Received: 27 January 2021 | Revised: 9 December 2021 | Accepted: 14 December 2021

DOI: 10.1002/sim.9313

RESEARCH ARTICLE

Statistics in Medicine WILEY

Conditional or unconditional logistic regression for frequency matched case-control design?

Fei Wan

Division of Public Health Sciences, Washington University in St. Louis, St. Louis, Missouri, USA

Correspondence
Fei Wan, Division of Public Health Sciences, Washington University in St. Louis, St. Louis, MO, USA.
Email: wan.fei@wustl.edu

Abstract

Frequency matching is commonly used in epidemiological case control studies to balance the distributions of the matching factors between the case and control groups and to improve the efficiency of case-control designs. Applied researchers have held a common opinion that unconditional logistic regression should be used to analyze frequency matched designs and conditional logistic regression is unnecessary. However, the justification of this view is unclear. To compare the performances of ULR and CLR in terms of simplicity, unbiasedness, and efficiency in a more intuitive way, we viewed frequency matching from the perspective of weighted sampling and derived the outcome models describing how the exposure and matching factors are associated with the outcome in the matched data separately in two scenarios: (1) only categorical variables are used for matching; (2) continuous variables are categorized for matching. In either scenario the derived outcome model is a logit model with stratum-specific intercepts. Correctly specified unconditional logistic regression can be more efficient than conditional logistic regression, particularly when continuous matching factors are used, whereas conditional logistic regression is a more practical approach because it is less dependent on modeling choices.

KEYWORDS

bias, case-control design, conditional logistic regression, frequency matching, unconditional logistic regression

1 | INTRODUCTION

Matching is commonly implemented in epidemiological case-control studies to balance the distributions of confounding variables between the case and control samples and improve the studies' efficiency, particularly when the matching factors are strong confounders.[1] Matching in case-control studies can be done in two different ways: individual matching or frequency matching. In an individually matched case-control study, each case is matched to one or multiple controls having similar matching factor values (eg, caliper matching to controls within ±5 years of the case's age). A main drawback of individual matching is that caliper matching on continuous factors could fail to find controls for every case and result in unmatched cases. Alternatively, frequency matching forms subgroups based on categorical or categorized matching factors and randomly selects controls in proportion to the number of cases from each subgroup. Thus, the distributions of matching factors are the same among cases and matched controls. While matching is intended to control for confounding, there is some disagreement on whether it can do so in case-control studies.[2]

Abbreviations: CLR, conditional logistic regression; ULR, unconditional logistic regression.

Applied researchers have held a common opinion that individually matched case-control designs should be analyzed using conditional logistic regression ("CLR") but unconditional logistic regression ("ULR") should be used for analyzing frequency-matched designs.[3] However, the literature has not been consistent with this conclusion.[4-7]

When matching cases and controls individually on continuous matching factors (eg, age) or nominal factors with many categories (eg, neighborhood or sibship), the number of matching pairs is usually high and the number of participants per matching pair is low. The ULR adjusting for many matched pairs as dummy regressors ("stratum specific intercepts") is not appropriate in this setting because maximum likelihood estimation can yield highly biased point estimates when the number of stratum specific intercepts is large. By contrast, CLR does not need to estimate stratum-specific intercepts and is a more robust choice.[4] To circumvent this limitation of ULR, controlling matching factors directly in ULR as regressors has been suggested as a more efficient alternative to CLR.[5] Wan et al[7] derived the outcome model for the individually matched design with continuous matching factors and demonstrated that fitting an ULR may need more complex modeling of continuous matching factors and model mis-specification can lead to biased estimates of the exposure effect. By contrast, CLR cancels out complex regression terms involving continuous matching factors in the likelihood function and is less dependent on modeling choices.

Since the number of participants per stratum is generally high for frequency-matched case control designs, some have argued that adjusted ULR, with matching factors included as covariates, is appropriate in frequency matching designs and the use of CLR is not necessary.[3] While this argument appears to address the limitations of the ULR that includes many matching pairs as dummy regressors, it does not directly clarify how we should adjust for matching factors in ULR. For example, is it adequate to only adjust for categorized matching factors in ULR or should we include these matching factors in their continuous forms? To examine this argument further and to compare ULR and CLR in a more intuitive way, we will first derive the outcome model describing how the exposure and matching factors are associated with the outcome in frequency matched designs. Individual matching in the case-control study can make the relationship between the outcome and continuous matching factors more complicated than their relationship in the source population.[7,8] It remains unclear how frequency matching may alter the relationship between the outcome and matching factors, which poses uncertainty on how we model matching factors appropriately.

To clarify these uncertainties in analyzing frequency matched designs, we derived the outcome models in the matched data separately in two different scenarios: (1) all matching factors are categorical variables; (2) some matching factors are continuous but categorized for matching. With the derived models, we assessed the relative complexity of using ULR verse CLR to estimate the exposure effect. We designed simulation studies to compare the potential biases and efficiency of various commonly used ULR and CLR models and to recommend the most appropriate analytic approach for the frequency matched design.

2 | ASSUMPTIONS AND METHOD

2.1 | Matching factors are categorical variables

Suppose we use the frequency matched case-control design to assess the association between a rare disease outcome Y ($Y = 1$ for a case; $Y = 0$ for a control) and a binary exposure variable E ($E = 1$ if exposed; $E = 0$ if not exposed) in the source population. For simplicity, we assume X_1 and X_2 denote two categorical confounding factors. Specifically, X_1 has I levels with each level denoted by $x_{1i}, \forall\, i = 1, 2, \dots I$. X_2 has J levels with each level denoted by $x_{2j}, \forall\, j = 1, 2, \dots J$. We can model X_1 using $I - 1$ dummy variables $\mathbf{D}_1 = (D_{11}, D_{12}, \dots, D_{1(I-1)})$, where D_{1i} is defined as follows:

$$D_{1i} = \begin{cases} 1, & \text{if } X_1 = x_{1(i+1)} \\ 0, & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots, I - 1,$$

and x_{11} is the reference level. Similarly, we can model X_2 using $J - 1$ dummy variables $\mathbf{D}_2 = (D_{21}, D_{22}, \dots, D_{2(J-1)})$, where D_{2j} is defined as follows:

$$D_{2j} = \begin{cases} 1, & \text{if } X_2 = x_{2(j+1)} \\ 0, & \text{otherwise} \end{cases}, \quad j = 1, 2, \dots, J - 1,$$

and x_{21} is the reference level.

We let $\mathbf{X} = \{X_1, X_2\}$ and further assume the following logit model that describes how E and $\mathbf{X}$ influence Y in the source population:

$$\begin{aligned}\text{logit}P(Y=1|E,\mathbf{X}) &= \text{logit}P(Y=1|E,\mathbf{D}_1,\mathbf{D}_2)\\ &= \beta_0 + \beta_1 E + \boldsymbol{\beta_2}\mathbf{D}_1 + \boldsymbol{\beta_3}\mathbf{D}_2 + \boldsymbol{\beta_4}\mathbf{D}_1\mathbf{D}_2\\ &= \beta_0 + \beta_1 E + \sum_{i=1}^{I-1}\beta_{2,i}D_{1i} + \sum_{j=1}^{J-1}\beta_{3,j}D_{2j} + \sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\beta_{4ij}D_{1i}D_{2j},\end{aligned} \tag{1}$$

where β_1 denotes the conditional exposure effect, $\boldsymbol{\beta_2} = (\beta_{21}, \beta_{22}, \dots, \beta_{2(I-1)})$ is $(I-1)\times 1$ coefficient vector for $\mathbf{D}_1$, $\boldsymbol{\beta_3} = (\beta_{31}, \beta_{32}, \dots, \beta_{3(J-1)})$ is $(J-1)\times 1$ coefficient vector for $\mathbf{D}_2$, and $\boldsymbol{\beta_4} = (\beta_{41}, \beta_{42}, \dots, \beta_{4(I-1)(J-1)})$ is $(I-1)(J-1)\times 1$ coefficient vector for the $\mathbf{D}_1$ by $\mathbf{D}_2$ interaction. Under the rare outcome assumption the logit model (1) can be approximated by a log-linear model.

The relationship between E and $\mathbf{X}$ in the source population is specified as follows:

$$\begin{aligned}\text{logit}P(E=1|\mathbf{X}) &= \text{logit}P(E=1|\mathbf{D}_1,\mathbf{D}_2)\\ &= \alpha_0 + \boldsymbol{\alpha_1}\mathbf{D}_1 + \boldsymbol{\alpha_2}\mathbf{D}_2 + \boldsymbol{\alpha_3}\mathbf{D}_1\mathbf{D}_2\\ &= \alpha_0 + \sum_{i=1}^{I-1}\alpha_{1,i}D_{1i} + \sum_{j=1}^{J-1}\alpha_{2,j}D_{2j} + \sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\alpha_{3ij}D_{1i}D_{2j},\end{aligned} \tag{2}$$

where $\boldsymbol{\alpha_1} = (\alpha_{11}, \alpha_{12}, \dots, \alpha_{1(I-1)})$ is the $(I-1)\times 1$ coefficient vector for $\mathbf{D}_1$, $\boldsymbol{\alpha_2} = (\alpha_{21}, \alpha_{22}, \dots, \alpha_{2(J-1)})$ is the $(J-1)\times 1$ coefficient vector for $\mathbf{D}_2$, and $\boldsymbol{\alpha_3} = (\alpha_{31}, \alpha_{32}, \dots, \alpha_{3(I-1)(J-1)})$ is the $(I-1)(J-1)\times 1$ coefficient vector for the $\mathbf{D}_1\mathbf{D}_2$ interaction.

Frequency matching forms $I \times J$ unique strata by each $\{X_1, X_2\}$ category combination, $i = 1, 2, \dots, I, j = 1, 2, \dots, J$. We denote the kth combination (x_{1i}, x_{2j}) by $\mathbf{x}_k$, $k = 1, 2, \dots, I \times J$. We let n_{1k} and n_{0k} denote the number of cases and controls in the stratum formed by $\mathbf{x}_k$. We have $n_{1k} \leq n_{0k}$ because the outcome is a rare disease. Next, we select all cases and controls in proportion to the number of cases in each stratum. Since cases are over-sampled and controls are under-sampled, the population that the frequency matched sample represents is different from the source population. Therefore, the outcome model for the matched data could be different from model (1) for the source population. We let S denote the selection process in the frequency matched design, in which $S = 1$ indicates that a subject is being randomly selected into the matched data and $S = 0$ indicates that this subject is not selected. The outcome model in the matched data can be derived as follows (Details in the Appendix A.1):

$$\begin{aligned}P(Y=1|E,\mathbf{X}=\mathbf{x}_k,S=1) &= \frac{1}{1+\frac{P(Y=1|X=x_k)}{P(Y=0|X=x_k)}e^{-\beta_0-\beta_1 E-\boldsymbol{\beta}_2\mathbf{D}_1-\boldsymbol{\beta}_3\mathbf{D}_2-\boldsymbol{\beta}_4\mathbf{D}_1\mathbf{D}_2}}\\ &= \frac{1}{1+e^{-c(\mathbf{x_k})-\beta_1 E}},\end{aligned} \tag{3}$$

or equivalently in a logit form,

$$\text{logit}P(Y=1|E,\mathbf{X}=\mathbf{x}_k,S=1) = c(\mathbf{x}_k) + \beta_1 E,$$

where $c(\mathbf{x}_k)$ is a complex stratum specific intercept term for each matching stratum. When the disease outcome is rare, $c(\mathbf{x}_k) \approx -\log\left(\frac{e^{\beta_1}-1}{1+e^{-\alpha_0-\alpha_1\mathbf{D}_1-\alpha_2\mathbf{D}_2-\alpha_3\mathbf{D}_1\mathbf{D}_2}}\right)$. $c(\mathbf{x}_k)$ is a nuisance term because it does not contain the exposure variable. $\frac{P(Y=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|\mathbf{X}=\mathbf{x}_k)}$ is actually the probability of randomly selecting n_{1k} controls from all n_{0k} controls in the matching stratum of $\mathbf{X} = \mathbf{x}_k$, whereas for cases, this probability is 1 because every case will be selected.

By comparing models (1) and (3), we can observe the following: (i) There is a new nuisance term $c(\mathbf{x}_k)$ in the logit outcome model (3) and $\frac{1}{1+e^{-\alpha_0-\alpha_1\mathbf{D}_1-\alpha_2\mathbf{D}_2-\alpha_3\mathbf{D}_1\mathbf{D}_2}}$ in $c(\mathbf{x}_k)$ comes from the exposure model (2). $c(\mathbf{x}_k)$ is introduced into model (3) through $\frac{P(Y=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|\mathbf{X}=\mathbf{x}_k)}$. (ii) All regressors from the outcome model (1) are canceled out in model (3), except the exposure variable. This cancellation is also attributable to $\frac{P(Y=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|\mathbf{X}=\mathbf{x}_k)}$ (details in Appendix A.1). The interpretation of (i) and (ii) is that matching in frequency matched case-control designs balances the distributions of $\mathbf{X}$ between cases and controls and thus ensures no “marginal” (unconditional) association between the matching factors and the outcome (ie, all regressors in model (1) are removed from model (3)). However, the case-control sampling in the design introduces a new term $c(\mathbf{x}_k)$. $\mathbf{X}$ is associated with E and model (3) shows $\mathbf{X}$ is “conditionally” associated with Y via $c(\mathbf{x}_k)$ conditioning on E in the matched data. Thus, $\mathbf{X}$ is confounder by definition in the matched data and still needs to be controlled for in the analysis.[9] Unadjusted ULR can result in an omitted-variable bias in the logistic regression model when $c(\mathbf{x}_k)$ is not accounted for. In general, unadjusted ULR tends to bias towards the null hypothesis and underestimates the true association between the exposure and the outcome. This bias is attributable to a mixture of confounding and the non-collapsibility of the odds ratio in a logit model.[10-12] Variability of $c(\mathbf{x}_k)$ determines the size of bias.

The outcome model (3) in frequency matched designs is an ULR with stratum-specific intercepts $c(\mathbf{x}_k)$. There are four different ways to model $c(\mathbf{x}_k)$:

Method 1: We can use dummy variables to model the stratum-specific effects. We first create dummy variables $T_t, t = 1, 2, \dots, I \times J - 1$ as follows:

$$T_t = \begin{cases} 1 & \text{if in the } \ t+1\text{th stratum formed by } \mathbf{x}_{t+1} \\ 0 & \text{otherwise} \end{cases},$$

where the stratum of $\mathbf{x}_1$ is the reference level. Model (3) can be re-parametrized as following:

$$\text{logit}P(Y = 1|E, T_t, S = 1) = \gamma_0 + \sum_{t}^{I\times J-1} \gamma_t T_t + \beta_1 E, \tag{4}$$

where $\gamma_0 = c(\mathbf{x}_1)$ and $\gamma_t = c(\mathbf{x}_{t+1}) - c(\mathbf{x}_1), t = 1, 2, \dots, I \times J - 1$.

Method 2: We can include dummy variables of the categorical matching factors $\mathbf{X}$ and their interaction directly as covariates as follows:

$$\begin{aligned} \text{logit}P(Y = 1|E, \mathbf{X}, S = 1) &= \text{logit}P(Y = 1|E, \mathbf{D}, S = 1) \\ &= \tilde{\beta}_0 + \beta_1 E + \sum_{i=1}^{I-1} \tilde{\gamma}_{2,i} D_{1i} + \sum_{j=1}^{J-1} \tilde{\gamma}_{3,j} D_{2j} + \sum_{i=1}^{I-1}\sum_{j=1}^{J-1} \tilde{\gamma}_{4,i,j} D_{1i} D_{2j}. \end{aligned} \tag{5}$$

Of note, $c(\mathbf{x}_k)$ can be expressed as a combination of $\tilde{\gamma}_0, \tilde{\gamma}_2$’s, and $\tilde{\gamma}_3$’s. For example, $\tilde{\gamma}_0 = c(\mathbf{x}_1), \tilde{\gamma}_{2,i} + \tilde{\gamma}_{3,j} + \tilde{\gamma}_{4,i,j} = c(\mathbf{x}_{t+1}) - c(\mathbf{x}_1), t = 1, 2, \dots, I \times J - 1; i = 1, 2, \dots, I - 1; j = 1, 2, \dots, J - 1$. Methods (1) and (2) are mathematically equivalent.

Method 3: If we ignore the interaction terms, we have the main effect model as follows:

$$\text{logit}P(Y = 1|E, \mathbf{X}, S = 1) = \tilde{\tilde{\gamma}}_0 + \tilde{\tilde{\beta}}_1 E + \sum_{i=1}^{I-1} \tilde{\tilde{\gamma}}_{2,i} D_{1i} + \sum_{j=1}^{J-1} \tilde{\tilde{\gamma}}_{3,j} D_{2j}. \tag{6}$$

If the population exposure model (2) contains the interaction terms, models (4) and (5) are correctly specified but model (6) is misspecified. If the interaction effect is not ignorable, we would expect that $\tilde{\tilde{\beta}}_1 \neq \beta_1$. If there exists no interact effect in model (1), all three models are unbiased in estimating β_1 but model(6) is more efficient because it does not estimate redundant interaction terms.

Method 4: We normally have no interest in estimating stratum-specific intercepts. CLR avoids the model misspecification problem by cancelling out the nuisance term $c(\mathbf{x}_k)$ in the likelihood function. In this case, we just fit a CLR including the exposure variable only.

2.2 | Continuous variables are categorized for matching

When matching factors are continuous, we need to categorize them first and then perform matching using categorized values. For simplicity, we will demonstrate using a single continuous confounder X. The outcome and exposure models (1) and (2) are simplified as:

$$\text{logit}P(Y = 1|E, X) = \beta_0 + \beta_1 E + f(X; \boldsymbol{\beta}_2),$$

and

$$\text{logit}P(E=1|X)=\alpha_0+g(X;\boldsymbol{\alpha}_1),$$

where $f(\cdot)$ and $g(\cdot)$ are arbitrary functions of X so that X can take arbitrary functional forms. For example, if X is only linearly associated with the outcome and the exposure in logit scale, $f(X;\boldsymbol{\beta}_2)=\beta_2X$ and $g(X;\boldsymbol{\alpha})=\alpha_1X$. If such associations are quadratic, $f(X;\boldsymbol{\beta}_2)=\beta_{21}X+\beta_{22}X^2$ and $g(X;\boldsymbol{\alpha}_1)=\alpha_{11}X+\alpha_{12}X^2$.

We categorize X into I distinct categories using $I-1$ knots x_i, $i=1,2,\dots,I-1$ such that $C=c_k$, where $c_k=[x_k,x_{k+1}), \forall\, k=1,2,\dots,K$. We can model C with $I-1$ dummy variables $\mathbf{D}=(D_1,D_2,\dots,D_{K-1})$ as follows:

$$D_j=\begin{cases}1, \text{ if } C=c_{j+1} \text{ or equivalently, } x_{j+1}\le X<x_{j+2}\\ 0, \text{ otherwise}\end{cases},$$

for $j=1,2,\dots,I-1$. $[x_1,x_2)$ or c_1 is the reference category.

Since we match cases and controls using categorized variable C, not continuous variable X, the probability of selecting a case or control with a given $x\in[x_{k+1},x_{k+2})$ is actually determined by c_k. As usual, the probability for selecting a case in any category does not depend on its exposure status. Every case will be selected and the probability of selecting a case with a given value x is

$$\begin{aligned}P(S=1|Y=1,E,X=x)&=P(S=1|Y=1,C=c_k)\\&=1.\end{aligned}$$

The probability of selecting a control with $x\in[x_k,x_{k+1}]$ is

$$\begin{aligned}P(S=1|Y=0,E,X=x)&=P(S=1|Y=0,C=c_k)\\&=\frac{P(Y=1|C=c_k)}{P(Y=0|C=c_k)}.\end{aligned}$$

To compute $P(Y|C)$, we need to compute $P(Y|E,C)$ first and thus we need to derive a population outcome model adjusting for E and C. When we replace X with dummy variables $\mathbf{D}$ in the outcome and exposure models, the exposure and outcome models become:

$$\text{logit}P(Y=1|E,\mathbf{D})=\tilde{\beta}_0+\tilde{\beta}_1E+\tilde{\beta}_2\mathbf{D},$$

where $\tilde{\beta}_2=(\tilde{\beta}_{21},\tilde{\beta}_{22},\dots,\tilde{\beta}_{2(I-1)})$, and

$$\text{logit}P(E=1|\mathbf{D})=\tilde{\alpha}_0+\tilde{\alpha}_1\mathbf{D},$$

where $\tilde{\alpha}_1=(\tilde{\alpha}_{21},\tilde{\alpha}_{22},\dots,\tilde{\alpha}_{2(I-1)})$

Next, we have

$$\frac{P(Y=1|C=c_k)}{P(Y=0|C=c_k)}=\left(\frac{e^{\tilde{\beta}_1}-1}{1+e^{-\tilde{\alpha}_0-\tilde{\alpha}_{1(k-1)}}}+1\right)e^{\tilde{\beta}_0+\tilde{\beta}_{2(k-1)}}.$$

Thus, the outcome model in the matched data is

$$\begin{aligned}P(Y=1|E,X=x,S=1)&=\frac{1}{1+\frac{P(S=1|Y=0,E,x)}{P(S=1|Y=1,E,x)}e^{-\beta_0-\beta_1E-f(x;\boldsymbol{\beta}_2)}}\\&=\frac{1}{1+\frac{P(Y=1|C=c_k)}{P(Y=0|C=c_k)}e^{-\beta_0-\beta_1E-f(x;\boldsymbol{\beta}_2)}}\\&=\frac{1}{1+e^{-c(k)-\beta_1E-f(X;\boldsymbol{\beta}_2)}}.\end{aligned}\tag{7}$$

Its equivalent logit form is

$$\text{logit}P(Y=1|E,X=x,S=1)=c(k)+\beta_1 E+f(X;\boldsymbol{\beta}_2).$$

We can derive the approximate expression for the stratum specific term $c(k)=-\log\left(\left(\frac{e^{\tilde{\beta}_1}-1}{1+e^{-\tilde{\alpha}_0-\tilde{\alpha}_{1(k-1)}}}+1\right)e^{\tilde{\beta}_0+\tilde{\beta}_{2(k-1)}}\right)+\beta_0$ (details in the Appendix A.2). We can make the following observations from the above derived results: (i) The outcome model (7) is a logit model with the stratum specific intercept term $c(\mathbf{x}_k)$; (ii) the matching variable X is still retained in the model (7) instead of being canceled out. The interpretation of (i) and (ii) is that when a continuous matching factor is categorized and matching is not exact, the case-control sampling introduces a complex nuisance term $c(k)$ but inexact matching can not cancel out the confounding term $f(X;\boldsymbol{\beta}_2)$.

Fitting a regular ULR including X could result in biased estimates of β_1 because this model only assumes a constant intercept and does not model $c(\mathbf{x}_k)$ properly. Fitting an unadjusted stratified model, which includes $c(k)$ only in ULR to accounts for stratum-specific intercepts, is also biased because this model fails to incorporate X as a regressor. This misspecified model could result in biased estimate of β_1 due to a mixture of confounding and non-collapsibility of the logistic regression model. The bias was determined by the size of $\boldsymbol{\beta}_2$. By contrast, a CLR with X as a regressor is a simpler way to estimate β_1 because it cancels out $c(k)$ and we only need to model X properly. It should be emphasized that if we do not model $f(X;\boldsymbol{\beta}_2)$ properly, the functional form of X, either CLR or ULR can still be biased. When there are more than one matching factor, the general form of the outcome model (7) in the frequency matched data follows the rules listed below:

(i) There is a term for stratum-specific intercepts. If categorical or categorized factors form K strata, the stratum specific intercept is generally expressed as $c(k)=-\log\left(\frac{P(Y=1|C=k)}{P(Y=0|C=k)}\right)+\beta_0,\ k=1,2,\ldots,K$. One way to model this term is to include categorical or categorized matching factors and their interactions as regressors.

(ii) Continuous matching factor needs to be included as covariate in its continuous form. It has the same functional form as in the population outcome model.

(iii) If the population outcome model contains categorical matching factors and their interactions, we do not need include them again as covariates in model (7) because these terms are already included in stratum-specific intercepts.

(iv) If the population outcome model has the interaction terms involving continuous matching factors, we need to include these interaction terms again in model (7).

(v) In more general scenarios in which the population outcome model (1) may include some confounders that are not used in matching and the interaction terms between the exposure and covariates, these terms will also be retained in the outcome models (3) or (7) for the matched data (details in Appendix A.3).

For example, when we have two matching factors X_1 (continuous) and X_2, and the outcome model in the source population is:

$$\text{logit}P(Y=1|E,X_1,X_2)=\beta_0+\beta_1 E+\beta_2 X_1+\beta_3 X_1^2+\beta_4 X_2+\beta_4 X_1 X_2,$$

If X_2 is continuous, all the regression terms involving X_1 and X_2 in the population outcome model are retained in the outcome model for the matched data. The outcome model in the matched data becomes

$$\text{logit}P(Y=1|E,X_1,X_2,S=1)=c(k)+\beta_1 E+\beta_2 X_1+\beta_3 X_1^2+\beta_4 X_2+\beta_4 X_1 X_2,$$

If X_2 is a binary dummy variable, its main effect term in the population outcome model will not be included in the new model because this effect is already included in the intercepts. Thus, the outcome model in the matched data becomes

$$\text{logit}P(Y=1|E,X_1,X_2,S=1)=c(k)+\beta_1 E+\beta_2 X_1+\beta_3 X_1^2+\beta_4 X_1 X_2.$$

3 | SIMULATION

3.1 | Simulation design

In this simulation study, we aimed to validate the derived theoretic results and to assess the potential biases of some commonly used ULR and CLR in analyzing frequency matched designs. We designed three separate simulation studies as follows:

3.1.1 | Assess the closed form expression of $c(\mathbf{x}_k)$ when matching factors are categorical variables

To examine the closed form expression of $c(\mathbf{x}_k)$ in model (3), we generated one discrete random variable $Z \sim P(Z = z) = \frac{1}{3}, z = 1, 2, 3$ and a Bernoulli random variable $X \sim Bernoulli(0.5)$. We created two dummy variables for Z, D_1, and D_2 with $z = 1$ as the reference level. We generated the exposure variable E and outcome variable Y using the following exposure and outcome models:

$$\text{logit}P(E = 1|X, D_1, D_2) = 0 + 0.928D_1 - 0.371D_2 - 0.5X - 0.6D_1X + 0.6D_2X,$$

and

$$\text{logit}P(Y = 1|E, X, D_1, D_2) = -4.5 + E + 0.894D_1 + 0.447D_2 - 0.5X.$$

Next, we did a cross-tab of the outcome and two confounders Z and X. For each combination of Z and X, we select all cases and randomly select equal number of controls. Last, we fit an ULR including matching strata as dummy variables and the exposure variable in the frequency-matched samples. We performed 10 000 simulations and 10 000 observations were generated for each simulation.

3.1.2 | Assess the derived outcome model when matching factors are continuous variables

To examine the closed-form expression of $c(\mathbf{x}_k)$ and functional form of continuous matching factor in model (7), we first generated the continuous confounder $Z \sim N(0, 5)$ and then categorized it into a discrete variable with six levels $(< -4; [-4, -2); [-2, 0); [0, 2); [2, 4); \geq 4)$. We next generated the exposure and outcome variables as follows:

$$\text{logit}P(E = 1|Z) = 2 - 0.8Z - 0.2Z^2,$$

and

$$\text{logit}P(Y = 1|E, Z) = \beta_0 + E + 0.2Z,$$

where $\beta_0 = -4$ for approximately 5% disease prevalence rate and $\beta_0 = -2.5$ for approximately 15% disease prevalence rate.

Next, we performed 10 000 simulations and generated 10 000 observations per simulation. In each simulation, we performed frequency matching based on categorized matching factor and fit an ULR including the matching strata as dummy variables in each matched sample. To get the estimates of $\tilde{\alpha}$'s and $\tilde{\beta}$'s, we generated 10 million observations and fit the exposure and outcome logit models including dummy variables for categorized Z.

3.1.3 | A comparison of ULR and CRL in different scenarios

To compare the performances of ULR vs CLR, we simulated the data under the following two settings:

(1) *Matching on categorical variables:* We first generated two discrete confounders: one discrete random variable $Z \sim P(Z = z) = \frac{1}{3}, z = 1, 2, 3$ and one binary random variable $X \sim Bernoulli(0.5)$. We created two dummy variables D_1 and

D_2 for Z with 1 as reference category. We then generated the binary exposure and outcome variables as follows:

$$\text{logit}P(E=1|X,D_1,D_2)=0+0.928D_1-0.371D_2-0.5X+\alpha_{31}D_1X+\alpha_{32}D_2X,$$

and

$$\text{logit}P(Y=1|E,X,D_1,D_2)=-4.5+0.8E+0.894D_1+0.447D_2+0.5X.$$

We set $(\alpha_{31},\alpha_{32})$ to be $(0,0)$, $(-0.8,0.8)$, $(-1.6,1.6)$ to represent different levels of interaction effects. Note that we allow the potential ZX interaction in the population exposure model. This is to induce the ZX interaction in the derived outcome model for the matched data even though there is no ZX interaction in the population outcome model.

(2) *Matching on categorized continuous variable:* We considered two different scenarios: (a) one categorical variable $X \sim$ *Bernoulli*(0.5) and one continuous variable $Z \sim N(0,5)$; (b) two continuous variables X and $Z \sim N(0,5)$. We generated the exposure and outcome variables using the following two models:

$$\text{logit}P(E=1|X,Z)=\alpha_0+\alpha_1Z+\alpha_2Z^2+\alpha_3X+\alpha_4ZX,$$

and

$$\text{logit}P(Y=1|E,X,Z)=\beta_0+\beta_1E+\beta_2Z+\beta_3Z^2+\beta_4X+\beta_5ZX.$$

For each setting, we set the regression coefficients as follows:

(1) $(\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4)=(0.5,-0.8,0,-0.5,0)$ and $(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)=(-4,0.8,0.2\ ,0,-0.2,0)$. Z and X are linear in the exposure and outcome models and there are no Z by X interaction terms in both models.

(2) $(\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4)=(0.5,-0.8,0.1,-0.5,0.2)$ and $(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)=(-4,0.8,0.2\ ,0.1,-0.2,0.1)$. Z and X have quadratic and interaction terms in the exposure and outcome models.

Values of β_0 were chosen to have less than 10% cases in all scenarios. For each scenario, we generated 20 000 observations for each sample and repeated 10 000 times. We analyzed each matched sample as follows:

(i) In the setting where Z and X are categorical variables, we first selected all cases and randomly selected the equal number of controls from each stratum formed by Z and X. We then fit unadjusted ULR ("ULR(E)"), ULR controlling for Z and X ("ULR(E,X,Z)"), and ULR including additional ZX interaction terms ("ULR(E,Z,X,ZX)"), unadjusted CLR using matched sets as stratifying variable ("CLR(E)").

(ii) In the setting in which X is discrete and Z is continuous, we grouped Z into six categories: $(-\infty,-4]$, $(-4,-2]$, $(-2,0]$, $(0,2]$, $(2,4]$, $(4,+\infty)$ when performing matching.

  (a) When Z has only a linear term but there is no ZX interaction in the outcome model, we fit ULR(E,Z,X), CLR(E), and CLR(E,Z). We also fit ULR(D_Z,X,E), which includes categorized Z, and the correctly specified model ULR(D_Z,X,D_ZX,E,Z), which uses D_Z,X,D_ZX to model stratum-specific intercepts and Z again to model its linear term.

  (b) When Z has a quadratic term and there is a ZX interaction in the outcome model, we fit ULR(E,Z,X), ULR(E,Z,Z^2,X,ZX), CLR(E), CLR(E,Z), CLR(E,Z,Z^2), and CLR(E,Z,Z^2,ZX). Next we fit two models using D_Z only, ULR(E,D_Z,X) and ULR(D_Z,X,D_ZX,E), and the correct model ULR(D_Z,X,D_ZX,E,Z,Z^2,ZX).

(iii) When X and Z are both continuous variables, we grouped both variables into four categories: $(-\infty,-3]$, $(-3,0]$, $(0,3]$, $(3,+\infty)$ when performing matching.

  (a) When Z has only a linear trend but there is no ZX interaction in the outcome model, we fit ULR(E,Z,X), CLR(E), and CLR(E,Z,X). We next fit ULR(D_Z,D_X,E), which includes categorized variables only, and the correct model ULR(D_Z,D_X,D_ZD_X,E,Z,X).

(b) When Z has a quadratic term and there is a ZX interaction in the outcome model, we fit ULR(E, Z, X), ULR(E, Z, Z^2, X, ZX), CLR(E), CLR(E, Z, X), CLR(E, Z, Z^2, X), and CLR(E, Z, Z^2, X, ZX). We next fit ULR(D_Z, D_X, E) and ULR(D_Z, D_X, D_ZD_X, E), and the correct model ULR$(D_Z, D_X, D_ZD_X, E, Z, Z^2, X, ZX)$.

For every ULR and CLR, we computed the averaged estimate of β_1, the model-based and empirical SEs, and root mean square errors.

3.2 | Simulation results

Tables 1 and 2 list the comparison results for the simulation studies outlined in Sections 3.1.1 and 3.1.2. We can observe that the values of each stratum specific parameter for the nuisance term computed via the formula and their estimates via ULR are very close when matching factors are categorical (Table 1), and when continuous matching factors are categorized (Table 2). Table 2 also shows that the coefficients of the other regression terms in model (7) are close to their estimates and the derived model is correct.

Table 3 compares the performances of ULRs and CLRs under three different settings:

- In the setting where both Z and X are categorical variables, CLR(E) and correctly specified ULR, which is ULR(E, Z, X) if there is no ZX interaction or ULR(E, Z, X, ZX) if there is interaction, are both unbiased. ULR(E) always underestimates β_1 (Scenarios 1-3). In presence of the interaction effect, ULR(E, Z, X) is misspecified without including ZX but the bias is minimal unless the interaction effect is very large (Scenario 3).
- In the setting where Z is continuous and X is binary (Scenarios 4 and 5), misspecifying either continuous matching factors or intercepts in ULRs are biased. When Z is linear and there is no ZX interaction in Scenario 4, CLR(E) is

TABLE 1 A comparison of analytic results and estimates of the nuisance term $c(\mathbf{X}_k)$ when matching factors are categorical

Stratum	Z	X	Analytic value	Averaged estimate
1	1	0	−0.6201145	−0.6260918
2	2	0	−0.8027216	−0.8098322
3	3	0	−0.5314581	−0.5332618
4	1	1	−0.5	−0.5061444
5	2	2	−0.5797767	−0.5832976
6	3	3	−0.5557626	−0.5608431

TABLE 2 A comparison of analytic results and estimates of the nuisance term and regression terms when continuous matching factor is categorized

Coefficient	Analytic value ($\beta_0 = -4$)	Averaged estimate ($\beta_0 = -4$)	Analytic value ($\beta_0 = -2.5$)	Averaged estimate ($\beta_0 = -2.5$)
Stratum $c(k)$				
$c(1)$	0.6647386	0.6713543	0.6754585	0.6871749
$c(2)$	−0.3730167	−0.3829793	0.3662463	0.3634882
$c(3)$	−0.7516625	−0.7794619	−0.7572519	−0.7537089
$c(4)$	−0.981386	−1.0020258	−0.9834139	−0.9609605
$c(5)$	−0.8037533	−0.8092525	−0.8119819	−0.7790259
$c(6)$	−1.471726	−1.4942983	−1.418277	−1.4191985
β_1	1	1.0076145	1	1.0017728
β_2	0.2	0.2017952	0.2	0.2003608

TABLE 3 Simulation results of comparing ULR and CLR in frequency matched design

Setting	β_1	Scenario	Model	Averaged estimate	Model *s.e*	Empirical *s.e*	Root MSE
$Z \sim P(Z = z) = \frac{1}{3}$	0.8	(1) No ZX	ULR(E)	0.7407	0.1414	0.1361	0.1484
$X \sim Bernoulli(0.5)$			ULR(E, Z, X)	0.8047	0.1479	0.1483	0.1484
			CLR	0.8014	0.1477	0.1494	0.1494
		(2) 0.8ZX	ULR(E)	0.7591	0.1408	0.1377	0.1437
			ULR(E, Z, X)	0.7844	0.1433	0.1423	0.1432
			ULR (E, Z, X, ZX)	0.8070	0.1456	0.1467	0.1468
			CLR(E)	0.8019	0.1452	0.1457	0.1457
		(3) 1.6 ZX	ULR(E)	0.7320	0.1397	0.1341	0.1504
			ULR(E, Z, X)	0.7382	0.1403	0.1354	0.1488
			ULR (E, Z, X, ZX)	0.8073	0.1473	0.1484	0.1486
			CLR(E)	0.8022	0.1468	0.1475	0.1475
$Z \sim N(0, 5)$	0.8	(4) Linear Z	ULR (E, Z, X)	0.9954	0.1743	0.1659	0.2563
$X \sim Bernoulli(0.5)$		No ZX	ULR (D_Z, X, E)	0.7144	0.1610	0.1615	0.1828
			ULR ($D_Z, X, D_Z X, E, Z$)	0.8102	0.1623	0.1631	0.1634
			CLR (E)	0.7130	0.1977	0.1995	0.2177
			CLR(E, Z)	0.8023	0.1984	0.1995	0.1996
		(5) Quadratic Z	ULR (E, Z, X)	0.4360	0.0792	0.0750	0.3713
		ZX	ULR(E, Z, Z^2, X, ZX)	0.7591	0.0891	0.0864	0.0956
			ULR (D_Z, X, E)	0.5652	0.0720	0.0719	0.2455
			ULR ($D_Z, X, D_Z X, E$)	0.5720	0.0724	0.0727	0.2393
			ULR($D_Z, X, D_Z X, E, Z, Z^2, ZX$)	0.8030	0.0760	0.0762	0.0762
			CLR(E)	0.5710	0.0887	0.0894	0.2458
			CLR(E, Z)	0.5883	0.0891	0.0890	0.2297
			CLR(E, Z, Z^2)	0.7983	0.0933	0.0940	0.0940
			CLR(E, Z, Z^2, ZX)	0.8012	0.0935	0.0941	0.0941
$Z \sim N(0, 5)$	0.8	(6) Linear Z	ULR (E, Z, X)	0.7530	0.1255	0.1199	0.1290
$X \sim N(0, 5)$		No ZX	ULR (D_Z, D_X, E)	0.6951	0.0971	0.0943	0.1411
			ULR($D_Z, D_X, D_Z D_X, E, Z, X$)	0.8056	0.1136	0.1140	0.1141
			CLR(E, X)	0.7271	0.1222	0.1229	0.1429
			CLR(E, Z, X)	0.8019	0.1390	0.1401	0.1401
		(7) Quadratic Z	ULR (E, Z, X)	−0.0956	0.1024	0.1001	0.9010
		ZX	ULR (E, Z, Z^2, X, ZX)	0.9917	0.1249	0.1267	0.2301
			ULR (D_Z, D_X, E)	0.2031	0.1077	0.1048	0.6060
			ULR ($D_Z, D_X, D_Z D_X, E$)	0.2163	0.1112	0.1116	0.5942
			ULR($D_Z, D_X, D_Z D_X, E, Z, Z^2, X, ZX$)	0.8076	0.1344	0.1345	0.1347
			CLR(E)	0.2151	0.1109	0.1109	0.5954
			CLR(E, Z, X)	−0.0121	0.12376	0.1315	0.8227
			CLR(E, Z, Z^2, X)	0.1304	0.1288	0.1458	0.6848
			CLR(E, Z, Z^2, X, ZX)	0.8040	0.1340	0.1345	0.1337

biased. CLR(E, Z), without including X, is unbiased. In Scenario 5 where Z has a quadratic effect and there is a ZX interaction, CLR(E, Z, Z^2, ZX) is unbiased. Not including Z^2 and ZX in CLR(E, Z) produces large bias. Omitting ZX only in CLR(E, Z, Z^2) has minor impact on bias.

- In the setting where both Z and X are continuous, we can observe similar patterns. In Scenario 6, ULR(E, Z, X) is biased because ULR wrongly fits a constant intercept. By contrast, CLR(E, Z, X) does not need to model stratum-specific intercepts. In Scenario 7, ULR has large bias when quadratic and interaction terms are not included. Even when we specify Z and X properly, ULR(E, Z, Z^2, X, ZX) is still biased. By contrast, CLR(E, Z, Z^2, X, ZX) is unbiased. However, not including the interaction and quadratic terms in CLR leads to large biases.
- In Scenarios 1-6, the model-based SEs of ULRs and CLRs are generally very close to their empirical SEs even when they are misspecified. This suggests model misspecification generally does not bias variance estimates in the frequency matched design. However, in Scenario 7, the model-based SEs of CLR(E, Z, X) and CLR(E, Z, Z^2, X) may underestimate true SEs when the quadratic or interaction term of Z is misspecified.
- In Scenarios 1-3 and 7, correctly specified ULR(E, Z, X, ZX) and CLR(E) have comparable variance estimates and root MSE. In Scenarios 4-6, correctly specified ULRs tend to have smaller variance estimates and root MSE than correctly specified CLRs.

The averaged estimates of coefficients of regressors involving continuous Z in Scenarios 5-7 are close to the true parameters (the Supplementary Table 2). Thus, the functional forms of matching factors derived in model (7) are validated.

4 | DATA APPLICATION

We reanalyzed the Environment and Genetics in Lung Cancer Etiology (EAGLE) study.[13] EAGLE is a population-based case-control study performed Italy to assess the association between exposure to outdoor particulate matter with aerodynamic diameter $\leq 10\ \mu$m (PM$_{10}$) and lung cancer risk. The study enrolled 2099 cases and 2120 controls. Cases and controls are frequency-matched for area of residence (five areas), gender, and five-year age classes in the range 35-79 years. The annual average PM$_{10}$ estimates at residence address in year 2000 are a surrogate of the etiologically relevant exposure occurring many years before cancer diagnosis. We categorized continuous PM$_{10}$ into a binary exposure variable (1 if higher than the median-47.76; 0 if lower than the median). We fit the following ULR and CLR models:

(i) Unadjusted ULR including PM$_{10}$ only;
(ii) ULR adjusting for matching strata as dummy regressors, with and without continuous age;
(iii) ULR adjusting for matching factors (gender, area of residence, age categories), with and without age. Comparing to ULR in (ii), interactions between matching factors are ignored.
(iv) ULR adjusting for matching strata, additional confounders such as education level (none, elementary, middle, high, university), and smoking variables including ever smoked cigarettes, mean-centered pack-years (linear, quadratic, and cubic components), years since quitting (categorical: 0 for never/current smokers; otherwise, 0.5-0.9, 1-1.9, 2-4.9, 5-9.9, 10-19.9, 20-29.9, or 30+ years), ever smoking of other types of tobacco (cigars, cigarillos, pipe), and ever exposed to environmental tobacco smoking (at home in childhood or in adult life at home or at workplace).
(v) Unadjusted CLR including PM$_{10}$ only.
(vi) CLR adjusting for continuous age.
(vii) CLR adjusting for continuous age and additional confounders.

The analysis results are presented in Table 4. Both ULR adjusting for matching strata, age, and all other confounders and CLR adjusting for age and all other confounders gave very similar estimates of both PM$_{10}$ and age (and their SEs). CLR adjusting for age and ULR adjusting for matching strata and age also produced very similar estimates of the exposure effect and SEs. These results are expected because ULRs adjusting for matching strata takes the interactions among matching factors into account and thus produce estimates equivalent to CLRs. The estimate of the exposure effect (0.150) using ULR adjusting for main effects terms of matching factors and age is larger than the estimate of the exposure effect (0.136) using ULR adjusting for matching strata (as dummies) and age. This difference could be explained by the fact that ULR adjusting for matching factors ignored the interactions among these matching factors. Unadjusted ULR has the smallest estimate because this model does not adjust for any confounders. Unadjusted CLR and CLR adjusting for age have very similar estimates of the exposure effect because age is relatively a weak confounder (estimate=0.037, OR=1.037).

TABLE 4 Analysis of the Environment and Genetics in Lung Cancer Etiology study

Models	Effect	Estimate (SE)	Odds ratio (95% CI)
Unadjusted ULR	PM10	0.029	1.029
		(0.068)	[0.901,1.176]
ULR+matching stratum	PM10	0.137	1.147
		(0.086)	[0.969,1.356]
ULR+matching factors	PM10	0.151	1.163
		(0.085)	[0.984,1.374]
ULR+matching stratum+Age	PM10	0.136	1.145
		(0.086)	[0.968,1.355]
	Age	0.042	1.043
		(0.024)	[0.994,1.094]
ULR+matching factors+Age	PM10	0.150	1.162
		(0.085)	[0.983,1.373]
	Age	0.037	1.037
		(0.024)	[0.990,1.087]
ULR+matching stratum+Age	PM10	0.193	1.213
+ all other confounders		(0.102)	[0.993, 1.483]
	Age	0.039	1.039
		(0.029)	[0.982,1.100]
Unadjusted CLR	PM10	0.135	1.145
		(0.085)	[0.969,1.353]
CLR+Age	PM10	0.135	1.143
		(0.085)	[0.968, 1.351]
	Age	0.041	1.042
		(0.024)	[0.994,1.092]
CLR+Age	PM10	0.192	1.211
+ all other confounders		(0.102)	[0.993,1.478]
	Age	0.037	1.037
		(0.029)	[0.981,1.097]

Including continuous age or not does not impact the estimates of the exposure effect in this case. Thus, correctly specified ULR and CLR produce equivalent estimates but fitting CLR is a simpler approach relative to ULR because CLR avoids the modeling of intercepts.

5 | DISCUSSION

ULR with matching factors included as covariates is commonly suggested for analyzing the frequency matched design. The justification is based on the argument that the number of matching strata is small relative to the number of subjects and the potential bias of fitting ULR with many nuisance parameters is minimized.[3] This recommendation may be reasonable when matching factors are categorical variables. When continuous matching factors are categorized, CLR could still be a more practical choice for applied researchers.

To determine which method, ULR or CLR, is a better choice, we derived the outcome model in the matched data by viewing frequency matching as a weighted sampling design because frequency matching over-samples cases and under-samples controls. With the derived outcome model, we can make a more informed decision on which analytic approach is a more practical choice.

When matching factors are all categorical variables, the outcome model for the matched data is an ULR including stratum-specific intercepts and the exposure. We can include the categorical matching factors to account for stratum-specific intercepts. When there is the interaction effect among matching factors in the population exposure model, ULR controlling for matching factors and their interactions is appropriate to use in the matched data. Otherwise, ULR without interaction is the correct model. By contrast, CLR is a simpler solution because it does not need to model stratum specific intercepts. In this case, our simulation results show that there is no significant efficiency gain with using ULR.

When continuous matching factors are categorized for matching, the outcome model for the matched data is an ULR. In this case, correct modeling requires: (i) the inclusion of the stratum-specific intercepts. We can include categorical or categorized matching factors to account for stratum-specific intercepts; (ii) modeling continuous matching factors correctly. Continuous matching factors need to be controlled for in the model again. Therefore, ULR only controlling for categorized matching factors is biased because it only accounts for intercepts, whereas ULR controlling for continuous matching factors only is also biased because it has a constant intercept. By contrast, CLR offers a simpler solution. In certain scenarios, ULR can provide more efficient estimates than CLR. It is also worthy of noting that when the number of nuisance stratum-specific intercepts increases, even correctly specified ULR becomes susceptible to large bias.[4]

Matching in a cohort study by the treatment or exposure status could remove the confounding effect of matching factors and should also make the estimates of the exposure effect less sensitive to particular outcome model specifications.[14] However, this is not true in case-control studies. We confirmed the previous conclusion that matching in case-control design not only fails to remove confounding but also add selection bias.[15] This selection bias can be controlled by the inclusion of stratum-specific intercepts. Matching in frequency matched designs makes the estimates of the exposure effect more sensitive to modeling choices. Thus, even though we conclude that CLR is a simpler and more practical choice for applied researchers than ULR when analyzing a frequency matched design, caution should still be taken even when fitting a CLR in the matched data. There are many well-established methods to assess the functional forms of continuous matching factors and the potential interaction effects involving continuous matching factors in logistic regression model.[16]

ACKNOWLEDGEMENTS

We thank Dr. Peter Imrey from Department of Quantitative Health Sciences at Cleveland Clinic for his critical review of the manuscript. We also thank the anonymous referees for their useful suggestions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at http://doi.org/10.1371/journal.pone.0203539

ORCID

Fei Wan https://orcid.org/0000-0002-8311-5553

REFERENCES

1. Thomas DC, Greenland S. The relative efficiencies of matched and independent sample designs for case-control studies. *J Chronic Dis*. 1983;36(10):685-697.
2. Rose S, Laan MJ. Estimation based on case-control designs with known prevalence probability. *Int J Biostat*. 2008;4(1):17.
3. Cheung YB. Analysis of matched case-control data. *J Clin Epidemiol*. 2003;56(8):814.
4. Breslow E, Day NE. *Statistical Methods in Cancer Research: Volume 1 - The Analysis of Case-Control Studies*. Lyon: IARC Scientific Publications; 1980.
5. Pearce N. Analysis of matched case-control studies. *BMJ*. 2016;352:i969.
6. Levin B, Paik MC. The unreasonable effectiveness of a biased logistic regression procedure in the analysis of pair-matched case-control studies. *J Stat Plann Infer*. 2001;96(2):371-385.
7. Wan F, Colditz G, Sutcliffe S. Matched versus unmatched analysis of matched case-control studies. *Am J Epidemiol*. 2021;109(9):1859-1866.
8. Greenland S. Partial and marginal matching in case-control studies. In: Moolgavkar SH, Prentice RL, eds. *Modern Statistical Methods in Chronic Disease Epidemiology*. New York, NY: Wiley; 1986:35-49.
9. VanderWeele TJ, Shpitser I. On the definition of a confounder. *Ann Stat*. 2013;41(1):196-220.
10. Wan F, Mitra N. An evaluation of bias in propensity score-adjusted nonlinear regression models. *Stat Methods Med Res*. 2018;27(3):846-862.
11. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci*. 1999;14(1):29-46.

12. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*. 1993;80(4):807-815.
13. Consonni D, Carugno M, De Matteis S, et al. Outdoor particulate matter (PM10) exposure and lung cancer risk in the EAGLE study. *PLoS One*. 2018;13(9).
14. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21.
15. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia, PA: Lippincott, Williams and Wilkins; 1998.
16. Harrell FE Jr. *Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer-Verlag; 2001.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Wan F. Conditional or unconditional logistic regression for frequency matched case-control design?. *Statistics in Medicine*. 2022;41(6):1023-1041. doi: 10.1002/sim.9313

APPENDIX A

A.1 Derivation of unconditional logit model in frequency matched data when matching factors are categorical variables

In a matched case control study, we match the cases and the same number of controls using confounder X_1 and X_2. X_1 has I distinct values $x_{1i}, i = 1, 2, \ldots, I$. X_2 has J distinct values $x_{2j}, j = 1, 2, \ldots, J$. Each distinct combination of values of X_1 and X_2 forms a stratum and there are $I \times J$ distinct strata. For example, the kth stratum formed by the values of $\mathbf{X} = (X_1, X_2)$ can be expressed as follows:

$$\mathbf{x}_k = (x_{1i}, x_{2j}), i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2, \ldots, I \times J,$$

We let n_{1k} and n_{0k} denote the number of cases and controls in the kth stratum formed by $\mathbf{x}_k, k = 1, 2, \ldots, I \times J$. $n_{1k} \leq n_{0k}$ because the outcome is a rare disease. When forming the frequency matched case-control sample, we need to make sure that the distributions of matching factors are balanced between the cases and controls by selecting a equal or a multiple of number of cases for controls in each stratum. Within this matching stratum, we let S denote the selection process with $S = 1$ for a subject being selected into the matched case control data and $S = 0$ for not being selected. Then, we can derive the conditional probability of having a disease outcome for a subject selected into the kth matching stratum, in which confounders of selected cases and controls are set to x_{1i} and x_{2j}, as follows:

$$\begin{aligned}
P(Y = 1|E, \mathbf{X} = \mathbf{x}_k, S = 1) &= \frac{P(Y = 1, E, \mathbf{X} = \mathbf{x}_k, S = 1)}{P(E, \mathbf{X} = \mathbf{x}_k, S = 1)} \\
&= \frac{P(Y = 1, E, \mathbf{X} = \mathbf{x}_k, S = 1)}{P(Y = 1, E, \mathbf{X} = \mathbf{x}_k, S = 1) + P(Y = 0, E, \mathbf{X} = \mathbf{x}_k, S = 1)} \\
&= \frac{1}{1 + \frac{P(Y=0,E,\mathbf{X}=\mathbf{x}_k,S=1)}{P(Y=1,E,\mathbf{X}=\mathbf{x}_k,S=1)}} \\
&= \frac{1}{1 + \frac{P(S=1|Y=0,E,\mathbf{X}=\mathbf{x}_k)P(Y=0|E,\mathbf{X}=\mathbf{x}_k)}{P(S=1|Y=1,E,\mathbf{X}=\mathbf{x}_k)P(Y=1|E,\mathbf{X}=\mathbf{x}_k)}}.
\end{aligned}$$

First, we have

$$\begin{aligned}
\frac{P(Y = 0|E, \mathbf{X})}{P(Y = 1|E, \mathbf{X})} &= \frac{P(Y = 0|E, \mathbf{D}_1, \mathbf{D}_2)}{P(Y = 1|E, \mathbf{D}_1, \mathbf{D}_2)} \\
&= e^{-\beta_0-\beta_1 E-\beta_2 \mathbf{D}_1-\beta_3 \mathbf{D}_2-\beta_3 \mathbf{D}_1 \mathbf{D}_2}.
\end{aligned}$$

Next, we need to show that

$$\frac{P(S=1|Y=0,E,\mathbf{X}=\mathbf{x}_k)}{P(S=1|Y=1,E,\mathbf{X}=\mathbf{x}_k)}=\frac{P(Y=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|\mathbf{X}=\mathbf{x}_k)}.$$

One intuitive way to establish this equality is that within each stratum we will select all cases (selection probability=1) and select the equal number of controls. Of note, this selection probability does not depend on subject's exposure status. Thus, we have

$$\begin{aligned}P(S=1|Y=1,E,\mathbf{X}=\mathbf{x}_k)&=P(S=1|Y=1,\mathbf{X}=\mathbf{x}_k)\\&=1,\end{aligned}$$

and

$$\begin{aligned}P(S=1|Y=0,E,\mathbf{X}=\mathbf{x}_k)&=P(S=1|Y=0,\mathbf{X}=\mathbf{x}_k)\\&=\frac{P(Y=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|\mathbf{X}=\mathbf{x}_k)}.\end{aligned}$$

The last equation holds because $P(S=1|Y=0,\mathbf{X}=\mathbf{x}_k)$ is estimated by $\frac{n_{1k}}{n_{0k}}$ (we select n_{1k} controls from a total of n_{0k} controls), which is equivalent to $\frac{n_{1k}/(n_{0k}+n_{1k})}{n_{0k}/(n_{0k}+n_{1k})}$. The proportion of cases in the stratum $n_{1k}/(n_{0k}+n_{1k})\overset{p}{\rightarrow}P(Y=1|\mathbf{X}=\mathbf{x_k})$ and the proportion of controls $n_{0k}/(n_{0k}+n_{1k})\overset{p}{\rightarrow}P(Y=0|\mathbf{X}=\mathbf{x_k})$ asymptotically as the sample size in this stratum increases. Thus, $\frac{n_{1k}}{n_{0k}}$ converges to both $P(S=1|Y=0,\mathbf{X}=\mathbf{x}_k)$ and $\frac{P(Y=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|\mathbf{X}=\mathbf{x}_k)}$ asymptotically and thus establishes the equality.

The alternative way is

$$\begin{aligned}\frac{P(S=1|Y=0,E,\mathbf{X}=\mathbf{x}_k)}{P(S=1|Y=1,E,\mathbf{X}=\mathbf{x}_k)}&=\frac{P(S=1|Y=0,\mathbf{X}=\mathbf{x}_k)}{P(S=1|Y=1,\mathbf{X}=\mathbf{x}_k)}\\&=\frac{P(Y=0|S=1,\mathbf{X}=\mathbf{x}_k)P(S=1,\mathbf{X}=\mathbf{x}_k)/P(Y=0|\mathbf{X}=\mathbf{x}_k)P(\mathbf{X}=\mathbf{x}_k)}{P(Y=1|S=1,\mathbf{X}=\mathbf{x}_k)P(S=1,\mathbf{X}=\mathbf{x}_k)/P(Y=1|\mathbf{X}=\mathbf{x}_k)P(\mathbf{X}=\mathbf{x}_k)}\\&=\frac{P(Y=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|\mathbf{X}=\mathbf{x}_k)}.\end{aligned}$$

Last equality holds because $P(Y=0|S=1,\mathbf{X}=\mathbf{x}_k)=P(Y=1|S=1,\mathbf{X}=\mathbf{x}_k)=1/2$ when we select equal number of controls and cases in each matching stratum.

We then have

$$\begin{aligned}\frac{P(Y=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|\mathbf{X}=\mathbf{x}_k)}&=\frac{P(Y=1,E=1|\mathbf{X}=\mathbf{x}_k)+P(Y=1,E=0|\mathbf{X}=\mathbf{x}_k)}{P(Y=0,E=1|\mathbf{X}=\mathbf{x}_k)+P(Y=0,E=0|\mathbf{X}=\mathbf{x}_k)}\\&=\frac{P(Y=1|E=1,\mathbf{X}=\mathbf{x}_k)P(E=1|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|E=1,\mathbf{X}=\mathbf{x}_k)P(E=1|\mathbf{X}=\mathbf{x}_k)+P(Y=0|E=0,\mathbf{X}=\mathbf{x}_k)P(E=0|\mathbf{X}=\mathbf{x}_k)}\\&\quad+\frac{P(Y=1|E=0,\mathbf{X}=\mathbf{x}_k)P(E=0|\mathbf{X}=\mathbf{x}_k)}{P(Y=0|E=1,\mathbf{X}=\mathbf{x}_k)P(E=1|\mathbf{X}=\mathbf{x}_k)+P(Y=0|E=0,\mathbf{X}=\mathbf{x}_k)P(E=0|\mathbf{X}=\mathbf{x}_k)}.\end{aligned}\tag{A1}$$

Note that $P(Y|E,\mathbf{X}=\mathbf{x}_k)$ and $P(E|X)$ can be expressed using both the outcome and the exposure models defined by Equations (1) and (2). It follows that

$$P(Y=1|E,\mathbf{X}=\mathbf{x_k},S=1)=\frac{1}{1+e^{c(\mathbf{x_k})-\beta_1E}},$$

where $c(\mathbf{x_k})$ is a very complex stratum specific term for each distinct value of $\mathbf{x_k}$. When the disease outcome is rare, the logit model (1) can be approximated by a log-linear model and we can derive a simpler form of $c(\mathbf{x_k})$. It follows that

$$\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}$$

$$\approx \frac{e^{\beta_1} \frac{e^{\alpha_0+\boldsymbol{\alpha}_1\mathbf{D}_1+\boldsymbol{\alpha}_2\mathbf{D}_2+\alpha_3\mathbf{D}_1\mathbf{D}_2}}{1+e^{\alpha_0+\boldsymbol{\alpha}_1\mathbf{D}_1+\boldsymbol{\alpha}_2\mathbf{D}_2+\boldsymbol{\alpha}_3\mathbf{D}_1\mathbf{D}_2}} + \frac{1}{1+e^{\alpha_0+\boldsymbol{\alpha}_1\mathbf{D}_1+\boldsymbol{\alpha}_2\mathbf{D}_2+\boldsymbol{\alpha}_3\mathbf{D}_1\mathbf{D}_2}}}{\frac{e^{\alpha_0+\alpha_1\mathbf{D}_1+\boldsymbol{\alpha}_2\mathbf{D}_2+\alpha_3\mathbf{D}_1\mathbf{D}_2}}{1+e^{\alpha_0+\alpha_1\mathbf{D}_1+\alpha_2\mathbf{D}_2+\alpha_3\mathbf{D}_1\mathbf{D}_2}} + \frac{1}{1+e^{\alpha_0+\boldsymbol{\alpha}_1\mathbf{D}_1+\boldsymbol{\alpha}_2\mathbf{D}_2+\boldsymbol{\alpha}_3\mathbf{D}_1\mathbf{D}_2}}} \times e^{\beta_0+\beta_2\mathbf{D}_1+\beta_3\mathbf{D}_2+\beta_3\mathbf{D}_1\mathbf{D}_2}.$$

Then we have $c(\mathbf{x_k}) = \log\left(\frac{e^{\beta_1}-1}{1+e^{-\alpha_0-\boldsymbol{\alpha}_1\mathbf{D}_1-\boldsymbol{\alpha}_2\mathbf{D}_2-\alpha_3\mathbf{D}_1\mathbf{D}_2}}\right)$. When $i = 1$ and $j = 1$, all dummies are 0, $c(\mathbf{x_k}) = \log\left(\frac{e^{\beta_1}-1}{1+e^{-\alpha_0}}\right)$; when $i = 1$ and $j > 1$, $c(\mathbf{x_k}) = \log\left(\frac{e^{\beta_1}-1}{1+e^{-\alpha_0-\alpha_{2,j-1}}}\right)$; when $i > 1$ and $j = 1$, $c(\mathbf{x_k}) = \log\left(\frac{e^{\beta_1}-1}{1+e^{-\alpha_0-\alpha_{1,i-1}}}\right)$; when $i > 1$ and $j > 1$, $c(\mathbf{x_k}) = \log\left(\frac{e^{\beta_1}-1}{1+e^{-\alpha_0-\alpha_{1,i-1}-\alpha_{2,j-1}-\alpha_{3,(i-1)(j-1)}}}\right)$.

A.2 Derivation of unconditional logit model in frequency matched data when the matching factor is a continuous variable

For simplicity, we will use single continuous confounder X to illustrate the potential bias from categorization in this scenario. We assume X is linear in the outcome model (1) and the exposure model (2). It follows that

$$\text{logit}P(Y = 1|E, X) = \beta_0 + \beta_1 E + f(X; \boldsymbol{\beta}_2),$$

and

$$\text{logit}P(E = 1|X) = \alpha_0 + g(X; \boldsymbol{\alpha}_1),$$

where $f(\cdot)$ and $g(\cdot)$ are arbitrary functions of X so that X can take arbitrary functional forms. For example, if X is only linearly associated with the outcome and the exposure in logit scale, $f(X; \boldsymbol{\beta}_2) = \beta_2 X$ and $g(X; \boldsymbol{\alpha}) = \alpha_1 X$. If such associations are quadratic, $f(X; \boldsymbol{\beta}_2) = \beta_{21}X + \beta_{22}X^2$ and $g(X; \boldsymbol{\alpha}_1) = \alpha_{11}X + \alpha_{12}X^2$.

We categorize X into K distinct intervals using $K + 1$ knots $x_i, i = 1, 2, \dots, K + 1$. We let $C = c_k \; \forall x_k \leq X < x_{k+1}, k = 1, 2, \dots, K$ denote the categorized variable. We can generate the following $I - 1$ dummy variables

$$D_{1j} = \begin{cases} 1, & \text{if } x_{j+1} \leq X < x_{j+2} \\ 0, & \text{otherwise} \end{cases},$$

for $j = 1, 2, \dots, I - 1$. $[x_1, x_2)$ or c_1 is the reference category.

To derive the form of logistic model in frequency matched data, we have

$$\begin{aligned} P(Y = 1|E, X, S = 1) &= \frac{1}{1 + \frac{P(S=1|Y=0,E,X)P(Y=0|E,X)}{P(S=1|Y=1,E,X)P(Y=1|E,X)}} \\ &= \frac{1}{1 + \frac{P(S=1|Y=0,E,X)}{P(S=1|Y=1,E,X)} e^{-\beta_0-\beta_1 E - f(X;\boldsymbol{\beta}_2)}}. \end{aligned}$$

As usual, the probability for a subject being selected into matched data does not depend on its exposure status. Cases and controls are matched on categorized variable. Every subject's probability of being selected is directly determined by the value of categorized variable C. For each categorized value c_k, every case having this value will be selected. It follows:

$$\begin{aligned} P(S = 1|Y = 1, E, X = x) &= P(S = 1|Y = 1, C = c_k) \\ &= 1. \end{aligned}$$

The probability of being selected for a control with $x \in [x_k, x_{k+1}]$ is

$$\begin{aligned} P(S = 1|Y = 0, E, X = x) &= P(S = 1|Y = 0, C = c_k) \\ &= \frac{P(Y = 1|C = c_k)}{P(Y = 0|C = c_k)}, \end{aligned}$$

$\frac{P(Y=1|C=c_k)}{P(Y=0|C=c_k)}$ actually can be presented by a logit model predicting the disease outcome using categorized variable C. Since we will replace $f(X;\boldsymbol{\beta}_2)$ with $\mathbf{D}$, the dummy variables of C, it involves the projection of $f(X;\boldsymbol{\beta}_2)$ onto E and $\mathbf{D}$. We let $f(X;\boldsymbol{\beta}_2) = \gamma_0 + \gamma_1 E + \boldsymbol{\gamma}_2\mathbf{D} + \epsilon$, ϵ represents some random error from this projection.

Then, the outcome model including E, $\mathbf{D}$, and ϵ becomes

$$\begin{aligned}\text{logit}P(Y=1|E,X) &= \text{logit}P(Y=1|E,\mathbf{D},\epsilon)\\ &= \beta_0 + \beta_1 E + (\gamma_0 + \gamma_1 E + \boldsymbol{\gamma}_2\mathbf{D} + \epsilon)\\ &= (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)E + \boldsymbol{\gamma}_2\mathbf{D} + \beta_2\epsilon.\end{aligned}$$

Next, when we omit unobservable ϵ from the model, we have

$$\text{logit}P(Y=1|E,\mathbf{D}) = \tilde{\beta}_0 + \tilde{\beta}_1 E + \tilde{\beta}_2\mathbf{D},$$

where $\tilde{\beta}_2 = (\tilde{\beta}_{21}, \tilde{\beta}_{22}, \dots, \tilde{\beta}_{2(I-1)})$. This change in regression coefficients is due to omitting a variable (ie, ϵ) from a non-collapsible logit model.[11,12]

Similarly,

$$\text{logit}P(E=1|\mathbf{D}) = \tilde{\alpha}_0 + \tilde{\alpha}_1\mathbf{D},$$

where $\tilde{\alpha}_1 = (\tilde{\alpha}_{21}, \tilde{\alpha}_{22}, \dots, \tilde{\alpha}_{2(I-1)})$.

Since $D_{i(k-1)} = 1$ if $x_k \leq X < x_{k+1}$ or $C = c_k$, we have

$$\begin{aligned}\frac{P(Y=1|C=c_k)}{P(Y=0|C=c_k)} &= \left(\frac{e^{\tilde{\beta}_1} - 1}{1 + e^{-\tilde{\alpha}_0 - \tilde{\alpha}_{1(k-1)}}} + 1\right) e^{\tilde{\beta}_0 + \tilde{\beta}_{2(k-1)}}\\ P(Y=1|E,X=x,S=1) &= \frac{1}{1 + \frac{P(S=1|Y=0,E,X)}{P(S=1|Y=1,E,X)} e^{-\beta_0 - \beta_1 E - f(X;\boldsymbol{\beta}_2)}}\\ &= \frac{1}{1 + e^{-c(k) - \beta_1 E - f(X;\boldsymbol{\beta}_2)}}.\end{aligned}$$

We can derive the approximate expression for the stratum specific term $c(k) = -\log\left(\left(\frac{e^{\tilde{\beta}_1}-1}{1+e^{-\tilde{\alpha}_0-\tilde{\alpha}_{1(k-1)}}} + 1\right) e^{\tilde{\beta}_0+\tilde{\beta}_{2(k-1)}}\right) + \beta_0$. Thus, we need to fit a stratified logistic regression including continuous confounder X as covariate in a frequency-matched case-control study when continuous matching factors are categorized.

If we have multiple matching factors,

$$\text{logit}P(Y=1|E,\mathbf{X}) = \beta_0 + \beta_1 E + f(\mathbf{X};\boldsymbol{\beta_2}),$$

where $\mathbf{X}$ is a vector of matching factors, $f(\cdot)$ denotes some arbitrary function (eg, quadratic form, interaction term). If all matching factors are continuous and their categorized values form K strata, the general form of the outcome model in the frequency matched design becomes

$$\text{logit}P(Y=1|E,\mathbf{X},S=1) = c(k) + \beta_1 E + f(\mathbf{X};\boldsymbol{\beta_2}),$$

where $c(k) = -\log\left(\frac{P(Y=1|C=k)}{P(Y=0|C=k)}\right) + \beta_0$, $k = 1, 2, \dots, K$. Thus, the outcome model in the frequency-matched data will take the same form as the outcome model in the source population except it has stratum-specific intercepts instead of a constant intercept. If some X's are categorical, their main effect terms and pairwise interaction terms do not need to be included because $c(k)$ contains all categorical or categorized matching factors and their interaction terms.

For example, we have two matching factors X_1 (continuous) and X_2. The outcome model in the source population is:

$$\text{logit}P(Y=1|E,X_1,X_2) = \beta_0 + \beta_1 E + \beta_2 X_1 + \beta_3 X_1^2 + \beta_4 X_2 + \beta_4 X_1 X_2,$$

If X_2 is continuous, the outcome model in the matched data becomes

$$\text{logit}P(Y = 1|E, X_1, X_2) = c(k) + \beta_1 E + \beta_2 X_1 + \beta_3 X_1^2 + \beta_4 X_2 + \beta_4 X_1 X_2,$$

If X_2 is a binary dummy variable, the outcome model in the matched data becomes

$$\text{logit}P(Y = 1|E, X_1, X_2) = c(k) + \beta_1 E + \beta_2 X_1 + \beta_3 X_1^2 + \beta_4 X_1 X_2,$$

The main effect term of X_2 is not included because $c(k)$ contains its main effect term.

A.3 Heterogeneous exposure effect and unmatched confounders

In the case that there is a heterogeneous exposure effect (eg, interaction) and there are unmatched confounders, $\mathbf{X}_1$ denotes the confounders used in matching and $\mathbf{X}_2$ denotes the unmatched confounders. The outcome model becomes

$$\text{logit}(P(Y = 1|E, \mathbf{X}_1, \mathbf{X}_2)) = \beta_0 + f(E, \mathbf{X}_1, \mathbf{X}_2),$$

where $f(\cdot)$ is an arbitrary function, which could contains interaction terms among E, $\mathbf{X}_1$, and $\mathbf{X}_2$.

$$\text{logit}(P(E = 1|E, \mathbf{X}_1, \mathbf{X}_2)) = \alpha_0 + g(E, \mathbf{X}_1, \mathbf{X}_2).$$

When $\mathbf{X}_1$ includes only categorical variables, we can derive the outcome model in the matched sample using the approach outlined in Appendix A.2 as

$$\begin{aligned}
&P(Y = 1|E, \mathbf{X}_1 = \mathbf{x}_{1k}, \mathbf{X}_2 = \mathbf{x}_{2k}, S = 1) \\
&= \frac{1}{1 + \frac{P(S=1|Y=0,E,\mathbf{X}_1=\mathbf{x}_{1k},\mathbf{X}_2=\mathbf{x}_{2k})P(Y=0|E,\mathbf{X}_1=\mathbf{x}_{1k},\mathbf{X}_2=\mathbf{x}_{2k})}{P(S=1|Y=1,E,\mathbf{X}_1=\mathbf{x}_{1k},\mathbf{X}_2=\mathbf{x}_{2k})P(Y=1|E,\mathbf{X}_1=\mathbf{x}_{1k},\mathbf{X}_2=\mathbf{x}_{2k})}} \\
&= \frac{1}{1 + \frac{P(S=1|Y=0,E,\mathbf{X}_1=\mathbf{x}_{1k})P(Y=0|E,\mathbf{X}_1=\mathbf{x}_{1k},\mathbf{X}_2=\mathbf{x}_{2k})}{P(S=1|Y=1,E,\mathbf{X}_1=\mathbf{x}_{1k})P(Y=1|E,\mathbf{X}_1=\mathbf{x}_{1k},\mathbf{X}_2=\mathbf{x}_{2k})}} \because \text{ } \mathbf{X}_2 \text{ is not used in matching and does not impact the selection probability} \\
&= \frac{1}{1 + e^{-c(k)-\beta_0-f(E,\mathbf{X}_1,\mathbf{X}_2)}}
\end{aligned}$$

$c(k)$ takes a very complex form.

Similarly, when $\mathbf{X}_1$ includes continuous variables, we can derive the outcome model in the matched sample using the approach outlined in Appendix A.3 as

$$P(Y = 1|E, \mathbf{X}_1 = \mathbf{x}_{1k}, \mathbf{X}_2 = \mathbf{x}_{2k}, S = 1) = \frac{1}{1 + e^{-c(k)-\beta_0-f(E,\mathbf{X}_1,\mathbf{X}_2)}}.$$

Some general rules:

(1) When the matching confounders $\mathbf{X}_1$ includes only categorical variables, the main effect and interaction terms of $\mathbf{X}_1$ in $f(E, \mathbf{X}_1, \mathbf{X}_2)$ do not need to be included because the stratum specific intercept term $c(k)$ already includes these terms. However, their interaction terms with E and $\mathbf{X}_2$ should be included.

(2) The main effect and interaction terms involving $\mathbf{X}_2$ needs to be included. If $\mathbf{X}_2$ contains continuous variables, their proper functional forms, same as their forms in the outcome model in the unmatched study population, should be used.

We designed a simulation study to validate the general rules above. We generated one discrete random variable $Z \sim P(Z = z) = \frac{1}{3}, z = 1, 2, 3$ and a Bernoulli random variable $X_1 \sim \textit{Bernoulli}(0.5)$ as matching factors. We created two dummy variables for Z, D_1, and D_2 with $z = 1$ as the reference level. We generated two additional normally distributed variables X_1 and $X_2 \sim N(0, 1)$. We generated the exposure variable E and outcome variable Y using the following exposure and outcome models:

$$\text{logit}P(E=1|X,D_1,D_2) = 0 + 0.928D_1 - 0.371D_2 - 0.5X_1 - 1.6D_1X_1 + 1.6D_2X_1 + 0.2X_2;$$

and

$$\begin{aligned}\text{logit}P(Y=1|E,X,D_1,D_2) = &-4.5 + 0.8E + 0.894D_1 + 0.447D_2 - 0.5X_1 - 0.1X_2 + 0.02X_2^2\\ &+ 0.1EX_1 - 0.05EX_3 + 0.2ED_1 + 0.05ED_2 + 0.01X_3^2;\end{aligned}$$

In this outcome model the exposure effect is not homogeneous. There are EX_1 and EX_3 interaction terms. Next, we did a cross-tab of the outcome and two confounders Z and X_1. For each combination of Z and X_1, we select all cases and randomly select equal number of controls. Last, we fit a CLR$(E, X_2, X_2^2, X_3^2, ED_1, ED_2, EX_1, EX_3)$ in frequency-matched samples to validate the regression terms. We performed 10 000 simulations and 10 000 observations were generated for each simulation. The simulation result is listed in the Supplemental Table 1. The averaged estimates of regression coefficients are the same as the regression coefficients in the population outcome model.