Statistica Sinica **17**(2007), 1643-1656

COMPARISON BETWEEN ESTIMATES OF THE POTENTIAL PROPORTION WITH AND WITHOUT STANDARDIZATION FOR A NON-CONFOUNDER

Xueli Wang[1,2], Zhi Geng[2], Qiang Zhao[2,3] and Qi Qiao[2]

[1]*Beijing University of Posts and Telecommunications,* [2]*Peking University and* [3]*Shandong Normal University*

Abstract: A covariate is not a confounder if it is not a risk factor to disease, or if it has the same distribution in the exposed and unexposed populations. Standardization for a confounder can reduce confounding bias, but that for a non-confounder cannot. A question argued by many authors asks whether or not standardization of a non-confounder can improve the precision of estimation. This paper discusses the hypothetical or potential proportion of individuals in the exposed population who would have developed the disease had they not been exposed. It is shown that the precision of estimation of the hypothetical proportion cannot usually be improved by using standardization for a non-confounder, no matter how one re-categorizes the non-confounder.

Key words and phrases: Adjustment, causal inference, confounder, confounding, potential-outcome model, precision, standardization.

1. Introduction

Causal effect of exposure on the rate of a disease in the exposed population can be measured by comparing the proportion of diseased individuals in the exposed population with the hypothetical or potential proportion of diseased individuals in the exposed population without exposure, the so-called potential-outcomes model (Neyman (1923), Rubin (1974), Holland (1986), Wickramaratne and Holford (1987) and Greenland, Robins and Pearl (1999)). Following the notation of Holland (1989), let E be an exposure with values e and $\bar{e}$ representing presence and absence, respectively, let D denote an observed binary outcome with values 1 and 0 denoting presence and absence of a disease, respectively, and let D_e and $D_{\bar{e}}$ be the outcomes under $E = e$ and $E = \bar{e}$, respectively. For an individual, we can observe only one outcome of D_e and $D_{\bar{e}}$, but the other is unobservable, hypothetical or potential. For example, consider smoking as exposure and lung cancer as outcome. We can observe only the outcome D_e for a smoker, and only the outcome $D_{\bar{e}}$ for a nonsmoker. So the

model is called a potential-outcome model. The hypothetical or potential proportion $P(D_{\bar{e}} = 1|E = e)$ represents the proportion of individuals in the exposed population who would have developed the disease even if they had not been exposed. Epidemiological studies focus on the exposure effect on the rate of a disease in the exposed population, and the effect can be assessed by comparing $P(D_e = 1|E = e)$ with $P(D_{\bar{e}} = 1|E = e)$. For example, $P(D_{\bar{e}} = 1|E = e)$ represents the proportion of diseased individuals if any person in the smoking population had never smoked. Then the causal effect of smoking on lung cancer in the smoking population can be assessed by comparing $P(D_e = 1|E = e)$ with $P(D_{\bar{e}} = 1|E = e)$. When the hypothetical proportion is estimated by choosing an unexposed or control population, confounding bias may arise from differences in risk between the exposed and unexposed populations that would exist even if exposure were entirely absent from both populations. To eliminate confounding bias, the populations may be stratified into subpopulations by using covariates, called confounders, and then the proportion of diseased individuals in the unexposed population is standardized or adjusted for the covariates by taking the exposed population as the standard population. Two necessary criteria for assessing a confounder were proposed by Miettinen and Cook (1981): if a covariate is a confounder, then

(a) it must be predictive of risk in the unexposed population, and
(b) it must have different distributions between the exposed and unexposed populations.

Adjustment for a confounder can reduce confounding bias, but that for a non-confounder cannot (Wickramaratne and Holford (1987), Greenland, Robins and Pearl (1999), Geng, Guo, Lau and Fung (2001) and Geng, Guo and Fung (2002)). An important question, argued by many authors, is whether or not adjustment for a non-confounder can improve the precision of estimation. Mantel and Haenszel (1959), Mantel (1989) and Gail (1986) pointed out that adjusting for covariates related to disease can improve the precision of estimation for regression analysis even if they have the same distribution between the exposed and unexposed populations. In the response to Mantel (1989), Wickramaratne and Holford (1989) illustrated that adjusting for a covariate decreases the precision of estimates for a linear logistic model, using hypothetical data for which the covariate is related to the response, but nearly unrelated to exposure status. Breslow and Day (1980) also addressed how stratification by non-confounders can increase the variability of the estimates of relative risk without eliminating any bias.

In this paper we discuss the hypothetical or potential proportion of individuals in the exposed population who would have developed the disease had they not been exposed. We prove that the precision of estimation of the hypothetical

proportion cannot be improved by using standardization for a non-confounder, no matter how one re-categorizes the non-confounder.

In Section 2, we introduce the potential-outcome model, confounding bias and exposure effects. Section 3 defines the crude estimate and the standardized estimate of the hypothetical proportion. In Section 4, we show expectations and variances of these estimates and prove that standardization for non-confounders decreases the precision of estimation. We give a discussion in Section 5, and all proofs are given in Appendix.

2. Confounding Bias, Confounder and Standardization

Consider the proportions of diseased in the exposed and the unexposed populations. If the exposed population is comparable with the unexposed population, that is, $P(D_{\bar{e}} = 1|E = e) = P(D_{\bar{e}} = 1|E = \bar{e})$, nonconfounding bias, then the average causal effect can be estimated by using a prima facie causal effect such as $P(D_e = 1|E = e) - P(D_{\bar{e}} = 1|E = \bar{e})$, an estimable quantity, and on first view appears to be the average causal effect (Holland (1989)). For example, the causal effect of smoking on lung cancer could be assessed by comparing the proportions of lung cancer in the smoking and nonsmoking populations were there no confounding bias.

When there is confounding bias, we try to stratify the populations by some covariates, called confounders, and then standardize the proportion of diseased for these covariates. For example, age is usually a confounder in epidemiological studies, where age is a risk factor and has different distributions between the exposed and unexposed populations. Let C be a covariate with possible values $1, \ldots, K$. This C is not an intermediate variable in a causal pathway from exposure to disease. It may also be considered as a composite covariate consisting of several covariates. Let $P(D_e = 1|E = e, C = k)$ and $P(D_{\bar{e}} = 1|E = \bar{e}, C = k)$ be the proportions of diseased in the exposed and unexposed subpopulations of $C = k$, respectively. Similarly, $P(D_{\bar{e}} = 1|E = e, C = k)$ is the hypothetical proportion in the exposed subpopulation of $C = k$. According to the internal standardization in epidemiology (Miettinen (1972), Kleinbaum, Kupper and Morgenstern (1982) and Rothman and Greenland (1998)), the standardized proportion $P_\Delta(D_{\bar{e}} = 1|E = \bar{e})$ obtained by adjusting the distribution of C in the unexposed population to that in the exposed population is

$$P_\Delta(D_{\bar{e}} = 1|E = \bar{e}) = \sum_{k=1}^{K} P(D_{\bar{e}} = 1|E = \bar{e}, C = k)P(C = k|E = e). \quad (1)$$

Let $A \perp\!\!\!\perp B|C$ denote conditional independence between A and B given C (Dawid (1979)). If $P(D_{\bar{e}} = 1|E = e, C = k) = P(D_{\bar{e}} = 1|E = \bar{e}, C = k)$ for all

k (i.e., $D_{\bar{e}} \perp\!\!\!\perp E|C$), we say that there is no confounding in the subpopulations, termed subpopulation nonconfounding (Wickramaratne and Holford (1987)). In this case, it can be shown that the hypothetical proportion $P(D_{\bar{e}} = 1|E = e)$ equals the standardized proportion $P_\Delta(D_{\bar{e}} = 1|E = \bar{e})$.

Under the assumption of the subpopulation nonconfounding, Wickramaratne and Holford (1987) showed that a sufficient condition for nonconfounding is

$(\bar{a})$ $D_{\bar{e}} \perp\!\!\!\perp C|E = \bar{e}$ or
$(\bar{b})$ $C \perp\!\!\!\perp E$.

If C satisfies the condition $(\bar{a})$ or $(\bar{b})$, then we have from (1) that $P_\Delta(D_{\bar{e}} = 1|E = \bar{e}) = P(D_{\bar{e}} = 1|E = \bar{e})$. This implies that confounding bias cannot be reduced by standardization for a factor C that satisfies $(\bar{a})$ or $(\bar{b})$: the confounding bias $P(D_{\bar{e}} = 1|E = e) - P_\Delta(D_{\bar{e}} = 1|E = \bar{e})$ obtained by adjusting for C equals the confounding bias $P(D_{\bar{e}} = 1|E = e) - P(D_{\bar{e}} = 1|E = \bar{e})$ without the adjustment. Note that conditions $(\bar{a})$ and $(\bar{b})$ are just the converse of Miettinen and Cook's criteria (a) and (b), respectively, and thus (a) and (b) can be used as necessary conditions for a confounder.

3. Estimates of Hypothetical Proportion

We have seen that standardization of a non-confounder cannot reduce confounding bias, also see Wickramaratne and Holford (1987), Greenland, Robins and Pearl (1999), Geng et al. (2001) and Geng, Guo and Fung (2002). In this section, we discuss whether standardization of a non-confounder can improve the precision of estimation of the hypothetical proportion. Let n_{ijk} denote the observed frequency for $D = i$, $E = j$ and $C = k$, and let n_{+jk} and n_{+j+} denote the marginal frequencies obtained by summing over the index corresponding to '+'. Assume that n_{ijk} for all i, j and k follow a multinomial distribution with parameters $P(D = i, E = j, C = k)$. In epidemiological studies, such as follow-up studies, sample sizes of exposed and unexposed individuals, n_{+e+} and $n_{+\bar{e}+}$, are fixed by design. Thus we assume $n_{+e+} \geq 1$ and $n_{+\bar{e}+} \geq 1$ are fixed by design. Given marginal frequencies $n_{+\bar{e}+}$ and n_{+e+}, then $\{n_{i\bar{e}k}$ for all i and $k\}$ and $\{n_{iek}$ for all i and $k\}$ are independent and have multinomial distributions with parameters $\{P(D = i, C = k|E = \bar{e})$ for all i and $k\}$ and $\{P(D = i, C = k|E = e)$ for all i and $k\}$, respectively. For simplicity, define $p_{jk} = P(D = 1|E = j, C = k)$, $q_{k|j} = P(C = k|E = j)$ and $r_j = P(D = 1|E = j)$ for $j = e$ and $\bar{e}$. Let the parameter of interes θ be the hypothetical proportion of diseased in the exposed population, $P(D_{\bar{e}} = 1|E = e)$.

Let $\Omega = \{\omega_1, \ldots, \omega_s\}$ for $s \geq 2$ be a partition of C's levels $\{1, \ldots, K\}$. Define $n_{ij\omega} = \sum_{k\in\omega} n_{ijk}$, $p_{\bar{e}\omega} = P(D = 1|E = \bar{e}, C \in \omega)$ and $q_{\omega|\bar{e}} = \sum_{k\in\omega} q_{k|\bar{e}}$. The

standardized estimate $\hat{\theta}_\Omega$ based on the stratification Ω is

$$\hat{\theta}_\Omega = \sum_{\omega \in \Omega} \hat{p}_{\bar{e}\omega} \hat{q}_{\omega|e},$$

where $\hat{p}_{\bar{e}\omega} = n_{1\bar{e}\omega}/n_{+\bar{e}\omega}$ and $\hat{q}_{\omega|e} = n_{+e\omega}/n_{+e+}$. In particular, for $\Omega = \{[1, \ldots, K]\}$, we pool all levels of C together and obtain the crude or marginal estimate of the hypothetical proportion θ, $\tilde{\theta} = n_{1\bar{e}+}/n_{+\bar{e}+}$; for $\Omega = \{[1], \ldots, [K]\}$, we obtain the standardized estimate of the hypothetical proportion θ, $\hat{\theta} = \sum_k \hat{p}_{\bar{e}k} \hat{q}_{k|e}$, where $\hat{p}_{\bar{e}k} = n_{1\bar{e}k}/n_{+\bar{e}k}$ and $\hat{q}_{k|e} = n_{+ek}/n_{+e+}$. Since $n_{+\bar{e}k}$ appears in the denominator, we define levels of C such that $n_{+\bar{e}k} \geq 1$ for all k.

Let Ω_1 and Ω_2 denote two stratifications. We say that stratification Ω_1 is cruder than stratification Ω_2, denoted as $\Omega_1 \succeq \Omega_2$, if for any $\omega_2 \in \Omega_2$, there exists an $\omega_1 \in \Omega_1$ such that $\omega_1 \supseteq \omega_2$. When C is a composite factor with several covariates, a stratification defined by a covariate set A is cruder than that by a covariate set B if A is a subset of B. For example, consider the covariates sex and age (e.g., grouped by every 10 years) for the example of lung cancer and smoking. Let Ω_1, Ω_2 and Ω_3 be stratifications defined by $B = \{sex, age\}$, $A = \{age\}$, and by every 20 years, respectively. Then Ω_1 is the finest stratification and Ω_3 is the crudest. $\hat{\theta}_{\Omega_1}$ is the standardized estimate obtained by adjusting for both sex and age, $\hat{\theta}_{\Omega_2}$ is one obtained by adjusting for age groups of every 10 years, and $\hat{\theta}_{\Omega_3}$ is one obtained by adjusting for age groups of every 20 years.

4. Expectation and Variances of Estimates

Under the assumption that n_{+e+} and $n_{+\bar{e}+}$ are fixed and $n_{+\bar{e}k} \geq 1$ for any k, we show that if C satisfies the condition $(\bar{a})$ or $(\bar{b})$, the standardization for C cannot reduce the bias of estimation, and it cannot usually improve the precision of estimation, no matter how to one recategorizes C.

Theorem 1. *If a factor C satisfies one of conditions $(\bar{a})$ and $(\bar{b})$, then the standardized estimate of the hypothetical proportion based on any stratification has the same expectation as the crude estimate, that is, $E(\hat{\theta}_\Omega) = E(\tilde{\theta})$ for all possible stratifications Ω. Under the assumption of subpopulation nonconfounding, the standardized estimate $\hat{\theta}_\Omega$ and the crude estimate $\tilde{\theta}$ are unbiased.*

Theorem 2. *If the condition $(\bar{a})$ holds, then* $\mathrm{Var}\,(\hat{\theta}_{\Omega_1}) \leq \mathrm{Var}\,(\hat{\theta}_{\Omega_2})$ *for any $\Omega_1 \succeq \Omega_2$.*

Suppose that $D_{\bar{e}} \perp\!\!\!\perp C | E = \bar{e}$ and C is a composite factor consisting of several covariates $C_1, \ldots, C_m$. Note that $D_{\bar{e}} \perp\!\!\!\perp C | E = \bar{e}$ implies $D_{\bar{e}} \perp\!\!\!\perp C_i | E = \bar{e}$ for each i, but the converse is not true. It can be seen from Theorem 2 that the precision

of standardized estimates can be improved by omitting any non-confounder C_i in C.

Theorem 3. *If condition* $(\bar{b})$ *holds (i.e.,* $C \perp\!\!\!\perp E$*), then the crude estimate* $\tilde{\theta}$ *has a smaller variance than the standardized estimate* $\hat{\theta}$ *if* $n_{+\bar{e}+} \geq n_{+e+}$.

The condition $n_{+\bar{e}+} \geq n_{+e+}$ in Theorem 3 is sensible. To show this, we give some examples in Table 1, for each of which we have $C \perp\!\!\!\perp E$, $n_{+\bar{e}+} < n_{+e+}$ and $K = 2$, but $\mathrm{Var}\,(\tilde{\theta}) > \mathrm{Var}\,(\hat{\theta})$, even for quite large n_{+e+} and $n_{+\bar{e}+}$.

Table 1. Some examples for $K = 2$, $n_{+e+} > n_{+\bar{e}+}$, $C \perp\!\!\!\perp E$, but $\mathrm{Var}\,(\tilde{\theta}) > \mathrm{Var}\,(\hat{\theta})$.

n_{+e+}	$n_{+\bar{e}+}$	$q_{1\|e} = q_{1\|\bar{e}}$	$q_{2\|e} = q_{2\|\bar{e}}$	$p_{\bar{e}1}$	$p_{\bar{e}2}$	$\mathrm{Var}\,(\hat{\theta})$	$\mathrm{Var}\,(\tilde{\theta})$
20	8	0.2	0.8	0.4	0.06	1.3893×10^{-2}	1.3952×10^{-2}
30	10	0.1	0.9	0.07	0.01	1.5626×10^{-3}	1.5744×10^{-3}
80	70	0.6	0.4	0.4	0.05	2.7440×10^{-3}	2.7486×10^{-3}
150	80	0.32	0.68	0.01	0.09	7.5297×10^{-4}	7.5316×10^{-4}
180	150	0.8	0.2	0.25	0.05	1.1051×10^{-3}	1.1060×10^{-3}
1,000	700	0.6	0.4	0.04	0.09	8.0538×10^{-5}	8.0571×10^{-5}
2,000	1,700	0.4	0.6	0.09	0.04	3.3165×10^{-5}	3.3176×10^{-5}

Theorem 3 implies that when the frequency of unexposed individuals is larger than that of exposed individuals, pooling all levels together improves (at least does not reduce) the precision of estimation if C satisfies condition $(\bar{b})$. Further more, the crudest estimate $\tilde{\theta}$ has the smallest variance among all standardized estimates $\hat{\theta}_\Omega$ since $C \perp\!\!\!\perp E$ still holds after pooling levels of C. Unlike Theorem 2, however, Theorem 3 cannot ensure that a cruder stratification has a smaller variance than a finer stratification, that is, it cannot ensure that $\mathrm{Var}\,(\hat{\theta}_{\Omega_1}) \leq \mathrm{Var}\,(\hat{\theta}_{\Omega_2})$ for any $\Omega_1 \succeq \Omega_2$. Since $C \perp\!\!\!\perp E$ still holds after pooling some levels of C together, the following result can be obtained immediately from Theorem 3.

Corollary 1. *Suppose that* $n_{+\bar{e}\omega} \geq n_{+e\omega}$ *for all* $\omega \in \Omega_1$. *If* $(\bar{b})$ *holds, then* $\mathrm{Var}\,(\hat{\theta}_{\Omega_1}) \leq \mathrm{Var}\,(\hat{\theta}_{\Omega_2})$ *for any* $\Omega_1 \succeq \Omega_2$.

The relative precision (RP) of the crude estimate $\tilde{\theta}$ to the standardized estimate $\hat{\theta}$ is defined as

$$RP(\tilde{\theta} \text{ to } \hat{\theta}) = \frac{[\mathrm{Var}\,(\tilde{\theta})]^{-1}}{[\mathrm{Var}\,(\hat{\theta})]^{-1}} = \frac{\mathrm{Var}\,(\hat{\theta})}{\mathrm{Var}\,(\tilde{\theta})}.$$

If the case in which $n_{+\bar{e}k}$ is zero is ignored, then we have from Stephan (1945) that to terms of order $n_{+\bar{e}+}^{-1}$,

$$E(\frac{1}{n_{+\bar{e}k}}) \approx \frac{1}{n_{+\bar{e}+}q_{k|\bar{e}}},$$

see also Cochran (1977, p.135). Substituting this into the equation (A.4) in the proof of Theorem 2, we can obtain the following result from $C \perp\!\!\!\perp E$.

Corollary 2. *If both* $D_{\bar{e}} \perp\!\!\!\perp C|E = \bar{e}$ *and* $C \perp\!\!\!\perp E$ *hold, then the relative precision* $RP(\tilde{\theta}$ *to* $\hat{\theta})$ *is approximately* $1 + (K-1)/n_{+e+}$.

If both $(\bar{a})$ and $(\bar{b})$ hold, from the definition of RP and Corollary 2, we can obtain the more general RP of $\hat{\theta}_{\Omega_1}$ to $\hat{\theta}_{\Omega_2}$ as

$$RP(\hat{\theta}_{\Omega_1} \text{ to } \hat{\theta}_{\Omega_2}) \approx \frac{n_{+e+} + K_2 - 1}{n_{+e+} + K_1 - 1},$$

where K_i denotes the number of levels of Ω_i.

5. Discussion

In this paper, we showed that standardization for non-confounders never reduces confounding bias, and it cannot usually improve the precision of estimation of the hypothetical proportion. These results are useful for design of epidemiological studies and data analysis. For a randomized design, the condition $(\bar{b})$ is satisfied, and thus standardization of the hypothetical proportion for covariates is unnecessary to reduce confounding bias and to improve the precision of estimate provided that the frequency of exposed individuals is not larger than that of unexposed individuals. In design of an observational study, a covariate C may be omitted without inducing bias or loss of precision if there is evidence from other studies which supports condition $(\bar{a})$ or $\{(\bar{b})$ and $n_{+\bar{e}+} \geq n_{+e+}\}$. In data analysis, we may omit C for simplification if there is evidence for condition $(\bar{a})$ or $\{(\bar{b})$ and $n_{+\bar{e}+} \geq n_{+e+}\}$ from observed data.

The studies considered in this paper are those with the numbers of exposed and unexposed individuals fixed. For the case-control studies in which the numbers of diseased and non-diseased individuals are fixed, it is impossible to use randomized treatment assignment, and thus subpopulation nonconfounding is dubious in most cases, see Holland and Rubin (1988). On the other hand, there is no information on the proportions $P(D_e = 1|E = e)$ and $P(D_{\bar{e}} = 1|E = \bar{e})$, and no estimate of the hypothetical proportion $P(D_{\bar{e}} = 1|E = e)$ in case-control studies.

We have only discussed standardized estimates of the hypothetical proportion with adjustment for discrete covariates. Robinson and Jewell (1991) discussed adjustment for continuous covariates in logistic regression models, and they showed asymptotically that adjustment for a continuous covariate C will lose the precision of estimates of parameters in logistic regression models if (i) $D \perp\!\!\!\perp C|E$ or (ii) $C \perp\!\!\!\perp E|D$. Note that condition (i) implies $(\bar{a})$, that condition (ii) is different to $(\bar{b})$, and that our results are exact but not asymptotic. Comparison between estimates of the risk ratio remains to be discussed.

Acknowledgements

We would like to thank the Co-Editor, an associate editor and two referees for their valuable comments and suggestions. This research was supported by NSFC, NBRP 2005CB523301 and MSRA.

Appendix

Proofs of Theorems and Corollary 1.

We first give the following lemmas which will be used in proofs of theorems.

Lemma 1. *When a set ω of C's levels is partitioned into subsets ω' and ω'',*

$$\frac{(n_{+e+}-1)q_{\omega'|e}^2+q_{\omega'|e}}{n_{+\bar{e}\omega'}}+\frac{(n_{+e+}-1)q_{\omega''|e}^2+q_{\omega''|e}}{n_{+\bar{e}\omega''}} \geq \frac{(n_{+e+}-1)(q_{\omega'|e}+q_{\omega''|e})^2+(q_{\omega'|e}+q_{\omega''|e})}{n_{+\bar{e}\omega'}+n_{+\bar{e}\omega''}}.$$

Proof. Moving the right hand side of the inequality to the left, we can get

$$\begin{aligned}&\frac{n_{+\bar{e}\omega''}(n_{+\bar{e}\omega'}+n_{+\bar{e}\omega''})[(n_{+e+}-1)q_{\omega'|e}^2+q_{\omega'|e}]}{n_{+\bar{e}\omega'}n_{+\bar{e}\omega''}(n_{+\bar{e}\omega'}+n_{+\bar{e}\omega''})}\\ &\quad+\frac{n_{+\bar{e}\omega'}(n_{+\bar{e}\omega'}+n_{+\bar{e}\omega''})[(n_{+e+}-1)q_{\omega''|e}^2+q_{\omega''|e}]}{n_{+\bar{e}\omega'}n_{+\bar{e}\omega''}(n_{+\bar{e}\omega'}+n_{+\bar{e}\omega''})}\\ &\quad-\frac{n_{+\bar{e}\omega'}n_{+\bar{e}\omega''}[(n_{+e+}-1)(q_{\omega'|e}+q_{\omega''|e})^2+(q_{\omega'|e}+q_{\omega''|e})]}{n_{+\bar{e}\omega'}n_{+\bar{e}\omega''}(n_{+\bar{e}\omega'}+n_{+\bar{e}\omega''})}.\end{aligned}$$

For the formula above, the denominators are the same, and the numerators are rewritten as

$$\begin{aligned}&n_{+\bar{e}\omega''}(n_{+\bar{e}\omega'}+n_{+\bar{e}\omega''})[(n_{+e+}-1)q_{\omega'|e}^2+q_{\omega'|e}]\\ &\qquad+n_{+\bar{e}\omega'}(n_{+\bar{e}\omega'}+n_{+\bar{e}\omega''})[(n_{+e+}-1)q_{\omega''|e}^2+q_{\omega''|e}]\\ &\qquad-n_{+\bar{e}\omega'}n_{+\bar{e}\omega''}[(n_{+e+}-1)(q_{\omega'|e}+q_{\omega''|e})^2+(q_{\omega'|e}+q_{\omega''|e})]\\ &=(n_{+e+}-1)n_{+\bar{e}\omega''}^2q_{\omega'|e}^2+(n_{+e+}-1)n_{+\bar{e}\omega'}^2q_{\omega''|e}^2+n_{+\bar{e}\omega''}^2q_{\omega'|e}+n_{+\bar{e}\omega'}^2q_{\omega''|e}\\ &\quad-2(n_{+e+}-1)n_{+\bar{e}\omega'}n_{+\bar{e}\omega''}q_{\omega'|e}q_{\omega''|e}\\ &\geq 2(n_{+e+}-1)n_{+\bar{e}\omega'}n_{+\bar{e}\omega''}q_{\omega'|e}q_{\omega''|e}+n_{+\bar{e}\omega''}^2q_{\omega'|e}+n_{+\bar{e}\omega'}^2q_{\omega''|e}\\ &\qquad-2(n_{+e+}-1)n_{+\bar{e}\omega'}n_{+\bar{e}\omega''}q_{\omega'|e}q_{\omega''|e}\\ &=n_{+\bar{e}\omega''}^2q_{\omega'|e}+n_{+\bar{e}\omega'}^2q_{\omega''|e}\geq 0.\end{aligned}$$

The lemma follows.

Lemma 2. *If $a_1 \geq \cdots \geq a_n$, $b_1 \leq \cdots \leq b_n$, and $p_1 + \cdots + p_n = 1$, where $p_i > 0$ for $i = 1, 2, \ldots, n$, then we have*

$$\Big(\sum_{i=1}^{n} p_i a_i\Big)\Big(\sum_{i=1}^{n} p_i b_i\Big) \geq \sum_{i=1}^{n} p_i a_i b_i.$$

Proof. Consider the p's as a probability measure on $\{1, \ldots, n\}$ and let random variables A and B take measure $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$.

Clearly $\mathrm{Cov}\,(A, B) \leq 0$, and the Lemma follows.

Lemma 3. *Suppose X has a binomial distribution with parameters $n > 0$ and $0 < p < 1$. Then*

$$E\Big(\frac{1}{X}\Big|0 \leq X \leq m\Big) \geq \frac{1}{np + 1 - p}. \tag{A.1}$$

Proof. First we prove (A.1) for $m = n$. Let X' be binomial variable with parameters $n - 1$ and p. From Lemma 2,

$$E(X' + 1)E\Big[\frac{1}{(X' + 1)^2}\Big] \geq E\Big(\frac{1}{X' + 1}\Big).$$

Dividing by $E(X' + 1) = (n - 1)p + 1$, the above inequality can be expressed as

$$\sum_{k=0}^{n-1} \frac{1}{(k+1)^2}\binom{n-1}{k} p^k (1-p)^{n-1-k} \geq \frac{\sum_{k=0}^{n-1} \frac{1}{k+1}\binom{n-1}{k} p^k (1-p)^{n-1-k}}{np + 1 - p}.$$

After the above summation over k from 0 to $n - 1$ is changed to that from 1 to n, we have

$$\sum_{k=1}^{n} \frac{1}{k^2}\binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \geq \frac{\sum_{k=1}^{n} \frac{1}{k}\binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}}{np + 1 - p}.$$

Multiplying both sides by np and noting that $\sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1 - (1-p)^n$, we get

$$\frac{\sum_{k=1}^{n} \frac{1}{k}\binom{n}{k} p^k (1-p)^{n-k}}{1 - (1-p)^n} \geq \frac{1}{np + 1 - p}.$$

Thus we have proved (A.1) when $m = n$.

Next, let X_m and X_n denote $[X|0 < X \leq m]$ and $[X|0 \leq X \leq n]$, respectively, where $m < n$. From (A.1), we need only show $E(1/X_m) \geq E(1/X_n)$. Since $P(X_m = k) = P(X_n = k)/P(X_n \leq m)$ for $0 < k \leq m$, we have

$$E\left(\frac{1}{X_m}\right) - E\left(\frac{1}{X_n}\right) = \sum_{k=1}^{m} \frac{1}{k} P(X_m = k) - \sum_{k=1}^{n} \frac{1}{k} P(X_n = k)$$

$$
\begin{aligned}
&= \sum_{k=1}^{m} \frac{1}{k} \frac{P(X_n = k)}{P(X_n \le m)} - \sum_{k=1}^{m} \frac{1}{k} P(X_n = k) - \sum_{k=m+1}^{n} \frac{1}{k} P(X_n = k) \\
&= \sum_{k=1}^{m} \frac{P(X_n = k)}{k} \frac{P(X_n > m)}{P(X_n \le m)} - \sum_{k=m+1}^{n} \frac{1}{k} P(X_n = k) \\
&\ge \sum_{k=1}^{m} \frac{P(X_n = k)}{m} \frac{P(X_n > m)}{P(X_n \le m)} - \sum_{k=m+1}^{n} \frac{1}{m} P(X_n = k) = 0.
\end{aligned}
$$

Thus we have proved (A.1).

Proof of Theorem 1. Given n_{+e+} and $n_{+\bar{e}+}$, $\hat{q}_{\omega|e}$ and $\hat{p}_{\bar{e}\omega}$ are conditionally independent. Thus we get that for any Ω,

$$
E(\hat{\theta}_\Omega) = \sum_{\omega \in \Omega} E(\hat{q}_{\omega|e} \hat{p}_{\bar{e}\omega}) = \sum_{\omega \in \Omega} E(\hat{q}_{\omega|e}) E(\hat{p}_{\bar{e}\omega}).
$$

For the first factor, it is obvious that $E(\hat{q}_{\omega|e}) = q_{\omega|e}$. For the second factor, we have

$$
\begin{aligned}
E(\hat{p}_{\bar{e}\omega}) &= E\left(\frac{n_{1\bar{e}\omega}}{n_{+\bar{e}\omega}}\right) = E\left[E\left(\frac{n_{1\bar{e}\omega}}{n_{+\bar{e}\omega}} | n_{+\bar{e}\omega}\right)\right] \\
&= E\left[\frac{E(n_{1\bar{e}\omega} | n_{+\bar{e}\omega})}{n_{+\bar{e}\omega}}\right] = E\left(\frac{n_{+\bar{e}\omega} p_{\bar{e}\omega}}{n_{+\bar{e}\omega}}\right) = E(p_{\bar{e}\omega}) = p_{\bar{e}\omega}.
\end{aligned}
$$

If the condition $(\bar{a})$ $D_{\bar{e}} \perp\!\!\!\perp C | E = \bar{e}$ holds, we have $E(\hat{\theta}_\Omega) = \sum_{\omega \in \Omega} q_{\omega|e} p_{\bar{e}\omega} = \sum_\omega q_{\omega|e} r_{\bar{e}} = r_{\bar{e}}$. If the condition $(\bar{b})$ $C \perp\!\!\!\perp E$ holds, we have $E(\hat{\theta}_\Omega) = \sum_\omega q_{\omega|e} p_{\bar{e}\omega} = \sum_\omega q_{\omega|\bar{e}} p_{\bar{e}\omega} = r_{\bar{e}}$. Thus in both cases, $E(\hat{\theta}_\Omega) = E(\tilde{\theta}) = r_{\bar{e}}$, where $\tilde{\theta}$ is a special $\tilde{\theta}_\Omega$ for $\Omega = \{[1, \ldots, K]\}$. Further, under the assumption of subpopulation nonconfounding, we have that $r_{\bar{e}} = \sum_k q_{k|e} p_{\bar{e}k} = \sum_k q_{k|e} P(D_{\bar{e}} = 1 | E = e, C = k) = \theta$.

Proof of Theorem 2. Because $D_{\bar{e}} \perp\!\!\!\perp C | E = \bar{e}$, we have that $p_{\bar{e}\omega} = r_{\bar{e}}$ and we write $p = p_{\bar{e}\omega} = r_{\bar{e}}$. Also for simplicity, we take $X = (n_{+\bar{e}\omega_k}, n_{+e\omega_k}, k = 1, \ldots, s)$, and $q_k = q_{\omega_k|e}, k = 1, \ldots, s$. Thus we obtain

$$
\begin{aligned}
\mathrm{Var}\,(\hat{\theta}_\Omega) &= \mathrm{Var}\Big(\sum_{k=1}^{s} \frac{n_{1\bar{e}\omega_k}}{n_{+\bar{e}\omega_k}} \frac{n_{+e\omega_k}}{n_{+e+}}\Big) \\
&= \mathrm{Var}\left[E\Big(\sum_{k=1}^{s} \frac{n_{1\bar{e}\omega_k}}{n_{+\bar{e}\omega_k}} \frac{n_{+e\omega_k}}{n_{+e+}} | X\Big)\right] + E\left[\mathrm{Var}\Big(\sum_{k=1}^{s} \frac{n_{1\bar{e}\omega_k}}{n_{+\bar{e}\omega_k}} \frac{n_{+e\omega_k}}{n_{+e+}} | X\Big)\right] \\
&= \mathrm{Var}\Big(\sum_{k=1}^{s} \frac{n_{+e\omega_k}}{n_{+e+}} p\Big) + E\left[\sum_{k=1}^{s} \Big(\frac{n_{+e\omega_k}}{n_{+e+}}\Big)^2 \frac{p(1-p)}{n_{+\bar{e}\omega_k}}\right]
\end{aligned}
$$

$$= \mathrm{Var}\,(p) + \sum_{k=1}^{s}\left[\frac{n_{+e+}q_k^2 + q_k(1-q_k)}{n_{+e+}}p(1-p)E\Big(\frac{1}{n_{+\bar{e}\omega_k}}\Big)\right]$$
$$= 0 + \frac{p(1-p)}{n_{+e+}}\sum_{k=1}^{s}\left[((n_{+e+}-1)q_k^2 + q_k)E\Big(\frac{1}{n_{+\bar{e}\omega_k}}\Big)\right]. \tag{A.2}$$

To prove Theorem 2, we need only show that if a set ω_k is further partitioned into two sets ω' and ω'', and thus $q_{\omega_k|j} = q_{\omega'|j} + q_{\omega''|j}$ for $j = e$ and $\bar{e}$, then we have

$$[(n_{+e+}-1)q_{\omega'|e}^2 + q_{\omega'|e}]E\left(\frac{1}{n_{+\bar{e}\omega'}}\right) + [(n_{+e+}-1)q_{\omega''|e}^2 + q_{\omega''|e}]E\left(\frac{1}{n_{+\bar{e}\omega''}}\right)$$
$$\geq [(n_{+e+}-1)(q_{\omega'|e} + q_{\omega''|e})^2 + (q_{\omega'|e} + q_{\omega''|e})]E\left(\frac{1}{n_{+\bar{e}\omega_k}}\right).$$

From Lemma 1, we get

$$\frac{(n_{+e+}-1)q_{\omega'|e}^2 + q_{\omega'|e}}{n_{+\bar{e}\omega'}} + \frac{(n_{+e+}-1)q_{\omega''|e}^2 + q_{\omega''|e}}{n_{+\bar{e}\omega''}}$$
$$\geq \frac{(n_{+e+}-1)(q_{\omega'|e} + q_{\omega''|e})^2 + (q_{\omega'|e} + q_{\omega''|e})}{n_{+\bar{e}\omega'} + n_{+\bar{e}\omega''}},$$

and thus we have proved Theorem 2.

Proof of Theorem 3. For simplicity, take $p_k = p_{\bar{e}k}$ for all k, and $X = (n_{+\bar{e}k}, n_{+ek}, k = 1, \cdots, K)$. By $C \perp\!\!\!\perp E$, we can write $q_k = q_{k|\bar{e}} = q_{k|e}$ for all k. Thus we have

$$\mathrm{Var}\,(\hat{\theta}) = \mathrm{Var}\Big(\sum_{k=1}^{K}\frac{n_{1\bar{e}k}}{n_{+\bar{e}k}}\frac{n_{+ek}}{n_{+e+}}\Big)$$
$$= \mathrm{Var}\left[E\Big(\sum_{k=1}^{K}\frac{n_{1\bar{e}k}}{n_{+\bar{e}k}}\frac{n_{+ek}}{n_{+e+}}|X\Big)\right] + E\left[\mathrm{Var}\Big(\sum_{k=1}^{K}\frac{n_{1\bar{e}k}}{n_{+\bar{e}k}}\frac{n_{+ek}}{n_{+e+}}|X\Big)\right]$$
$$= \mathrm{Var}\Big(\sum_{k=1}^{K}\frac{n_{+ek}}{n_{+e+}}p_k\Big) + \sum_{k=1}^{K}\left[\frac{n_{+e+}q_k^2 + q_k(1-q_k)}{n_{+e+}}p_k(1-p_k)E\Big(\frac{1}{n_{+\bar{e}k}}\Big)\right].$$

The first term can be expressed as

$$\mathrm{Var}\Big(\sum_{k=1}^{K}\frac{n_{+ek}}{n_{+e+}}p_k\Big) = \sum_{k=1}^{K}\mathrm{Var}\Big(\frac{n_{+ek}}{n_{+e+}}p_k\Big) + \sum_{k\neq l}\mathrm{Cov}\Big(\frac{n_{+ek}}{n_{+e+}}p_k, \frac{n_{+el}}{n_{+e+}}p_l\Big)$$
$$= \frac{1}{n_{+e+}}\left[\sum_{k=1}^{K}p_k^2q_k(1-q_k) - \sum_{k\neq l}q_kq_lp_kp_l\right] \geq 0.$$

For $\tilde{\theta}$, we have

$$\begin{aligned}\mathrm{Var}\,(\tilde{\theta}) &= \frac{r_{\bar{e}}(1-r_{\bar{e}})}{n_{+\bar{e}+}} = \frac{\sum_{k=1}^{K} p_k q_k (1-\sum_{k=1}^{K} p_k q_k)}{n_{+\bar{e}+}} \\ &= \frac{1}{n_{+\bar{e}+}}\left[\sum_{k=1}^{K} p_k^2 q_k(1-q_k) - \sum_{k\neq l} q_k q_l p_k p_l\right] + \frac{1}{n_{+\bar{e}+}}\sum_{k=1}^{K} q_k p_k (1-p_k).\end{aligned}$$

Comparing the above equations of $\mathrm{Var}\,(\hat{\theta})$ and $\mathrm{Var}\,(\tilde{\theta})$, the first item of $\mathrm{Var}\,(\hat{\theta})$ is larger than the first item of $\mathrm{Var}\,(\tilde{\theta})$ for $n_{+\bar{e}+} \geq n_{+e+}$. Thus we need only show that for all k,

$$\frac{n_{+e+}q_k^2 + q_k(1-q_k)}{n_{+e+}} p_k(1-p_k) E\left(\frac{1}{n_{+\bar{e}k}}\right) \geq \frac{q_k p_k (1-p_k)}{n_{+\bar{e}+}}.$$

Dividing both sides by $q_k p_k (1-p_k)$, this amounts to

$$\frac{n_{+e+}q_k + 1 - q_k}{n_{+e+}} E\left(\frac{1}{n_{+\bar{e}k}}\right) \geq \frac{1}{n_{+\bar{e}+}}.$$

From (A.1) and $n_{+\bar{e}+} \geq n_{+e+}$, we have

$$\begin{aligned}\frac{n_{+e+}q_k + 1 - q_k}{n_{+e+}} E\left(\frac{1}{n_{+\bar{e}k}}\right) &\geq \frac{n_{+\bar{e}+}q_k + 1 - q_k}{n_{+\bar{e}+}} E\left(\frac{1}{n_{+\bar{e}k}}\right) \\ &\geq \frac{n_{+\bar{e}+}q_k + 1 - q_k}{n_{+\bar{e}+}} \frac{1}{n_{+\bar{e}+}q_k + 1 - q_k} = \frac{1}{n_{+\bar{e}+}}.\end{aligned}$$

Thus, we proved that $\mathrm{Var}\,(\hat{\theta}) \geq \mathrm{Var}\,(\tilde{\theta})$ when $n_{+\bar{e}+} \geq n_{+e+}$.

Proof of Corollary 1. Since $\Omega_1 \succeq \Omega_2$, for any $\omega_k \in \Omega_1$ there exist $\omega_{k1}, \ldots, \omega_{kn_k} \in \Omega_2$ such that $\omega_k = \cup_{j=1}^{n_k} \omega_{kj}$. We write

$$\begin{aligned}\hat{\theta}_{\Omega_1} &= \sum_{k=1}^{K} \frac{n_{1\bar{e}\omega_k}}{n_{+\bar{e}\omega_k}} \frac{n_{+e\omega_k}}{n_{+e+}}, \\ \hat{\theta}_{\Omega_2} &= \sum_{k=1}^{K}\sum_{j=1}^{n_k} \frac{n_{1\bar{e}\omega_{kj}}}{n_{+\bar{e}\omega_{kj}}} \frac{n_{+e\omega_{kj}}}{n_{+e+}} = \sum_{k=1}^{K} \frac{n_{+e\omega_k}}{n_{+e+}} \sum_{j=1}^{n_k} \frac{n_{1\bar{e}\omega_{kj}}}{n_{+\bar{e}\omega_{kj}}} \frac{n_{+e\omega_{kj}}}{n_{+e\omega_k}}.\end{aligned}$$

According to Theorem 3, we have

$$\begin{aligned}\mathrm{Var}\,(\hat{\theta}_{\Omega_1}) &= \mathrm{Var}\,[E(\hat{\theta}_{\Omega_1} \mid X)] + E[\mathrm{Var}\,(\hat{\theta}_{\Omega_1} \mid X)] \\ &= \mathrm{Var}\left(\sum_{k=1}^{K} \frac{n_{+e\omega_k}}{n_{+e+}} p_k\right) + E\left[\sum_{k=1}^{K} \left(\frac{n_{+e\omega_k}}{n_{+e+}}\right)^2 \mathrm{Var}\left(\frac{n_{1\bar{e}\omega_k}}{n_{+\bar{e}\omega_k}} \middle| X\right)\right]\end{aligned}$$

$$\leq \mathrm{Var}\Big(\sum_{k=1}^{K}\frac{n_{+e\omega_k}}{n_{+e+}}p_k\Big)+E\Big[\sum_{k=1}^{K}\Big(\frac{n_{+e\omega_k}}{n_{+e+}}\Big)^2\mathrm{Var}\Big(\sum_{j=1}^{n_k}\frac{n_{1\bar{e}\omega_{kj}}}{n_{+\bar{e}\omega_{kj}}}\frac{n_{+e\omega_{kj}}}{n_{+e\omega_k}}\Big|X\Big)\Big]$$
$$= \mathrm{Var}\,[E(\hat{\theta}_{\Omega_2} \mid X)]+E[\mathrm{Var}\,(\hat{\theta}_{\Omega_2} \mid X)] = \mathrm{Var}\,(\hat{\theta}_{\Omega_2}),$$

where X and p_k have the same definitions as those in the proof of Theorem 3. Thus we have proved Corollary 1.

References

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research. Vol. I, The Analysis of Case-control Studies*. Lyon: IARC.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd edition. Wiley, New York.

Dawid, A. P. (1979). Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B* **41**, 1-31.

Gail, M. H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In *Modern Statistical Methods in Chronic Disease Epidemiology* (Edited by S. H. Moolgavkar and R. L. Prentice), 3-18. Wiley, New York.

Geng, Z., Guo, J. and Fung, W. K. (2002). Criteria for confounders in epidemiological studies. *J. Roy. Statist. Soc. Ser. B* **64**, 3-15.

Geng, Z., Guo, J. H., Lau, T. S. and Fung, W. K. (2001). Confounding, homogeneity and collapsibility for causal effects in epidemiologic studies. *Statist. Sinica* **11**, 63-75.

Greenland, S., Robins, J. M. and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statist. Sci.* **14**, 29-46.

Holland, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81**, 945-970.

Holland, P. W. (1989). Reader reactions: confounding in epidemiologic studies. *Biometrics* **45**, 1310-1316.

Holland, P. W. and Rubin, D. B. (1988). Causal inference in retrospective studies. *Evaluation Rev.* **12**, 203-231.

Kleinbaum, D. G., Kupper, L. L. and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Van Nostrand Reinhold, New York.

Mantel, N. (1989). Confouding in epidemiologic studies. *Biometrics* **45**, 1317-1318.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancel Inst.* **22**, 719-748.

Miettinen, O. S. (1972). Standardization of risk ratios. *Amer. J. Epidemiol.* **96**, 383-388.

Miettinen, O. S. and Cook, E. F. (1981). Confounding: Essence and detection. *Amer. J. Epidemiol.* **114**, 593-603.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section (In Polish), *Roczniki Nauk Roiniczych*, Tom X, 1-51, [English translation of excerpts by D. Dabrowska and T. Speed with Discussion in *Statist. Sci.* **5**, 463-480].

Robinson, L.D. and Jewell, N. P. (1991). Some surprising results about covariate adjusting in logistic regression models. *Internat. Statist. Rev.* **58**, 227-240.

Rothman, K. J. and Greenland, S. (1998). *Modern Epidemiology*. 2nd edition. Lippincott-Raven Publishers, Philadelphia.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educational Psychology* **66**, 688-701.

Stephan, F. F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoulli variate. *Ann. Math. Statist.* **16**, 50-61.

Wickramaratne, P. J. and Holford, T. R. (1987). Confounding in epidemiologic studies: the adequacy of the control groups as a measure of confounding. *Biometrics* **43**, 751-65.

Wickramaratne, P. J. and Holford, T. R. (1989). Confouding in epidemiologic studies. Response. *Biometrics* **45**, 1319-1322.

School of Sciences, Beijing University of Posts and Telecommunications, Beijing 100876, China.

E-mail: wangxl@math.pku.edu.cn

School of Mathematical Sciences, Peking University, Beijing 100871, China.

E-mail: zgeng@math.pku.edu.cn

School of Mathematical Science, Shandong Normal University, Jinan 250014, China.

E-mail: zhq@math.pku.edu.cn

School of Mathematical Sciences, Peking University, Beijing 100871, China.

E-mail: qiaoqibella@eyou.com

(Received September 2004; accepted February 2006)