Statistics
in Medicine

Research Article

Received 17 October 2012, Accepted 20 May 2013 Published online 12 June 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5879

Ignoring the matching variables in cohort studies – when is it valid and why?

Arvid Sjölander[a*†] **and Sander Greenland**[b]

In observational studies of the effect of an exposure on an outcome, the exposure–outcome association is usually confounded by other causes of the outcome (potential confounders). One common method to increase efficiency is to match the study on potential confounders. Matched case-control studies are relatively common and well covered by the literature. Matched cohort studies are less common but do sometimes occur. It is often argued that it is valid to ignore the matching variables, in the analysis of matched cohort data. In this paper, we provide analyses delineating the scope and limits of this argument. We discuss why the argument does not carry over to effect estimation in matched case-control studies, although it does carry over to null-hypothesis testing. We also show how the argument does not extend to matched cohort studies when one adjusts for additional confounders in the analysis. Ignoring the matching variables can sometimes reduce variance, even though this is not guaranteed. We investigate the trade-off between bias and variance in deciding whether adjustment for matching factors is advisable. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: case-control studies; cohort studies; confounding; matching; stratification

1. Introduction

In observational studies of the effect of an exposure on an outcome, the exposure–outcome association is usually confounded by other causes of the outcome (potential confounders). There are several traditional strategies to adjust for potential confounders in the analysis, for example, stratification and outcome regression modeling. These methods account for measured confounders by comparing the exposed and unexposed within levels of the confounders (conditioning on confounders). More recent methods such as propensity scoring and inverse-probability weighting balance the confounder distributions across levels of the exposure. If there is a strong association between the confounders and the exposure, these analytic strategies are often inefficient relative to what could be accomplished using design strategies.

One common method to increase design efficiency is to select subjects matched on potential confounders for which the need for adjustment is fairly certain such as age and sex. Matched case-control studies are typically constructed so that for each case a fixed number k of controls are selected, having the same levels of the matching variables as the case (1:k matching). Such balanced matching forces the ratio of cases to controls to be constant across all strata of the matching variables, resulting in unconditional independence of the matching variables and the outcome but does not obviate the need to adjust for the matching factors because the variables and the outcome may remain dependent conditional on exposure [1–7]. Matched cohort studies are less common but do sometimes occur and follow a very different (if superficially parallel) strategy; typically, for each exposed subject, a fixed number of unexposed subjects are selected, at the same levels of the matching variables as the exposed. Such balanced matching forces the ratio of exposed to unexposed to be constant across all strata of the matching variables, resulting in independence of the matching variables and the exposure.

[a]*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*
[b]*Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, U.S.A.*
**Correspondence to: Arvid Sjölander, Nobels väg 12A, 171 77, Stockholm, Sweden.*
†*E-mail: arvid.sjolander@ki.se*

This independence may obviate the need to adjust for the matching variables if no other adjustments are needed, because independence of a variable and exposure will violate an oft-cited necessary condition for a variable to be a confounder, namely that it be associated with exposure [7, chap. 11]. It can also provide greater robustness to model-form misspecification if adjustment for the matching factors is needed [8]. Nonetheless, there seems to be some divergence among textbooks on how the matching variables should be treated in cohort studies. Woodward [6] stated without explanation that 'When a case-control study (or indeed, any other type of study) is matched, the analysis *must* take account of the matching'. In contrast, Jewell [5] argued that because the association between the matching variables and the exposure is broken by cohort matching, the 'sample data cannot show confounding', and so 'the pooled data created after breaking the matches and combining yields a valid estimate'. Similarly, Newman [4] stated that in a matched-pairs cohort study 'the distribution of matching variables is the same in exposed and unexposed cohorts, and consequently these variables are eliminated as sources of confounding.' Rothman *et al*. [7, chap. 11] illustrated by examples that if the exposure–outcome risk ratio is constant across levels of the matching variables, then a 'pooled' analysis produces this constant risk ratio. They concluded, in line with Jewell [5] that

> In a cohort study without competing risks or losses to follow up, no additional action is required to control for confounding of the point estimate by the matching factors, because matching unexposed to exposed prevents an association between exposure and the matching factors.

These quotes raise several questions, not simply of who is correct but also of what precisely is meant and how generally the comments apply. For example, Jewell [5] says that ignoring the matching variables yields a 'valid estimate' but what target parameter is being estimated? Similarly, Rothman *et al*. [7, chap. 11] considered the case when the risk ratio is constant across levels of the matching variables, but do the conclusions hold if it varies, and to what parameter do they apply? Finally, all four books considered the case in which there are no additional confounders or selection factors, but what happens if additional adjustments are needed? As Rothman *et al*. [7, chap. 11] noted, the simple conclusions quoted previously can break down if there is censoring (competing risks or losses).

The aim of this paper is to answer these questions. In Section 2 we show that ignoring the matching variables in the analysis of cohort studies yields a population causal effect, in the absence of additional confounders. Which population this causal effect applies to depends on the matching scheme; when the unexposed are matched to the exposed, this causal effect applies to the exposed subpopulation. When there are additional uncontrolled bias sources (such as unmeasured confounders), ignoring the matching variables yields a standardized measure of the exposure–outcome association. The distribution to which this association is standardized is that of the group to which subjects are being matched. For example, when the unexposed are matched to the exposed, the standard distribution is that of the matching variables in the exposed. We show that this result holds regardless of whether the effect is constant across levels of the matching variables. In this sense, it is indeed valid to ignore the matching variables in the analysis of cohort studies, as long as no additional confounder adjustments are made. In Section 3, we discuss why this argument does not carry over to case-control studies. In Section 4, we show that ignoring the matching variables in a cohort study does not yield a causal effect if there are additional confounders, *even if the additional confounders are adjusted for in the analysis*. Thus, when one adjusts for additional confounders, it is not valid to to ignore the matching variables in the analysis. Even if bias is introduced by ignoring the matching variable, this bias could be outweighed by variance reduction. We note though that ignoring the matching variables can also increase variance. In Sections 5 and 6, we investigate the trade-off between bias and variance for linear models and logistic models, respectively. In Section 7, we discuss the impact of model misspecification. In Section 8, we present a real data example.

We note that our concern here is with analysis choices given a matched design has been used, rather than with comparing matched and unmatched designs. There is a large literature on the latter comparison, with overviews and citations available in various textbooks [3–7].

2. No additional confounders

To illustrate basic concepts, consider the hypothetical target population of 100,000 individuals given in Table I, stratified by sex. In this population there is a marginal (unconditional with respect to sex) association between the exposure and the outcome; the marginal risk ratio is $\{(4972+7843)/(16,256+8788)\}\{(7154+2756)/(70895+4061)\} = 3.87$. Sex is not balanced across exposure levels; $16,256/(16,256+8788) = 65\%$ of the exposed are male, whereas $70,895/(70,895+4061) = 95\%$

Table I. Source population stratified by sex.

	Men		Women	
	Exposed	Unexposed	Exposed	Unexposed
Cases	4972	7154	7843	2756
Total	16256	70895	8788	4061

Table II. Cohort matched on sex.

	Men		Women	
	Exposed	Unexposed	Exposed	Unexposed
Cases	497	164	784	596
Total	1626	1626	879	879

of the unexposed are male. Thus, one may suspect that sex is a confounder for the exposure–outcome association. Indeed, when we stratify on sex, the association is attenuated. For men, the risk ratio is $(4972/16,256)/(7154/70,895) = 3.03$, and for women the risk ratio is $(7843/8788)(2756/4061) = 1.31$.

Suppose that a cohort study draws an exposed cohort from the exposed target population and matches an unexposed cohort to the exposed cohort on sex. If 10% of the exposed in the target population is included in the cohort study and these subjects are selected independently of sex, we expect about 1626 men and 879 women in the exposed cohort. Table II displays the expected results, where sex is balanced across exposure levels.

Suppose that we ignore sex in the analysis, and compute the marginal (over sex) risk ratio from Table II. We then obtain $(497 + 784)(164 + 596) = 1.69$. To give the marginal risk ratio in Table II a precise interpretation, we introduce some notation. Let Y be the case status ($Y = 1$ if case, $Y = 0$ else), let X be the exposure status ($X = 1$ for exposed, $X = 0$ for unexposed), and let M be sex ($M = 1$ for men, $M = 0$ for women). Let $p(\cdot)$ and $p^*(\cdot)$ refer to proportions in the target population and in the matched cohort, respectively. For instance, $p(M = 1|X = 0) = 70,895/(70,895 + 4061) = 0.95$ and $p^*(M = 1|X = 0) = 1626/(1626 + 879) = 0.65$. In this notation, the marginal risk ratio in Table II can be written as

$$\begin{aligned}
&\frac{p^*(Y=1|X=1)}{p^*(Y=1|X=0)} \\
&= \frac{\overbrace{p^*(Y=1|X=1,M=1)}^{497/1626}\overbrace{p^*(M=1|X=1)}^{1626/2505} + \overbrace{p^*(Y=1|X=1,M=0)}^{784/879}\overbrace{p^*(M=0|X=1)}^{879/2505}}{\underbrace{p^*(Y=1|X=0,M=1)}_{164/1626}\underbrace{p^*(M=1|X=0)}_{1626/2505} + \underbrace{p^*(Y=1|X=0,M=0)}_{596/879}\underbrace{p^*(M=0|X=0)}_{879/2505}} \\
&= \frac{p(Y=1|X=1,M=1)p^*(M=1) + p(Y=1|X=1,M=0)p^*(M=0)}{p(Y=1|X=0,M=1)p^*(M=1) + p(Y=1|X=0,M=0)p^*(M=0)},
\end{aligned} \tag{1}$$

where we have used the facts that the proportion of cases within each combination of exposure and sex is the same in the matched cohort as in the target population, so that $p^*(Y = y|X = x, M = m) = p(Y = y|X = x, M = m)$, and that sex and exposure are independent in the matched cohort, so that $p^*(M = m|X = x) = p^*(M = m)$. The right-hand side of Eq. (1) is a standardized risk ratio, where the standardization is over the distribution of M in the matched cohort. To proceed, suppose that there is no confounder for the association between X and Y, except M. We can then interpret $p(Y = 1|X = x, M = m)$ as the proportion of cases that we would observe, if hypothetically all individuals with $M = m$ would receive exposure level $X = x$. In standard potential-outcome notation [9, 10]: $p(Y = 1|X = x, M = m) = p(Y_x = 1|M = m)$. We may thus rewrite the right-hand side of (1) as

$$\frac{p(Y_1=1|M=1)p^*(M=1) + p(Y_1=1|M=0)p^*(M=0)}{p(Y_0=1|M=1)p^*(M=1) + p(Y_0=1|M=0)p^*(M=0)} = \frac{p^*(Y_1=1)}{p^*(Y_0=1)},$$

which is the causal risk ratio of X on Y, in a population where sex is distributed as in the matched cohort. In our matching scheme, the unexposed were matched to the exposed. Thus, the sex distribution in the matched cohort is identical to the sex distribution among the exposed in the target population, so that $p^*(M = m) = p(M = m|X = 1)$. It follows that the marginal risk ratio in the matched cohort is the causal effect in that cohort, in which sex is distributed as in the exposed in the target population (provided of course there are no uncontrolled sources of bias). This causal parameter is traditionally known as the effect standardized to the exposed population. Our result is analogous to the fact that propensity-score matching of unexposed to exposed followed by conventional analysis estimates the effect of exposure on the exposed cohort (again assuming there are no uncontrolled biases) [11].

There are other matching schemes besides matching unexposed to the exposed. For instance, one could match the exposed to the unexposed, in which case the marginal risk ratio in the matched cohort equals the (counterfactual) effect exposure would have had on on the unexposed cohort, in the absence of additional confounders (traditionally known as the effect of exposure standardized to the unexposed). Matched pairs also arise naturally in many designs. For instance, a cohort of exposure-discordant twin pairs can be viewed as being matched on the large set of factors that twins have in common (e.g., maternal lifestyle during pregnancy and, for monozygotic twins, genetic make-up). Thus, absent uncontrolled bias, the marginal risk ratio among the exposure-discordant twin pairs equals the effect of exposure on exposed members of exposure-discordant twins. Because the unexposed members have the same covariate distribution, this is also the effect exposure would have on the unexposed, assuming exposure in these twins is independent of modifiers of the risk ratio.

We have considered nonparametric estimation of risk ratios for binary variables, but the argument holds for any measure of association (e.g., odds ratios and hazard ratios), for any type of data (e.g., categorical and continuous), and for parametric and semiparametric estimation. For instance, the estimated hazard ratio obtained by fitting a Cox proportional hazards model to a matched cohort, ignoring the matching variables, is a consistent estimator of the causal hazard ratio in a population where the matching variables are distributed as in the matched cohort, provided that the model is correct, censoring is noninformative conditional on exposure and matching variables, and that the matching variables are the only confounders. The difference is that the odds ratio and hazard ratio are noncollapsible, meaning that they may change upon stratification by a perfectly balanced (and thus nonconfounding) factor, yet the unstratified (marginal) ratio is still unbiased for the population effect standardized to the exposed [7, 12].

In practice, one would rarely believe that the matching variables are the only confounders. Even if there are other confounders, Eq. (1) is still valid, so that the marginal risk ratio in the matched cohort can still be interpreted as a standardized risk ratio (a ratio of risks averaged over the matching variables). The utility of this summary is debatable, however, due to its lack of correspondence to a causal effect.

3. Case-control studies

It is frequently stated that the matching variables cannot be ignored in the analysis of case-control studies [1–3, 5–7]. The standard argument against doing so is that even if the exposure–outcome odds ratio is constant across levels of the matching variables, ignoring the matching variables will not produce this constant odds ratio [3, 7]. This argument is not entirely satisfactory however. To see this, consider the hypothetical population in Table I. The odds ratios for men and women are equal to $\{4972/(16,256-4972)\}/\{7154/(70,895-7154)\} = 3.93$ and $\{7843/(8788-7843)\}/\{2756/(4061-2756)\} = 3.93$, respectively, so we can say that the conditional (age specific) odds ratio is virtually constant. Nonetheless, the marginal odds ratio in the matched cohort (Table II) is much smaller: $\{(497+784)/(1626-497+879-784)\}/\{(164+596)/(1626-164+879-596)\} = 2.40$. Thus, following the standard argument previously, it would not be valid to ignore the matching variables in the analysis of cohort studies either, when focusing on odds ratios. But we know from Section 2 that this is not the case; ignoring the matching variables in cohort studies is equally valid regardless of what measure of effect we use. The flaw in the argument is that the discrepancy between the conditional odds ratio and the marginal odds ratio in the matched cohort does not reflect confounding in the sense of bias in estimating the marginal causal odds ratio. It instead reflects that the marginal and conditional causal odds ratio differ despite independence of the exposure and the matching factors (noncollapsibility without confounding [12]).

Does this mean that it is equally valid to ignore the matching variables in the analysis of case-control studies as in the analysis of cohort studies? No, by analogy with Eq. (1), the marginal odds ratio in a

matched case-control study can be written as

$$\frac{p^*(Y=1|X=1)p^*(Y=0|X=0)}{p^*(Y=0|X=1)p^*(Y=1|X=0)} = \frac{\sum_m p^*(Y=1|X=1,M=m)p^*(M=m|X=1)\sum_m p^*(Y=0|X=0,M=m)p^*(M=m|X=0)}{\sum_m p^*(Y=0|X=1,M=m)p^*(M=m|X=1)\sum_m p^*(Y=1|X=0,M=m)p^*(M=m|X=0)}. \quad (2)$$

The expression cannot be simplified further, however, because the matching variable M is not independent of the exposure X in the case-control sample. Thus, $p^*(M=m|X) \neq p^*(M=m)$, which implies that expression will not generally be equal to the unconditional causal odds ratio

$$\frac{p^*(Y_1=1)p^*(Y_0=0)}{p^*(Y_1=0)p^*(Y_0=1)},$$

even in the absence of additional confounders.

We note that by symmetry we can also write the odds ratio in (2) as

$$\begin{aligned}\frac{p^*(X=1|Y=1)p^*(X=0|Y=0)}{p^*(X=0|Y=1)p^*(X=1|Y=0)} &= \frac{\sum_m p^*(X=1|Y=1,M=m)p^*(M=m|Y=1)\sum_m p^*(X=0|Y=0,M=m)p^*(M=m|Y=0)}{\sum_m p^*(X=0|Y=1,M=m)p^*(M=m|Y=1)\sum_m p^*(X=1|Y=0,M=m)p^*(M=m|Y=0)} \\ &= \frac{\sum_m p(X=1|Y=1,M=m)p(M=m|Y=1)\sum_m p(X=0|Y=0,M=m)p(M=m|Y=1)}{\sum_m p(X=0|Y=1,M=m)p(M=m|Y=1)\sum_m p(X=1|Y=0,M=m)p(M=m|Y=1)},\end{aligned} \quad (3)$$

where the second equality follows from the fact that in case-control studies, matching to the cases implies that $p^*(X=x|Y=y,M=m) = p(X=x|Y=y,M=m)$ and $p^*(M=m|Y=y) = p(M=m|Y=1)$. From (3), we see that whenever Y and X are conditionally independent in the target population, given M (e.g., when M is the only confounder and X has no causal effect on Y), the marginal (over M) odds ratio in the matched cohort will be equal to 1. Thus, even though it is generally difficult to interpret the magnitude of the marginal odds ratio, it provides a valid test of the causal null hypothesis in the absence of additional confounders. We note that to obtain a correct p-value, it is necessary to take any marginal correlation among matched subjects (within the matching strata) into account, for example, by properly adjusting the denominator of the Wald test statistic or by conditioning on the matching variables.

4. Additional adjustment variables in matched cohorts

We now consider matched cohorts with additional covariates that we wish to adjust for in the analysis. Consider Table III, which displays the population in Table I further stratified by age (young/old). Table I can be obtained by collapsing Table III over age. Within strata defined by age and sex jointly, there is virtually no association between the exposure and the outcome (i.e., all conditional exposure–outcome

Table III. Source population stratified by age and sex.

	Men		Women	
Young	Exposed	Unexposed	Exposed	Unexposed
Cases	4374	446	54	349
Total	10029	1022	217	1412
Old				
	Men		Women	
	Exposed	Unexposed	Exposed	Unexposed
Cases	598	6708	7789	2407
Total	6227	69873	8571	2649

Table IV. Cohort matched on sex and subsequently stratified by age.

	Men		Women	
Young	Exposed	Unexposed	Exposed	Unexposed
Cases	437	10	5	75
Total	1003	23	22	306
Old				
	Men		Women	
	Exposed	Unexposed	Exposed	Unexposed
Cases	60	154	779	521
Total	623	1603	857	573

risk ratios, given age and sex, are almost identical to 1. Thus, if age and sex are the only confounders, there is virtually no causal effect of the exposure on the outcome in this population.

Consider now Table IV, which displays the cohort obtained by matching on sex, as described in Section 2, and then stratified on age. Table II is obtained by collapsing Table IV over age. Suppose that we ignore sex in the analysis and compute the marginal (over sex) risk ratio for young and old separately from Table IV. For the young, the risk ratio is $\{(437+5)/(1003+22)\}/\{(10+75)/(23+306)\} = 1.67$, and for the old, the risk ratio is $\{(60+779)/(623+857)\}/\{(154+521)/(1603+573)\} = 1.83$. Thus, even if there is no effect of the exposure on the outcome in the population, and there are no confounders apart from age and sex, there is a strong exposure–outcome association within age groups. This example illustrates that it is usually not valid to ignore the matching variables, when adjusting for additional confounders.

To understand what goes wrong, note that in Table IV, sex is not balanced across exposure groups, conditional on age. For instance, in the stratum of young individuals, $1003/(1003+22) = 98\%$ of the exposed are male, whereas only $23/(23+306) = 7\%$ of the unexposed are male. Thus, even though we have matched on sex, sex may still be a confounder in the age subgroups.

It can be shown (Appendix A) that it is valid to ignore the matching variables when adjusting for additional covariates if any of the following criteria hold in the target population:

(1) The unmatched covariates are conditionally independent of the exposure, given the matching variables.
(2) The unmatched covariates are conditionally independent of the matching variables, given the exposure.
(3) The outcome is conditionally independent of the matching variables, given the exposure and the unmatched covariates.

Unfortunately, all three criteria will usually be violated when the unmatched covariates are confounders given the matching variables and the matching variables are confounders given the unmatched covariates. To see this, consider the causal diagram [10] in Figure 1, which displays a simple causal structure between the exposure X, the outcome Y, sex (the matching variable), and age (an unmatched covariate). The first criterion is violated because age has a causal influence on X. The second criterion is violated because both sex and age have a causal influence on X. Thus, by conditioning on X (a collider), sex and age become associated [7, 10, 13–15]. The third criterion is violated because sex has a causal influence on Y. It is easy to show that the same argument can be made when the matching variables, and the unmatched covariates do not necessarily have causal influences on both X and Y but lie on back-door paths between X and Y [10].

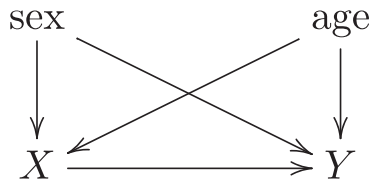

Figure 1. Causal diagram.

The previous concerns about additional covariate adjustment would be moot if balanced propensity-score matching is carried out with all measured confounders accounted for in score construction (as in high-dimensional scoring; [16]). In that case, one could arguably ignore the matching for point estimation, although the question of validity of variance estimates would arise [17].

5. Linear models

We have shown that ignoring the matching variables in a cohort study does not yield a causal effect if there are additional confounders, even if the additional confounders are adjusted for in the analysis. Even if bias is introduced by ignoring the matching variable, this bias could be outweighed by variance reduction. We note though that ignoring the matching variables can also increase variance. In this section we investigate the trade-off between bias and variance for linear models.

5.1. Bias

Let M, C, and Y be a continuous matching variable, a continuous covariate, and a continuous outcome, respectively, and let X be a binary (0/1) exposure. Suppose that in the target population we have that

$$M|X = 1 \sim N(0, 1) \tag{4}$$

$$C|X, M \sim N\left(\theta_{C0} + \theta_{CX}X + \theta_{CM}M, \sigma_C^2\right) \tag{5}$$

$$Y|X, C, M \sim N\left(\theta_0 + \theta_{YX}X + \theta_{YC}C + \theta_{YM}M, \sigma_Y^2\right), \tag{6}$$

where we assume that $\theta_{C0} = \theta_{Y0} = 0$ and $\sigma_C^2 = \sigma_Y^2 = 1$ for simplicity. Suppose that we construct a matched cohort from this population, by matching unexposed ($X = 0$) to exposed ($X = 1$) on M. To investigate the consequences of ignoring M, we note (Appendix B) that

$$Y|X, C \sim N\left(\theta_{Y0}^* + \theta_{YX}^*X + \theta_{YC}^*C, \sigma_Y^{2*}\right) \tag{7}$$

where

$$\theta_{YX}^* = \theta_{YX} - \frac{\theta_{CX}\theta_{CM}\theta_{YM}}{\sigma_C^2 + \theta_{CM}^2}. \tag{8}$$

Thus, even when M is ignored (marginalized over), the mean of Y is still linear in X and C, but the regression coefficient for X is attenuated (biased) by the second term on the right-hand side of (8). To relate this bias term to the criteria listed in Section 4, we note that the first criterion holds if and only if $\theta_{CX} = 0$, the second criterion holds if and only if $\theta_{CM} = 0$, and the third criterion holds if and only if $\theta_{YM} = 0$. When either $\theta_{CX} = 0$ or $\theta_{CM} = 0$ or $\theta_{YM} = 0$, the bias term equals 0, so that $\theta_{YX}^* = \theta_{YX}$.

5.2. Bias versus variance.

Ignoring the matching variables can sometimes reduce variance, even though this is not guaranteed. Thus, even if bias is introduced by ignoring the matching variable, this bias could be outweighed by variance reduction. To investigate the trade-off between bias and variance, we compared two analyses. The first analysis, which we refer to as 'conditional', takes M into accounts by fitting the model in (6) to the matched cohort by using maximum likelihood. The second analysis, which we refer to as 'marginal', ignores M by fitting the model in (7) to the matched cohort by using maximum likelihood. We note that under the population defined by Eqns (4) through (6), both the conditional analysis model and marginal analysis model are correctly specified, so model misspecification is not an issue here.

To cover a wide range of scenarios, we treated the parameter vector $\theta = (\theta_{CX}, \theta_{CM}, \theta_{YX}, \theta_{YC}, \theta_{YM})$ as random. We assumed the elements of θ to be independent, and following either a uniform distribution on [0, 1], an exponential distribution with rate equal to 2 (that is, with mean equal to $1/2 = 0.5$), or a normal distribution with mean and standard deviation equal to 0.5 and 0.25, respectively. For each of these three distributions we generated 1,000,000 samples of θ. For each sample of θ we generated a 1:1 matched cohort sample of $n = 200$ pairs from the population defined by Eqns (4) through (6). Each matched cohort sample was analyzed with both the conditional model (6) and the marginal model (7).

Table V. Summary statistics from 1,000,000 simulations. $\hat{\theta}_{YX}$ is the ML-estimate of θ_{YX} in the conditional analysis, $\hat{\theta}^*_{YX}$ is the ML-estimate of θ^*_{YX} in the marginal analysis, $V = E\{\text{var}(\hat{\theta}_{YX}|\theta_{YX})\}$, $B^2 = E\{\text{Bias}^2(\hat{\theta}_{YX}|\theta_{YX})\}$, $D^2 = (\hat{\theta}_{YX} - \theta_{YX})^2$, $\text{MSE}(\hat{\theta}_{YX}) = E(D^2) = V + B^2$, $V^* = E\{\text{var}(\hat{\theta}^*_{YX}|\theta_{YX})\}$, $B^{*2} = E\{\text{Bias}^2(\hat{\theta}^*_{YX}|\theta_{YX})\}$, $D^{*2} = (\hat{\theta}^*_{YX} - \theta_{YX})^2$, $\text{MSE}(\hat{\theta}^*_{YX}) = E(D^{*2}) = V^* + B^{*2}$.

$f(\theta)$	B^2	V^2	$\text{MSE}(\hat{\theta}_{YX})$	B^{*2}	V^{*2}	$\text{MSE}(\hat{\theta}^*_{YX})$	$\Pr(D^{*2} < D^2)$
$U(0,1)$	0.00	0.01	0.01	0.01	0.02	0.03	0.36
$Exp(2)$	0.00	0.01	0.01	0.01	0.03	0.04	0.40
$N(0.5, 0.25^2)$	0.00	0.01	0.01	0.01	0.02	0.03	0.34

Table VI. Estimated regression coefficients in the logistic regression model for $D^{*2} < D^2$.

$f(\theta)$	k_{CX}	k_{CM}	k_{YX}	k_{YC}	k_{YM}
$U(0,1)$	−1.08	−0.83	0.00	0.00	−1.22
exp(2)	−0.54	−0.29	0.00	0.00	−0.71
$N(0.5, 0.25^2)$	−1.14	−0.82	0.00	0.02	−1.26

To quantify the trade-off between bias and variance, we used the mean squared error (MSE), which for a generic parameter β, and its estimate T, is defined as

$$E(T-\beta)^2 = E\{\text{var}(T|\beta)\} + E\{\text{Bias}^2(T|\beta)\}, \tag{9}$$

where $\text{Bias}(T|\beta) = E(T-\beta|\beta)$. We note that in frequentist paradigm, the target parameter is considered fixed so that the expectation on the LHS of (9) is only taken over the distribution of T, and the RHS simplifies to $\text{var}(T|\beta) + \text{Bias}^2(T|\beta)$. In our simulation, the target parameter θ_{YX} is considered random, so that the expectation on the LHS of (9) is taken over the joint distribution of θ_{YX} and its estimate. Let $\hat{\theta}_{YX}$ denote the ML-estimate of θ_{YX} in the conditional analysis and let $\hat{\theta}^*_{YX}$ denote the ML-estimate of θ^*_{YX} in the marginal analysis. Define $V = E\{\text{var}(\hat{\theta}_{YX}|\theta_{YX})\}$, $B^2 = E\{\text{Bias}^2(\hat{\theta}_{YX}|\theta_{YX})\}$, and $D^2 = (\hat{\theta}_{YX} - \theta_{YX})^2$, so that $\text{MSE}(\hat{\theta}_{YX}) = E(D^2) = V + B^2$. Similarly, define $V^* = E\{\text{var}(\hat{\theta}^*_{YX}|\theta_{YX})\}$, $B^{*2} = E\{\text{Bias}^2(\hat{\theta}^*_{YX}|\theta_{YX})\}$, and $D^{*2} = (\hat{\theta}^*_{YX} - \theta_{YX})^2$, so that $\text{MSE}(\hat{\theta}^*_{YX}) = E(D^{*2}) = V^* + B^{*2}$.

Table V displays the empirical values of B^2, V, $\text{MSE}(\hat{\theta}_{YX})$, B^{*2}, V^{*2}, $\text{MSE}(\hat{\theta}^*_{YX})$, and $\Pr(D^{*2} < D^2)$, obtained from the 1,000,000 simulated samples, for each distribution of θ separately. We observe that $\text{MSE}(\hat{\theta}_{YX}) < \text{MSE}(\hat{\theta}^*_{YX})$ for all distributions of θ. $\Pr(D^{*2} < D^2)$ is relatively small but not very close to 0. This indicates that 'on average' we increase MSE by ignoring the matching variables in linear models.

To characterize the situations in which D^{*2} can be expected to be smaller than D^2, we regressed the indicator of $D^{*2} < D^2$ on θ by fitting the logistic regression model

$$\text{logit}\{\Pr(D^{*2} < D^2|\theta)\} = k_0 + k_{CX}\theta_{CX} + k_{CM}\theta_{CM} + k_{YX}\theta_{YX} + k_{YC}\theta_{YC} + k_{YM}\theta_{YM}$$

to the simulated data. Table VI displays the estimated regression coefficients, for each distribution of θ separately. We observe that k_{CX}, k_{CM}, and k_{YM} are negative and quite large in magnitude for all distributions of θ, except for the exponential distribution in which case k_{CM} is quite close to 0. k_{YX} and k_{YC} are equal to 0, or very close to 0, for all distributions of θ. This indicates that ignoring the matching variables is most likely to be beneficial, in terms of MSE, in situations where the association is weak between either C and X, conditional on M, between C and M, conditional on X, or between Y and M, conditional on (X, C), that is, when the either of the three criteria listed in Section 4 is approximately met. However, as discussed in Section 4, this is precisely the situation where adjustment for M and/or C is not needed.

6. Logistic models

Although linear models are mathematically convenient, logistic models are more common in epidemiologic research. In this section, we provide some simulations results for logistic models. Toward this end,

let X, M, C, and Y be a binary (0/1) exposure, matching variable, covariate, and outcome, respectively. Suppose that in the target population we have that

$$\begin{aligned} \Pr(M=1|X=1) &= 0.5 \\ \text{logit}\{\Pr(C=1|X,M)\} &= \theta_{C0}+\theta_{CX}X+\theta_{CM}M \\ \text{logit}\{\Pr(Y=1|X,C,M)\} &= \theta_{Y0}+\theta_{YX}X+\theta_{YC}C+\theta_{YM}M, \end{aligned} \tag{10}$$

where we assume that $\theta_{C0}=\theta_{Y0}=0$ for simplicity. Suppose that we construct a matched cohort from this population, by matching unexposed ($X=0$) to exposed ($X=1$) on M.

It is less straight forward to make a comparison between a conditional analyses and a marginal analysis in the logistic model than in a linear model. The logistic model for Y, conditional on X, C, and M, does not translate into a logistic model for Y, when marginalized over M. In other words, the first-order linear-logistic model form is generally not collapsible over covariates in the model even when those covariates are marginally independent of the remaining covariates. In particular, although the conditional X-Y log odds ratio, given C and M, is constant across levels of C and M ($=\theta_{YX}$), the conditional X-Y log odds ratio, given C, is not constant across levels of C. To bypass these problems and enable a comparison between a conditional analysis and a marginal analysis we focus on the standardized log odds ratios

$$\phi=\log\left[\frac{E^*\{\Pr(Y=1|X=1,C,M)\}E^*\{\Pr(Y=0|X=0,C,M)\}}{E^*\{\Pr(Y=0|X=1,C,M)\}E^*\{\Pr(Y=1|X=0,C,M)\}}\right] \tag{11}$$

and

$$\phi^*=\log\left[\frac{E^*\{\Pr^*(Y=1|X=1,C)\}E^*\{\Pr^*(Y=0|X=0,C)\}}{E^*\{\Pr^*(Y=0|X=1,C)\}E^*\{\Pr^*(Y=1|X=0,C)\}}\right]. \tag{12}$$

In (11), the expectations are taken over the joint distribution of (C,M) induced by the matched sampling scheme. In (12), the expectations are taken over the marginal (over M) distribution of C induced by the matched sampling scheme. If C and M are the only confounders, then ϕ is the causal effect in the exposed. ϕ^* is equal to ϕ when either of the three criteria listed in Section 4 holds (i.e., when $\theta_{CX}=0$, when $\theta_{CM}=0$, or when $\theta_{YM}=0$), but is otherwise generally not equal to ϕ, and does not generally have a causal interpretation. We define $\hat{\phi}$ as the the conditional analysis estimate obtained by replacing $\Pr(Y=1|X=x,C,M)$ in (11) with the parametric ML-estimate on the basis of model (10) and subsequently averaging over the sample distribution of (C,M). We define $\hat{\phi}^*$ as the the marginal analysis estimate obtained by replacing $\Pr^*(Y=1|X=x,C)$ in (12) with the corresponding nonparametric sample proportion and subsequently averaging over the sample distribution of C.

We generated 1,000,000 matched cohort samples and calculated summary statistics in the same fashion as for the linear model described in Section 5.2. Tables VII and VIII display the results. We observe that $\text{MSE}(\hat{\phi})=\text{MSE}(\hat{\phi}^*)$ and $\Pr(D^{*2}<D^2)$ is quite close to 0.5, for all distributions of θ.

Table VII. Summary statistics from 1,000,000 simulations. $\hat{\theta}_{YX}$ is the ML-estimate of θ_{YX} in the conditional analysis, $\hat{\theta}^*_{YX}$ is the ML-estimate of θ^*_{YX} in the marginal analysis, $V=E\{\text{var}(\hat{\theta}_{YX}|\theta_{YX})\}$, $B^2=E\{\text{Bias}^2(\hat{\theta}_{YX}|\theta_{YX})\}$, $D^2=(\hat{\theta}_{YX}-\theta_{YX})^2$, $\text{MSE}(\hat{\theta}_{YX})=E(D^2)=V+B^2$, $V^*=E\{\text{var}(\hat{\theta}^*_{YX}|\theta_{YX})\}$, $B^{*2}=E\{\text{Bias}^2(\hat{\theta}^*_{YX}|\theta_{YX})\}$, $D^{*2}=(\hat{\theta}^*_{YX}-\theta_{YX})^2$, $\text{MSE}(\hat{\theta}^*_{YX})=E(D^{*2})=V^*+B^{*2}$.

$f(\theta)$	B^2	V^2	$\text{MSE}(\hat{\theta}_{YX})$	B^{*2}	V^{*2}	$\text{MSE}(\hat{\theta}^*_{YX})$	$\Pr(D^{*2}<D^2)$
$U(0,1)$	0.00	0.05	0.05	0.00	0.05	0.05	0.47
xp(2)	0.00	0.05	0.05	0.00	0.05	0.05	0.48
$N(0.5,0.25^2)$	0.00	0.05	0.05	0.00	0.05	0.05	0.47

Table VIII. Estimated regression coefficients in the logistic regression model for $D^{*2}<D^2$.

$f(\theta)$	k_{CX}	k_{CM}	k_{YX}	k_{YC}	k_{YM}
$U(0,1)$	−0.14	−0.11	−0.08	−0.04	−0.17
$Exp(2)$	−0.11	−0.05	−0.04	−0.01	−0.11
$N(0.5,0.25^2)$	−0.15	−0.13	−0.09	−0.04	−0.18

This indicates that 'on average' it may not make a big difference, in terms of MSE, if the matching variable is ignored in logistic models. The estimated regression coefficients in the logistic regression for $D^{*2} < D^2$ against θ are all negative, which indicates that ignoring the matching variables is most likely to be beneficial, in terms of MSE, in situations where the association is weak between either C and X, conditional on M, between C and M, conditional on X, between Y and X, conditional on (C, M), between Y and C, conditional on (X, M), or between Y and M, conditional on (X, C).

We end this section by noting that there have been several matched-data log odds-ratio estimators that average marginal and conditional or full likelihood estimators with the objective of reducing MSE [18, 19], but these have yet to see extensive use or implementation.

7. Model-form misspecification

So far we have assumed that all models are correctly specified. The full likelihood analysis requires specification of $E(Y|X, C, M)$, whereas the marginal analysis (i.e., ignoring the matching variables) requires specification of $E^*(Y|X, C)$. Because the latter mean function is of a lower dimension than the former, it may be easier to approximate the latter by using simple forms. Thus, one could argue that the marginal analysis may typically be less prone to model misspecification bias than the full likelihood analysis. An alternative analysis is conditional maximum likelihood. In this analysis, we typically postulate the model

$$g\{E(Y|X, C, M)\} = \alpha + \beta X + \gamma C + h(M),$$

where $g(\cdot)$ is a link function and $h(\cdot)$ is an unspecified function of M. Because M is by definition constant within each pair, this model is equivalent to

$$g\{E(Y|X, C, \text{set } i)\} = \alpha_i + \beta X + \gamma C,$$

where α_i is an unspecified set-specific intercept. If $g(\cdot)$ is the canonical link in the distribution for Y (e.g., the identity link for normal Y and logit link for Bernoulli Y), then the α_i's can be eliminated by partial or conditional likelihood arguments [20, 21]. For a logit link function, this leads to conditional logistic regression [2, 3]. The conditional likelihood analysis makes fewer modeling assumptions than the full likelihood analysis, and is thus less prone to model misspecification bias. Although it may produce less precise estimates than a full likelihood analysis in which $h(M)$ is completely parameterized, there is evidence that it will often have lower MSE when allowance is made for possible misspecification [8] .

8. Real data example

In this section, we present a real data example, borrowed from Carlsson *et al*. [22]. These authors carried out a study to investigate if the association between body mass index (BMI) and mortality can be explained by genetic factors. Their data comprise 44,258 same-sex twins, both monozygotic (MZ) and dizygotic (DZ), who filled in a questionnaire in 1967 or 1972 on lifestyle factors, health, height, and weight. Mortality was determined by linkage to the National Causes of Death Registry for the years 1972–2004. The authors stratified the twins on zygosity and sex, and adjusted for smoking in the statistical analysis. They observed an association of BMI with mortality when analyzing data as a cohort of unrelated individuals by using ordinary (i.e., non-stratified) Cox regression.

We reanalyzed these data as follows. BMI was dichotomized into BMI $\leqslant 25$ (unexposed) and BMI>25 (exposed), and a cohort was constructed from the exposure-discordant twin pairs (2944 DZ twin pairs and 1119 MZ twin pairs). As discussed in Section 2, this cohort can be viewed as being matched on all factors that are shared within each twin pair; we use U to denote this set of shared factors. By design of the data set, U includes sex but does not include BMI or smoking. Table IX summarizes the data. The association between BMI and mortality was estimated in DZ and MZ twins separately by using four different models: (i) ordinary Cox proportional hazards (PH) model without any covariate adjustments; (ii) ordinary Cox PH model adjusted for smoking (dichotomized as ever/never); (iii) stratified (on twin pair) Cox PH model without any covariate adjustment; and (iv) stratified Cox PH model adjusted for smoking. Models (i) and (ii) ignore (marginalize over) U, whereas models (iii) and (iv) condition on U. Model (i) gives a marginal (over U) causal hazard ratio if there are no confounders in the target population (i.e., the population of all twins pairs, both exposure discordant and exposure concordant) except U. Model (ii) generally induces bias by simultaneously marginalizing over U and adjusting for smoking,

Table IX. Number of individuals in the dataset for each combination of zygosity/body mass index (BMI)/smoking/death during follow up (1972–2004).

		Never smoking		Smoking		
Zygosity	BMI	No death	Death	No death	Death	Sum
DZ	$\leqslant 25$	760	720	844	620	2944
DZ	>25	738	857	740	609	2944
MZ	$\leqslant 25$	268	293	291	267	1119
MZ	>25	289	314	268	248	1119

DZ, dizygotic; MZ, monozygotic.

Table X. Analysis results. Models: (i) ordinary Cox PH model without any covariate adjustments; (ii) ordinary Cox PH model adjusted for smoking; (iii) stratified Cox PH model without any covariate adjustment; and (iv) stratified Cox PH model adjusted for smoking.

Stratum	Model	Stratified on pair	Adjusted for smoking	HR	95% CL
DZ	i	No	No	1.20	1.12, 1.30
DZ	ii	No	Yes	1.23	1.14, 1.32
DZ	iii	Yes	No	1.21	1.10, 1.33
DZ	iv	Yes	Yes	1.23	1.12, 1.35
MZ	i	No	No	1.01	0.90, 1.14
MZ	ii	No	Yes	1.03	0.91, 1.16
MZ	iii	Yes	No	1.07	0.92, 1.24
MZ	iv	Yes	Yes	1.08	0.93, 1.26

HR, hazard ratio; DZ, dizygotic; MZ, monozygotic; CL, confidence limits.

and does not generally give any causal effect (see Section 4). Model (iii) gives a conditional (on U) causal hazard ratio if there are no confounders in the target population except U. Model (iv) gives a conditional (on U) causal hazard ratio if there are no confounders in the target population except U and smoking.

Table X displays the estimated hazard ratio (HR) together with 95% confidence limits (CL) for DZ and MZ twins and each model (i/ii/iii/iv). Smoking adjustment has very little impact, eliminating concern that its adjustment is necessary or harmful. Nonetheless, all the DZ estimates suggest about a 20% elevated mortality rate among those with elevated BMI versus those without. In contrast, the MZ estimates are close to the null, which may reflect the much more extensive control of genetic factors inherent in the MZ analysis, although the conditional estimates are slightly elevated. The results are consistent with the hypothesis that the association between BMI and mortality is largely due to shared (environmental and genetic) confounding. We caution, however, that this result does not rule out a BMI effect and that a more biologically founded analysis would examine adjustment for other variables including height [23] and the use of flexible continuous parameterizations for BMI and smoking.

9. Conclusion

We have shown how ignoring the matching variables in a cohort study with matching of the unexposed to the exposed yields the effect of exposure on the exposed population, in the absence of additional confounders or other bias sources. We have further shown, however, that ignoring the matching variables in a cohort study can leave bias if there are additional confounders, even with adjustment for the additional confounders. This bias can be avoided by adjusting for the matching variables.

For linear and logistic models, we have shown that the bias induced by ignoring the matching variables can be outweighed by variance reduction, but this is not guaranteed. For linear models we observed that on average we tend to increase MSE when ignoring the matching variable. For logistic models we observed no major impact on average of ignoring the matching variable. Thus, as a tentative heuristic we would recommend some adjustment for the matching variables when analyzing matched cohort data, although a matched analyses *per se* is not needed. Caution is warranted though by noting that the impact of cohort matching can vary considerably with the choice of target parameter and underlying model

[24]. Also, when there is censoring (competing risks and loss to follow-up), the balance produced by the matching may not be preserved during follow-up, which makes it necessary to account for the matching variables in the analysis, even in the absence of additional confounders [7, chap. 11].

Finally, as is customary in focused methodologic studies, for the most part we have assumed that there are no sources of bias other than those under discussion. If, however, there are important unmeasured confounders or other uncontrolled bias sources such as measurement error, the biases discussed here may be of secondary concern.

Appendix A

Suppose that C and M are the only confounders for exposure X and outcome Y, so that $p(Y = y|X = x, C, M) = p(Y_x = y|C, M)$. We then have that

$$\begin{aligned} E^*(Y|X = x, C = c) &= E^*\{E^*(Y|X = x, C = c, M)|X = x, C = c\} \\ &= E^*\{E(Y|X = x, C = c, M)|X = x, C = c\} \\ &= E^*\{E(Y_x|C = c, M)|X = x, C = c\}. \end{aligned} \tag{13}$$

Generally, a contrast between $E^*(Y|X = 1, C)$ and $E^*(Y|X = 0, C)$ cannot be interpreted as a causal effect of X on Y because the outer average in (13) is taken over a distribution of M that depends on exposure level x. The contrast can only be interpreted as a causal effect if either $E(Y|X, C, M)$ does not depend on M or $p^*(M|X, C)$ does not depend on X. In the first case, the third criterion listed in Section 4 holds. To investigate the second case, note that

$$\begin{aligned} p^*(M = m|X = x, C = c) &= \frac{p^*(C = c|X = x, M = m)p^*(M = m|X = x)}{E^*\{p^*(C = c|X = x, M)|X = x\}} \\ &= \frac{p(C = c|X = x, M = m)p(M = m|X = 1)}{E\{p(C = c|X = x, M)|X = 1\}}, \end{aligned}$$

which depends on x unless either of the first two criteria listed in Section 4 holds.

Appendix B

It follows from standard theory that $Y|^*X, C$ is normally distributed with mean given by

$$\begin{aligned} E^*(Y|X, C) &= E^*\{E^*(Y|X, C, M)|X, C\} \\ &= E^*\{E(Y|X, C, M)|X, C\} \\ &= \theta_{Y0} + \theta_{YX}X + \theta_{YC}C + \theta_{YM}E^*(M|X, C). \end{aligned}$$

We have that

$$E^*(M|X, C) = E^*(M|X) + \frac{\text{cov}^*(C, M|X)}{\text{var}^*(C|X)}\{E^*(C|X) - C\}$$

Standard calculations give that $E^*(M|X) = 0$, $\text{cov}^*(C, M|X) = \theta_{CM}$, $\text{var}^*(C|X) = \sigma_C^2 + \theta_{CM}^2$, $E^*(C|X) = \theta_{C0} + \theta_{CX}X$, so that

$$E^*(M|X, C) = \frac{\theta_{CM}}{\sigma_C^2 + \theta_{CM}^2}(\theta_{C0} + \theta_{CX}X - C)$$

and

$$E^*(Y|X, C) = \theta_{Y0}^* + \theta_{YX}^*X + \theta_{YC}^*C,$$

where θ_{YX}^* is given by (8).

Acknowledgements

Arvid Sjölander is supported by the Swedish Research Council [340-2012-6007].

References

1. Seigel D, Greenhouse S. Validity in estimating relative risk in case-control studies. *Journal of Chronic Diseases* 1973; **26**(4):219.
2. Breslow N, Day N. *Statistical Methods in Cancer Research. Vol. 1. The analysis of Case-control Studies.* Lyon: IARC/WHO, 1980.
3. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford University Press: New York, 1993.
4. Newman S. *Biostatistical methods in epidemiology*. Wiley: New York, 2001.
5. Jewell N. *Statistics for Epidemiology*. CRC Press, 2004.
6. Woodward M. *Epidemiology: Study Design and Data Analysis*. CRC Press, 2005.
7. Rothman K, Greenland S, Lash T. *Modern Epidemiology*, 3rd. Lippincott Williams & Wilkins, 2008.
8. Waernbaum I. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine* 2012; **31**(15):1572–1581.
9. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 1974; **66**(5):688–701.
10. Pearl J. *Causality: Models, Reasoning, and Inference*, 2nd. Cambridge University Press, 2009.
11. Kurth T, Walker A, Glynn R, Chan K, Gaziano J, Berger K, Robins J. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 2006; **163**(3):262–270.
12. Greenland S, Robins J, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**(1):29–46.
13. Cole S, Platt R, Schisterman E, Chu H, Westreich D, Richardson D, Poole C. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* 2010; **39**(2):417–420.
14. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003; **14**(3):300–306.
15. Greenland S, Pearl J. Adjustments and their consequences – collapsibility analysis using graphical models. *International Statistical Review* 2011; **79**(3):401–426.
16. Schneeweiss S, Rassen J, Glynn R, Avorn J, Mogun H, Brookhart M. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009; **20**(4):512–522.
17. Williamson E, Morley R, Lucas A, Carpenter J. Variance estimation for stratified propensity score estimators. *Statistics in Medicine* 2012; **31**(15):1617–1632.
18. Liang K, Zeger S. On the use of concordant pairs in matched case-control studies. *Biometrics* 1988; **44**(4):1145–1156.
19. Kalish L. Reducing mean squared error in the analysis of pair-matched case-control studies. *Biometrics* 1990; **46**(2): 493–499.
20. Cox D. Partial likelihood. *Biometrika* 1975; **62**(2):269–276.
21. Neuhaus J, McCulloch C. Separating between-and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006; **68**(5):859–872.
22. Carlsson S, Andersson T, de Faire U, Lichtenstein P, Michaëlsson K, Ahlbom A. Body mass index and mortality: is the association explained by genetic factors?. *Epidemiology* 2011; **22**(1):98–103.
23. Michels KB, Greenland S, Rosner BA. Does body mass index adequately capture the relation of body composition and body size to health outcomes?. *American Journal of Epidemiology* 1998; **147**(2):167–172.
24. Greenland S, Morgenstern H. Matching and efficiency in cohort studies. *American Journal of Epidemiology* 1990; **131**(1):151–159.