Statistical Science
2012, Vol. 27, No. 3, 395–411
DOI: 10.1214/12-STS390

Analysis of 1:1 Matched Cohort Studies and Twin Studies, with Binary Exposures and Binary Outcomes

Arvid Sjölander, Anna L. V. Johansson, Cecilia Lundholm, Daniel Altman, Catarina Almqvist and Yudi Pawitan

Abstract. To improve confounder adjustments, observational studies are often matched on potential confounders. While matched case-control studies are common and well covered in the literature, our focus here is on matched cohort studies, which are less common and sparsely discussed in the literature. Matched data also arise naturally in twin studies, as a cohort of exposure–discordant twins can be viewed as being matched on a large number of potential confounders. The analysis of twin studies will be given special attention. We give an overview of various analysis methods for matched cohort studies with binary exposures and binary outcomes. In particular, our aim is to answer the following questions: (1) What are the target parameters in the common analysis methods? (2) What are the underlying assumptions in these methods? (3) How do the methods compare in terms of statistical power?

Key words and phrases: Cohort studies, likelihood, matching.

Arvid Sjölander is Ph.D. Student, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, Sweden (e-mail: arvid.sjolander@ki.se). Anna L. V. Johansson is Statistician, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, Sweden (e-mail: anna.johansson@ki.se). Cecilia Lundholm is Statistician, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, Sweden (e-mail: cecilia.lundholm@ki.se). Daniel Altman is Associate Professor, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, Sweden, and Associate Professor, Division of Obstetrics and Gynecology, Department of Clinical Sciences, Danderyd Hospital, Karolinska Institutet, Stockholm, Sweden (e-mail: daniel.altman@ki.se). Catarina Almqvist is Associate Professor, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, Sweden, and Associate Professor, Astrid Lindgren Children's Hospital and Department of Woman and Child Health, Karolinska Institutet, Stockholm, Sweden (e-mail: catarina.almqvist@ki.se). Yudi Pawitan is Professor, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, Sweden (e-mail: yudi.pawitan@ki.se).

1. INTRODUCTION

A common goal of epidemiological research is to estimate the causal effect of a particular exposure on a particular outcome. The common tool is an observational study, utilizing, for example, hospital data, cohort data or health register data. In observational studies, the exposure-outcome association is invariably confounded by factors that induce spurious (i.e., non-causal) associations. For example, age may confound an exposure-outcome association if older people are more often exposed and more likely to develop the outcome. Without adjustment for age, that is, if the confounding influence by age is not accounted for in the analysis, there may be an association of exposure and outcome, even in the absence of a causal effect. Hence, the exposure-outcome association cannot, in general, be given a causal interpretation, unless all confounders are properly adjusted for.

There are several strategies to adjust for potential confounders in the analysis, for example, stratification or regression modeling. Essentially, these methods solve the problem of confounding by comparing

the exposed and unexposed within levels of the confounders, thus balancing the confounders across levels of the exposure and comparing "like with like." If there is a strong association between the confounders and the exposure, or between the confounders and the outcome, these strategies are often inefficient. In particular, some strata may contain few exposed subjects or few cases (i.e., subjects that developed the outcome); the lack of balance may lead to unstable estimates for these strata.

One common method to increase the efficiency is to match the study on potential confounders. For example, matched case-control studies are constructed so that for each case, a fixed number of controls are selected, having the same confounder levels as the case. When each case is matched to one control, we say that the study is 1:1 matched. In case-control studies, matching forces the ratio of cases to controls to be constant across all strata of the matched factors, which implies that the association between the confounders and the outcome is broken. Matched case-control studies are commonplace, and well covered in the literature (e.g., Breslow and Day, 1980; Jewell, 2004; Woodward, 2005). A matched cohort study can be constructed in a similar fashion; for each exposed subject, a fixed number of unexposed subjects are selected, having the same confounder levels as the exposed. In cohort studies, matching forces the ratio of exposed to unexposed to be constant across all strata of the matched factors, which implies that the association between the confounders and the exposure is broken. Matched cohort studies are relatively rare, and the literature is sparse and typically rather brief (e.g., Cummings et al., 2003). The reason, we believe, is mainly due to available data sources. Matched cohort studies are suitable for situations where a researcher has access to large population data sources with exposure information.

Matched data also arise naturally in twin studies. By nature, a large number of potential confounders are shared (i.e., having constant levels) within each twin pair, for example, genetic factors, maternal uterine environment, gestational age, etc. It follows that a cohort of exposure–discordant twin pairs (i.e., pairs in which one of the twins is exposed, and the other twin is unexposed) can be viewed as being 1:1 matched on all shared confounders. In such a cohort there is no association between the shared confounders and the exposure. An attractive feature of twin studies is that the shared confounders often include factors which are normally very difficult to match on, or even to measure. For example, monozygotic twins have identical genes and can thus can be viewed as being matched on the whole genome. However, a twin study is not simply a special case of a regular 1:1 matched cohort study; whereas the latter only contains exposure–discordant pairs, the former also contains pairs which are concordant in the exposure. Because of their unique and attractive properties, twin studies will be given special attention in this paper.

The aim of this paper is to give a detailed overview of different analysis methods for matched cohort studies with binary exposures and binary outcomes. In particular, our aim is to answer the following questions: (1) What are the target parameters in the common analysis methods? (2) What are the underlying assumptions in these methods? (3) How do the methods compare in terms of statistical power?

We illustrate the methods with two examples. The first example is a register-based study on the effect of hysterectomy on the risk for cardiovascular disease (CVD) in Swedish women (Ingelsson et al., 2010). The study is matched on birth year, year of hysterectomy and county of residence at year of hysterectomy, so that for each hysterectomized woman (exposed), three non-hysterectomized women at same age and year were selected from the general population. The second study is a population-based twin study of the association between fetal growth and childhood asthma (Örtqvist et al., 2009).

The paper is organized as follows. In Section 2 we review the concepts of marginalization, conditioning and standardization. In Section 3 we define a matched cohort study. In Section 4 we describe the most common analysis methods for matched cohorts. These methods can also be used to analyze the exposure–discordant pairs in twin studies. In Section 5 we demonstrate how these methods can be adapted for inclusion of the exposure–concordant pairs in twin studies as well. In Section 6 we carry out a simulation study. In Section 7 we provide the two illustrating examples. We will restrict our attention to 1:1 matching, and we will not consider additional covariate adjustments. Extensions to other matching schemes and adjustments for additional covariates are discussed in Section 8.

2. MARGINALIZATION, CONDITIONING AND STANDARDIZATION

We first establish the notations and briefly review the concepts of marginalization, conditioning and standardization, which are crucial for the understanding of

matching and confounder adjustment. More thorough discussions can be found in standard epidemiological textbooks (e.g., Rothman et al., 2008). Let X denote the binary exposure of interest (0/1), let Y denote the binary outcome of interest (0/1) and let Z denote a set of potential confounders for the association between X and Y. We use $\Pr(\cdot)$ generically for both probabilities (population proportions) and densities, and we use $E(\cdot)$ for expected value (population average). We use $V_1 \perp V_2|V_3$ as shorthand for "V_1 and V_2 conditionally independent, given V_3." We use (log) odds ratios to quantify the X–Y association. Other possible options would be risk differences or risk ratios. There are two reasons for focusing on odds ratios. First, regression models for odds ratios can be conveniently fitted without restrictions; see Section 4.1.1. Second, in applied scenarios, it is often desirable to make results comparable with case control studies, in which only odds ratios are estimable.

An unadjusted analysis targets the marginal (over Z) association between X and Y, for example, through the marginal odds ratio

$$OR_m = \frac{\Pr(Y=1|X=1)\Pr(Y=0|X=0)}{\Pr(Y=0|X=1)\Pr(Y=1|X=0)}.$$

We define $\psi_m = \log(OR_m)$. In the presence of confounders Z, OR_m fails to have a causal interpretation. In particular, it may differ from 1 in the absence of a causal effect.

The influence of Z can be eliminated by conditioning on Z, as in the conditional odds ratio

$$OR_c(Z) = \frac{\Pr(Y=1|X=1,Z)\Pr(Y=0|X=0,Z)}{\Pr(Y=0|X=1,Z)\Pr(Y=1|X=0,Z)}.$$

The conditional odds ratio $OR_c(Z)$ depends, in general, on Z. If Z is the only confounder for the X–Y association, then $OR_c(Z)$ can be interpreted as the conditional causal effect of X on Y, given Z, on the odds ratio scale. If there are additional confounders, then $OR_c(Z)$ has no causal interpretation.

$OR_c(Z)$ is a subpopulation (i.e., Z-specific) effect. The effect for the whole population can be obtained through standardization. The standardized probability of $Y=1$ given $X=x$, is given by

$$E_Z\{\Pr(Y=1|X=x,Z)\}, \tag{1}$$

where we have used subindex Z to highlight that the expectation is taken over the marginal distribution $\Pr(Z)$. We emphasize that the expression in (1) is not, in general, equal to $E_{Z|X=x}\{\Pr(Y=1|X=x,Z)|X=x\} = \Pr(Y=1|X=x)$, which is the marginal (unadjusted) probability of $Y=1$, given $X=x$. If Z is the only confounder, then $E_Z\{\Pr(Y=1|X=x,Z)\}$ can be interpreted as the hypothetical (counterfactual) probability of $Y=1$, had everybody attained level $X=x$ in the source population (Hernán and Robins, 2006). $\Pr(Y=1|X=x,Z)$ can be standardized to any proper distribution $\Pr^*(Z)$, not necessarily equal to $\Pr(Z)$. We let $E^*_Z(V)$ denote the expected value of V, where the expectation is taken over $\Pr^*(Z)$. If Z is the only confounder, then $E^*_Z\{\Pr(Y=1|X=x,Z)\}$ can be interpreted as the hypothetical (counterfactual) probability of $Y=1$, had everybody attained level $X=x$ in the fictitious population where Z follows the distribution $\Pr^*(Z)$. A standardized odds ratio is constructed as

$$OR_s = \frac{E\{\Pr(Y=1|X=1,Z)\}E\{\Pr(Y=0|X=0,Z)\}}{E\{\Pr(Y=0|X=1,Z)\}E\{\Pr(Y=1|X=0,Z)\}}.$$

We define $\psi_s = \log(OR_s)$. In (1), $\Pr(Y=1|X=x,Z)$ is standardized to $\Pr(Z)$, that is, the distribution of Z in the source population. In order to keep the notation simple, we use OR_s and ψ_s, even if $\Pr(Z)$ is replaced by $\Pr^*(Z)$, and we let it be clear from the context which distribution of Z these parameters are standardized to. If Z is the only confounder, then OR_s can be interpreted as the causal effect of X on Y in the source/fictitious population, on the odds ratio scale. We emphasize that although the numerical values of OR_s and ψ_s may depend heavily on which distribution of Z they are standardized to, they are always, by construction, adjusted for Z.

In general, there is no ordering in the magnitudes of $OR_c(Z)$, and OR_s. An interesting special case occurs when $OR_c(Z)$ is constant across levels of Z, that is,

$$\log\{OR_c(Z)\} = \psi_c. \tag{2}$$

It can be shown (Neuhaus et al., 1991) that $|\psi_c| \geq |\psi_s|$.

In general, there is no ordering in the magnitudes of OR_m and $OR_c(Z)$, or of OR_m and OR_s; confounding by Z can both inflate or deflate the association between X and Y. There are a few special cases though. If $Y \perp Z|X$, then $\Pr(Y=1|X,Z) = \Pr(Y=1|X)$ which implies that $OR_m = OR_c(Z) = OR_s$ for all Z and all standardization distributions $\Pr^*(Z)$. This would happen if the true causal structure between X, Y and Z is as in Figure 1. If $X \perp Z$, then $\Pr(Z|X) = \Pr(Z)$ which implies that $OR_m = OR_s$ for the particular distribution $\Pr(Z)$, that is, the distribution of Z in the source population. This would happen if the true causal structure is as in Figure 2.

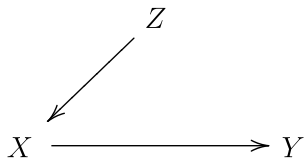

FIG. 1. *A causal structure for which $Y \perp Z|X$.*

We note that in Figures 1 and 2, Z is not a confounder, and OR_m can be given a causal interpretation. Thus, for these scenarios, adjusting for Z is not necessary for causal inference. We further note that the structure in Figure 2 does not render OR_m equal to $OR_c(Z)$, even if $OR_c(Z)$ is constant across levels of Z. This is a consequence of the noncollapsibility of the odds ratio. For a more thorough discussion on (non)collapsibility and the special properties of odds ratios, we refer the reader to Greenland et al. (1999).

3. MATCHED COHORT STUDIES

3.1 Design

A cohort study that is 1:1 matched on Z consists of n pairs of observations, each pair consisting of one exposed subject ($X = 1$) and one unexposed subject ($X = 0$). The pairs are constructed so that the two subjects within each pair have the same level of confounder Z; that is, Z may vary between pairs, but not within pairs. Thus, Z is equally distributed among exposed and unexposed in the matched cohort. The outcome Y is assumed to be recorded for each subject. Ignoring Z, the paired data can be conveniently represented as in Table 1. In practice, 1:1 matched pairs are typically constructed by first drawing an exposed person from the whole population, then drawing an unexposed person with an equal or similar level of confounder Z; we refer to this sampling scheme as *exposure-driven* matching.

We note that in twin studies Z is not directly observed, but should be interpreted as all the unobserved factors that are common within a twin pair.

3.2 Likelihood Construction

Before discussing the various analysis methods, we construct the likelihood for the observed data. Let

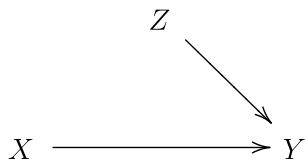

FIG. 2. *A causal structure for which $X \perp Z$.*

Z_i denote the common value of Z for pair i, $i \in \{1, 2, \ldots, n\}$. Let Y_i^0 and Y_i^1 denote the outcome Y for the unexposed ($X = 0$) and the exposed ($X = 1$) subject in pair i, respectively. The matched data consists of n i.i.d. observations (Y_i^0, Y_i^1, Z_i). We suppress the index i when not needed, so that Y^x denotes Y for the subject with $X = x$, $x \in (0, 1)$, within an arbitrary pair. We use $\Pr(Y = y, X = x, Z = z)$ to denote the population probability of $(Y = y, X = x, Z = z)$, and we will use $\Pr^*(Y^0 = y^0, Y^1 = y^1, Z = z)$ to denote the probability for $(Y^0 = y^0, Y^1 = y^1, Z = z)$ induced by the matched sampling scheme. Under exposure-driven matching, the design implies that

$$\Pr^*(Y^x = y^x|Z) = \Pr(Y = y^x|X = x, Z) \tag{3a}$$

and

$$\Pr^*(Y^0 = y^0, Y^1 = y^1|Z) = \Pr^*(Y^0 = y^0|Z)\Pr^*(Y^1 = y^1|Z). \tag{3b}$$

Equation (3a) "ties" the induced distribution to the source population distribution, thus allowing for samples from the former to be used for inference on the latter. Equation (3b) determines the correlation structure of the data, which is crucial for correct standard error computations. In twin studies, (3a) and (3b) do not necessarily hold (see Section 5), but are assumed throughout the paper.

The induced marginal distribution of Z is determined by the type of matching. Under exposure-driven matching, the induced marginal distribution of Z equals the source population distribution of Z among the exposed, that is, $\Pr^*(Z) = \Pr(Z|X = 1)$. In twin studies restricted to the exposure–discordant pairs, we have that $\Pr^*(Z) = \Pr(Z|\text{discordant in } X)$.

When Z is observed (as in regular matched studies), the likelihood contribution for pair i is

$$\Pr^*(Y_i^0 = y_i^0, Y_i^1 = y_i^1, Z_i) = \prod_{x=0}^{1} \Pr(Y = y_i^x|X = x, Z_i)\Pr^*(Z_i),$$

so that the likelihood for the whole data set becomes equal to

$$\prod_{i=1}^{n}\prod_{x=0}^{1} \Pr(Y = y_i^x|X = x, Z_i)\Pr^*(Z_i).$$

When Z is unobserved (as in twin studies), the likelihood contribution for pair i is

$$E^*_{Z_i}\{\Pr^*(Y_i^0 = y_i^0, Y_i^1 = y_i^1|Z_i)\} = E^*_{Z_i}\left\{\prod_{x=0}^{1} \Pr(Y = y_i^x|X = x, Z_i)\right\},$$

TABLE 1
Crude summary of matched 1:1 *cohort data*

	Unexposed pair member ($X = 0$)		Totals
	Event ($Y = 1$)	**No event ($Y = 0$)**	
Exposed pair member ($X = 1$)			
Event ($Y = 1$)	T	U	$T + U$
No event ($Y = 0$)	V	W	$V + W$
Totals	$T + V$	$U + W$	n

so that the the likelihood for the whole data set becomes equal to

$$\prod_{i=1}^{n} E^*_{Z_i}\left\{\prod_{x=0}^{1} \Pr(Y = y_i^x | X = x, Z_i)\right\}.$$

We note that marginally (over Z), Y^0 and Y^1 are associated through the common value of Z; the stronger conditional association between Y and Z, given X, the stronger marginal association between Y^0 and Y^1.

4. ANALYSIS METHODS

In this section we describe and compare the most common analysis methods for matched cohorts. We emphasize that all these methods can in principle be used to analyze the exposure–discordant pairs in twin studies as well. However, the explicit regression model (Section 4.1) requires Z to be observed, which is typically not the case in twin studies.

4.1 Regression Model Explicitly Involving Z

A straightforward way to adjust for Z is to fit a regression model for Y, given X and Z, for example,

$$\text{logit}\{\Pr(Y = 1|X, Z; \psi_c, \gamma)\} = b(Z; \gamma) + \psi_c X, \tag{4}$$

where $b(Z; \gamma)$ is an explicitly specified parametric function of Z, typically a linear function $\gamma^T Z$ for continuous Z. We refer to a regression model for Y, given X and Z, as "explicit." Under model (4), $\log\{OR_c(Z)\} = \psi_c$, so that the condition in (2) is met. This restriction is not crucial though; in principle we can add arbitrary interaction terms between X and any of the components of Z. Maximum likelihood estimates (MLEs) of (ψ_c, γ) are obtained by maximizing the conditional (given Z) likelihood

$$\begin{aligned}&\prod_{i=1}^{n} \Pr^*(Y_i^0 = y_i^0, Y_i^1 = y_i^1 | Z_i)\\ &= \prod_{i=1}^{n}\prod_{x=0}^{1} \Pr(Y = y_i^x | X = x, Z_i; \psi_c, \gamma),\end{aligned} \tag{5}$$

where the equality follows from (3a) and (3b). If (3b) is violated, then Y^1 and Y^0 are not conditionally independent, given Z, and the right-hand side of (5) is not a proper likelihood. However, if (3a) holds (and model (4) is correct), then each separate term $\Pr(Y = y_i^x | X = x, Z_i; \psi_c, \gamma)$ in (5) equals the true marginal (over Y_i^{1-x}) likelihood $\Pr(Y_i^x = y_i^x | Z_i)$. It follows that the obtained estimate of ψ_c is consistent under (3a), regardless of whether (3b) holds or not.

4.1.1 *Disadvantages.*

(1) If Z is high dimensional, it may be difficult to well specify the function $b(Z; \gamma)$.

(2) If Z is not directly observed, as in twin studies, explicit specification of $b(Z; \gamma)$ is not possible.

(3) In principle, explicit regression models can be adapted for risk differences and risk ratios, by using identity links or the log links, respectively. However, absolute risks and logarithms thereof are, unlike log odds, restricted to ranges $(0, 1)$ and $(0, \infty)$, respectively. Thus, models utilizing identity links or log links have to be fitted under these restrictions, which can be rather inconvenient, or they may produce estimates which are outside the supported ranges.

4.2 Conditional Logistic Regression

Conditional logistic regression mitigates the problems with an explicit specification of $b(Z; \gamma)$. In conditional logistic regression, the function $b(Z; \gamma)$ in (4) is replaced with a scalar pair-specific parameter b:

$$\text{logit}\{\Pr(Y = 1|X, Z)\} = b + \psi_c X. \tag{6}$$

Nothing is assumed about b, and thus the risk for model misspecification in $b(Z; \gamma)$ is avoided. A MLE of ψ_c is obtained by conditioning on $Y_i^0 + Y_i^1$, for each pair i, and maximizing the resulting conditional likelihood, which under (3a) and (3b) is given by

$$\prod_{i: y_i^0 \neq y_i^1} \frac{e^{\psi_c y_i^1}}{1 + e^{\psi_c}}. \tag{7}$$

Since the conditional likelihood (7) does not involve b (or Z), it can be used, even if Z is not directly observed, as in twin studies. The MLE of ψ_c obtained by maximizing (7) is given by

$$\hat{\psi}_{c.clr} = \log(U/V), \tag{8}$$

with standard error $s.e.\{\hat{\psi}_{c.clr}\} = \sqrt{U^{-1} + V^{-1}}$.

4.2.1 *Disadvantages.*

(1) The constant odds ratio assumption (2) is crucial in conditional logistic regression. If an interaction term is included between b and X in model (6), then b cannot be eliminated by conditioning arguments. If (2) is violated, then $\hat{\psi}_{c.clr}$ converges to a weighted average of the Z-specific odds ratios; see Section 4.4.

(2) $\hat{\psi}_{c.clr}$ is generally inconsistent if (3b) is violated. There is an important exception. Define the null hypothesis

$$\mathrm{H}_0: \quad (2) \text{ holds, with } \psi_c = 0. \tag{9}$$

In Appendix B we show that $\hat{\psi}_{c.clr}$ converges to 0 under H_0 and (3a), regardless of whether (3b) holds or not.

(3) Conditional logistic regression cannot be used for other measures of association than the log odds ratio, since for other links than the logit link, b cannot be eliminated by conditioning arguments.

4.3 Mixed Model

In the mixed model approach, b is assumed to be random, with a specified parametric distribution $\Pr^*(b;\theta)$. The MLE of (ψ_c, θ) is obtained by maximizing the marginal (over b) likelihood

$$\prod_{i=1}^{n} E^*_{Z_i}\{\Pr^*(Y_i^0 = y_i^0, Y_i^1 = y_i^1 | Z_i)\}$$
$$= \prod_{i=1}^{n} E^*_{b_i}\left[\left\{\prod_{x=0}^{1} \Pr(Y = y_i^x | X = x, b_i; \psi_c)\right\}; \theta\right], \tag{10}$$

where the equality follows from (3a) and (3b), and the expectation on the right-hand side is taken over $\Pr^*(b;\theta)$. Neuhaus et al. (1994) showed that the mixed model estimate of ψ_c is identical to $\hat{\psi}_{c.clr}$, under mild conditions. This implies that the two methods are equally efficient, and that the mixed model is robust against misspecification of $\Pr^*(b;\theta)$.

4.3.1 *Disadvantages.*

(1) The constant odds ratio assumption (2) is crucial in the mixed model. Neuhaus et al. (1994) showed that the mixed model is saturated, under mild conditions, so that an interaction term between b and X would lead to identifiability problems.

(2) The mixed model estimate of ψ_c is generally inconsistent if (3b) is violated.

(3) In principle, the mixed model can be adapted for risk differences and risk ratios, by using identity links or the log links, respectively. In practice, these adaptations require that the model is fitted under restrictions, or it may produce estimates outside the supported ranges.

(4) Explicit maximization of the likelihood in (10) requires numerical techniques. This makes the method less transparent and relatively computer-intensive.

4.4 Exposure–Discordant Crude Analysis

The methods described in Sections 4.1–4.3 all target the conditional odds ratio, $OR_c(Z)$. Matched data can also be used to estimate a standardized odds ratio. Let n_{yx} denote the number of subjects in the sample with $Y = y$ and $X = x$, so that $n_{00} = U + W$, $n_{01} = V + W$, $n_{10} = V + T$ and $n_{11} = U + T$. Under (3a) we have that $\Pr^*(Y^x = y^x) = E^*_Z\{\Pr(Y = y^x | X = x, Z)\}$, that is, $\Pr^*(Y^x = y^x)$ equals the probability of $Y = y^x$ given $X = x$, standardized to $\Pr^*(Z)$. Thus, under (3a) a consistent estimate of ψ_s is given by the crude log odds ratio

$$\hat{\psi}_{s.crude} = \log\left(\frac{n_{11}n_{00}}{n_{01}n_{10}}\right). \tag{11}$$

The standard error of $\hat{\psi}_{s.crude}$ (see Appendix A) is given by

$$\sqrt{n_{11}^{-1} + n_{01}^{-1} + n_{10}^{-1} + n_{00}^{-1} - 2n\frac{nT - n_{11}n_{10}}{n_{11}n_{00}n_{01}n_{10}}}. \tag{12}$$

The first four terms under the square root sign can be recognized from the usual standard error formula for a log odds ratio, and the fifth term is an adjustment for non-i.i.d. observations.

We remind the reader that the interpretation of ψ_s depends on what distribution of Z that ψ_s is standardized to. Under exposure-driven matching, $\Pr^*(Z) = \Pr(Z|X = 1)$ so that ψ_s is standardized to the distribution of Z among the exposed. In a twin study, $\Pr^*(Z) = \Pr(Z|\text{discordant in } X)$ so that ψ_s is standardized to the distribution of Z among the exposure–discordant pairs.

4.4.1 *Advantages.* One potential disadvantage of the exposure–discordant crude analysis is that it estimates a parameter that is rather nonstandard. In the simple scenario that we consider (i.e., 1:1 matching and no additional covariate adjustments) the exposure–discordant crude analysis does not suffer from any of the other disadvantages listed in Sections 4.1–4.3. The relative advantages of the exposure–discordant crude analysis are threefold:

(1) The exposure–discordant crude analysis relies on fewer assumptions than the other methods. Specifically, it does not rely on assumptions (2) and (3b).

(2) The exposure–discordant crude analysis is computationally simple.

(3) In the exposure–discordant crude analysis, the standardized probabilities $\Pr^*(Y^0 = 1)$ and $\Pr^*(Y^1 = 1)$ can be estimated separately, and can subsequently be used to construct any standardized measure of the X–Y association, for example, risk difference or risk ratio. For this reason, the exposure–discordant crude analysis easily extends to nonbinary outcomes as well. For survival outcomes, for instance, an exposure–discordant crude analysis can be used to produce standardized Kaplan–Meier curves.

4.4.2 *A closer comparison with conditional logistic regression.* Because ψ_s and ψ_c are different parameters, it is not meaningful to compare the methods in Sections 4.1–4.3 with the exposure–discordant crude analysis in terms of efficiency of estimates. However, we can make a meaningful comparison in terms of statistical power. Define the null hypothesis

$$\mathrm{H}_0^*: \quad \psi_s = 0. \tag{13}$$

It is easy to show that H_0 in (9) implies H_0^*, regardless of whether (3a) and (3b) hold or not. If both (3a) and (3b) hold, then a Wald test of H_0 is based on the statistic $T_c = \hat{\psi}_{c.clr}/s.e.(\hat{\psi}_{c.clr})$. If (3a) holds, then a Wald test of H_0^* is based on the statistic $T_s = \hat{\psi}_{s.crude}/s.e.(\hat{\psi}_{s.crude})$. In Appendix B we show that T_c and T_s are asymptotically equal. It immediately follows that the two Wald tests have the same asymptotic power, for any fixed alternative.

One potential argument against the exposure–discordant crude analysis is that it does not inform us about the exposure effect in the source population. Under exposure-driven matching (and no confounders apart from Z), ψ_s is a causal effect in a fictitious population where Z is distributed as among the exposed. In a twin study restricted to the exposure–discordant pairs (and no confounders apart from Z), ψ_s is a causal effect in a fictitious population where Z is distributed as among the exposure–discordant pairs. The effect in these fictitious populations may differ from the effect in the source population, and it is not always obvious whether these fictitious population effects are relevant targets for inference. However, a closer examination shows that a similar argument can be used against the methods that target ψ_c as well, and in particular against conditional logistic regression. Conditional logistic regression relies on the constant odds ratio assumption (2). This is a very strong assumption, which in any real scenario is most likely violated, to some extent. Regardless of whether (2) holds or not, $\hat{\psi}_{c.clr}$ converges to

$$\begin{aligned}
&\log\left\{\frac{\Pr^*(Y^1=1,Y^0=0)}{\Pr^*(Y^0=1,Y^1=0)}\right\}\\
&\quad=\log\left[\frac{E_Z^*\{\Pr^*(Y^1=1,Y^0=0|Z)\}}{E_Z^*\{\Pr^*(Y^0=1,Y^1=0|Z)\}}\right]\\
&\quad\overset{(3a),(3b)}{=}\log\left[\frac{E_Z^*\{\Pr(Y=1|X=1,Z)\Pr(Y=0|X=0,Z)\}}{E_Z^*\{\Pr(Y=1|X=0,Z)\Pr(Y=0|X=1,Z)\}}\right]\\
&\quad=\log[E_Z^*\{W(Z)OR_c(Z)\}],
\end{aligned} \tag{14}$$

where

$$W(Z) = \frac{\Pr(Y=1|X=0,Z)\Pr(Y=0|X=1,Z)}{E_Z^*\{\Pr(Y=1|X=0,Z)\Pr(Y=0|X=1,Z)\}}.$$

In (14), the average is taken over $\Pr^*(Z)$, that is, the same distribution of Z as being standardized to in the exposure–discordant crude analysis. Thus, if (2) is violated, then conditional logistic regression does not inform the analyst about exposure effects outside the fictitious population characterized by $\Pr^*(Z)$, to any wider extent than the exposure–discordant crude analysis. Furthermore, whereas ψ_s has a clear interpretation as a population causal effect (when there are no confounders except Z), the weighted average in (14) does not have any such simple interpretation.

An analyst is always at the liberty to assume a priori that (2) holds. But equally well, the analyst may assume that the effect in the fictitious population, characterized by $\Pr^*(Z)$, is equal to the effect in the source population, characterized by $\Pr(Z)$. Neither of these assumptions is stronger than the other, since neither of them implies the other. Furthermore, with paired data and Z being unobserved (as in twin studies), these assumptions are both untestable.

Although our focus is on cohort studies, we end this section by making a comparison with case control studies. A matched case control study is designed analogously to a matched cohort study, but the roles of exposure and outcome are "switched" in the sampling scheme; see Section 1. Thus, in a matched case control study the crude sample log odds ratio consistently

estimates the standardized log odds ratio

$$\log\left[\frac{E_Z^*\{\Pr(X=1|Y=1,Z)\}E_Z^*\{\Pr(X=0|Y=0,Z)\}}{E_Z^*\{\Pr(X=0|Y=1,Z)\}E_Z^*\{\Pr(X=1|Y=0,Z)\}}\right], \tag{15}$$

where $\Pr^*(Z) = \Pr(Z|Y=1)$. In contrast to conditional odds ratios, standardized odds ratios are not symmetrical. That is, the log odds ratio in (15), in which X appears to the left of the conditioning sign, cannot be written as ψ_s, in which X appears to the right of the conditioning sign. Hence, the log odds ratio in (15) has no simple interpretation as a causal effect of X on Y on the log odds ratio scale, even if there are no confounders apart from Z.

5. ANALYSIS OF TWIN DATA

In contrast to a regular 1:1 matched cohort study, a twin cohort also contains pairs that are concordant in the exposure. In this section we describe three common methods to incorporate the exposure–concordant pairs in the analysis.

To deal with twin studies we extend the notation slightly. Let X_{ij} and Y_{ij} denote X and Y for twin j in pair i, $j \in (1,2)$. We suppress the index i when not needed, so that X_j and Y_j denote X and Y for twin j, $j \in (1,2)$, within an arbitrary pair i. As before, Z_i represents all the unobserved factors that are common within a twin pair. As discussed in Section 1, the exposure–discordant pairs in a twin cohort can be viewed as a 1:1 matched cohort. However, some care must be taken. All methods discussed in Section 4 rely on assumption (3a), and conditional logistic regression (Section 4.2) and mixed models (Section 4.3) rely in addition on assumption (3b). For an exposure–discordant twin pair we have that

$$\begin{aligned}&\Pr^*(Y^0=y^0,Y^1=y^1|Z)\\ &\quad=\Pr(Y_j=y^0,Y_{j'}=y^1|X_j=0,X_{j'}=1,Z).\end{aligned} \tag{16}$$

The right-hand side of (16) can be factorized into $\Pr(Y_j=y^0|X_j=0,Z)\Pr(Y_{j'}=y^1|X_{j'}=1,Z)$ if

$$Y_j \perp X_{j'}|(X_j,Z) \tag{17a}$$

and

$$Y_1 \perp Y_2|(X_1,X_2,Z). \tag{17b}$$

Thus, the analogs to (3a) and (3b) for twin data are given by (17a) and (17b), respectively. Under (17a), (3a) holds, so that the explicit model (Section 4.1) and the exposure–discordant crude analysis (4.4) are valid when applied to the exposure–discordant pairs. We note though that it is typically not possible to fit an explicit model to twin data, since Z is typically unobserved. If, in addition, (17b) holds, then (3b) holds as well, and all methods in Section 4 are valid when applied to the exposure–discordant pairs.

Potentially, (17a) could be violated if $X_{j'}$ has a causal effect on Y_j, that is, if the exposure for one twin affects the outcome for the other twin. Similarly, (17b) could be violated if $Y_{j'}$ has a causal effect on Y_j, that is, if the outcome of one twin affects the outcome for the other twin.

5.1 All-Pair Crude Analysis

Let r_{yx} denote the number of subjects in the full (i.e., both exposure–concordant and exposure–discordant pairs) sample with $Y=y$ and $X=x$. One simple way to make use of all twin pairs in the analysis is to compute the crude sample log odds ratio

$$\hat{\psi}_{m.crude} = \log\left(\frac{r_{11}r_{00}}{r_{01}r_{10}}\right), \tag{18}$$

which consistently estimates the marginal log odds ratio ψ_m. Thus, unlike the exposure–discordant crude analysis (Section 4.4), the all-pair crude analysis does not adjust for confounding by Z. The standard error of $\hat{\psi}_{m.crude}$ is rather complicated, due to the paired nature of the data. In Appendix A we provide an analytic expression for the standard error. We note that the standard error can also be computed numerically, through Generalized Estimating Equation (GEE) procedures, which are implemented in most common statistical softwares.

5.2 Decomposition into Within- and Between-Effects

In twin studies with continuous exposures and outcomes, a popular regression model is

$$\begin{aligned}E(Y_j|X_j,X_{j'}) &= \beta_0+\beta_{\mathrm{W}}(X_j-\bar{X})+\beta_{\mathrm{B}}\bar{X}\\ &= \beta_0+\beta_{\mathrm{W}}X_j+\beta'_{\mathrm{B}}\bar{X},\end{aligned} \tag{19}$$

with $\bar{X}=\frac{X_1+X_2}{2}$ and $\beta'_{\mathrm{B}}=\beta_{\mathrm{B}}-\beta_{\mathrm{W}}$ (Carlin et al., 2005). In (19), the pair-specific mean $\bar{X}$ is thought of as conveying information about the confounders Z, which are not observed, but constant within each pair. Thus, the parameter β_{B} is thought of as quantifying the strength of confounding, a "between effect," and the parameter β_{W} is thought of as quantifying the adjusted

X–Y association, a "within effect." When X and Y are binary, a natural analog to (19) is

$$\begin{aligned}&\operatorname{logit}\{\Pr(Y_j=1|X_j,X_{j'})\}\\&\quad=\beta_0+\beta_{\mathrm{W}}X_j+\beta_{\mathrm{B}}'\bar{X}.\end{aligned}\tag{20}$$

To see the connection with the methods described in this paper, note that

$$\begin{aligned}\beta_{\mathrm{W}}&=\operatorname{logit}\{\Pr(Y_j=1|X_j=1,X_{j'}=0)\}\\&\quad-\operatorname{logit}\{\Pr(Y_j=1|X_j=0,X_{j'}=1)\}\\&=\operatorname{logit}[E\{\Pr(Y_j=1|X_j=1,X_{j'}=0,Z)|\\&\qquad X_j=1,X_{j'}=0\}]\\&\quad-\operatorname{logit}[E\{\Pr(Y_j=1|X_j=0,X_{j'}=1,Z)|\\&\qquad X_j=0,X_{j'}=1\}]\\&=\operatorname{logit}[E^*\{\Pr(Y_j=1|X_j=1,Z)\}]\\&\quad-\operatorname{logit}[E^*\{\Pr(Y_j=1|X_j=0,Z)\}]\\&=\psi_s,\end{aligned}$$

where $\Pr^*(Z)=\Pr(Z|X_1\neq X_2)$, and the third equality follows from assumption (17a). Thus, the within-effect β_W is identical to the log odds ratio standardized to the distribution of Z among the exposure–discordant pairs. This argument shows that the decomposition into within- and between-effects is a legitimate method for binary exposures, which was questioned by Carlin et al. (2005).

When X is binary, $\bar{X}$ can only take values 0, 0.5 and 1. Thus, it is feasible to replace the linear term $\beta_0+\beta_{\mathrm{B}}'\bar{X}$ in (20) with one parameter for each level of $\bar{X}$, that is,

$$\operatorname{logit}\{\Pr(Y_j=1|X_j,X_{j'})\}=\beta_{\mathrm{W}}X_j+m(\bar{X}),\tag{21}$$

with

$$\begin{aligned}m(\bar{X})&=\beta_0\mathbf{1}(\bar{X}=0)\\&\quad+\beta_{0.5}\mathbf{1}(\bar{X}=0.5)+\beta_1\mathbf{1}(\bar{X}=1).\end{aligned}\tag{22}$$

It is easy to show that the model in (21) is saturated (i.e., imposes no restrictions on $\Pr(Y_j|X_1,X_2)$, which implies that the MLE of β_{W} based on (21) is identical to the crude sample log odds ratio in (11).

5.3 Mixed Model

The model in (6) can be fitted to all pairs, assuming a parametric distribution of b indexed with θ. Parameter estimates are obtained by maximizing the marginal (over b) likelihood

$$\begin{aligned}&\prod_{i=1}^{n}E^*_{Z_i|X_{i1},X_{i2}}\{\Pr(Y_{i1}=y_{i1},Y_{i2}=y_{i2}|X_{i1},X_{i2},Z_i)|\\&\qquad X_{i1},X_{i2}\}\\&=\prod_{i=1}^{n}E^*_{b_i|X_{i1},X_{i2}}\left[\left\{\prod_{j=1}^{2}\Pr(Y_{ij}=y_{ij}|X_{ij},b_i;\psi_c)\right\}\Big|\right.\\&\qquad\left.X_{i1},X_{i2};\theta\right].\end{aligned}\tag{23}$$

This approach, however, is associated with a severe problem which is often overlooked. Typically, the distribution of b is specified to not depend on (X_1,X_2), for example, a normal distribution with fixed but unspecified mean and variance. However, from the expression in (23) it is clear that this procedure only produces a proper likelihood under the additional assumption that $b\perp(X_1,X_2)$. In standard textbooks, this assumption is often stated without justification or interpretation (e.g., Fitzmaurice et al., 2004, page 329). Since b is supposed to represent the potential confounders Z, we would not generally expect that $b\perp(X_1,X_2)$. Indeed, if Z (and thus b) is independent of (X_1,X_2), it cannot be a confounder, and there is no need to adjust for Z in the first place. We note that in matched cohort studies, (X_1,X_2) is constant and equal to $(0,1)$ for all pairs, so that an association between b and (X_1,X_2) is ruled out by design. When b is associated with (X_1,X_2), the aforementioned procedure can yield severely biased estimates (Neuhaus and Kalbfleisch, 1998; Neuhaus and McCulloch, 2006). In general, the proper marginal likelihood is obtained by averaging over a specified distribution $\Pr(b|X_1,X_2)$ for each pair. This procedure can be very computer intensive, and cannot be carried out with standard software. As noted by Neuhaus and Kalbfleisch (1998) and Neuhaus and McCulloch (2006), there is a simple solution to this problem. Suppose that given (X_1,X_2), b has a normal distribution where the mean, but not the variance, depends on (X_1,X_2). Without loss of generality, we can formulate this as

$$b=d+m(\bar{X}),\tag{24}$$

where $m(\bar{X})$ is defined in (22) and $d|X_1,X_2\sim N(0,\sigma^2)$. Under (24), model (6) translates to

$$\operatorname{logit}\{\Pr(Y_j=1|X_j,Z)\}=d+\psi_cX_j+m(\bar{X}),\tag{25}$$

where $d\perp(X_1,X_2)$ by construction. The model in (25) can be fitted with standard mixed model software. By comparing the model in (25) with the model

in (21), we see that the solution proposed by Neuhaus and Kalbfleisch (1998) and Neuhaus and McCulloch (2006) can be thought of as combining a mixed model with a within-between decomposition.

Neuhaus and Kalbfleisch (1998) and Neuhaus and McCulloch (2006) observed that for various scenarios, the estimate of ψ_c obtained by combining a mixed model with a within-between decomposition is nearly identical to $\hat{\psi}_{c.clr}$. Neuhaus and McCulloch (2006) gave a theoretical motivation for this observation. We note that there are situations when the two estimates may differ; see Brumback et al. (2010) for an example.

6. SIMULATIONS

6.1 Part I: Efficiency and Power

In this section we compare the performance of the methods described in Sections 4 and 5, in terms of efficiency and power. To enable a fair comparison, we analyze the simulated data so that all assumptions hold, for each method respectively. In these simulations, twin pairs were generated. We emphasize that this simulation scheme covers matched data as well, since the exposure–discordant twin pairs can be viewed as a matched cohort. For each twin pair, the random variables (X_1, X_2, b, Y_1, Y_2) were generated from the model

$$
(26)\quad \begin{cases} \dfrac{\Pr(X_1=1|X_2=0)}{\Pr(X_1=0|X_2=0)} = \dfrac{\Pr(X_2=1|X_1=0)}{\Pr(X_2=0|X_1=0)} \\ \quad = \rho = \dfrac{1}{2}, \\ \dfrac{\Pr(X_1=1, X_2=1)\Pr(X_1=0, X_2=0)}{\Pr(X_1=1, X_2=0)\Pr(X_1=1, X_2=0)} = \phi, \\ b|X_1, X_2 \sim N\{\theta\bar{X}, 1\}, \\ Y_1 \perp Y_2|(X_1, X_2, b), \\ Y_j \perp X_{j'}|(X_j, b), \\ \text{logit}\{\Pr(Y_j|X_j, b)\} = b + \psi_c X_j. \end{cases}
$$

We highlight a few aspects of the model in (26):

(1) Under model (26), assumptions (2), (17a), (17b) and (24) all hold.

(2) The restriction $\Pr(X_1 = 1|X_2 = 0) = \Pr(X_2 = 1|X_1 = 0)$ in the first row of (26) follows by symmetry.

(3) It may appear natural to first specify a marginal distribution of b, then specify a conditional distribution of (X_1, X_2), given b. The reason for doing it the other way around is twofold. First, it allows us to directly control the rate of exposure-discordance through ϕ. Second, it allows us to easily formulate the distribution of b given (X_1, X_2) in such a way that (24) holds.

(4) It follows from results in Chen (2007) that the joint distribution of (X_1, X_2) is completely defined by ρ and ϕ. It also follows that ρ and ϕ are variation independent (i.e., the value of ρ does not restrict the value of ϕ, and vice versa).

(5) The values of ϕ and θ determine the degree of conditional association of X_1 and X_2, given b. It can be shown (see Appendix C) that for $\theta = 2\sqrt{\log(\phi)}$, $X_1 \perp X_2|b$. For convenience, we have used $\theta = 2\sqrt{\log(\phi)}$ throughout. We note though that none of the methods presented relies on this restriction.

In the first set of simulations, we used $\phi = 4$ and $\psi_c = 0$, that is, the data were generated under H_0 in (9). For these values, $\psi_s = 0$ and $\psi_m = 1.28$, which implies a severe degree of confounding. Further, $\Pr(X_1 \neq X_2) = 0.33$, and $\Pr(X_1 \neq X_2, Y_1 \neq Y_2) = 0.11$. We generated 5000 samples, each of size $n = 2000$. Each sample was analyzed with 6 different methods:

(1) Explicit regression model $\text{logit}\{\Pr(Y = 1|X, b)\} = \gamma_0 + \gamma_1 b + \psi_c X$ (Section 4.1). We remind the reader that for twin data, b (or rather, Z) is typically unobserved, which rules out the use of an explicit model. For a regular matched cohort, the explicit model is a viable choice. Thus, the model was only fitted to the exposure–discordant pairs.

(2) Conditional logistic regression (Section 4.2).

(3) Mixed model fitted to the exposure–discordant pairs (Section 4.3). We used the model $\Pr(Y = 1|X, b) = b + \psi_c X$, with $b|X_1 \neq X_2 \sim N(\theta, \sigma^2)$.

(4) Exposure–discordant crude analysis (Section 4.4).

(5) All pair crude analysis (Section 5.1).

(6) Mixed model fitted to all pairs (Section 5.3). We used the model $\Pr(Y = 1|X, b) = b + \psi_c X$, with $b|X_1, X_2 \sim N(\theta\bar{X}, \sigma^2)$.

Table 2 displays the mean (over samples) point estimate, the empirical standard error and the mean theoretical standard error for each analysis, respectively. We note that all methods yield virtually unbiased estimates of their target parameters. For all methods the mean theoretical standard error is identical to the empirical standard error, to the second decimal.

To compare the methods in terms of their power to reject H_0, we carried out a second set of simulations. We used $\phi = 4$ and varied ψ_c over the range $(0, 0.6)$. For each value of ψ_c, we drew 5000 samples of 2000 pairs each. Each sample was analyzed using methods 1, 2, 3, 4, 6. Figure 3 displays the empirical rejection probability (i.e., the power) for a Wald test at 5% significance level, for each method as a function of ψ_c.

TABLE 2
Simulation results for $\psi_c = 0, \phi = 4$

Analysis method	Target parameter	Mean est	Emp s.e.	Th s.e.
1. Explicit	$\psi_c = 0$	0.00	0.13	0.13
2. Cond log reg	$\psi_c = 0$	0.00	0.13	0.13
3. Mixed discordant	$\psi_c = 0$	0.00	0.13	0.13
4. Crude discordant	$\psi_s = 0$	0.00	0.11	0.11
5. Crude all	$\psi_m = 1.28$	1.28	0.08	0.08
6. Mixed all	$\psi_c = 0$	0.00	0.12	0.12

We observe that the all methods have almost identical power, for the simulated scenarios.

In a third set of simulations, we used $\psi_c = 0.4$ and varied ϕ over the range $(4, 22)$. These values correspond to the range $(0.33, 0.13)$ for $\Pr(X_1 \neq X_2)$, and the range $(0.11, 0.03)$ for $\Pr(X_1 \neq X_2, Y_1 \neq Y_2)$. For each value of ϕ, we drew 5000 samples of 2000 pairs each. Each sample was analyzed using methods 1, 2, 3, 4, 6. Figure 4 displays the power for each method as a function of ϕ. Again, we observe that there is almost no difference between the methods, in terms of power, even when the discordance rate is very low.

Some care must be taken when interpreting power curves. In small samples, parameter estimates can be biased, which may lead to an increased probability of rejection, both under the alternative hypothesis and under the null hypothesis. Thus, an increased power under the alternative hypothesis may come at the cost of a violated significance level under the null hypothesis. Figure 3 shows that the nominal significance level ($= 5\%$ at $\psi_c = 0$) is preserved for all methods when $\phi = 4$. To confirm that the nominal significance level is preserved across the range $\phi \in (4, 22)$, which generated the power curves in Figure 4, we carried out a

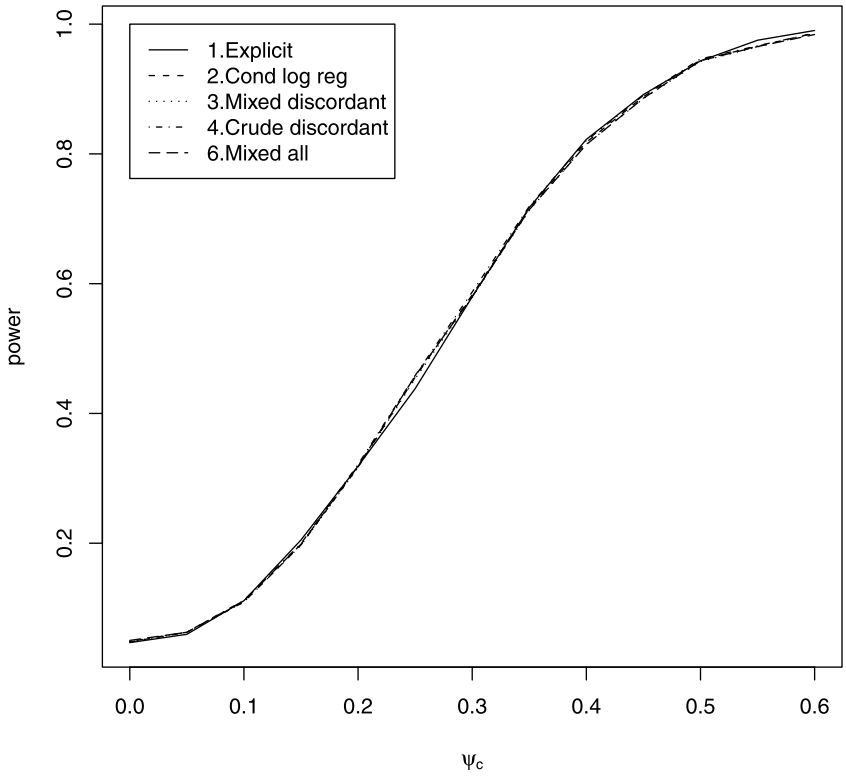

FIG. 3. *Simulation results for* $\psi_c \in (0, 0.6), \phi = 4$.

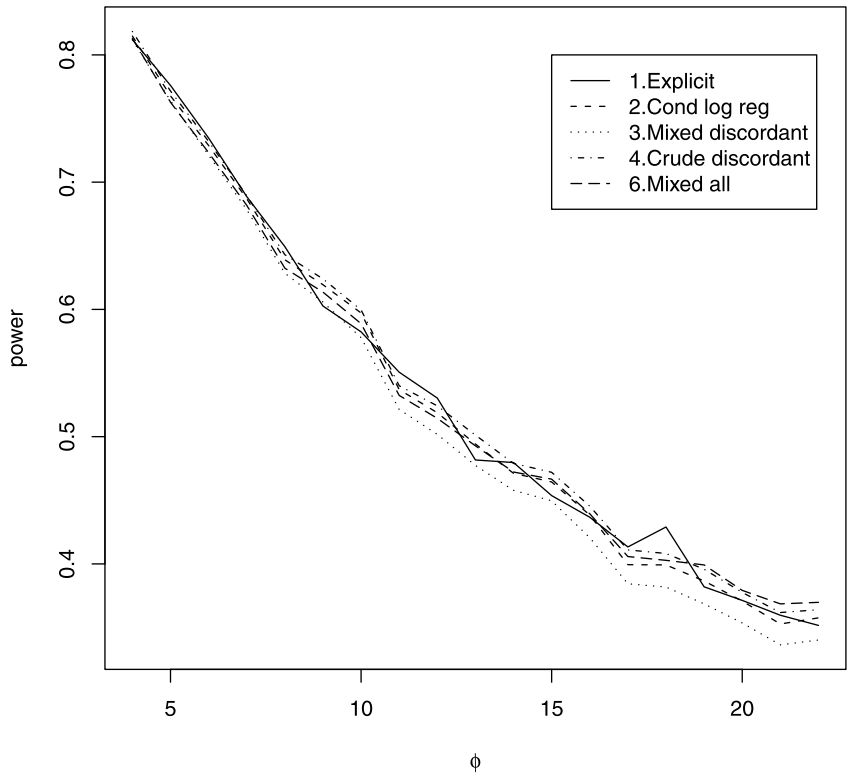

FIG. 4. *Simulation results for* $\psi_c = 0.4$, $\phi \in (4, 22)$.

fourth set of simulations, using $\psi_c = 0$ and varying ϕ over the range $(4, 22)$. For each value of ϕ, we drew 5000 samples of 2000 pairs each. Each sample was analyzed using methods 1, 2, 3, 4, 6. Figure 5 displays the rejection probability for each method as a function of ϕ. We observe that the rejection probability is close to 0.05, for all methods and all values of ϕ in the simulated range.

Table 2 and Figure 3 indicate that methods 1–4, and 6 are unbiased under the null hypothesis. Additional simultions have confirmed that the methods are unbiased under various alternative hypotheses as well (data not shown).

6.2 Part II: Sensitivity to Underlying Assumptions.

In this section we demonstrate through examples that the explicit model, conditional logistic regression and the mixed model, can yield biased estimates, if their underlying assumptions are violated.

We first consider the assumption that $b \perp (X_1, X_2)$, which is often made for mixed models; see Section 5.3. Toward this end we reanalyzed the 5000 simulated samples which generated Table 2, now fitting the mixed model $\Pr(Y = 1|X, b) = b + \psi_c X$ to all pairs, with $b|X_1, X_2 \sim N(\theta, \sigma^2)$. We obtained a mean estimate of ψ_c equal to 1.32, which is indeed biased as an estimate of the true value $\psi_c = 0$. We note that this mean estimate is very close to the $\hat{\psi}_{m.crude}$ $(= 1.28)$ in Table 2. This further demonstrates that ignoring the association between b and (X_1, X_2) produces an estimate which is not adjusted for Z.

Next, we consider the independence assumption (3b)/(17b), which is a prerequisite for conditional logistic regression and mixed models. Toward this end we consider a simple scenario for which

$$(Y_1, Y_2) \perp Z|(X_1, X_2), \tag{27}$$

so that $\psi_c = \psi_s = \psi_m$; see Section 2. We define

$$\begin{cases} \Pr(Y_j = 1|Y_{j'} = 0, X_j = 1, X_{j'} = 0) = p, \\ \Pr(Y_j = 1|Y_{j'} = 0, X_j = 0, X_{j'} = 1) = q, \\ \Pr(Y_j = 1, Y_{j'} = 1|X_j = 1, X_{j'} = 0) \\ \quad \cdot \Pr(Y_j = 0, Y_{j'} = 0|X_j = 1, X_{j'} = 0) \\ \quad /\big(\Pr(Y_j = 0, Y_{j'} = 1|X_j = 1, X_{j'} = 0) \\ \quad \cdot \Pr(Y_j = 1, Y_{j'} = 0|X_j = 1, X_{j'} = 0)\big) = c. \end{cases} \tag{28}$$

It follows from results in Chen (2007) that the joint distribution of Y_j and $Y_{j'}$ among the exposure–discordant pairs, $\Pr(Y_j, Y_{j'}|X_j = 1, X_{j'} = 0)$, is completely defined by the variation independent parameters p, q and c. c quantifies the degree of deviation from (17b);

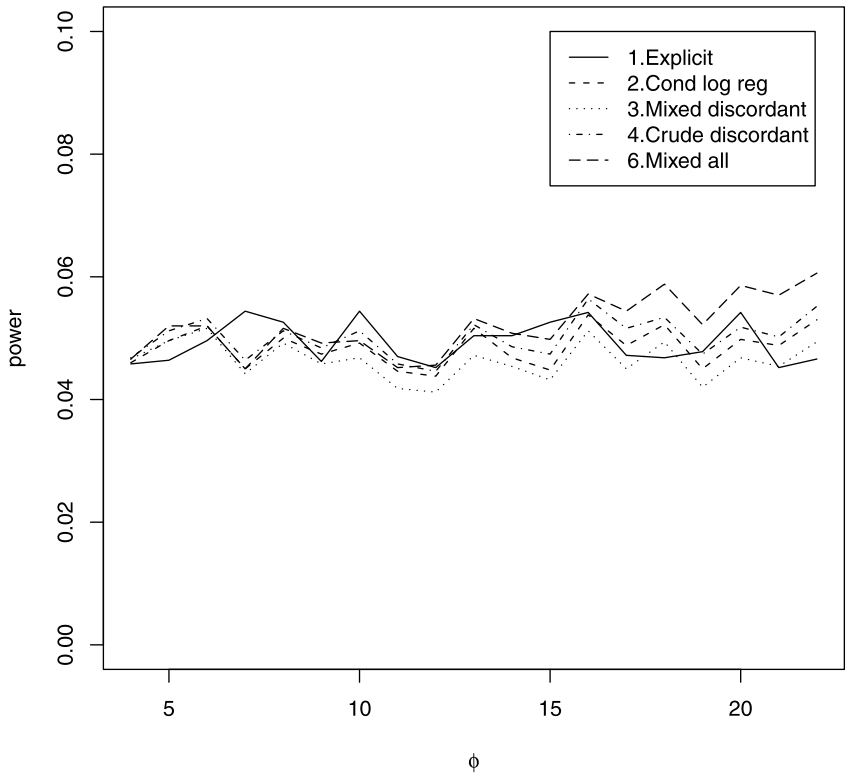

FIG. 5. *Simulation results for* $\psi_c = 0$, $\phi \in (4, 22)$.

in particular, (17b) is violated when $c \neq 1$. It is easy to show that assumption (17a) is logically compatible with all joint values of (p, q, c). Thus, we proceed by assuming that (17a) holds, so that the exposure–discordant crude analysis consistently estimates $\psi_s = \psi_c$. Combining (27) and (28), and using results in Chen (2007), gives that $\hat{\psi}_{c.clr}$ converges to

$$\log\left\{\frac{p(1-q)}{q(1-p)}\right\}, \tag{29}$$

whereas the true value of $\psi_c (= \psi_s = \psi_m)$ is given by

$$\log\left\{\frac{p(1-q)}{q(1-p)}\right\} + \log\left\{\frac{1-q+qc}{1-p+pc}\right\}. \tag{30}$$

Thus, the true value of ψ_c depends on the association between Y_1 and Y_2 through the second term in (30), whereas the asymptotic limit of $\hat{\psi}_{c.clr}$ does not. We used $p = 0.3$, $q = 0.1$, and $c = 4$. For these values, $\psi_c = 0.97$, whereas the asymptotic limit of $\hat{\psi}_{c.clr}$ equals 1.35, for conditional logistic regression. We generated 5000 samples, each consisting of $n = 2000$ exposure–discordant twin pairs. For each pair, the random variables (Y_1, Y_2) were generated from the model in (28). Each sample was analyzed with conditional logistic regression (method 2), the mixed model (method 3) and the exposure–discordant crude analysis (method 4). For these methods, we obtained an average estimate of ψ_c equal to 1.35, 1.26 and 0.97, respectively. Thus, both conditional logistic regression and the mixed model produced biased estimates, whereas the exposure–discordant crude analysis estimate was unbiased.

Next, we consider misspecification of the function $b(Z; \gamma)$, in the explicit model. We generated 5000 samples, each consisting of $n = 2000$ twin pairs. For each twin pair, the random variables (Z, X_1, X_2, Y_1, Y_2) were generated from the model

$$\begin{cases} Z = (V, W), \\ V \perp W, \\ V \sim N(0, 1), \\ W \sim \text{Ber}(0.5), \\ X_1 \perp X_2 | Z, \\ \text{logit}\{\Pr(X_j = 1|Z)\} \\ \quad = \alpha_0 + \alpha_1 V + \alpha_2 W + \alpha_3 V W, \\ Y_1 \perp Y_2 | (X_1, X_2, b), \\ Y_j \perp X_{j'} | (X_j, b), \\ \text{logit}\{\Pr(Y_j = 1|X_j, Z)\} = b(Z; \gamma) + \psi_c X_j, \\ b(Z; \gamma) = \gamma_0 + \gamma_1 V + \gamma_2 W + \gamma_3 V W, \end{cases} \tag{31}$$

with $\alpha_0 = 2$, $\alpha_1 = \alpha_2 = 1$, $\alpha_3 = -1.5$, $\gamma_0 = -2$, $\gamma_1 = \gamma_2 = -1$, $\gamma_3 = 1.5$, $\psi_c = 1.3$. Each sample

was analyzed with the misspecified explicit model $\text{logit}\{\Pr(Y_j = 1|X_j, Z)\} = \gamma_0 + \gamma_1 V + \gamma_2 W + \psi_c X_j$. We obtained an average estimate of ψ_c equal to 0.69, which is severly biased.

Finally, we consider the assumption that the random effect $b(Z;\gamma)$ is normally distributed, which is commonly made for mixed models. Toward this end we reanalyzed the 5000 samples generated from model (31), now fitting the mixed model $\Pr(Y = 1|X, b) = b + \psi_c X$ to the exposure–discordant pairs, with $b|X_1 \neq X_2 \sim N(\theta, \sigma^2)$. Under the data generating model, the conditional distribution of $b(Z;\gamma)$, given $X_1 \neq X_2$ is rather complicated, and, in particular, not normal. We obtained an average estimate of ψ_c equal to 1.30, which is identical to the true value, to the second decimal. This finding supports the theoretical results in Neuhaus et al. (1994), which state that the mixed model is robust against the normal random effect assumption.

7. REAL DATA EXAMPLES

7.1 Matched Cohort Data

The first example is taken from a matched cohort study that aimed to investigate the effect of hysterectomy on risk for CVD (Ingelsson et al., 2010). A common surgery among perimenopausal women, hysterectomy is often performed on benign indications, but its long-term consequences are not fully understood. The study is based on the Swedish Inpatient Register, where all women who underwent hysterectomy between January 1973 and December 2003 (227,389 individuals) were identified. For each hysterectomized woman, three women who never had hysterectomy were randomly selected from the Register of Total Population. The three unexposed women were individually matched to the exposed woman by birth year, year of hysterectomy, and county of residence at year of hysterectomy.

Information on CVD status was obtained from the Inpatient Register and information of follow up through record linkage to the Cause of Death Register, Emigration Register and Cancer Register. To avoid bias from CVD events occurring in relation to the hysterectomy surgery, the exposed women started their risk time from 30 days after hysterectomy; they were then followed until CVD, heart failure, cervical, corpus or ovarian cancer, death, emigration or end of study (Dec 31, 2003). Similarly, unexposed women started their risk time 30 days after the date of matching, that is, the date of hysterectomy of the corresponding exposed woman. For further details on the study, see Ingelsson et al. (2010).

TABLE 3
Analysis results for the 1:1 *matched subset of the hysterectomy-CVD data*

Analysis method	Target parameter	Point est	95% CI
1. Explicit	ψ_c	0.03	−0.02, 0.08
2. Cond log reg	ψ_c	0.03	−0.02, 0.08
3. Mixed discordant	ψ_c	0.03	−0.02, 0.08
4. Crude discordant	ψ_s	0.03	−0.02, 0.07

In the current analysis we focus on 1:1 matched studies with binary outcomes. We constructed a binary outcome by defining $Y = 1$ for women who developed CVD during follow-up, and $Y = 0$ for the remaining women. We constructed a 1:1 matched sample by matching each exposed woman to one unexposed woman, which was randomly selected from the three unexposed women in the same set. After the exclusions described above, we ended up with 52,814 1:1 matched pairs, of which 6712 were discordant in both the exposure and the outcome. The data were analyzed with methods 1–4 described in Section 6. For method 1 we used the explicit model $\text{logit}\{\Pr(Y = 1|Z, X)\} = \gamma_0 + \gamma_1[\text{birth year}] + \gamma_2[\text{year at hysterectomy}] + \gamma_3[\text{county}] + \psi_c X$, where γ_3 is a factor parameter with one level for each county.

Table 3 displays the results. For all three methods, there is a significant (at 5% level) association between hysterectomy and CVD. The point estimates obtained by conditional logistic regression and exposure–discordant crude analysis are almost identical, whereas the point estimate obtained from the mixed model is twice as large. According to theory (Neuhaus et al., 1994) we would expect the mixed model estimate to be identical to the estimate obtained from conditional logistic regression. Indeed, methods 1–4 all give identical estimates to the second decimal.

Although our focus is on 1:1 matching, all methods in this paper generalize directly to m:n matching (see Section 8). Table 4 displays the results when the whole 1:3 matched data is analyzed, using methods 1–4 described in Section 6.

7.2 Twin Data

The second example is from a twin study of the association between fetal growth and asthma (Örtqvist et al., 2009). Several studies have shown that there is an association between asthma and low birth weight. This association could potentially be explained by a causal effect of impaired fetal growth on asthma, but may also

TABLE 4
Analysis results for the full 1:3 *matched hysterectomy-CVD data*

Analysis method	Target parameter	Point est	95% CI
1. Explicit	ψ_c	0.06	0.02, 0.09
2. Cond log reg	ψ_c	0.06	0.02, 0.09
3. Mixed discordant	ψ_c	0.06	0.02, 0.09
4. Crude discordant	ψ_s	0.05	0.02, 0.09

be explained by confounding factors. In particular, gestational age is correlated with both birth weight and asthma, and may confound the birth weight-asthma association (Örtqvist et al., 2009). Twins provide an excellent opportunity to separate the causal effect of birth weight from the confounding effect of gestational age, and at the same time adjust for other shared familial factors.

All twins born in Sweden in June 1992 to June 1998 were identified through the Swedish Twin Register at the age of 9 or 12 years. Information on asthma and zygosity was collected in telephone interviews with their parents. Birth weight was retrieved from the Medical Birth Register (MFR). Of the 15,808 eligible twins 69% (10,918 individuals) had information on asthma and could also be securely linked to the MFR. In total, there were 3107 MZ pairs. 1087 pairs were discordant in birth weight (exposure), where discordance was defined as a difference greater than 400 grams or 15%, and 175 pairs were discordant on both birth weight and asthma (outcome).

The data were analyzed using methods 2–6 described in Section 6. Table 5 displays the results. The estimates obtained from conditional logistic regression and the exposure–discordant crude analysis are both smaller than estimate obtained from the all-pair crude analysis. This finding suggests that the birth weight-asthma association is inflated by shared confounding. Methods 2, 3 and 6 gave very similar results, as predicted by theory (Neuhaus et al., 1994; Neuhaus and Kalbfleisch, 1998).

TABLE 5
Analysis results for the birth weight-asthma twin data

Analysis method	Target parameter	Point est	95% CI
2. Cond log reg	ψ_c	0.29	−0.01, 0.59
3. Mixed discordant	ψ_c	0.29	−0.01, 0.59
4. Crude discordant	ψ_s	0.18	−0.01, 0.37
5. Crude all	ψ_m	0.33	0.16, 0.50
6. Mixed all	ψ_c	0.30	0.00, 0.60

8. DISCUSSION

We have given an overview of the most common analysis methods for matched cohort studies. We have identified the target parameters in each method, outlined the underlying assumptions and compared the methods in terms of statistical power. The analysis methods that we have considered do not estimate the same parameter; the exposure–discordant crude analysis and the within–between model estimate a standardized odds ratio, whereas the explicit method, conditional logistic regression, and the mixed model estimate a conditional odds ratio. Thus, the choice between these methods should primarily be guided by the research question being asked. In addition, it is also important to consider the statistical power, underlying assumptions, computer intensity and flexibility of the methods. Theoretical arguments suggest that when all underlying assumptions hold, all methods that we have considered have the same statistical power. This was confirmed in our simulation study. In terms of underlying assumptions, the methods differ significantly. The exposure–discordant crude analysis relies on fewer assumptions than the other methods. In terms of computer intensity, the mixed model requires numerical optimization, and is far more time consuming than the other methods. In terms of flexibility, all methods, except the exposure–discordant crude analysis, most naturally target odds ratios. The exposure–discordant crude analysis however, can easily be used to target any measure of the exposure-outcome association.

We have considered 1:1 matching. Frequently, m:n matching is employed, that is, each set is constructed by matching m exposed subjects to n unexposed subjects. All methods in this paper generalize directly to m:n matching. Specifically, the underlying assumptions and the interpretation of the target parameters remains the same under m:n matching. We conjecture that many of the theoretical properties that we have derived for 1:1 matching carry over to m:n matching as well, for example, the asymptotic equivalence in terms of power. However, a stringent treatment of m:n matching is more difficult. For instance, under violation of (2) the probability limit of $\hat{\psi}_{c.clr}$ has no longer an analytic expression, which hampers a theoretical comparison with the exposure–discordant crude analysis. Comparing the methods under m:n is a topic for future research.

In practice, it is often desirable to adjust the analysis for additional covariates which are not matched on. In the model-based methods (i.e., all methods except

the exposure–discordant crude analysis), adjustment for additional covariates can easily be accomplished by adding the covariates as a regressor in the model. It is not obvious though, how to adjust for additional covariates in the exposure–discordant crude analysis. Extensions of the exposure–discordant crude analysis for additional covariate adjustments is a topic for future research.

APPENDIX A

Define $p^x = \Pr(Y = 1|X = x)$, $q = \Pr(X = 1)$, $q^{00} = \Pr(X_1 = X_2 = 0)$, $q^{11} = \Pr(X_1 = X_2 = 1)$, $q^d = \Pr(X_1 \neq X_2)$, $c^{00} = \operatorname{cov}(Y_1, Y_2|X_1 = X_2 = 0)$, $c^{11} = \operatorname{cov}(Y_1, Y_2|X_1 = X_2 = 1)$, $c^d = \operatorname{cov}(Y_1, Y_2|X_1 \neq X_2)$, $\psi_0 = \operatorname{logit}(p_0)$, $\psi_m = \operatorname{logit}(p_1) - \operatorname{logit}(p^0)$ and $\psi = (\psi_0, \psi_m)^T$. $\hat{\psi}_{m.crude}$ in (18) can be expressed as the second element of the solution to $\sum_i U_i(\psi) = 0$, where

$$U_i(\psi) = \begin{Bmatrix} (1 - X_{i1})(Y_{i1} - p^0) + (1 - X_{i2})(Y_{i2} - p^0) \\ X_{i1}(Y_{i1} - p^1) + X_{i2}(Y_{i2} - p^1) \end{Bmatrix}.$$

It follows from standard theory that $n^{1/2}(\hat{\psi} - \psi)$ is asympotically normal with mean 0 and variance

$$\left[E\left\{\frac{\partial U_i(\psi)}{\partial \psi^T}\right\}\right]^{-1} \operatorname{var}\{U_i(\psi)\}\left[\left[E\left\{\frac{\partial U_i(\psi)}{\partial \psi^T}\right\}\right]^{-1}\right]^T,$$

where, after some algebra,

$$E\left\{\frac{\partial U_i(\psi)}{\partial \psi^T}\right\} = \begin{pmatrix} -2p^0(1-p^0) & 0 \\ -2p^1(1-p^1) & -2p^1(1-p^1) \end{pmatrix}$$

and

$$\operatorname{var}\{U_i(\psi)\} = \begin{pmatrix} 2(1-q)p^0(1-p^0) + q^{00}c^{00} & q^d c^d \\ q^d c^d & 2qp^1(1-p^1) + q^{11}c^{11} \end{pmatrix}.$$

After additional algebra, the asymptotic variance for $n^{1/2}(\hat{\psi}_{m.crude} - \psi_m)$ is obtained as

$$\begin{aligned} &\frac{1}{2(1-q)p^0(1-p^0)} + \frac{1}{2qp^1(1-p^1)} \\ &\quad + \frac{q^{00}c^{00}}{4\{p^0(1-p^0)\}^2} + \frac{q^{11}c^{11}}{4\{p^1(1-p^1)\}^2} \\ &\quad - \frac{q^d c^d}{2q(1-q)p^0(1-p^0)p^1(1-p^1)}. \end{aligned} \tag{32}$$

Replacing the population parameters in (32) with their sample counterparts gives the standard error for $\hat{\psi}_{m.crude}$.

To derive the standard error formula in (12) we note that a regular 1:1 matched cohort can be obtained by setting $q = 0.5$, $q^{00} = q^{11} = 0$ and $q^d = 1$. The expression in (32) then simplifies to

$$\begin{aligned} &\frac{1}{p^0(1-p^0)} + \frac{1}{p^1(1-p^1)} \\ &\quad - \frac{2c^d}{p^0(1-p^0)p^1(1-p^1)}. \end{aligned} \tag{33}$$

Replacing the population parameters in (33) with their sample counterparts gives the standard error formula in (12).

APPENDIX B

Define $\psi_c^\dagger = \log\{\frac{\Pr^*(Y^1=1,Y^0=0)}{\Pr^*(Y^0=1,Y^1=0)}\}$, $\mathrm{H}_c^\dagger : \psi_c^\dagger = 0$, $\psi_s^\dagger = \log\{\frac{\Pr^*(Y^1=1)\Pr^*(Y^0=0)}{\Pr^*(Y^1=0)\Pr^*(Y^0=1)}\}$, $\mathrm{H}_s^\dagger : \psi_s^\dagger = 0$. $\mathrm{H}_c^\dagger$ can be tested using the likelihood ratio test (LRT) statistic

$$T_{c,LR}^\dagger = -2\log\left\{\frac{\sup_{\mathrm{H}_c^\dagger}(p_{00}^W p_{01}^U p_{10}^V p_{11}^T)}{\sup(p_{00}^W p_{01}^U p_{10}^V p_{11}^T)}\right\},$$

and $\mathrm{H}_s^\dagger$ can be tested using the LRT statistic

$$T_{s,LR}^\dagger = -2\log\left\{\frac{\sup_{\mathrm{H}_s^\dagger}(p_{00}^W p_{01}^U p_{10}^V p_{11}^T)}{\sup(p_{00}^W p_{01}^U p_{10}^V p_{11}^T)}\right\},$$

where $p_{y^0y^1} = \Pr^*(Y^0 = y^0, Y^1 = y^1)$, and the suprema are taken under the restrictions $0 < p_{y^0y^1} < 1$ and $\sum_{y^0y^1} p_{y^0y^1} = 1$. Regardless of whether (2), (3a) and (3b) hold or not, $\hat{\psi}_{c.clr}$ and $\hat{\psi}_{s.crude}$ are the nonparametric MLEs of $\psi_c^\dagger$ and $\psi_s^\dagger$, respectively. Thus, $T_{c,LR}^\dagger$ and T_c are asymptotically equal, and $T_{s,LR}^\dagger$ and T_s are asymptotically equal. It is easy to show that $\mathrm{H}_c^\dagger$ and $\mathrm{H}_s^\dagger$ are equivalent (i.e., $\mathrm{H}_c^\dagger$ holds if and only if $\mathrm{H}_s^\dagger$ holds), which implies that $T_{c,LR}^\dagger$ and $T_{s,LR}^\dagger$ are identical, which then in turn implies that T_c and T_s are asymptotically equal.

It is easy to show that H_0 and (3a) together imply $\mathrm{H}_s^\dagger$, and thus also $\mathrm{H}_c^\dagger$. Because $\hat{\psi}_{c.clr}$ converges to $\psi_c^\dagger$, it then follows that $\hat{\psi}_{c.clr}$ converges to 0 under H_0 and (3a).

APPENDIX C

Under (26), we have that

$$\begin{aligned} &\Pr(X_1, X_2, b) \\ &\quad = \frac{1}{\sqrt{2\pi}} e^{\{b-\theta\bar{X}\}^2/2} \Pr(X_1, X_2) \\ &\quad = h(X_1, b)h(X_2, b)e^{-\theta^2 X_1 X_2/4} \Pr(X_1, X_2), \end{aligned}$$

for some function $h(\cdot,\cdot)$. $X_1 \perp X_2|b$ now implies that

$$e^{-\theta^2 X_1 X_2/4}\Pr(X_1, X_2) = k(X_1)k(X_2)$$

for some function $k(\cdot)$, which in turn implies that $\theta = 2\sqrt{\log(\phi)}$.

ACKNOWLEDGMENT

Arvid Sjölander acknowledges financial support from The Swedish Research Council (2008-5375).

REFERENCES

BRESLOW, N. E. and DAY, N. E. (1980). Statistical methods in cancer research. Volume 1—The analysis of case control studies. IARC Scientific Publications No 32.

BRUMBACK, B. A., DAILEY, A. B., BRUMBACK, L. C., LIVINGSTON, M. D. and HE, Z. (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statist. Probab. Lett.* **80** 1650–1654. MR2684013

CARLIN, J. B., GURRIN, L. C., STERNE, J. A. C., MORLEY, R. and DWYER, T. (2005). Regression models for twin studies: A critical review. *International Journal of Epidemiology* **34**, 1089–1099.

CHEN, H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics* **63** 413–421. MR2370799

CUMMINGS, P., MCKNIGHT, B. and GREENLAND, S. (2003). Matched cohort methods for injury research. *Epidemiol Rev.* **25** 43–50.

FITZMAURICE, G. M., LAIRD, N. M. and WARE, J. H. (2004). *Applied Longitudinal Analysis*. Wiley, Hoboken, NJ. MR2063401

GREENLAND, S., ROBINS, J. M. and PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statist. Sci.* **14** 29–46.

HERNÁN, M. A. and ROBINS, J. M. (2006). Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **60** 578–586.

INGELSSON, E., LUNDHOLM, C., JOHANSSON, A. L. and ALTMAN, D. (2010). Hysterectomy and risk of cardiovascular disease: A population based cohort study. *European Heart Journal* doi:10.1093/eurheartj/ehq477.

JEWELL, N. P. (2004). *Statistics for Epidemiology*. Chapman & Hall/CRC Press, Boca Raton, FL.

NEUHAUS, J. M., KALBFLEISCH, J. D. and HAUCK, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int. Stat. Rev.* **59** 25–35.

NEUHAUS, J. M., KALBFLEISCH, J. D. and HAUCK, W. W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Canad. J. Statist.* **22** 139–148. MR1271451

NEUHAUS, J. M. and KALBFLEISCH, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54** 638–645.

NEUHAUS, J. M. and MCCULLOCH, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 859–872. MR2301298

ÖRTQVIST, A. K., LUNDHOLM, C., CARLSTRÖM, E., LICHTENSTEIN, P., CNATTINGIUS, S. and ALMQVIST, C. (2009). Familial factors do not confound the association between birth weight and childhood asthma. *Pediatrics* **124** e737–43.

ROTHMAN, K. J., GREENLAND, S. and LASH, T. L. (2008). *Modern Epidemiology*. Lippincott Williams and Wilkins, Philadelphia, PA.

WOODWARD, M. (2005). *Epidemiology*: *Study Design and Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2154361