for the nine covariates (results not shown; Rose and van der Laan 2009). There were consistent increases in efficiency when the association between W_i and Y was high (W_3, W_6, and W_9), when comparing matched to independent. Results when the association with W_i and Y was medium (W_2, W_5, and W_8) were not entirely consistent, although covariates W_5 and W_8 did show increases in efficiency for the matched

Table 14.1 Simulated covariates

		Y		
	Association	Weak	Medium	Strong
	Weak	W_1	W_2	W_3
A	Medium	W_4	W_5	W_6
	Strong	W_7	W_8	W_9

Table 14.2 Simulated covariates: probabilities

W_i	$P_{X,0}(Y = 1 \mid W_i = 1, Z, A)$	$P_{X,0}(Y = 1 \mid W_i = 0, Z, A)$
W_1	0.039	0.021
W_2	0.013	0.049
W_3	0.003	0.060
W_4	0.021	0.040
W_5	0.013	0.047
W_6	0.003	0.061
W_7	0.040	0.023
W_8	0.013	0.046
W_9	0.004	0.066

Table 14.3 Simulation 1: MSE is mean squared error, RE is relative efficiency, and nC is number of cases

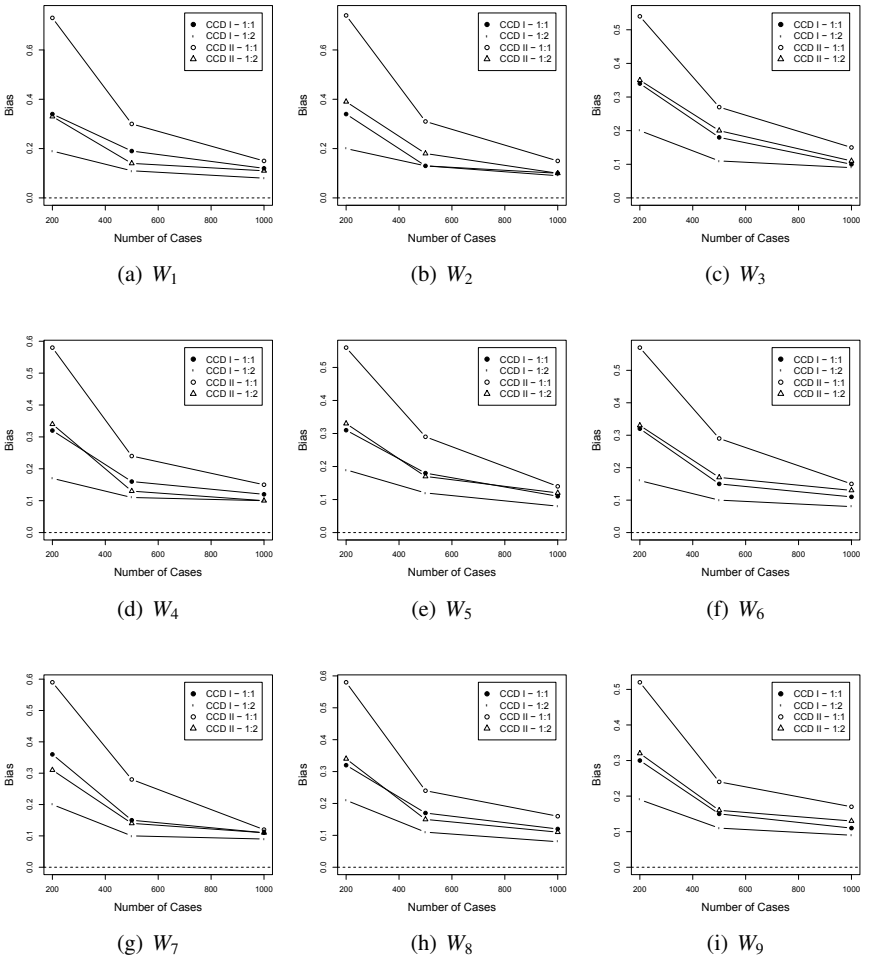
		1:1			**1:2**		
	nC	200	500	1000	200	500	1000
W_1	Matched MSE	2.67	0.77	0.30	0.98	0.32	0.14
	Independent RE	1.09	1.05	1.03	0.97	0.97	1.00
W_2	Matched MSE	2.63	0.70	0.33	1.07	0.40	0.15
	Independent RE	1.01	0.93	1.18	1.00	1.21	1.07
W_3	Matched MSE	1.95	0.59	0.23	0.93	0.29	0.13
	Independent RE	0.80	0.78	0.79	0.90	0.88	1.00
W_4	Matched MSE	2.20	0.64	0.30	1.05	0.32	0.14
	Independent RE	0.77	1.07	1.11	1.00	0.94	0.93
W_5	Matched MSE	2.10	0.61	0.28	0.98	0.30	0.14
	Independent RE	0.82	0.80	0.93	0.91	0.83	1.00
W_6	Matched MSE	2.28	0.61	0.24	0.92	0.27	0.12
	Independent RE	0.74	0.97	0.80	0.95	0.84	0.86
W_7	Matched MSE	2.55	0.69	0.30	1.08	0.32	0.16
	Independent RE	1.11	0.96	1.00	0.98	1.00	1.23
W_8	Matched MSE	2.00	0.61	0.22	0.86	0.25	0.11
	Independent RE	0.78	0.88	0.76	0.90	0.78	0.85
W_9	Matched MSE	1.77	0.58	0.24	0.71	0.24	0.12
	Independent RE	0.72	0.91	0.77	0.63	0.75	0.92

design for all or nearly all sample sizes. These results are in line with the consensus found in the literature: that matching may produce gains in efficiency when the distribution of the matching variable differs drastically between the cases and the controls. Efficiency results for the odds ratio can be seen in Table 14.3.

Simulation 2. The second simulation study was designed to address less ideal more common situations where control information is discarded. Controls were sampled from the population of controls in simulation 1 until a match on covariate W_i was found for each case. Nonmatches were returned to the population of controls. The number of total controls sampled to find sufficient matches was recorded for each simulation. This was the number of randomly sampled controls that was used for the corresponding independent case-control simulation. The mean number of controls sampled to achieve 1:1 and 1:2 matching at each sample size is noted in Table 14.4 as nCo. For example, in order to obtain 200 controls matched on covariate W_1 in a 1:1 design, an average of 404 controls had to be sampled from the population. Thus,

Table 14.4 Simulation 2: MSE is mean squared error, RE is relative efficiency, nC is number of cases, and nCo is mean number of controls for the independent case-control design

		1:1			**1:2**		
	nC	200	500	1000	200	500	1000
	nCo	404	1006	2010	804	2011	4026
W_1	Matched MSE	2.90	0.76	0.28	1.00	0.27	0.14
	Independent RE	2.89	2.24	2.14	2.12	1.70	2.16
	nCo	404	1009	2016	808	2016	4031
W_2	Matched MSE	2.91	0.77	0.30	1.15	0.36	0.16
	Independent RE	2.91	2.72	2.13	2.32	2.21	2.49
	nCo	406	1016	2033	812	2034	4065
W_3	Matched MSE	1.99	0.48	0.22	0.84	0.28	0.11
	Independent RE	1.82	1.43	1.65	1.81	1.78	1.85
	nCo	403	1006	2010	806	2012	4023
W_4	Matched MSE	2.47	0.67	0.29	1.09	0.28	0.13
	Independent RE	2.38	2.09	2.20	2.29	1.91	2.03
	nCo	406	1010	2019	810	2019	4040
W_5	Matched MSE	2.41	0.63	0.25	0.92	0.29	0.12
	Independent RE	2.24	2.00	1.92	1.95	1.89	2.10
	nCo	411	1025	2046	819	2045	4094
W_6	Matched MSE	2.08	0.64	0.23	0.88	0.27	0.13
	Independent RE	2.13	1.99	1.69	1.92	1.70	2.23
	nCo	402	1001	2000	801	1999	4000
W_7	Matched MSE	2.71	0.72	0.30	1.09	0.34	0.15
	Independent RE	2.54	2.42	2.18	2.19	2.25	2.18
	nCo	407	1014	2028	811	2027	4055
W_8	Matched MSE	2.28	0.56	0.23	0.97	0.25	0.11
	Independent RE	2.35	1.76	1.71	1.99	1.59	1.68
	nCo	413	1030	2059	824	2061	4121
W_9	Matched MSE	1.97	0.54	0.22	0.80	0.26	0.12
	Independent RE	1.91	1.77	1.69	1.62	1.69	1.84



(a) W_1 (b) W_2 (c) W_3

(d) W_4 (e) W_5 (f) W_6

(g) W_7 (h) W_8 (i) W_9

Fig. 14.1 Simulation 2 bias. CCD I is "Case-Control Design I" referring to the independent case-control design and CCD II is "Case-Control Design II" referring to the matched case-control design

an average of 404 controls were used in the corresponding independent case-control design.

CCW-TMLE was performed for both designs with correctly specified case-control-weighted logistic regression estimators for the exposure mechanism and conditional mean of Y given A and W. The independent design outperformed the matched design with respect to efficiency and bias for all sample sizes and both 1:1 and 1:2 matching. This was not surprising given the mean number of controls in each of the independent unmatched designs was, on average, about two times the number of controls for the matched design. Additionally, as association between W_i

and Y increased, there was a trend that the number of controls necessary for complete matching also increased. A similar trend between A and W_i was not apparent. Bias results do not vary greatly with association between W_i and A or Y. Efficiency results can be seen in Table 14.4. Bias results are displayed in Fig. 14.1.

14.4 Discussion

The main benefit of a matched case-control study design is a potential increase in efficiency. However, an increase in efficiency is not automatic. If one decides to implement a matched case-control study design, selection of the matching variable is crucial. In practice, it may be difficult to ascertain the strength of the association between the matching variable, the exposure of interest, and the outcome. Our simulations confirmed the consensus in the existing literature: that in situations where the distribution of the matching covariate is drastically different between the case and control populations, matching may provide an increase in efficiency. Our simulations indicated that $P_{X,0}(Y = 1 \mid W_i = 1, Z, A)$, for matching variable W_i and covariate vector Z, may need to be very small for an increase in efficiency using a matched design. These results were true, however, only for simulations where *no control subjects were discarded*; it is very common for matched study designs to discard controls (Freedman 1950; Cochran 1965; Billewicz 1965; McKinlay 1977). We showed that in practical situations (e.g., when controls are discarded), an unmatched design is likely to be a more efficient, less biased study design choice.

14.5 Notes and Further Reading

There is a collection of literature devoted to the topic of individual matching in case-control study designs, and discussion of the advantages and disadvantages of matching goes back more than 40 years. While some literature cites the purpose of matching as improving validity, later publications (Kupper et al. 1981; Rothman and Greenland 1998) demonstrate that matching has a greater impact on efficiency over validity. Costanza (1995) notes that matching on confounders in case-control studies does nothing to remove the confounding. Similarly, Rothman and Greenland (1998) discuss that matching cannot control confounding in case-control study designs but can, in fact, introduce bias. Methodologists in the literature stress that it is often possible and preferred for confounders to be *adjusted for* in the analysis instead of matching in case-control designs (Schlesselman 1982; Vandenbrouke et al. 2007).

Matching has a substantial impact on the study sample; most notably, it creates a sample of controls that is not representative of exposure in the population or the population as a whole. The effect of the matching variable can no longer be studied directly, and the exposure frequency in the control sample will be shifted towards that of the cases (Rothman and Greenland 1998).

Matched sampling leads to a balanced number of cases and controls across the levels of the selected matching variables. This balance can reduce the variance in the parameter of interest, which improves statistical efficiency. A study with a randomly selected control group may yield some strata with an imbalance of cases and controls. It is important to add, however, that matching in case-control studies can lead to gains *or* losses in efficiency (Kupper et al. 1981; Rothman and Greenland 1998). Matching variables are chosen a priori on the belief that they confound the relationship between exposure and disease. If controls are matched to cases based on a variable that is not a true confounder, this can impact efficiency. For example, if the matching variable is associated not with disease but with the exposure, this will increase the variance of the estimator compared to an unmatched design. Here, the matching leads to larger numbers of exposure-concordant case-control pairs, which are not informative in the analysis, leading to an increase in variance. If the matching variable is only associated with disease, there is often a loss of efficiency as well (Schlesselman 1982). If the matching variable is along the causal pathway between disease and exposure, then matching will contribute bias that cannot be removed in the analysis (Vandenbrouke et al. 2007). The number of matching variables should also be reduced to as few as possible. As the number of matching variables grows, the cases and controls will become increasingly similar with respect to the exposure of interest, and the study may produce a spurious result or provide no information (Breslow and Day 1980). Additionally, when matching on more than one variable, matching variables should not be strongly correlated with each other (Schlesselman 1982). This chapter was adapted from Rose and van der Laan (2009). We refer readers to this paper for additional discussion of the implications of individually matched designs.

Cochran (1953) demonstrates the efficiency of matched designs. However, as noted by McKinlay (1977), Cochran's result can be misleading. Comparisons between matched and unmatched study designs are often made with *equal* sample sizes and no other method of covariate adjustment. In a matched design, controls may be discarded if they do not match a particular case on the variable or variables of interest. Multiple controls may be discarded per case, depending on the variables of interest (Freedman 1950; Cochran 1965; McKinlay 1977). In many cases, if the discarded controls were available to be rejected in the matched study, they would be available for an unmatched design in the same investigation (Billewicz 1965; McKinlay 1977). Therefore, it is often more appropriate to compare the efficiencies of matched case-control studies of size n to randomly selected case-control studies of size n+*number of discarded controls*.

The predominant method of analysis in individually matched case-control studies is conditional logistic regression in a parametric statistical model. The logistic regression model for matched case-control studies differs from unmatched studies in that it allows the intercept to vary among the matched units of cases and controls. The matching variable is not included in the model (Breslow et al. 1978; Holford et al. 1978; Breslow and Day 1980; Schlesselman 1982). In order to estimate an effect of exposure A with conditional logistic regression, the case and control must be discordant on A. Rothman and Greenland (1998) and Greenland (2004) demonstrate

the use of standardization in case-control studies, which estimate marginal effects with population or person-time averaging.