Comput. Math. Applic. Vol. 14, No. 9–12, pp. 869–916, 1987
Printed in Great Britain.

0097-4943/87 $3.00 + 0.00

THE FOUNDATIONS OF CONFOUNDING IN EPIDEMIOLOGY

J. M. ROBINS
Occupational Health Program, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, U.S.A.

H. MORGENSTERN
Division of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90024, U.S.A.

Abstract—A statistically coherent view of confounding motivated by the controversy over the proper control of confounding in the presence of prior knowledge is presented. Confounding by a covariate C in the presence of data on C is distinguished from confounding in the absence of data on C. A covariate C is defined to be a nonconfounder in the absence of data on C if the population parameter of interest can be unbiasedly estimated (asymptotically) absent data on C. Under this definition, C may be a confounder for some parameters of interest and a nonconfounder for others. If C is a confounder for a parameter of interest that has a causal interpretation, we call C a causal confounder. When data on C are available, C is defined to be a nonconfounder for a particular parameter of interest if and only if inference on the parameter of interest does not depend on the data through C. Bayesians, frequentists and pure likelihoodists will in general agree on the prior knowledge necessary to render C a nonconfounder. In particular C will in general be a nonconfounder precisely when the crude data ignoring C are S-sufficient for the parameter of interest. The intuitive view held by many practicing epidemiologists that confounding by C represents a bias of the unadjusted crude estimator is in a sense correct provided inference is performed conditional on approximate ancillary statistics that measure the degree to which associations in the data differ due to sampling variability from those population associations known *a priori*.

1. INTRODUCTION

In epidemiologic studies of the effect of an exposure E on a disease D, the marginal association of E with D, often called by epidemiologists the crude association, may fail to reflect a causal association due to "confounding" by one or more covariates C. On occasion the investigator has prior knowledge concerning the associations of various covariates with either E or D in the source population. The proper control of confounding in the presence of prior knowledge has been a controversial subject in epidemiology; see, for example, Miettinen and Cook [1], Day *et al.* [2] and Breslow and Day [3]. Much of the reason for this lack of consensus is that confounding remains an ill-defined concept. For example, in the most comprehensive article on confounding to date, Miettinen and Cook [1] develop "first principles" of confounding based on inductive, intuitive considerations of various examples. In that paper, Miettinen and Cook state that "confounding implies a need and desire to replace the 'crude' estimate of the effect by one that has been adjusted for the covariate at issue". Yet they fail to give a general definition of an "appropriate adjusted estimate". We shall try to offer a formal foundation for their intuitions.

It is helpful to distinguish confounding in the absence of data on C from confounding given data on C in order to capture the way a number of epidemiologists use the word confounding. (Note that implicit in Miettinen and Cook's definition of confounding is the assumption that data on C have been obtained so that an adjusted estimate is available.) For example, in a randomized trial in which a risk factor C is not measured, a number of epidemiologists regard the crude estimator of effect as unconfounded. This reflects the fact that in a randomized study, C is as likely to be positively as negatively associated with exposure in the data at hand. If data on C have not been recorded for data analysis, we would feel neither the "need nor desire" to adjust the crude estimate, for in which direction would we adjust it? But if data on the risk factor C become available and, by chance, a large E–C association exists in the data, then the crude estimate of effect is now regarded as confounded, since we would desire to replace the crude estimate by an estimate adjusted for C [1]. Therefore, C was a nonconfounder absent data on C but a confounder given data on C.

C.A.M.W.A. 14/9-12—O

In Sections 2–4, we develop from first principles a theory of confounding in the absence of data on a covariate C. We define in Section 4 a covariate C to be a confounder for a particular parameter of interest in the absence of data on C if and only if the population parameter of interest *cannot* be unbiasedly estimated (asymptotically) in the absence of data on C. Under this definition, C may be a confounder under one sampling design and a nonconfounder under another. Furthermore, C may be a confounder for some parameters of interest and a nonconfounder for others. Miettinen and Cook's contrary claim that definitions of confounding should be independent of the parameter of interest reflects their decision to restrict attention to causal parameters that compare the observed outcomes of an exposed group to the outcomes that would have been observed in this group in the absence of exposure. If C is a confounder for a causal parameter, C will be called a causal confounder. From Section 4 onward, we suppose that the study population has been sampled from a large near-infinite superpopulation and that the causal parameter of interest is a parameter of the superpopulation. The motivation for this superpopulation model is presented in Section 3. In Section 4 we consider the prior knowledge necessary to render C a nonconfounder absent data on C for the study designs and effect parameters most commonly employed in epidemiologic research on the etiology of disease.

In Section 3, we show that under a deterministic outcome model, standard binomial intervals may fail to cover the causal parameter of interest at the nominal rate even in the absence of bias when, following Miettinen and Cook [1], we take the causal parameters of interest to be parameters associated with the observed study population (regardless of whether the observed study population was sampled from a large superpopulation). New interval estimators are proposed that, in this setting, improve upon the performance of standard "binomial intervals".

Informally in Section 5 and formally in Section 8, we present a consistent statistical interpretation of Miettinen and Cook's intuitive concept of confounding given data on C. We consider the implications for inference of discovering, when we have collected data on C, that associations of C with exposure and/or disease in the data differ from those associations known to hold in the population. Following Miettinen and Cook, we define C to be a *confounder in the data at hand* if the crude estimate of the parameter of interest differs from an appropriate adjusted estimate. We demonstrate that the requirement that Miettinen and Cook's appropriate adjusted estimators be asymptotically efficient is necessary in general to reach their intuitive, inductive decisions. Therefore, from their intuitive point of view, inefficient estimators, even when unbiased, require adjustment. Our position that the intuitive epidemiological concept of confounding depends on efficiency considerations is heretical. Most epidemiologists assume that a crude estimate is confounded by a covariate C if the crude estimate does not equal the observed value of any "intuitively unbiased" estimator in the data at hand. For example, Miettinen and Cook describe the "confounding bias" of a crude estimate in their recent paper. We will show that Miettinen and Cook appear to regard as "intuitively unbiased" estimators that can center large-sample confidence intervals, i.e. estimators that are locally uniformly asymptotically unbiased. Furthermore, when conditioned on approximate ancillary statistics that measure the degree to which associations observed in the data differ due to sampling variability from our *a priori* knowledge of the corresponding population associations, inefficient unconditionally (asymptotically) unbiased estimators become conditionally (asymptotically) biased. Conversely unconditionally efficient estimators are conditionally (asymptotically) unbiased. Thus, the two approaches to confounding via efficiency and bias can be unified through the recognition that epidemiologists (such as Miettinen and Cook) are implicitly conditioning on certain approximate ancillary statistics. In Section 5(G), we show that, when data on C are unavailable, the above definition of a "confounder in the data at hand" reduces to the previous definition of "a confounder in the absence of data on C."

Although Miettinen and Cook restricted their attention to point estimation, most investigators are interested in confidence intervals as well. In fact large-sample Wald-type confidence intervals (i.e. intervals constructed from the maximum likelihood estimate and its estimated standard error) are the most common method of summarizing the inferences drawn from a given data set in present day epidemiologic practice. If we require that frequentist inference be performed conditional on ancillary statistics and furthermore that, as suggested by Buehler [4], continuous perturbations in model specification result in continuous perturbations in the inferences drawn from a given data set, then in accord with the intuition of epidemiologists, inference based on Wald-type confidence

intervals accurate to $O(N^{-3/2})$ in variance [5] would require that appropriate adjusted estimators be viewed as conditionally asymptotically unbiased rather than as unconditionally efficient. We show that long before Efron and Hinkley [5] formally considered conditional inference accurate to $O(N^{-3/2})$ in variance, analysts of epidemiologic data were intuitively requiring that their conditional inference be accurate to $O(N^{-3/2})$ in variance.

Since the intuitive epidemiologic concept of bias corresponds closely to the formal statistical concept of asymptotic bias, the above statistical interpretation of confounding in terms of intuitive (i.e. asymptotic) bias had to be based on asymptotic (i.e. large sample theory). In Section 6 an exact theory of confounding applicable to small samples or sparse data is developed. In the presence of data on C, we define C to be an (exact) *nonconfounder given data on C* if and only if inference on the parameter of interest does not depend on the data through C for all possible data outcomes. We show in Section 6 that Bayesians, frequentists, and pure likelihoodists will in general agree on the prior knowledge necessary to render C a nonconfounder given data on C in both follow-up and case-control studies. Specifically C will in general be a nonconfounder given data on C exactly when the crude data ignoring C comprise a cut [6] whose marginal distribution depends on the parameter of interest. We then investigate the prior knowledge necessary to make C a nonconfounder given C for the study designs and effect parameters considered in Section 4. Of particular concern is the role played by the rare disease assumption in case-control studies. In Section 6, in the interest of concreteness and notational simplicity, E, C and D will represent dichotomous random variables. Section 7 extends the results of Section 6 to the general case when E, C and D are arbitrary random variables.

Throughout the paper, to help focus our discussion, we shall consider the validity of the following principles endorsed by Miettinen and Cook:

1. If (a) in a case-control study the exposure E and a covariate C are known to be unassociated in the source population or (b) if in either a follow-up or case-control study C is known *a priori* not to be a risk factor for disease in the unexposed population, then the crude estimate of the parameter of interest ignoring data on C is appropriate irrespective of associations observed in the data.
2. In a follow-up or case-control study if C is known *a priori* to be a risk factor for disease, then even if C does not appear to be a risk factor in the data, the crude estimate is in general inappropriate. In a case-control study if E and C are known to be associated in the source population, then even if E and C are unassociated in the data among the controls, the crude estimate is in general inappropriate.
3. In a follow-up study if no E–C association exists in the source population from which the data were sampled, but an E–C association exists in the data, a stratified estimate of the parameter of interest, adjusting for C, is required.
4. In a follow-up study if no E–C association exists in the data, the crude estimate is appropriate irrespective of prior knowledge.

We will demonstrate that principle 1 holds for case-control studies only if we add to supposition (a) of principle 1 the suppositions that (1) the disease is rare over the follow-up period and (2) the sampled controls constitute only a small fraction of the potential controls. Principle 4 requires modification.

2. CONFOUNDING WITHOUT SAMPLING VARIABILITY

2(A). *Descriptive Epidemiology*

Consider an unmatched follow-up study in which at the start of follow-up each of the N study subjects is either exposed (written E or E_1) or unexposed (E_0 or $\bar{E}$). Measurements are taken at start of follow-up on a dichotomous covariate with levels C_1 (C) and C_0 ($\bar{C}$) and, at end of follow-up, on disease status with levels D (or D_1) and $\bar{D}$ (or D_0). The investigator observes the empirical distribution of E, C and D. For example, $p(E|C)$ is the proportion of subjects with covariate level C who are exposed. $p(D|E,\bar{C})$ is the proportion of exposed subjects at level $\bar{C}$ who develop disease. $\text{cRD} = p(D|E) - p(D|\bar{E})$ is the crude risk difference. Such observable quantities are the parameters of interest for descriptive epidemiology.

2(B). *A Deterministic Model Useful in Etiologic Research*

In etiologic research, on the other hand, the parameters of interest are intrinsically unobservable (i.e. nonidentifiable). For example, Miettinen and Cook suggest that causal parameters of interest in etiologic research should be expressed in terms of comparisons (e.g., the difference) between the observed number of cases occurring in the exposed group (O) and the number of cases that would have been observed in the exposed group had that group been unexposed (i.e. the *ex*pected number of cases, EX).

Specifically, we define $O/N_E - \mathrm{EX}/N_E$ to be the *causal risk difference in the exposed*, where N_E is the number of exposed subjects. O/N_E is the observable parameter $p(D \mid E)$. EX is unobservable, since the outcome of exposed subjects when unexposed cannot be observed.

Note that in order for EX to be a well defined number, it is necessary to entertain a deterministic model in which each study subject is one of four possible types according to their response to the presence and absence of exposure. Letting 1 indicate disease occurs and 0 indicate disease does not occur over the study period, we can tabulate the types in the following table (see Greenland and Robins [7]):

"Common" description of type	Exposed	Unexposed
Type 1. No effect (individual "doomed")	1	1
Type 2. Exposure causative (individual susceptible)	1	0
Type 3. Exposure preventive (individual susceptible)	0	1
Type 4. No effect (individual immune to disease)	0	0

We define N_j, N_{jE} and $N_{jE\bar{C}}$ to be respectively the total number of subjects, the number of exposed subjects, and the number of exposed subjects in stratum $\bar{C}$ who are of type j with $j \in \{1, 2, 3, 4\}$. Similarly, $P_{jE} = N_{jE}/N_E$ is the proportion of exposed subjects of type j. Also, $O = N_{1E} + N_{2E}$, $\mathrm{EX} = N_{1E} + N_{3E}$, $p(D \mid \bar{E}) = (N_{1\bar{E}} + N_{3\bar{E}})/N_{\bar{E}} = P_{1\bar{E}} + P_{3\bar{E}}$, and $(O/N_E - \mathrm{EX}/N_E) = P_{2E} - P_{3E}$.

2(C). *Confounding*

Following Miettinen and Cook [1] and Greenland and Robins [7], we say *there is no confounding for the causal risk difference in the exposed* if and only if the crude risk difference equals the causal risk difference in the exposed. This condition holds if and only if $p(D \mid E_0) = \mathrm{EX}/N_E$ (that is, the empirical rate of disease in the unexposed equals the rate of disease that would have been observed in the exposed had they been unexposed). Since EX is an unobservable, we can never know whether confounding exists. Under this definition, we can have confounding without needing to refer to any covariate that is the *cause* of that confounding (i.e. we have confounding without confounders [7]).

2(D). *Confounding by a Covariate* C

Suppose there is confounding for the causal risk difference in the exposed. We now define circumstances under which it is legitimate to say that the confounding is due to a covariate C. To do so, we shall need the following definitions.

Definition

We say there is no confounding for the *stratum-specific causal risk differences in the exposed* if, in each stratum of a covariate, the rate of disease in the unexposed stratum members equals the rate of disease that would have been observed in the exposed stratum members had they been unexposed. [That is, if for $i \in \{0, 1\}$ $(O_{C_i} - E_{C_i})/N_{EC_i} = \mathrm{RD}_{C_i}$. Equivalently, $\mathrm{EX}_{C_i}/N_{EC_i} = p(D \mid E, C_i)$. Here $\mathrm{RD}_{C_i} \equiv p(D \mid E, C_i) - p(D \mid \bar{E}, C_i)$. N_{EC_i} and O_{C_i} are respectively the total number of subjects and the number of diseased subjects who are (E, C_i), and EX_{C_i} is the number of those N_{EC_i} subjects who would have been diseased if unexposed.]

We shall say that the question "Is C a confounder for the causal risk difference in the exposed?" is meaningful (i.e. is nonvacuous) only if there is no confounding for the stratum-specific causal risk difference in the exposed.

This definition was motivated by the observation that if there is confounding for the stratum-specific causal risk differences in the exposed then, in general, even when data on E, C and D are available, the parameter of interest, $O/N_E - \mathrm{EX}/N_E$, will not equal any observable parameter.

On the other hand, if there is no confounding for the stratum-specific causal risk differences in the exposed, it follows, as a mathematical theorem, that $O/N_E - \mathrm{EX}/N_E$ equals the observable internally standardized morbidity difference (sMD), i.e. the (internally) standardized risk difference with weights taken from the exposed population, specifically,

$$\mathrm{sMD} \equiv w_{1E}\mathrm{RD}_C + (1 - w_{1E})\mathrm{RD}_{\bar{C}},$$

with $w_{1E} = N_{EC}/(N_{EC} + N_{E\bar{C}})$.

It follows that, given there is no confounding for the stratum-specific causal risk differences in the exposed, C is a confounder for the causal risk difference in the exposed if and only if $\mathrm{cRD} \neq \mathrm{sMD} = (O - \mathrm{EX})/N_E$. When confounding by C exists, the causal risk difference in the exposed can be computed when data on C are present (since one can compute the sMD), but not when data on C are absent (i.e. when data on C have not been recorded for data analysis). An alternative approach would have been to define C to be a confounder for the causal difference in the exposed whenever $\mathrm{cRD} \neq \mathrm{sMD}$. Under this alternative approach, when $\mathrm{cRD} = \mathrm{sMD} \neq (O - \mathrm{EX})/N_E$, we would say that even though C is not a confounder, there must exist confounding within levels of C by some other risk factor.

2(E). Confounding for the Causal Risk Difference

Miettinen and Cook restricted their consideration to comparisons between O and EX. In fact, other causal parameters may be of interest. For example, one may be interested in comparing the total number of cases that would have been observed if the entire population had been exposed $(N_1 + N_2)$ to the total number of cases that would have been observed if the entire population had been unexposed $(N_1 + N_3)$. Specifically, define $[(N_1 + N_2) - (N_1 + N_3)]/N = (N_2 - N_3)/N = P_2 - P_3$ to be the population causal risk difference.

By analogy to our definitions in Section 2(C), we say that there is *no confounding for the (population) causal risk difference if and only if* $\mathrm{cRD} = (N_2 - N_3)/N$.

Definition

There is no confounding for the stratum-specific causal risk differences if, for $i \in \{0, 1\}$, $\mathrm{RD}_{C_i} = (N_{2C_i} - N_{3C_i})/N_{C_i}$, where e.g., N_{2C_i} is the number of Type 2 subjects in stratum C_i.

Further we say that the question "Is C a confounder for the causal risk difference?" is meaningful only if there is no confounding for the stratum-specific causal risk differences. When there is no confounding for the stratum-specific causal risk differences, then it follows, as a mathematical theorem, that $(N_2 - N_3)/N$ is the observable standardized risk difference with weights taken from the entire population (which we denote sRD), where

$$\mathrm{sRD} \equiv w_1\mathrm{RD}_C + (1 - w_1)\mathrm{RD}_{\bar{C}} \quad \text{and} \quad w_1 = \frac{N_C}{N_C + N_{\bar{C}}}.$$

Given that there is no confounding for the stratum-specific causal risk differences, we say that C is a confounder for the causal risk difference if and only if $\mathrm{cRD} \neq \mathrm{sRD} = (N_2 - N_3)/N$.

2(F). Conditions for Nonconfounding

It is easy to show algebraically that $\mathrm{sMD} = \mathrm{cRD}$ if either C is not an (observable) risk factor among the unexposed, or E and C are unassociated, that is if either $\mathrm{OR}_{DC|\bar{E}} = 1$ or $\mathrm{OR}_{EC} = 1$. ($\mathrm{OR}_{DC|\bar{E}}$ is the observable odds ratio comparing D and C among the unexposed.) Similarly, $\mathrm{sRD} = \mathrm{cRD}$ if either C is not a risk factor in both the exposed and unexposed or E and C are unassociated, i.e. if either $\mathrm{OR}_{DC|\bar{E}} = 1 = \mathrm{OR}_{DC|E}$ or if $\mathrm{OR}_{EC} = 1$.

2(G). A Comparison with Miettinen and Cook

Miettinen and Cook state that C is a nonconfounder if either C is unassociated with exposure or C is a nonrisk factor among the unexposed. Our results show that Miettinen and Cook were implicitly assuming that (1) the causal parameter of interest is the causal risk difference in the exposed (or some other function of O and EX such as causal risk ratio among the exposed) and (2) there is no confounding for the stratum-specific causal risk differences in the exposed. Thus,

Table 1

	C		$\bar{C}$		Crude	
	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
D	30	100	120	50	150	150
$\bar{D}$	—	—	—	—	—	—
Total	200	400	400	200	600	600
	$RD_C = -10/100$		$RD_{\bar{C}} = 5/100$		cRD = 0	

as the example below demonstrates, Miettinen and Cook's concept of confounding lumps together two distinct issues: (1) what are the causal parameters of interest, and (2) what restrictions must exist on the associations of C with E or D in the population so that the causal parameter of interest equals the crude parameter. In contrast, we have allowed conditions for confounding to depend on the particular parameter of interest, keeping selection of the causal parameter of interest distinct from the issue of whether C is a "confounder" for the chosen parameter of interest.

Example

For the population represented in Table 1, suppose there is no confounding for the stratum-specific causal risk differences or the stratum-specific causal risk differences in the exposed. Then there is no confounding for the causal risk difference in the exposed since sMD = cRD (because $OR_{DC|\bar{E}} = 1$). Note that exposure is protective in stratum C. Now suppose we wish to consider a public health intervention, but that our options are limited by the fact that C-status is too expensive to measure routinely. Then, our options would be (1) no intervention, (2) expose everyone and (3) prevent all exposure. Since cRD = sMD = 0, it follows that the public health impact of options (1) and (3) would be the same. To compare options (2) and (3), a causal parameter of interest would be the causal risk difference, i.e. the sRD. But $sRD = 0.5(-0.1) + 0.5(0.05) = -0.025 \neq cRD = 0$. Thus the optimal public health decision is to expose everyone, and C is a confounder for the causal risk difference because $OR_{DC|E} \neq 1$.

2(H). Summary: Confounding for Causal and Observable Parameters

We summarize the approach to confounding given in Sections 2(A)–(G) in the following steps. Each step is illustrated (in parentheses) by a specific example:

(1) define the causal parameter of interest, (e.g. the causal risk difference $P_2 - P_3$);
(2) decide whether one is willing to make a nonidentifiable (i.e. untestable) assumption necessary to equate the causal parameter of interest to an observable parameter (e.g. the assumption of no confounding for the stratum-specific causal risk differences, so that $P_2 - P_3 = sRD$). If not, the question of whether a covariate C is a confounder for the causal parameter of interest is vacuous. If one is willing to make such an assumption, then;
(3) determine whether the (now) observable parameter of interest can be computed from data on E, C and D (e.g. sRD can be computed from data on E, C and D). If not, the question of whether C is a confounder for this parameter is vacuous;
(4) if the observable parameter of interest is computable from data on E, C and D, define C to be a nonconfounder for this parameter if and only if the observable parameter is computable from data on E and D alone (e.g. if and only if sRD = cRD);
(5) determine the conditions that must be satisfied by the associations of the covariate with exposure and/or disease in the study population in order for C to be a nonconfounder (e.g. $OR_{DC|E} = OR_{DC|\bar{E}} = 1$ or $OR_{EC} = 1$ implies sRD = cRD).

An investigator may be interested in an observable parameter that has no causal interpretation. For example, a life insurance company which is interested in selling policies to 65 y olds might wish to compare the probability of death before age 80 among 65 y olds who exercise regularly (the exposed) to that of 65 y olds who do not when controlling for cigarette smoking history regardless of whether any difference in life expectancy (when adjusted for smoking) can be causally attributed

to the effect of exercise. If the stratum-specific risk difference is constant over levels of smoking, then sMD = sRD is the observable parameter of interest, and we can determine whether smoking is a confounder for this observable parameter by carrying out steps 3–5 above.

In summary, to determine whether a covariate C is a confounder for a particular observable parameter, carry out steps 3–5. To determine whether it is a causal confounder, carry out steps 1–5.

We can extend the above definitions to define "confounding by a covariate C controlling for a covariate F". Specifically if the parameter of interest is computable given data on E, C, D and F, then we define C to be a confounder (nonconfounder) controlling for F if the parameter of interest cannot (can) be computed from data on E, D and F.

3. THE INTRODUCTION OF SAMPLING VARIABILITY

3(A). A Superpopulation Model

Suppose in the epidemiologic study described in Section 2(A), 30 of 60 exposed subjects develop disease. Then, according to the deterministic model described in Sections 2(A) and 2(B), $p(D \mid E)$ is known to be exactly 0.5. There can be no sampling variability since (1) each exposed subject's outcome is predetermined and (2) we have not assumed that our study population is a sample drawn from a larger population.

In contrast, the standard approach taken by most epidemiologists would be to report a 95% binomial confidence interval of $0.5 \pm 1.96\sqrt{(0.5)(0.5)/60}$ for the unknown parameter $p(D \mid E)$. The model implicit in the standard approach is as follows.

Superpopulation model. The study population has been sampled at random from a near infinite superpopulation. The parameter $p(D \mid E)$ is the proportion of exposed subjects in the superpopulation who become diseased and $\tilde{p}(D \mid E) = 30/60$ is the proportion of the (sampled) exposed study subjects who become diseased. Under this sampling model, the number of diseased exposed study subjects is a binomial random variable (upon hypothetical resamplings of 60 exposed study subjects from the superpopulation).

An alternative model that would (1) give binomial confidence intervals for $p(D \mid E)$ but (2) not require the introduction of a hypothetical superpopulation would be a model in which (a) each exposed study subject's outcome was a Bernoulli random variable and (b) the probability of developing disease over the study interval was the same for all exposed study subjects. $p(D \mid E)$ would be that common probability.

If the probability of developing disease varied among study subjects, the number of exposed study subjects who become diseased would not be a binomial random variable. Therefore, we reject this alternative model as biologically implausible since the model implies that there is no between-individual variation in any risk factor for disease.

In summary, any epidemiologist who (1) reports a binomial confidence interval for $p(D \mid E)$ and (2) who acknowledges that there exists between-individual variation in risk must be implicitly assuming (1) the study subjects were sampled from a near-infinite superpopulation and (2) that all inferences are concerned with the parameters of that superpopulation.

Since, in most epidemiologic studies, study subjects have not been randomly sampled from any near-infinite superpopulation, the superpopulation model is a fiction. Why then is the model routinely used? We offer a possible reason.

An investigator often wishes to generalize his or her findings from the study population to some large target population. For example, an investigator, who is giving consideration to recommending a public health intervention, would hope to have studied a population that is representative of the population of potential recipients of the intervention. The simplest possible model is to consider the study population to be a random sample of a larger population of subjects who are potential recipients of the intervention. The use of a hypothetical superpopulation model can sometimes be justified on subjective Bayesian grounds by use of DeFinetti's theorem [8]. In general an investigator should not entertain a superpopulation model if (1) the population of potential recipients is believed to differ from the study population on unmeasured disease risk factors to an extent that cannot be accounted for by sampling variability or (2) the size of the pool of potential recipients is not much larger than the size of the study population.

3(B). Randomization: an Alternative Sampling Model

Miettinen and Cook make it clear that in a follow-up study they wish to treat (a) disease outcomes as deterministic and (b) the causal parameters associated with the observed study population as the causal parameters of interest (regardless of whether that study population was sampled from a larger target population).

In this set-up, it is unlikely that any investigator would be willing to assume that the distribution of risk factors was so well balanced across exposure groups that there is no confounding for the causal risk difference or the causal risk difference in the exposed. [More generally, even when data on various risk factors have been obtained, it is unlikely that an investigator would be willing to assume the distribution of unmeasured risk factors for disease is so well balanced across exposure groups within levels of measured risk factors that there would be no residual confounding for the causal risk difference or the causal risk difference in the exposed after controlling for the measured factors, i.e. an investigator would not be willing to assume that $\text{sMD} = (O - \text{EX})/N_E$ or $\text{sRD} = (N_2 - N_3)/N$.] Nonetheless, an investigator might be willing to make subjective statements such as "although there may be some small association of risk factors with exposure, I do not believe such associations are systematic".

We can give a formal meaning to such a subjective statement by making the assumption that nature assigned exposure to N_E randomly chosen study subjects and left the remaining $N_{\bar{E}}$ subjects unexposed. N_E and $N_{\bar{E}}$ are considered fixed by design. That is, the observed study was a randomized trial performed by nature. Then the value of the empirical risk difference, $\tilde{\text{c}}\text{RD}$, depends on the particular N_E subjects who received exposure. Therefore, in hypothetical rerandomizations, $\tilde{\text{c}}\text{RD}$ has a well defined distribution. [The $\tilde{}$ over cRD is used to stress that we are now regarding cRD as a random variable with a distribution defined by hypothetical rerandomizations. The value of $\tilde{\text{c}}\text{RD}$ that we observed reflects the (random) exposure assignment that actually occurred.] In particular, if we define (using the notation in Copas [9]) $m_1 = (N_1 + N_2)/N$, $m_0 = (N_1 + N_3)/N$ and $\beta = (N_2 + N_3)/N$, then $E(\tilde{\text{c}}\text{RD}) = m_1 - m_0$. $m_1 - m_0$ is the causal risk difference. Thus, although in a randomized trial, in general $\tilde{\text{c}}\text{RD} \neq m_1 - m_0$, nonetheless $E(\tilde{\text{c}}\text{RD}) = m_1 - m_0$. Furthermore, $\tilde{\text{c}}\text{RD}$ is asymptotically normal with asymptotic variance,

$$N\,\text{Var}^A(\tilde{\text{c}}\text{RD}) = \frac{m_1(1-m_1)}{P_E} + \frac{m_0(1-m_0)}{P_{\bar{E}}} + R, \tag{1}$$

where $P_E = N_E/N$ and

$$R = m_0 + m_1 - 2m_0m_1 - m_1(1-m_1) - m_0(1-m_0) - \beta, \tag{2}$$

where, by its definition, the range of possible β is $|m_0 - m_1| \leqslant \beta \leqslant \min(m_0 + m_1, 2 - m_0 - m_1)$. For a proof see Appendix A.

Lemma 3.1

For all possible β, $R \leqslant 0$. The inequality is strict except when $\beta = 0$ and $m_1 = m_0$.

Proof. See Appendix A.

m_1 and m_0 can be unbiasedly and consistently estimated by $\tilde{m}_1 = \tilde{p}(D \mid E)$ and $\tilde{m}_0 = \tilde{p}(D \mid \bar{E})$ where $\tilde{p}(D \mid \bar{E})$ is the proportion of unexposed subjects observed to have disease. Unfortunately since β is unidentifiable, $N\,\text{Var}^A(\tilde{\text{c}}\text{RD})$ cannot be consistently estimated [9]. Nonetheless, we can set conservative large sample confidence intervals (i.e. confidence intervals that are guaranteed to cover at least their nominal rate) for $m_1 - m_0$ by deriving a consistent estimator, say $N\,\tilde{\text{V}}\text{ar}^A(\tilde{\text{c}}\text{RD})$, of a number that is at least as large as $N\,\text{Var}^A(\tilde{\text{c}}\text{RD})$. This can be accomplished if, in equation (2), we replace β by a consistent estimator of its minimum possible value, that is, by $|\tilde{m}_1 - \tilde{m}_0|$ and, in both equations (1) and (2), we replace m_1 and m_0 by their empirical estimates $\tilde{m}_1$ and $\tilde{m}_0$. On simplifying, we have that if $\tilde{m}_1 - \tilde{m}_0 \geqslant 0$,

$$N\,\tilde{\text{V}}\text{ar}^A(\tilde{\text{c}}\text{RD}) = \frac{\tilde{m}_1(1-\tilde{m}_1)}{P_E} + \frac{\tilde{m}_0(1-\tilde{m}_0)}{P_{\bar{E}}} + (2\tilde{m}_0 - \tilde{m}_1)(1-\tilde{m}_1) - \tilde{m}_0(1-\tilde{m}_0). \tag{3}$$

If $\tilde{\text{c}}\text{RD}$ was, as in the superpopulation model, the difference of two binomial proportions, its variance would be estimated by the first two terms on the right side of the equation (3). It then

follows from Lemma 3.1 that, when the causal risk difference is nonzero, the usual binomial confidence intervals for the causal risk difference in a randomized trial will be unnecessarily conservative since, taken together, the last two terms in equation (3) are negative if $\tilde{m}_1 - \tilde{m}_0 \neq 0$. In contrast, the usual binomial test of the null hypothesis $m_1 = m_0$ need not be conservative [since, in equation (1), $R = 0$ when $m_1 = m_0$ and $\beta = 0$].

Example 3.1

If

$$N_E = 100, \quad N_{\bar{E}} = 100, \quad \tilde{m}_1 = \frac{40}{100}, \tilde{m}_0 = \frac{15}{100},$$

then "binomial" 95% confidence intervals for the causal risk difference, $m_1 - m_0$, are

$$0.25 \pm 1.96 \sqrt{\frac{(0.4)(0.6) + (0.85)(0.15)}{100}} = 0.250 \pm 0.118$$

while conservative 95% confidence intervals based on equation (3) are

$$0.25 \pm 1.96\left[\frac{(0.4)(0.6) + (0.85)(0.15)}{100} + \frac{[2(0.15) - (0.4)](1 - 0.4) - 0.15(0.85)}{200}\right]^{1/2} = 0.250 \pm 0.102.$$

Thus, the binomial intervals are 16% wider than necessary.

Nonetheless, before one begins to routinely report confidence intervals for the causal risk difference in randomized trials based on equation (3), two caveats are in order. First, in moderate (as opposed to large) samples, it will happen frequently that "Wald" type intervals based on $\tilde{c}\text{RD} \pm 1.96 \sqrt{\tilde{\text{Var}}^A(\tilde{c}\text{RD})}$ will exclude the null even when standard X^2 or Fisher exact tests for $m_1 = m_0$ do not "reject" at the 5% level. In such cases, one should not "reject" the null hypothesis. (The problem with Wald confidence intervals is that $\tilde{\text{Var}}^A(\tilde{c}\text{RD})$ is evaluated at $\tilde{m}_1 - \tilde{m}_0$ rather than at the null hypothesis $m_1 = m_0$.)

Second, confidence intervals based on equation (3) are valid only if a deterministic model for disease outcome is correct. If outcomes were, in fact, stochastic, the causal risk difference would be defined to be

$$\left\{\sum_{i=1}^{N} [p(D \mid E, i) - p(D \mid \bar{E}, i)]\right\}/N,$$

where, for example, $p(D \mid E, i)$ is the probability that subject i will become diseased if exposed. In this stochastic world, $\tilde{c}\text{RD}$ is still unbiased for the causal risk difference in hypothetical rerandomizations but $\text{Var}^A(\tilde{c}\text{RD})$ is not given by equation (1). [For example, if $p(D \mid E, i)$ and $p(D \mid \bar{E}, i)$ did not depend on i, $\text{Var}^A(\tilde{c}\text{RD})$ is given by the usual binomial variance.] One cannot empirically determine whether outcomes are stochastic or deterministic.

Suppose now, following Miettinen and Cook, that our interest is in the causal risk difference in the exposed study population, and we have assumed that (a) disease outcomes are deterministic and (b) nature conducts a randomized trial. Then, the causal risk difference in the exposed is a random variable (and not a population parameter) since O and EX both depend on the particular set of subjects who were exposed. For a random variable, the analog of a confidence interval is a prediction interval. Specifically, a prediction interval is rule that gives for any data realization an interval. This interval may or may not contain the unobservable random variable $(O - \text{EX})/N_E$ associated with that realization. If, in hypothetical rerandomizations, 95% of realizations produce an interval that includes (the realization-specific) $(O - \text{EX})/N_E$, we have, by definition, a 95% prediction interval for $(O - \text{EX})/N_E$.

In Appendix A we show that a valid 95% large sample prediction interval for $(O - \text{EX})/N_E$ is

$$\tilde{c}\text{RD} \pm 1.96 \sqrt{\frac{\tilde{m}_0(1 - \tilde{m}_0)N}{N_E N_{\bar{E}}}}. \tag{4}$$

Example 3.2

Given the data in Example 3.1, we compute a prediction interval of

$$0.250 \pm 1.96\sqrt{\frac{(0.15)(0.85)200}{(100)^2}} = 0.250 \pm 0.099$$

under the randomization distribution. Thus, the usual "binomial" interval (computed in Example 3.1) is 18% too wide.

Note: equation (4) can be written as

$$\tilde{c}\text{RD} \pm 1.96\left[\frac{\tilde{m}_0(1-\tilde{m}_0)}{N_E} + \frac{\tilde{m}_0(1-\tilde{m}_0)}{N_{\bar{E}}}\right]^{1/2}. \tag{4a}$$

For certain states of nature, "binomial intervals" for the causal risk difference in the exposed (in contrast to "binomial intervals" for the causal risk difference) may be anti-conservative, that is they may fail to attain their nominal coverage rates. To see this, note that the prediction interval given in equation (4a) differs from the usual "binomial interval" only in that the second term under the square root sign in equation (4a) is $[\tilde{m}_0(1-\tilde{m}_0)]/N_E$ rather than $[\tilde{m}_1(1-\tilde{m}_1)]/N_E$. Thus, if $[\tilde{m}_1(1-\tilde{m}_1)]/N_E$ is greater than $[\tilde{m}_0(1-\tilde{m}_0)]/N_E$ (as in our example) binomial confidence intervals will be conservative. If $[\tilde{m}_1(1-\tilde{m}_1)]/N_E$ is less than $[\tilde{m}_0(1-\tilde{m}_0)]/N_E$, binomial confidence intervals will be anti-conservative. (This follows from the fact that, in large samples, the prediction interval given by equation (4a) covers the causal risk difference in the exposed in precisely 95% of hypothetical rerandomizations.) In fact, in the limit as $m_1 \to 1$ and $N_E/N_{\bar{E}} \to 0$, the rate at which the usual "binomial interval" covers the random variable $(O\text{-EX})/N_E$ approaches 0!

We can extend this randomization model to include measured covariates as follows. If data on a dichotomous covariate C has been obtained, and one believes that within levels of C there is no systematic association of exposure with unmeasured risk factors, then we could assume that, for $i \in \{0, 1\}$, nature exposed at random N_{EC_i} subjects and left $N_{\bar{E}C_i}$ subjects unexposed.

3(C). Relationships Between the Superpopulation and Randomization Models

One might find it surprising that the variance of the crude risk difference in the deterministic randomization model is less than the binomial variance, even asymptotically. Here we try to make it intuitively clear why this is so by, in a sense, imbedding our randomization model within the superpopulation model.

Subjective statements such as "exposure is not systematically associated with other risk factors in the observed study population" can be represented in our superpopulation model by assuming that $P_{jE} = P_{j\bar{E}}, j \in \{1, 2, 3, 4\}$ (i.e. the distribution of the four types are the same in the exposed and unexposed). Under this assumption there exists no confounding in the superpopulation. Even so, we will, in general, have confounding for both the "causal risk difference" and the "causal risk difference in the exposed" in the observed study population, since $\tilde{P}_{jE} = \tilde{P}_{j\bar{E}}, j \in \{1, 2, 3, 4\}$ will be false due to chance associations of exposure with risk factors in the sample. (We have put a $\tilde{}$ over the parameters of the observed study population to emphasize that these parameters are random variables under hypothetical resamplings of the superpopulation.) Suppose an investigator accepts the superpopulation model as a sampling model, but his interest lies in the causal risk difference of the observed study population (i.e. $[\tilde{P}_2 - \tilde{P}_3]$) rather than the causal risk difference of the superpopulation $(P_2 - P_3)$. $(\tilde{P}_2 - \tilde{P}_3)$ and $(P_2 - P_3)$ will, in general, differ from one another. This reflects the fact that $(\tilde{P}_2 - \tilde{P}_3)$ is a random variable with nonzero variance since its value will depend on the particular N_E exposed and $N_{\bar{E}}$ unexposed subjects sampled. Consider the set of hypothetical resamplings of the superpopulation in which the vector $(\tilde{N}_1, \tilde{N}_2, \tilde{N}_3, \tilde{N}_4)$ is fixed at its observed value, but $(\tilde{N}_{1E}, \tilde{N}_{2E}, \tilde{N}_{3E}, \tilde{N}_{4E})$ is not fixed. Then, the *conditional* distribution of this latter vector, given the former vector, is the same as the distribution of the vector $(\tilde{N}_{1E}, \tilde{N}_{2E}, \tilde{N}_{3E}, \tilde{N}_{4E})$ in a randomized trial with parameters equal to $(\tilde{N}_1, \tilde{N}_2, \tilde{N}_3, \tilde{N}_4)$. That is, we have generated the randomization distribution as a conditional distribution within the superpopulation model. It follows that the *conditional* expectation of the empirical crude risk difference $E[\tilde{c}\text{RD} | (\tilde{N}_1, \tilde{N}_2, \tilde{N}_3, \tilde{N}_4)]$ is $(\tilde{P}_2 - \tilde{P}_3)$ and the *conditional* variance is given by equation (1) where, m_1, m_0 and β now will depend on the particular superpopulation subjects that were sampled. Now,

it is a well known identity that

$$\mathrm{Var}(\tilde{c}\mathrm{RD}) = \mathrm{Var}\{E[\tilde{c}\mathrm{RD}|(\tilde{N}_1, \tilde{N}_2, \tilde{N}_3, \tilde{N}_4)]\} + E\{\mathrm{Var}\{\tilde{c}\mathrm{RD}|(\tilde{N}_1, \tilde{N}_2, \tilde{N}_3, \tilde{N}_4)]\}.$$

We have previously seen that, unconditionally, $\mathrm{Var}(\tilde{c}\mathrm{RD})$ is the usual binomial variance for the risk difference $\tilde{c}\mathrm{RD}$. Furthermore, as discussed above, $\mathrm{Var}[(\tilde{P}_2 - \tilde{P}_3)]$ is nonzero and of the same order as $\mathrm{Var}(\tilde{c}\mathrm{RD})$. It follows that on average $\mathrm{Var}[\tilde{c}\mathrm{RD}|(\tilde{N}_1, \tilde{N}_2, \tilde{N}_3, \tilde{N}_4)]$ [given by equation (1)] must be less than the binomial variance.

Thus, a valid *conditional* confidence interval for $\tilde{P}_2 - \tilde{P}_3$ is given by $\tilde{c}\mathrm{RD} \pm 1.96\,\tilde{\mathrm{Var}}^A(\tilde{c}\mathrm{RD})$ [as defined in equation (3)]. It follows that this interval is also an unconditional 95% prediction interval for the random variable $\tilde{P}_2 - \tilde{P}_3$. On the other hand, the unconditional confidence interval for $P_2 - P_3$ (which is the usual binomial interval) is wider because of the uncertainty in $P_2 - P_3$ given knowledge of $\tilde{P}_2 - \tilde{P}_3$. Thus, an investigator who (1) accepts a deterministic superpopulation model but (2) has interest in the causal risk difference and/or the causal risk difference in the exposed of the (sampled) study population should report (unconditional) prediction intervals for the causal risk difference as $\tilde{c}\mathrm{RD} \pm 1.96\,\tilde{\mathrm{Var}}^A(\tilde{c}\mathrm{RD})$ and for the causal risk difference in the exposed as given in equation (4).

3(D). Confounding for Superpopulation Parameters

In this subsection we suppose that (1) the study population has been *randomly* sampled from a near infinite superpopulation, (2) the parameter of interest is a parameter of the superpopulation and (3) we could compute the value of the parameter of interest if we had data on E, C and D for all members of the superpopulation. We shall say that *there is no confounding by a covariate* C *in the absence of data on* C *for the parameter of interest when there is no confounding in the superpopulation* (i.e. *when the superpopulation parameter of interest could be computed from data on* E *and* D *for each superpopulation member*). Therefore, there is no confounding in the absence of data on C for the causal risk difference if $\mathrm{cRD} = P_2 - P_3 = \mathrm{sRD}$ in the superpopulation even when, due to chance exposure-covariate associations in the sample, the crude risk difference in the data does not equal the (superpopulation) parameter of interest, i.e. $\tilde{c}\mathrm{RD} \neq P_2 - P_3 = \mathrm{sRD} = \mathrm{cRD}$. Nonetheless, when there is no confounding for the causal risk difference in the absence of data on C, $E[\tilde{c}\mathrm{RD}] = P_2 - P_3$. In fact *the above definition implies that* C *is a nonconfounder in the absence of data on* C *for the superpopulation parameter of interest* (*e.g. the causal risk difference*) *precisely when the crude estimator* (*e.g.* $\tilde{c}\mathrm{RD}$) *is locally uniformly asymptotically unbiased for that parameter; that is, if* C *is a nonconfounder in the absence of data on* C, *we can unbiasedly estimate* (*asymptotically*) *the parameter of interest in the absence of data on* C. For reasons discussed in Section 4(A) and in Definition 3 of Appendix B we have only required the crude estimator to be locally uniformly asymptotically unbiased for the parameter of interest. [Henceforth, unless stated otherwise, the term "unbiased" will mean locally uniformly asymptotically unbiased. See Appendix B.]

In the following section, we slightly modify this definition of confounding in the absence of data on C in order to allow for *nonrandom* sampling of the superpopulation as would occur, for example, in a matched case-control study.

4. CONFOUNDING IN THE ABSENCE OF DATA ON C

Throughout the remainder of the paper, we suppose that the study population has been sampled (although not necessarily randomly) from a superpopulation and that the parameter of interest is a parameter of that superpopulation. [In order to justify binomial confidence intervals in experimental studies, the subjects entered in a randomized trial can be considered a random sample from a superpopulation in which $OR_{EC} = 1$ for all covariates C.]

In this section, we assume that data on the covariate C have not been recorded for data analysis. We define C *to be a nonconfounder in the absence of data on* C if and only if, given the available prior knowledge, the study design, and the sampling scheme, the superpopulation "parameter of interest" can be unbiasedly estimated from the crude data (i.e. the large sample expectation of some crude estimator equals the population parameter of interest; see Definition 3 in Appendix B). We will require that the parameter of interest can be unbiasedly estimated whenever data on C are available, since otherwise the above definition would often be vacuous.

If a crude parameter is defined to be any population parameter that can be unbiasedly estimated from the crude data given the available prior knowledge, the study design and the sampling scheme (even if nonrandom), then C will be a nonconfounder in the absence of data on C if and only if the parameter of interest equals a crude parameter.

4(A). Parameters of Interest in Cumulative Incidence Follow-up Studies

As in Section 3, we consider a study of the effect of a binary exposure factor $(E, \bar{E})$ on the risk of a disease in a large (super)population over a specified follow-up period. Lost to follow-up and death from causes other than the disease of interest are assumed to be negligible. Each member of the population either does or does not have a factor C. For each individual, the levels of E and C remain constant over the follow-up period. (It follows that if exposure occurred exactly at start of follow-up, then C cannot be an intermediate variable on the causal pathway from exposure to disease.) We now consider the conditions under which C is a confounder for the effect parameter and study designs commonly used in epidemiologic research.

First consider a follow-up study in which some fixed number of exposed and unexposed subjects are sampled at random from the superpopulation. In Sections 2 and 3 we have considered conditions for confounding for the superpopulation *causal risk difference in the exposed.* Define O/EX to be the causal risk ratio in the exposed. In the absence of confounding for the stratum-specific causal risk ratios in the exposed superpopulation [i.e. $\mathrm{EX}_{C_i} = p(D \mid \bar{E}, C_i)$ for $i \in \{0, 1\}$], O/EX equals the estimable superpopulation sMR (i.e. the internally standardized morbidity ratio), where

$$\mathrm{sMR} = \frac{p(D \mid E)}{p(\bar{C} \mid E)p(D \mid \bar{E}, \bar{C}) + p(C \mid E)p(D \mid \bar{E}, C)}.$$

$\mathrm{sMR} = \mathrm{cRR} \equiv p(D \mid E)/p(D \mid \bar{E})$ if and only if, in the superpopulation, $\mathrm{OR}_{EC} = 1$ or $\mathrm{OR}_{DC|\bar{E}} = 1$. Since the crude risk ratio estimator $\tilde{\mathrm{c}}\mathrm{RR} = \tilde{p}(D \mid E)/\tilde{p}(D \mid \bar{E})$ is locally uniformly asymptotically unbiased for cRR (although the exact expectation of $\tilde{\mathrm{c}}\mathrm{RR}$ is infinite), these conditions for no confounding for the sMR are the same as those for the sMD. Throughout the remainder of the paper, we concentrate on the sMR rather than the sMD as the parameter of interest because the sMR, in contrast to the sMD, can be estimated from case-control data under the rare disease assumption.

4(B). Odds Ratios and Case-control Studies

In both case-control and follow-up studies epidemiologists commonly report stratum-specific odds ratios as effect parameters. In case-control studies, this reflects the fact that risk ratios and differences cannot be estimated without knowledge of the control sampling fraction. In follow-up studies, this reflects the recent popularity of logistic regression models. In case-control studies, the internally standardized odds ratio, sOR [10], is often chosen as a summary measure of interest when the exposure disease odds ratio is not constant across levels of C, where

$$\mathrm{sOR} = \frac{p(E \mid D)}{p(\bar{E}, C \mid D)\dfrac{p(E \mid \bar{D}, C)}{p(\bar{E} \mid \bar{D}, C)} + p(\bar{E}, \bar{C} \mid D)\dfrac{p(E \mid \bar{D}, \bar{C})}{p(\bar{E} \mid \bar{D}, \bar{C})}}.$$

sOR is a weighted average of the stratum specific odds ratios. In an unmatched case-control study, C is a nonconfounder for the sOR (or for a common odds ratio) if and only if C is not a risk factor in the unexposed or E and C are unassociated in those nondiseased at end of follow-up since.

$$\mathrm{sOR} = \mathrm{cOR} \quad \text{if and only if} \quad \mathrm{OR}_{EC|\bar{D}} = 1 \quad \text{or} \quad \mathrm{OR}_{DC|\bar{E}} = 1$$

(where, for example, cOR is the crude exposure disease odds ratio in the super population). Here we have used the fact that in an unmatched case control study $\tilde{\mathrm{c}}\mathrm{OR}$, the empirical crude odds ratio estimator, is asymptotically unbiased for cOR. Note if C is a risk factor in the unexposed and $\mathrm{OR}_{EC} = 1$, then $\mathrm{OR}_{EC|\bar{D}} \neq 1$, and thus C will be a confounder for the sOR or for a common odds ratio.

If the disease is rare $\mathrm{sOR} \approx \mathrm{sMR}$ [10] (where in this paper the use of the expressions "the disease is rare" or "under the rare disease assumption" always implies that at each *joint* level of E and

C, the cumulative incidence of disease over the entire follow-up period is small). Since the sOR, in contrast to the sMR, has no clear-cut epidemiologic meaning, we concur with Miettinen and Cook in their implicit rejection of the sOR as the parameter of interest when the disease is not rare. If the disease is rare, then whenever E and C are unassociated in the population ($OR_{EC} = 1$), E and C will be nearly unassociated among those nondiseased at the end of follow-up (i.e. $OR_{EC|\bar{D}} \approx 1$). Thus in unmatched cumulative incidence studies in which the sMR is parameter of interest, and C is at 2 levels, C is a nonconfounder in the absence of data on C in either case-control studies (under the rare disease assumption) or in follow-up studies if and only if C is not a risk factor for disease in the unexposed ($OR_{DC|\bar{E}} = 1$) or E and C are unassociated in the superpopulation. These are well known results [11].

In a matched case-control study in which controls are matched to cases on level of C, if C is not a risk factor for disease among the unexposed but E and C are associated among the nondiseased in the population then even if the disease is rare, $sOR = cOR \neq cOR_{matched}$ where $cOR_{matched}$ is the crude parameter estimated by the crude OR estimator from matched data [1] (cOR is the superpopulation crude odds ratio and is the crude parameter estimated by $\check{c}OR$ in an unmatched study); thus the parameter of interest, sOR, does not equal the crude parameter defined by the sampling scheme. Therefore C is a confounder for the sOR in the absence of data on C (that is, when C status is not recorded for data analysis) in such a matched study and inference on sOR cannot be performed from the crude data.

In general, the decision not to measure a covariate C in an observational (i.e. nonexperimental) study with sMR as parameter of interest requires an investigator to have prior knowledge that C is a causal nonconfounder absent data on C (or at least prior knowledge that the discrepancy between cRR and sMR is biologically unimportant). If C is at more than 2 levels or continuous, then it is possible that cOR = sOR or cRR = sMR (and thus C is a nonconfounder for the sOR or sMR) even if C is both a risk factor for disease in the unexposed, and E and C are associated in the population, provided the population associations between C, E and D satisfy certain complex restrictions (see Section 7). Although of theoretical interest, in practise an investigator would be very unlikely to have prior knowledge about such a complicated association. On the other hand, prior knowledge that C is not a risk factor for disease, (e.g. C is preferred shirt color) or not related to certain environmental exposures (e.g. C is a genetic risk factor) is commonplace.

4(C). Cumulative Incidence Case-control Study of a Common Disease with the Overall Probability of Disease Known or Estimable

Suppose the disease is not rare and therefore $OR_{EC|\bar{D}} \tilde{\neq} OR_{EC}$ and $sMR \tilde{\neq} sOR$ but the probability of disease in the target population is known *a priori* from other studies or the sampling fraction in cases and controls is known. Then sMR and cRR can still be unbiasedly estimated given data on C from case-control data using Bayes rule [12]. Absent data on C, cRR can be estimated. Thus, sMR can be unbiasedly estimated if and only if $cRR = sMR \Leftrightarrow OR_{EC} = 1$ or $OR_{DC|\bar{E}} = 1$ is known *a priori* (i.e. the conditions for nonconfounding are the same as in a follow-up study). On the other hand, if the disease is not rare and neither the probability of disease nor the control sampling fraction is known, then, by our definition, the question of whether C is a confounder for the sMR is vacuous, since the sMR cannot be unbiasedly estimated even were data on C available.

4(D). Other Common Study Designs and Effect Parameters

4(D).1. Incidence density ratio in a steady state population

In a case-control study within a stable (steady state) population with the internally standardized incidence density ratio (sIDR) *as parameter of interest*, sOR = sIDR *without the rare disease assumption* [13]. In the absence of matching on C, $\check{c}OR$ is asymptotically unbiased for cOR. Therefore, the exposure-covariate condition for nonconfounding absent data on C ($OR_{EC|\bar{D}} = 1$) implies E and C are unassociated at all times in the population at risk without reference to any rare disease assumption. The disease covariate condition for nonconfounding is, as before, $OR_{DC|\bar{E}} = 1$. These conditions for nonconfounding remain unchanged when the incidence density ratio is known to be constant over strata.

4(D).2. Incidence density ratio in fixed cohort studies

Consider a fixed cohort study where time of disease onset is recorded and, within each level of C, the incidence density ratio (hazard ratio) is constant over time and levels of C. If C is not a risk factor for disease in the population, C is a nonconfounder for the incidence density ratio (IDR) absent data on C (since IDR = cIDR) in both a follow-up and time-matched case-control study without need for the rare disease assumption [14], where cIDR is the crude hazard ratio. In a time-matched case-control study the controls associated with a given case are sampled from individuals at risk at the time of disease onset of the case. But if C is a risk factor, then even when E and C are unassociated in the population at the start of follow-up, if the disease is not rare, the E–C association changes over time, cIDR is not constant over time, and cIDR $\neq$ IDR; thus surprisingly C would effectively be a confounder for the IDR in the absence of data on C in both a follow-up and a time-matched case-control study.

The above result concerning the IDR is made less disturbing when it is recognized that even in a randomized trial the IDR often fails to have a causal interpretation and thus C is not a confounder for a causal parameter. For example, in a hypothetical study of individuals randomized at age 20 to either cigarette smoking or no cigarette smoking, it is quite possible that although smoking may shorten each individual's life span, the 65 y old smokers reflect a (survivor) population of relatively resistant individuals with greater life expectancy than 65 y old nonsmokers. As an extreme example it might be that individuals can only get lung cancer if they possess a (unmeasured) genetic allele X, that all smokers who possessed allele X have died by age 65, but some 65 y old nonsmokers with allele X are still living. Thus, the comparison of lung cancer incidence in smokers and nonsmokers at age 65 is not causal since it fails to control for the risk factor-allele X, even though in a large trial allele X was almost certainly balanced by randomization at start of follow-up. Since a similar lack of comparability on X will arise within each stratum of a dichotomous measured covariate C, the stratum-specific IDR will, in general, also fail to have a causal interpretation.

The above suggests that, as is common practice in the randomized trial literature, causal parameters should be defined in terms of cumulative incidences rather than in terms of incidence densities (or conditional survivals). If so, then if E and C are unassociated in the population at start of follow-up, C will be a nonconfounder absent data on C for parameters such as the sMR, which depend only on the cumulative incidence. Note that one should, in general, compare the cumulative incidence of disease in exposed and unexposed individuals at all times after start of follow-up and not just at end of follow-up (i.e. one should compare entire mortality or survival curves).

More formally, in a deterministic model, we say that the cIDR at time t is confounded if it fails to equal the causal parameter defined as the limit as $\Delta t \to 0$ of $1/\Delta t$ times the causal risk ratio with follow-up through $t + \Delta t$ for the subset of study subjects who would survive to t regardless of exposure. This causal parameter would equal the cIDR at t in a large randomized trial (so that there is no confounding at start of follow-up) if (1) the conditional probability of death among exposed subjects between t and $t + \Delta t$ equals the same conditional probability among the subset of exposed subjects who would survive to t both when exposed and unexposed, and (2) condition (1) holds with "unexposed" replacing "exposed". Condition (2) fails in the above example, since, although some unexposed subjects alive at 65 have allele X, no unexposed subject alive at 65 both has allele X and would be alive at 65 if exposed to cigarettes.

5. CONFOUNDING GIVEN DATA ON C

5(A). Definition of Confounding Given Data on C

In this section we consider the implications for statistical inference on the causal parameter of interest of discovering, when we have collected data on C, that the associations of C with exposure and/or disease in our sample differ from those associations known to hold in the superpopulation. As usual, we assume that the parameter of interest can be unbiasedly estimated from data on E, C and D.

In the most comprehensive article on confounding to date, Miettinen and Cook develop "first principles" of confounding based on inductive, intuitive considerations of various examples [1]. In that paper, Miettinen and Cook state that "confounding implies a need and desire to replace the 'crude' estimate of the effect by one that has been adjusted for the covariate at issue". Implicit in this definition is the assumption that data on the covariate have been recorded for data analysis so that an estimate adjusted for the covariate can be computed.

As an example, suppose in a follow-up study that, by chance, an E–C association exists in the study data even though it is known *a priori* that no E–C association exists in the superpopulation. Then if, in the data, C is a risk factor in the unexposed, the crude and stratified estimate of the sMR will differ. Miettinen and Cook (see principle 3 in Section 1) regard the stratified estimate as appropriate. They regard the crude estimate as being confounded and in need of adjustment even though, as we have seen in Section 4, because there is no population E–C association, the crude estimator is a valid unbiased estimator of the sMR in the *absence* of data on C.

In accord with Miettinen and Cook, we define C to be a point estimate nonconfounder in the data at hand, if and only if the crude estimate of the parameter of interest equals the observed value of an *appropriate* adjusted estimator. Miettinen and Cook fail to give a general definition of an appropriate adjusted estimate. We formally define an appropriate estimator to be an efficient Fisher* consistent estimator. We demonstrate that the requirement that appropriate estimators* be efficient* estimators is necessary, in general, to reach Miettinen and Cook's intuitive inductive decisions. (Heuristic definitions of statistical terms marked with an asterisk are given in Appendix B.) Because we call for a comparison of the crude estimate with an appropriate estimate in the data at hand, it may appear that our definition of confounding is similar to earlier purely data-driven definitions. This is in fact not the case because, as shown in Examples A and B below, efficient estimators explicitly incorporate any prior knowledge that the investigator may have concerning the population parameters. The need for Fisher consistency is discussed in Section 5(P).

The position that the intuitive epidemiological concept of confounding depends on efficiency considerations is quite heretical. Many epidemiologists seem to assume that a crude estimate is confounded by a covariate C if the crude estimate does not equal the observed value of any "intuitively unbiased" estimator in the data at hand [1, 10, 15, 16]. For example, Miettinen and Cook [1] describe the "confounding bias" of a crude estimate. Estimators are "intuitively unbiased" if they can serve as the center of a large-sample* confidence interval* for the parameter of interest (i.e. if they are locally uniformly asymptotically unbiased*). In this paper an estimator will be referred to as biased if and only if it cannot center large-sample confidence intervals (CI). By introducing a new statistical concept into epidemiology—approximate ancillarity [5]—we can view those estimators requiring adjustment (inefficient estimators) to be biased in repeated trials, conditional on the observed values of any approximate ancillary statistics. *Approximate ancillaries measure the degree to which associations observed in the data differ due to sampling variability from our* a priori *knowledge of the corresponding population associations. Thus, appropriate estimators can be defined either as unconditionally efficient or as unbiased conditional on the observed values of all approximate ancillary statistics.*

Miettinen and Cook only concern themselves with the issue of whether the crude point estimate of effect is appropriate. But frequentist inference concerning a parameter of interest is best summarized in a confidence interval*. We therefore define C to be a confidence interval nonconfounder in the data at hand if and only if inference on the parameter of interest does not depend on the data through C, i.e. if and only if the observed crude confidence interval equals an appropriate observed confidence interval*. An appropriate confidence interval can be constructed from a conditionally unbiased (i.e. appropriate) estimator and a conditionally unbiased estimator of the estimator's conditional standard error. In this paper, confounding when not preceded by "CI", will refer to point estimate confounding. We show in Example F that ignoring C in the analysis when the crude point estimate equals an appropriate estimate, but the observed crude confidence interval is inappropriate, may lead to serious misinterpretations of the data. Furthermore, if large-sample confidence intervals and tests (p-values) are of interest in addition to point estimates, a unified approach would require that CI confounding (in contrast to point estimate confounding) by viewed (as in the above definition) in terms of bias conditional on approximate

ancillary statistics rather than in terms of unconditional efficiency. In particular, in Example C below, we show that an approach based on unconditional efficiency can lead to the reporting of inappropriate confidence intervals.

Furthermore, in Example C, we will show that even if, based on prior knowledge, C is a nonconfounder in the absence of data on C, if data on C (and therefore on an approximate ancillary) subsequently become available, C can become a confounder. Confounding, as we have defined it, thus depends not only on the available prior knowledge, but also on the available data. In fact in Section 5(G) we demonstrate that, when data on C are unavailable, the definition of confounding given in this subsection reduces to the definition of confounding in the absence of data on C as given in Section 4.

Any statistical interpretation of confounding in terms of intuitive (i.e. asymptotic) bias must, by definition, be based on large-sample* (i.e. asymptotic) theory. In this section, our interest is in offering a statistical interpretation of the epidemiologist's intuitive concept of confounding. An "exact" theory of confounding applicable to small samples or sparse data* is developed in Section 6.

5(B). Prior Knowledge vs Prior Belief

Central to this paper is the explicit incorporation into our inference of prior knowledge of various population parameters. Prior knowledge must be carefully distinguished from prior belief. Prior knowledge will refer to correct knowledge concerning a state of nature (i.e. a population parameter) obtained from outside the study. It need not be prior temporally. It may be derived from previous empirical studies, from biological theories, or from previous actual physical randomization. The decision as to what prior knowledge exists in a given study is (except in the case of randomization) a "subject matter" issue [1], and depends on the knowledge of the investigator.

Although Bayesian statistical inference explicitly incorporates prior information, in this paper we largely restrict ourselves to inference based on outcomes in hypothetical repeat trials (frequentist inference), since frequentist confidence intervals are the present standard method of reporting epidemiologic results. Bayesian inference, in which all parameters are assumed to have prior probability distributions, overcomes the need for our prior information to be correct concerning a state of nature, by defining probability as the subjective beliefs of an individual. Frequentist inference based on confidence intervals, on the other hand, requires that prior knowledge about nature used in constructing confidence intervals be correct. Since, in frequentist statistics, population parameters are considered fixed unknown constants and not random variables (with the exceptions of empirical Bayes and random effects models), the concept of a prior (nonpersonal) probability distribution for a parameter does not make sense. For example, we might have prior knowledge that a parameter lies in an interval (2,4) but not that its prior probability distribution (as a fact of nature) is uniform on (2,4).

To make explicit the above distinction between prior belief of an individual and correct prior knowledge concerning nature consider routine data analysis in regression packages incorporating backward elimination (see Ref. [17]). In such analyses a test is performed to determine whether an estimate of the covariate's parameter (i.e. coefficient) differs significantly from zero. If not, the parameter is set to zero; otherwise, it is set equal to its estimate from the full model. An investigator using such a package has in effect a personal prior probability distribution for the covariate's parameter that assumes the parameter is probably near zero, but allows a small probability that the parameter may take on any value. Thus such an investigator does not have prior knowledge since all parameter values are considered possible. If the investigator believed that the value of the covariate's parameter was more likely 10 than 0, he or she presumably would have tested to see whether the estimate differs significantly from 10, and, if not, set the parameter to 10.

Once the parameter is set to zero by the backward elimination program, it is then handled in the subsequent analysis as if the investigator had prior knowledge that the parameter was zero. For example, using the model with the covariate's parameter set to zero the confidence interval for the exposure parameter of interest will cover the true parameter of interest 95% of the time only if in fact the covariate's parameter is zero. A 95% confidence interval, by definition, must cover the true parameter 95% of the time for all possible value of the other parameters, but if the covariate is eliminated, zero is considered the only possible values for the covariate's coefficient.

How can a test lead from a prior belief that a parameter can take on any value to absolute knowledge it is zero? Obviously, it cannot. For example, such a test may have little power to detect a true value of two for the covariate's parameter. Only true prior knowledge rather than prior belief is the basis for ignoring covariates in construction of confidence intervals.

The usefulness of backward elimination procedures is for point estimation, not for confidence intervals. For example, if one's prior beliefs are nearly correct, then the backward elimination estimator, although in general biased [see Section 8(E)], will yield a smaller mean square error (i.e. be closer on average to the true parameter) in small and moderate sized samples than the estimator from the full model, and will do only slightly worse than the full model estimator if one's prior beliefs were mistaken. But biased estimators cannot be used to center confidence intervals.

It may be argued that we never have prior knowledge, but only prior beliefs, i.e. extreme data results will always make us revise our "prior knowledge" and thus our "knowledge" was actually only strong belief. If so, in purely frequentist framework, we could never exclude any of the often numerous measured covariates from the analysis for the construction of confidence intervals, and we would thus have no power to detect biologically relevant alternatives to the null. For the present, we will assume that those prior beliefs used in data analysis are of sufficient strength that, given the sample size of our study, no outcome (no matter how extreme), would make us revise those beliefs to any large degree. Thus the investigator acts inferentially as if in the possession of true prior knowledge. (This latter assumption is relaxed in Robins and Greenland [17].) Similarly, whenever an investigator chooses not to obtain data on a covariate in an observational study, the investigator acts inferentially as if in the possession of true prior knowledge that the covariate is a nonconfounder absent data on C. A central theme of this paper is the implication for data analysis and study design of inference performed as if in the possession of prior knowledge. Note that, by our definition, C would be considered a confounder by a particular investigator even when, as a fact of nature, the crude estimator was unbiased, if that investigator lacked prior knowledge of the crude estimator's unbiasedness. Therefore, what we call a "confounder", others might choose to call a "potential confounder".

5(C). *Confounding as Efficiency*

Example A. Prior knowledge that C *is not a risk factor in the population in a follow-up study*

Imagine C was a covariate representing social security number ending in 5 or greater vs $\bar{C}$ (4 or less), and Table 2 was the raw data from a follow-up study of the effect of E on D with the sMR as the population parameter of interest. When considering follow-up (i.e. prospective) studies, the rows of Table 2 represent disease status and the columns exposure status. For case-control studies, the rows represent exposure status and the columns disease status. Most epidemiologists would assume *a priori* that social security number is not a risk factor for disease in the unexposed population, (i.e. $OR_{DC|\bar{E}} = 1$) and suggest the collapsed analysis be used even though a D–C association exists in the data among the unexposed. (Miettinen and Cook's Appendix [1] that purports to demonstrate that a covariate such as social security number could still be a confounder is not pertinent here, but is discussed in detail following Example C.)

Since $OR_{DC|\bar{E}} = 1$, sMR = cRR in the population. Thus both the crude c̃RR and the usual stratified estimator (the unrestricted maximum likelihood estimator—ũsMR) are unbiased for sMR. Note that $\tilde{u}sMR = \frac{30}{45} \neq \tilde{c}RR = 1$ in these data because due to sampling variability $\tilde{u}OR_{DC|\bar{E}} = 7 \neq OR_{DC|\bar{E}} = 1$. If we sampled the entire population, given that our prior knowledge is correct, this discrepancy could not occur.

Why would many epidemiologists prefer c̃RR to ũsMR? Heuristically, in the absence of prior knowledge, the raw data give the "best unbiased estimates" (the unrestricted maximum likelihood

Table 2

Case control		C		$\bar{C}$		Total	
		D	$\bar{D}$	D	$\bar{D}$	D	$\bar{D}$
	Prospective	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
E	D	20	20	10	10	30	30
$\bar{E}$	$\bar{D}$	60	20	30	70	90	90
Total		80	40	40	80	120	120

Table 3

Case control		C		$\bar{C}$		Total	
		D	$\bar{D}$	D	$\bar{D}$	D	$\bar{D}$
	Prospective	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
E	D	20	10	10	20	30	30
$\bar{E}$	$\bar{D}$	60	30	30	60	90	90
Total		80	40	40	80	120	120

estimator) of the unknown population proportions, and thus $\tilde{u}$sMR calculated from the raw data is the "best estimate" of sMR. But given prior knowledge $OR_{DC|\bar{E}} = 1$ the raw data proportions are inconsistent with the known population proportions. Our "best unbiased estimate" of the population proportions (called the restricted maximum likelihood estimate) would instead be calculated from a fitted table (Table 3) that is as "close" to the observed data as possible subject to the restriction $OR_{DC|\bar{E}} = 1$ in the fitted table. $\tilde{r}$sMR (the restricted maximum likelihood estimator of sMR) is the stratified estimate from the fitted table and is our "best" unbiased estimate of sMR given our prior knowledge $OR_{DC|\bar{E}} = 1$. But for all data outcomes $\tilde{r}$sMR = $\tilde{c}$RR if $OR_{DC|\bar{E}} = 1$, as exemplified in Table 3. From a formal statistical viewpoint, since both $\tilde{u}$sMR and $\tilde{r}$sMR are unbiased and thus appropriate for centering confidence intervals, the choice of $\tilde{r}$sMR over $\tilde{u}$sMR by epidemiologists must be based on efficiency considerations, i.e. the fact that the variance of $\tilde{r}$sMR is less than $\tilde{u}$sMR.

$\tilde{r}$sMR has smaller variance than $\tilde{u}$sMR since unrestricted maximum likelihood estimation estimates all population parameters in the model simultaneously (e.g. $\tilde{u}$sMR, $\tilde{u}OR_{DC|\bar{E}}$, etc.). Each parameter that is estimated uses up some of the information in the data. In restricted maximum likelihood estimation, $OR_{DC|\bar{E}}$ is not estimated, but rather set equal to the true known value. Thus more data information, in general, is available for precise estimation of the parameter of interest. The usual (unrestricted) maximum likelihood estimator is, in general, efficient only in the absence of prior knowledge. Of course if our prior knowledge was incorrect, $\tilde{r}$sMR is biased and thus inappropriate.

Inefficient estimation is equivalent to ignoring informative aspects of the data. For instance, in Example A the difficulty with $\tilde{u}$sMR is that it estimates the probability of D given C and $\bar{E}$ $[p(D \mid C, \bar{E})]$ as 20/40 based on only 40 observations, ignoring that the *a priori* knowledge $p(D \mid C, \bar{E}) = p(D \mid \bar{C}, \bar{E})$ would allow estimation of $p(D \mid C, \bar{E})$ to be based on 120 observations. Thus the best estimate of $p(D \mid C, \bar{E})$ is 30/120 [the restricted maximum likelihood estimator of $p(D \mid C, \bar{E})$], as in the fitted table.

Example B. Confounding given prior knowledge that C *is a risk factor*

Our Example B will be a follow-up study equivalent to Miettinen and Cook's Example 2 [1] from their 1981 paper in which E is identified with tolbutamide, C with age (at two levels), and D with death from cardiac disease, except we assume that instead of a randomized trial, our data are a random sample from a large superpopulation with $OR_{EC} \neq 1$ and with sMR as parameter of interest. Following Miettinen and Cook, we assume that age is known *a priori* to be related to disease among the unexposed, where, for the sake of argument, we assume this relationship is known precisely to be $OR_{DC|\bar{E}} = 7$.

The data in Table 3 were observed. Note in Table 3, a positive E–C association exists in the data. Although $\tilde{c}$RR is unconditionally biased since cRR $\neq$ sMR (and thus C is a confounder in the absence of data on C), nevertheless, in the data at hand, the crude estimate equals the usual adjusted estimate, i.e. $\tilde{u}$sMR = $\tilde{c}$RR = 1. But the fitted table using the prior knowledge that $OR_{DC|\bar{E}} = 7$ is precisely Table 2. Therefore, $\tilde{r}$sMR = 30/45. Now if our choice was $\tilde{r}$sMR over $\tilde{u}$sMR when $OR_{DC|\bar{E}} = 1$ *a priori*, the same choice is required when $OR_{DC|\bar{E}} = 7$ *a priori*. Thus, although we have no apparent confounding in the data, (i.e. $\tilde{u}$sMR = $\tilde{c}$RR), $\tilde{c}$RR = 1 must be adjusted to $\tilde{r}$sMR = 30/45. C is therefore a confounder in the presence of data on C, in agreement with Principle 2 of the Introduction. (The statistical package GLIM 3 [18] can often be used to calculate the restricted maximum likelihood estimator if the *a priori* known parameter is set equal to its known value in the model using the offset option.)

In practice it is much more common to have exact prior knowledge that $OR_{DC|\bar{E}} = 1$ as in Example A than $OR_{DC|\bar{E}} = 7$. But often, as for age and heart disease, previous studies often might show $OR_{DC|\bar{E}}$ lies within some interval, e.g. (3, 7). Given such an interval prior restriction, the $\tilde{r}$sMR would be obtained from one of three possible fitted tables calculated as above, where, in the fitted table:

(1) set $OR_{DC|\bar{E}} = 3$ if $\tilde{u}OR_{DC|\bar{E}} < 3$ in the data;
(2) set $OR_{DC|\bar{E}} = 7$ if $\tilde{u}OR_{DC|\bar{E}} > 7$;
(3) set $OR_{DC|\bar{E}} = \tilde{u}OR_{DC|\bar{E}}$ if $3 \leqslant \tilde{u}OR_{DC|\bar{E}} \leqslant 7$.

5(D). *Confounding as Bias Conditional on Approximate Ancillary Statistics*

How can decisions concerning confounding, e.g. the choice of r̃sMR over ũsMR in Examples A and B, be considered by some epidemiologists to be a bias rather than an efficiency issue? Consider Example B again, with $OR_{DC|\bar{E}} = 7$ known *a priori*, but $\tilde{u}OR_{DC|\bar{E}} = 1$; an unusual outcome has occurred. In repeated trials ũsMR is unbiased for sMR, where such trials, of course, include trials in which $\tilde{u}OR_{DC|\bar{E}} - OR_{DC|\bar{E}}$ is greater as well as less than zero.

But why should outcomes (such as $\tilde{u}OR_{DC|\bar{E}} - OR_{DC|\bar{E}} > 0$) that might have occurred, but did not occur in the actual study, influence our inference on sMR. Rather one might explicitly (as Miettinen and Cook do implicitly) restrict themselves to those subsets of all hypothetical repetitions in which the *difference* between the observed $DC|\bar{E}$ association and the known population $DC|\bar{E}$ association is the same as in the actual study. Let $\tilde{A}$ equal

$$\frac{\tilde{u}OR_{DC|\bar{E}} - 7}{\sqrt{\tilde{V}ar(\tilde{u}OR_{DC|\bar{E}})}},$$

where the denominator is an estimate of the variance of $\tilde{u}OR_{DC|\bar{E}}$ and $OR_{DC|\bar{E}} = 7$ is known *a priori*. $\tilde{A}$ is a standardized measure of the above difference, and is in large samples normally distributed with mean zero and variance one. Thus, $\tilde{A} = -2$ means that due to sampling variability, the observed data association is two standard deviations below its expected known value of 7. In hypothetical repetitions with $\tilde{A} = -2$ the observed proportion of unexposed subjects at level C who are diseased will systematically underestimate the corresponding population proportion, and likewise the observed proportion of unexposed subjects at level $\bar{C}$ who are diseased will systematically overestimate the corresponding population proportion. Algebra then shows that if there exists a positive E–C association, then ũsMR systematically overestimates sMR (as Miettinen and Cook suggest). Thus although ũsMR is unbiased over all the hypothetical trials, it is biased over the set of trials where $\tilde{A} = -2$ and is unsuitable for centering conditional confidence intervals, even though ũsMR remains consistent* conditional on $\tilde{A} = -2$.

A statistic such as $\tilde{A}$ which in large samples is $N(0, 1)$ and thus whose asymptotic distribution does not depend on unknown parameters, is called an approximate ancillary [5]. To observe $\tilde{A}$ required both prior knowledge that $OR_{DC|\bar{E}} = 7$ and data on $\tilde{u}OR_{DC|\bar{E}}$ (i.e. data on C).

Conditional on approximate ancillaries, unconditional inefficient estimators are, in general, biased, while efficient estimators such as r̃sMR are unbiased [see Section 8(B)]. Thus confounding in the presence of prior knowledge can be viewed either as an efficiency issue over all hypothetical repetitions or as bias if hypothetical repetitions are restricted to those with the same value of $\tilde{A}$ as the observed data. In fact ũsMR and r̃sMR have identical variance over the sets of trials where $\tilde{A} = -2$ [see Section 8(B)]. Thus conditionally it is only bias that distinguishes the statistical properties of ũsMR from r̃sMR.

Epidemiologists (such as Miettinen and Cook) who interpret confounding as a form of bias are implicitly imagining the same study repeated over and over again with the same outlier in terms of the $DC|\bar{E}$ association, and realize the "experiment" is biased. Until recently statisticians had not considered approximate ancillarity [5]. In fact $\tilde{A}$ is not even a function of the minimal sufficient statistic*, so it is no wonder statisticians did not consider conditional inference in Example B [see Section 8(D)].

5(E). *Conditional Precision and the Need for the Concept of Ancillarity*

Approximate ancillarity is an extension of the concept of exact ancillarity. To appreciate why the concept of conditioning on an ancillary statistic (introduced by Fisher [19]) is essential to frequentist statistics, consider the following problem adapted from Cox and Hinkley [20]: a fair coin is flipped. If heads is observed, a micrometer is used to measure an object of interest; if tails, a yardstick is used. The micrometer has a variance of 1 mm^2. The yardstick has a variance of 400 mm^2. The coin is flipped; heads appears; the micrometer is used; and the measurement is 50 mm.

Is the rod significantly longer than 45 mm? If we calculate the standard deviation of the measurement over the hypothetical repetitions of the entire experiment including the flip of the

coin, then the standard deviation is

$$\sqrt{\frac{400+1}{2}} \approx 14.1 \text{ mm},$$

since the yardstick is used 1/2 of the time. And thus 50 mm is well within 2 standard deviations of 45 mm. If we calculate the standard deviation over hypothetical repetitions conditional on obtaining heads and measuring with the micrometer in each trial, then the standard deviation is 1 mm, and the rod is deemed significantly longer than 45 mm. Common sense demands that we condition, since the fact that a different measuring instrument could have been used but was not seems irrelevant to the interpretation of the measurement. Thus we are conditioning on heads, the observed value of the coin toss in the experiment. The outcome heads is the ancillary statistic.

Note the similarities to our approximate ancillary. The probability of getting heads, like the probability of $\tilde{A} = -2$, does not depend on the parameter of interest. The observed heads tells us on which side of the average expected outcome our data lie—i.e. to the heads side of the two possible coin outcomes—similar to $\tilde{A}$. The only difference is that heads is an exact ancillary since the probability of observing heads does not depend on the parameter of interest even in small samples, while $\tilde{A}$ is only an approximate ancillary, since its distribution is independent of the parameter of interest only in large samples. In summary if the standard error of the restricted maximum likelihood estimator conditional on an ancillary differs from the unconditional standard error, conditional inference is required. The restricted maximum likelihood estimator is essentially identical conditionally and unconditionally.

In Examples A and B, the standard error of r̃sMR is the same unconditionally and conditional on $\tilde{A}$. Thus it is a matter of philosophical preference whether one views the choice of r̃sMR over ũsMR for unconditional efficiency considerations or conditional bias considerations since for large-sample tests and confidence intervals it makes no difference. Most statisticians philosophically prefer reduction of the data to the minimal sufficient statistic*, and thus choose the efficiency point of view.

In Example C below, the unconditional standard error of the restricted maximum likelihood estimator differs from the standard error conditional on the approximate ancillary, and thus conditioning on the approximate ancillary is essential (as in our measurement example above), if tests or confidence intervals are of interest [see Section 8(D)]. Thus bias conditional on approximate ancillaries, unlike unconditional efficiency, can provide a single unified approach to confidence interval confounding. Furthermore, the estimated variance of the restricted maximum likelihood estimator conditional on any approximate ancillary is obtained from the inverse of the observed information matrix, while the estimated unconditional variance is obtained from the inverse of the estimated expected information [5]. Technically these results refer to large sample inference accurate to order $N^{-3/2}$ in variance [see Section 8(C)]. The observed information is the matrix of second partial derivatives of the log likelihood evaluated at the restricted maximum likelihood estimator. The estimated expected information is the unconditional expectation of the matrix of second derivatives of the log-likelihood evaluated at the restricted maximum likelihood estimator. Thus, the appropriate 95% confidence intervals mentioned in Section 5(A) are to be formed from r̃sMR $\pm$ 1.96 times the standard error derived from observed information. If an investigator reports the restricted maximum likelihood estimator and its estimated variance obtained from the observed information matrix, then the crude estimate and observed crude confidence interval will be automatically reported where appropriate and not otherwise.

5(F). Confounding in a Follow-up Study Given No E–C Association in the Population

Example C

Consider crude data from a follow-up study (Table 4) in which 120 exposed and unexposed individuals were randomly sampled from a near infinite target population with sMR as parameter of interest. E and C are known to be unassociated in the target population and C is a suspected risk factor for disease. Since $\text{OR}_{EC} = 1$, cRR equals sMR. Therefore, c̃RR = 24/42 is appropriate for inference on sMR absent data on C. We now show that even though absent data on C one has performed perfectly valid inference based on c̃RR and its crude confidence interval, this inference is inadequate when additional data on C become available.

Table 4

	E	$\bar{E}$
D	24	42
$\bar{D}$	96	78
Total	120	120

Table 5

	C		$\bar{C}$		Total	
	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
D	4	2	20	40	24	42
$\bar{D}$	76	38	20	40	96	78
Total	80	40	40	80	120	120

Assume after results based on $\check{c}RR$ are published, a colleague provides data on C that had been collected from the study subjects at the time of the study but not coded onto the original data tapes. The complete data is given in Table 5.

Note that prior information on the E–C association in the target population does not influence the estimate of the stratum specific risk ratios and thus

$$\check{r}RR_c = \check{r}RR_{\bar{c}} = 1 = \tilde{u}RR_c = \tilde{u}RR_{\bar{c}}$$

and therefore

$$\check{r}sMR = 1.$$

Nevertheless, $\check{r}sMR \neq \check{c}RR$ because by a sampling fluke a large E–C association exists in the data ($\tilde{u}OR_{EC} = 4$). If the entire target population had been sampled, this discrepancy would not have occurred. In general, if C is a risk factor, the variability of $\check{r}sMR$ in large samples is less than that of $\check{c}RR$ (see Theorem 2 in Appendix C), and thus we choose $\check{r}sMR$ over $\check{c}RR$ for reasons of efficiency.

In addition

$$\tilde{B} = \frac{\tilde{u}OR_{EC} - 1}{\tilde{V}ar(\tilde{u}OR_{EC})}$$

is an approximate ancillary that measures the strength of the E–C association in the data. $\check{c}RR$ is biased conditional on $\tilde{B}$ while $\check{r}sMR$ is unbiased (see Appendix C) and thus we also choose $\check{r}sMR$ over $\check{c}RR$ in general for reasons of conditional bias. Furthermore unlike in Examples A and B, the variance of the appropriate adjusted estimator $\check{r}sMR$ differs depending on whether it is calculated over all hypothetical repetitions or only over those with the same value of $\tilde{B}$ [see Section 8(C)]. By the previous discussion of ancillarity, the conditional variance (derived from the observed information) is required for confidence interval construction. If we viewed confounding as only a matter of unconditional efficiency, and thus used the unconditional standard error of $\check{r}sMR$, confidence intervals for sMR would not depend on the E–C association in the data. This is in direct contrast to our intuition. For example, if the observed rates of disease at each joint level of E and C remained unchanged but the E–C association had been even more marked than in Table 5, e.g. the number of individuals who were $C\bar{E}$ or $E\bar{C}$ had been 8, say, rather than 40, it is intuitively clear and confirmed using the observed information (the estimator of the conditional standard error) that our estimates of $p(D \mid C, \bar{E})$, $p(D \mid \bar{C}, E)$, and sMR would be less precise.

If no E–C association had existed in the data (i.e. $\tilde{B} = 0$), then $\check{r}sMR = \tilde{u}sMR = \check{c}RR$, and the crude estimator is appropriate for all data outcomes for which $\tilde{B} = 0$. But the variance of $\check{c}RR$ must be calculated conditionally because in general when C is a risk factor, $\mathrm{Var}(\check{c}RR \mid \tilde{B} = 0) = \mathrm{Var}(\check{r}sMR \mid \tilde{B} = 0) < \mathrm{Var}\check{c}RR$ (see Theorem 2 in Appendix C). Thus the usual crude confidence interval based on $\check{V}ar\check{c}RR$ will give falsely wide conditional confidence intervals, i.e., the confidence interval will cover the true parameter over 95% of the time in repeat trials in which $\tilde{B} = 0$. Thus once data on C have been acquired, our inference must be based on $\check{r}sMR$ and its conditional standard error.

5(G). Confounding in the Absence of Data on C *Revisited*

Even when we have prior knowledge of the association of C with E or D in the population, if data on C are unavailable, we cannot observe the degree to which the association of C with E or D in the data differs from the known population association. In other words, the value of the approximate ancillary associated with the available prior knowledge will be unobserved. Thus, absent data on C, our definition of confounding given in Section 5(A) becomes "C will be a nonconfounder in the data at hand if and only if the crude estimate equals the observed value of

an unconditionally unbiased estimator for the parameter of interest". But absent data on C, the only possible candidate for that unbiased estimator is the crude estimator itself. Thus, absent data on C, C will be a nonconfounder in the data at hand if and only if C is a nonconfounder for all possible data outcomes, i.e., *the crude estimator is unconditionally unbiased. But this is exactly the definition of confounding in the absence of data on* C *given in Section 4.* In this case, the crude confidence interval will be appropriate as well, and furthermore the crude estimator will be the efficient estimator based on the observed data.

5(H). A Critique of Conditioning on Unobserved Ancillaries

Some epidemiologists [1], in disagreement with our approach, implicitly suggest that a proper approach to confounding requires that we consider, at least hypothetically, conditioning on unobserved approximate ancillaries. We will show Miettinen and Cook implicitly considered such an approach in their appendix. In this section we critique the idea of conditioning on unobserved ancillaries. In the next subsection we apply our critique to answer a question raised by Miettinen and Cook in their Appendix. An approach to confounding that conditions on unobserved ancillary statistics would interpret Example C as proof that given $\mathrm{OR}_{EC} = 1$, C can still be a "confounder absent data on C". One was simply unable to assess the strength of this "confounding" by C until the data on C were observed. This would be a very unhelpful definition of "confounding in the absence of data on C", in that it tends to suggest that inference on the sMR should not be performed from the crude data even if $\mathrm{OR}_{EC} = 1$ is known *a priori*, since, if C was not observed, it is always possible that unknown to the investigator a strong E–C association might exist in the data. The validity of all randomized trials would then dissolve. A better point of view is to realize that given $\mathrm{OR}_{EC} = 1$, the price we pay for not observing C and thus $\tilde{B}$ is reflected in the increased length on average of the crude confidence interval compared to that of the stratified confidence interval based on r̃sMR. The greater length of the crude confidence interval reflects precisely our additional uncertainty about sMR because, unknown to us, a strong E–C association might exist in the data. In fact, asymptotically

$$\mathrm{Var}(\tilde{\mathrm{c}}\mathrm{RR}) = \mathrm{Var}[E(\tilde{\mathrm{c}}\mathrm{RR} - \mathrm{cRR} \mid \tilde{B})] + E[\mathrm{Var}\tilde{\mathrm{c}}\mathrm{RR} \mid \tilde{B}],$$

where $E(\tilde{\mathrm{c}}\mathrm{RR} - \mathrm{cRR} \mid \tilde{B})$ is the asymptotic conditional bias of c̃RR due to the unknown E–C association in the data represented by $\tilde{B}$. The above formula demonstrates that uncertainty concerning the magnitude of the conditional bias of c̃RR explicitly contributes to the variability of c̃RR around sMR. When we report a crude estimate, we know from the outset that in 5% of samples our estimate is far from sMR (i.e. more than 2 standard errors) due in part to data sets with a strong but unobserved E–C association. But even so, absent data on C, we have neither the "need nor desire" to adjust c̃RR, for absent data on $\tilde{B}$ in which direction would we adjust it?

Instead if, as we recommend, one considers conditioning only on observed ancillaries, then absent data on C and therefore on $\tilde{B}$ as well, we have no ancillary to condition on. Thus, absent data on C but given $\mathrm{OR}_{EC} = 1$, c̃RR is unbiased conditional on all observed ancillaries, and is therefore appropriate for all data outcomes. Alternatively, from the point of view of efficiency, c̃RR is the efficient estimator absent data on C and thus is appropriate. But, given data on C, c̃RR is both inefficient compared to r̃sMR and biased conditional on $\tilde{B}$.

5(I). Resolution of a Question Posed by Miettinen and Cook in Their Appendix

By using an argument analogous to that given in the last subsection, we demonstrate that the concern expressed by Miettinen and Cook in Example B of their Appendix (that the restricted maximum likelihood estimator might, in some cases, not be the appropriate estimator) is unfounded, even when interest is in O/EX of the sampled population in the deterministic superpopulation model discussed in Section 3. In their Example B, Miettinen and Cook contend that in a follow-up study even if a covariate is known *a priori* not to be a risk factor in the unexposed (e.g. C as defined in Example A) it would be a confounder if by chance it is associated with the outcomes that would have been observed if all subjects had been unexposed. To see this, define approximate ancillaries, $\tilde{A}_1$, $\tilde{A}_2$ that measure the degree to which the DC association in the unexposed sample, and in the exposed sample when unexposed respectively in the data differs from its *a priori* known value of 1 in the superpopulation. To observe $\tilde{A}_2$ we must observe the outcomes

Table 6

	C		$\bar{C}$		Total	
	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
D	1	4	2	4	3	8
$\bar{D}$						
Total	8	8	4	4	12	12

Table 7

	C		$\bar{C}$		Total	
	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
D	1	2	2	4	3	8
$\bar{D}$						
Total	8	8	4	4	12	12

of the exposed when unexposed. Although we can calculate $\tilde{A}_1$ given data on C, data on $\tilde{A}_2$ will of course never be observed. We cannot condition on an unobserved ancillary statistic $\tilde{A}_2$. Conditional on only $\tilde{A}_1$, the observed ancillary, the unbiased efficient estimator is $\check{c}$RR as in Example A. The variance of $\check{c}$RR reflects in part our uncertainty about O/EX due to lack of knowledge of $\tilde{A}_2$. Absent data on $\tilde{A}_2$ we have neither the need nor desire to adjust $\check{c}$RR, for in which direction would we adjust it? Since $\check{c}$RR is appropriate, C is a nonconfounder given data on C (but absent data on the outcomes of the exposed when unexposed).

5(J). The Failure of Miettinen and Cook's Principle 4

In Example C, if the risk ratio was also known *a priori* to be constant over levels of C, then even if no E–C association exists in the data, $\check{r}$RR (the restricted maximum likelihood estimator incorporating this further prior restriction) may not equal $\check{c}$RR. Although this contradicts Miettinen and Cook's principle that, if in a follow-up study no E–C association exists in the data, then C is not a confounder (Principle 4), Miettinen and Cook did not consider prior restrictions on the variability of the risk ratio over strata, but rather only on the $DC\,|\,\bar{E}$ and E–C associations. Nonetheless, Principle 4 still fails even within the class of restrictions they considered. For example, given the data from a follow-up study in Table 6, $\check{c}\mathrm{RR} = 3/8$, and no E–C association exists in the data. Nevertheless given the prior knowledge $\mathrm{RR}_{DC|\bar{E}} = 2/8$, the fitted table is Table 7. Therefore, $\check{r}\mathrm{sMR} = (1/2) \neq (3/8) = \check{c}\mathrm{RR}$.

Since we feel use of the restricted maximum likelihood estimator as the appropriate estimator is both necessary for a unified and consistent approach to confounding given data on C, and is implicit in the Miettinen and Cook approach, Principle 4 should be abandoned. (In the above example if our prior knowledge was in terms of $\mathrm{OR}_{DC|\bar{E}}$ rather than $\mathrm{RR}_{DC|\bar{E}}$, then $\check{c}\mathrm{RR} = \check{r}\mathrm{sMR}$ whenever no E–C association exists in the data.)

Principles 1–4 do not consider the implications of simultaneous prior knowledge that $\mathrm{OR}_{DC|\bar{E}} = 1$ and $\mathrm{OR}_{EC} = 1$ in a follow-up study. If an E–C association exists in the data, $\check{c}\mathrm{RR} \neq \check{r}\mathrm{sMR}$, where $\check{r}\mathrm{sMR}$ is the restricted maximum likelihood estimator incorporating both restrictions, and thus C is a confounder in the presence of data on C. Nonetheless, in contrast to Examples B and C above, the absolute but not relative magnitudes of the confounding (i.e. the difference between $\check{c}$RR and $\check{r}\mathrm{sMR}$) in the above counterexamples to Principles 1 and 4 remain bounded for all data outcomes. Furthermore, without substantial calculation, even the direction of the bias is not apparent. Therefore, the intuition of Miettinen and Cook had some basis for overlooking such counterexamples.

5(K). Mean Square Error, Confidence Intervals and Confounding

Example C also provides a clear demonstration that in Miettinen and Cook's as well as our own view of confounding, appropriate estimators are estimators that, given the available prior knowledge, can center large-sample conditional confidence intervals (i.e. conditionally unbiased estimators) irrespective of whether an investigator with particular prior beliefs might prefer a different confounded estimator for a decision problem based on mean square error loss.

Suppose in Example C, C was a covariate about which the investigator held the prior belief that $\mathrm{RR}_{DC|\bar{E}}$ was probably in the interval (1.01, 1.1) but might possibly lie outside that interval, and suppose also the risk ratio was known to be constant over levels of C. Although in large samples the conditional mean square error of $\check{r}$RR is usually less than the conditional mean square error of the conditionally biased $\check{c}$RR, nevertheless for data sets with a strong chance E–C association, $\check{c}$RR will be closer on average to the true risk ratio (i.e. have smaller mean square error) than will $\check{r}$RR if in fact $1.01 < \mathrm{RR}_{DC|\bar{E}} < 1.1$ where the average is computed over hypothetical repetitions in which either the E–C margin or $\tilde{B}$ is fixed at its observed values. This result will not hold if $\mathrm{RR}_{DC|\bar{E}}$

lies far outside the interval. Thus for point estimation, an investigator with the above prior beliefs would select the confounded (i.e. conditionally biased) $\check{c}RR$ over $\check{r}RR$.

Furthermore even for the investigator interested in mean square error loss in light of his prior beliefs, the restricted maximum likelihood estimator treating $1.01 \leqslant RR_{DC|\bar{E}} < 1.1$ as known *a priori* would probably be preferred over either $\check{c}RR$ or $\check{r}RR$, although it, too, like the $\check{c}RR$ is inappropriate for centering confidence intervals in the absence of true prior knowledge that $RR_{DC|\bar{E}}$ lies in the above interval.

5(L). *Case-control Studies Given Data on* C

Example D. Steady state incidence density case-control study

Consider the case-control data in Table 2 where incident cases and controls have been sampled from a stable population with sIDR as the parameter of interest.

If E and C are known *a priori* to be unassociated in those at risk, then $OR_{EC|\bar{D}} = 1$ and cOR = sIDR = sOR. Both the usual stratified estimator $\tilde{u}sOR$ and $\check{c}OR$ unbiasedly estimate sIDR over all hypothetical repetitions. Yet

$$\tilde{u}sOR = \frac{20 + 10}{60\left(\frac{20}{20}\right) + 30\left(\frac{10}{70}\right)} = \frac{21}{45} \neq \check{c}OR = 1$$

in the data at hand. Which is the appropriate estimator?

The fitted table utilizing the information $OR_{EC|\bar{D}} = 1$ is just Table 3. $\check{r}sOR$ calculated from Table 3 equals $\check{c}OR$. In fact $\check{r}sOR = \check{c}OR$ for all possible data outcomes, and $\check{c}OR = \check{r}sOR$ has variance strictly less than $\tilde{u}sOR$ if C is a risk factor for disease (see Appendix D). Furthermore, conditional on

$$\tilde{A} = \frac{\tilde{u}OR_{EC|\bar{D}} - 1}{\tilde{V}ar(\tilde{u}OR_{EC|\bar{D}})}$$

(an approximate ancillary given $OR_{EC|\bar{D}} = 1$), if $\tilde{A} \neq 0$, $\tilde{u}sOR$ is biased, while $\check{r}sOR = \check{c}OR$ is unbiased. $\check{c}OR$ remains the efficient estimator even if controls had been matched to cases on level of C.

If prior knowledge that $OR_{DC|\bar{E}} = 1$ rather than $OR_{EC|\bar{D}} = 1$ was available, then again cOR = sOR and in an unmatched (but not in a matched) study $\check{c}OR = \check{r}sOR$. Thus, in both case-control and follow-up studies, it is prior knowledge of the conditional independence of C and E or D (i.e. $OR_{EC|\bar{D}} = 1$ or $OR_{DC|\bar{E}} = 1$) that justifies the crude analysis irrespective of associations in the data. Prior knowledge of the marginal independence of C and E ($OR_{EC} = 1$), as in Example C, requires a stratified analysis.

If the odds ratio is known to be constant over strata, in small samples confidence intervals should be exact and in sparse data point estimates should be based on the conditional (matched logistic) maximum likelihood estimator [21]. If either $OR_{EC|\bar{D}} = 1$ or $OR_{DC|\bar{E}} = 1$, the exact confidence interval and conditional maximum likelihood estimator from the crude table derived by conditioning on the total number of exposed individuals in the study are superior in terms of confidence interval length and sampling variability respectively to the exact confidence interval and conditional maximum likelihood estimator from the stratified analysis derived from conditioning on the full margins [see Section 6(C)].

Example E. Cumulative incidence case-control study given data on C

If the data in Table 2 had arisen from a cumulative incidence type case-control study, any prior knowledge concerning a lack of an E–C association is usually of the form that E and C are unassociated in the source population ($OR_{EC} = 1$) rather than of the form $OR_{EC|\bar{D}} = 1$. Nonetheless if $OR_{EC} = 1$ and if the disease is rare, then $OR_{EC|\bar{D}} \approx 1$. Thus, although the available prior knowledge is $OR_{EC} = 1$, if the use of the approximation $OR_{EC|\bar{D}} = 1$ is valid, the above results for the sIDR demonstrate that the crude analysis would be appropriate irrespective of associations observed in the data. Given data on C, if the disease is rare, the approximation $OR_{EC|\bar{D}} = 1$ can

be used when the number of controls sampled represents a small fraction of potential controls (e.g. controls are sampled from a near infinite population).

The importance of this condition is illustrated by consideration of a case-control study within a cohort. Consider a double blind clinical trial involving 10,000 subjects followed for 3 y. The efficacy of a certain drug (C) was tested for its ability to prevent first occurrence of a disease (D). Subjects were randomized into treated (C) and placebo ($\bar{C}$) groups so that each subject had a 33% chance of getting the drug. Unfortunately, the results showed that the drug significantly increased the risk of the disease. At the end of the trial, the investigators were told of a new hypothesis linking another exposure (E) with the same disease. To test this hypothesis in their study population, the investigators conducted a nested case control study, comparing all 120 observed cases of D with an equal number of noncases randomly sampled from the total cohort. Exposure histories were obtained from all 240 subjects (but no others), and the results are given in Table 2.

Given $\mathrm{OR}_{EC} = 1$, the ability to estimate the probability of disease in the population as 120/10,000 allows valid estimation of cRR = sMR by Bayes theorem from both the crude data alone and the full data. This remains true regardless of the rarity of the disease [12]. [Given data on C, if the disease had not been rare (so that $\mathrm{OR}_{EC|\bar{D}} \not\approx 1$ even though $\mathrm{OR}_{EC} = 1$), then r̃sMR, the efficient estimator calculated via Bayes theorem that uses data on C, can differ substantially from either c̃OR or the crude estimator of cRR = sMR calculated from the crude data via Bayes theorem. Thus, C is a confounder given data on C, demonstrating that Principle 1 of the Introduction is false in this instance.] If, as in our data, the disease is rare and only a small fraction of potential controls have been sampled, then c̃OR and its crude variance will be excellent approximations (although not exactly equal) to r̃sMR and its variance [see Section 6(F)], and inference based on the crude data is appropriate.

To demonstrate this result, we note that if the disease is rare in the data, the controls to a good approximation may be considered a random sample from the entire population of 10,000. Although a strong E–C association exists in the nondiseased controls ($\tilde{u}\mathrm{OR}_{EC|\bar{D}} = 7$), a strong E–C association in a random sample from a randomized population provides no information on the unsampled individuals. To see this intuitively, imagine that the random sample is selected prior to randomization and furthermore that the sampled individuals are randomized to C or $\bar{C}$ by the flip of a coin an hour before the unsampled individuals. Thus with a high probability almost no E–C association exists in the 9760 unsampled potential controls, and thus almost no E–C association in the total group of 9880 potential controls (because 120 is such a small fraction of the total number of eligible controls). Although a test for no E–C association in the 120 nondiseased controls rejects the null hypothesis at the $p < 0.01$ level, this is only evidence for a rare outcome under the null hypothesis; it is not evidence that the null hypothesis of almost no E–C association among the 9880 potential controls is false. Thus, $\mathrm{OR}_{EC|\bar{D}} = 1$ is a much better estimate than $\mathrm{OR}_{EC|\bar{D}} = 7$ of the association in the entire cohort of 10,000 and the appropriate fitted table is nearly Table 3. Therefore, r̃sOR ≅ c̃OR. But r̃sOR ≅ r̃sMR if the disease is rare.

On the other hand, if we take more controls per case, it is no longer necessarily true that r̃sMR ≅ c̃OR. In the limit we would take all potential controls, and observe exactly the E–C association in the total population of 10,000 (i.e. we would have the precise equivalent of the full follow-up study). When all controls are sampled, with high probability no E–C association will be present in the 10,000 due to randomization. In such an event, given that the disease is rare, estimates and confidence intervals for sMR (or equivalently sOR) will be nearly the same in the stratified and collapsed analysis, e.g. r̃sMR ≅ c̃OR ≅ c̃RR [see Section 6(E)]. But the one time in 100,000,000,000 that chance randomization failure produced a large E–C association in the 10,000, we find that r̃sMR ≇ c̃OR ≅ c̃RR. Since we have the equivalent of a follow-up study when all controls are sampled, the proper analysis in such an event is the stratified estimate, r̃sMR, as in Example C above.

5(M). Fixed Cohort Study with Incidence Density Ratio as Parameter of Interest Given Data on C

If C is known *a priori* not to be a risk factor, then the most efficient estimator of the (stratum-specific) incidence density ratio (assumed constant over time and level of C) in follow-up or time- matched case control studies will depend only on the crude data, and so C will be a nonconfounder. If $\mathrm{OR}_{EC} = 1$ is known *a priori* at start of follow-up and C is a risk factor, and if

the cumulative incidence of disease is not small, data on C, if available, will be used to efficiently estimate the incidence density ratio in both follow-up studies and time matched case-control studies. Thus C will be a confounder for the incidence density ratio given C in both the follow-up and case-control study, in conflict with Principle 1 of the Introduction.

On the other hand in a time-matched case control study in which only a small fraction of potential controls have been sampled, if the cumulative incidence of disease at each joint level of E and C is small (and no significant nondisease related loss to follow-up occurs) over the study period, then given $\mathrm{OR}_{EC} = 1$ in the superpopulation, $\tilde{c}\mathrm{OR}$ and its associated crude confidence intervals obtained by collapsing over both time and level of C will to an excellent approximation give the appropriate adjusted point estimates and confidence intervals for the incidence density ratio. Thus C will not be a confounder for the incidence density ratio in a time-matched case control study if $\mathrm{OR}_{EC} = 1$ at start of follow up is known *a priori*, the disease is rare, and a small fraction of potential controls has been sampled.

5(N). Confounding Absent Prior Knowledge—the Failure of the Change in Estimate Criterion

Example F. Cumulative incidence case-control study in the absence of prior knowledge

We have justified the crude analysis in case-control studies when we had prior knowledge of either no E–$C \mid \bar{D}$ or no D–$C \mid \bar{E}$ association in the population irrespective of associations found in the data. Does the presence of "no confounding in the data" justify the crude analysis in the absence of prior knowledge?

Consider the data in Table 8 obtained from a case-control study with OR known constant over strata. Note $\tilde{u}\mathrm{OR} = \tilde{c}\mathrm{OR} = 4$ because no disease-covariate association exists in the data conditional on $\bar{E}$ (i.e. $\tilde{u}\mathrm{OR}_{DC|\bar{E}} = 1$), and the odds ratio is constant over strata in the data.

Suppose we fit a logistic regression model with D as dependent variable:

$$\ln\left[\frac{p(D \mid E, C)}{1 - p(D \mid E, C)}\right] = B_0 + B_e E + B_c C,$$

where $E = 1$ if E, $E = 0$ if $\bar{E}$, $C = 1$ if C, and $C = 0$ if $\bar{C}$; then $\tilde{B}_e = \ln 4 = \ln \tilde{u}\mathrm{OR}_{ED|\bar{C}}$ and $\tilde{B}_c = \ln 1 = \ln \tilde{u}\mathrm{OR}_{DC|\bar{E}} = 0$.

Many authors consider this a case of "no confounding in the data" and suggest the collapsed analysis. The reasons generally stated for this view include:

1. The estimate of the odds ratio does not change on collapsing or equivalently on dropping C from the logistic regression.
2. In logistic regression, a variable with $\tilde{B}_c = 0$ should be dropped.
3. If the null hypothesis $B_c = 0$ is not rejected at the $p < 0.05$ level, drop C from the regression, i.e. perform backward elimination.

Although the backward elimination has been widely criticized both earlier in this paper and elsewhere, the first and second reasons are still often advocated.

Our object in epidemiologic studies is in general to find a point estimate of the parameter of interest and confidence interval. Miettinen and Cook's concept of confounding only discusses whether the crude point estimate of effect requires adjustment. In our example $\tilde{c}\mathrm{OR} = \tilde{u}\mathrm{OR}$ so the crude estimate equals the appropriate stratified estimate in the data at hand (although not in all possible data outcomes), and thus does not require adjustment.

However, the usual confidence interval for the odds ratio calculated from the collapsed table based on $\ln(\tilde{c}\mathrm{OR})$ is (1.43, 10.1), while the confidence interval from the stratified analysis, based on $\ln(\tilde{u}\mathrm{OR})$ and its standard error, is (0.66, 23). Exact confidence intervals give similar results. Many data analysts look with pleasure on the improvement in precision (narrowing of the confidence interval) obtained by dropping the apparent "nonconfounder" C. An insignificant result has become significant, and thus we can publish a positive result by dropping such a non-confounder.

But within sampling variation our data could have been sampled from a population with the expected values given in Table 9, where $\mathrm{OR}_{DC|\bar{E}} = 49/16$, the common $\mathrm{OR} = \mathrm{OR}_{ED|\bar{C}} = 1$, $\mathrm{cOR} = (81/51)^2 \neq$ the common OR.

Table 8

	C		$\bar{C}$		Total	
	D	$\bar{D}$	D	$\bar{D}$	D	$\bar{D}$
E	20	10	2	1	22	11
$\bar{E}$	1	2	10	20	11	22
Total	21	12	12	21	33	33

Table 9

	C		$\bar{C}$		Total	
	D	$\bar{D}$	D	$\bar{D}$	D	$\bar{D}$
E	$\frac{77}{4}$	11	1	$\frac{7}{4}$	$\frac{81}{4}$	$\frac{51}{4}$
$\bar{E}$	$\frac{7}{4}$	1	11	$\frac{77}{4}$	$\frac{51}{4}$	$\frac{81}{4}$
Total	21	12	12	21	33	33

In fact if in the above D was lung cancer, C was mild cigarette smoking, and E was yellow fingers, we would infer from the collapsed analysis that yellow fingers is a risk factor for cancer. The stratified analysis correctly shows the data cannot discriminate between the effect of cigarettes and yellow fingers, since confidence intervals for both estimates include the null value [CI for $\mathrm{OR}_{DC|\bar{E}}$ *is* (0.17, 5.9)]. Note that either E or C does determine D since, although neither $\tilde{B}_e$ nor $\tilde{B}_c$ is significant in the presence of the other, the hypothesis that both $\tilde{B}_c = 0$ and $\tilde{B}_e = 0$ is rejected at $p < 0.01$. This is a typical example of multi-collinearity in regression. E and C are highly correlated ($\tilde{\mathrm{u}}\mathrm{OR}_{EC|\bar{D}} = 100$) and thus the data are insufficient to distinguish their independent effects, although clearly one or the other or both have an effect. Only prior knowledge that the effect of either E or C is negligible can allow precise estimation of the other. Thus, dropping C from the model is equivalent to an *a priori* decision that C is not a risk factor for disease. In the absence of such knowledge, the stratified confidence intervals are telling us all that the data can tell us.

The crude confidence interval gives a confidence interval for cOR. Thus it is not a confidence interval for the common odds ratio (the parameter of interest) unless cOR equals the common odds ratio. But cOR = common OR if and only if $\mathrm{OR}_{EC|\bar{D}} = 1$ or $\mathrm{OR}_{DC|\bar{E}} = 1$ in the population. Therefore we must ask "what is the evidence that $\mathrm{OR}_{EC|\bar{D}} = 1$ or $\mathrm{OR}_{DC|\bar{E}} = 1$?" if we wish to report the crude confidence interval. Our only evidence is that $\tilde{\mathrm{u}}\mathrm{OR}_{DC|\bar{E}} = 1$. But, since the confidence interval for $\mathrm{OR}_{DC|\bar{E}}$ is (0.17, 5.9), $\mathrm{OR}_{DC|\bar{E}}$ could easily within sampling variability be 4 rather than 1. Thus without prior knowledge we have no justification for treating $\mathrm{OR}_{DC|\bar{E}}$ as if it were 1.

If on the other hand prior knowledge (e.g. $\mathrm{OR}_{DC|\bar{E}} = 49/16$) was available, using GLIM 3 [18] we can calculate both the restricted maximum likelihood estimator and its estimated standard error (the observed and estimated expected information are identical in this example). If a logistic regression program is used, prior knowledge of $\mathrm{OR}_{EC|\bar{D}}$ can only be incorporated if E is the dependent variable, since a logistic model with D as dependent variable conditions on the E–C association in the data (and thus $\mathrm{OR}_{EC|\bar{D}}$ is not part of the model). Conversely prior knowledge of $\mathrm{OR}_{DC|\bar{E}}$ requires D as the dependent variable. Log linear models can incorporate either type of prior knowledge. If $\mathrm{OR}_{DC|\bar{E}}$ is known only to lie in the interval (2, 4), the restricted maximum likelihood estimator can be obtained as shown in Example B above. Confidence intervals incorporating the prior interval restriction on $\mathrm{OR}_{DC|\bar{E}}$ are narrower on average than the unrestricted confidence interval; such confidence intervals can be calculated using either Efron's bootstrap or jackknife techniques [22].

5(*O*). *Confidence Interval Confounding*

If we plan to report both a point estimate and a confidence interval, we should broaden the definition of confounding to guard against the temptation to incorrectly drop C (as in the above example). Specifically, C *is a confidence interval* (*CI*) *nonconfounder* if and only if the crude point estimate of the effect equals an appropriate adjusted estimate and the crude observed confidence interval (calculated from the crude estimate and its crude estimated standard error) equals an appropriate adjusted observed confidence interval (calculated from the observed values of the restricted maximum likelihood estimator and an estimator of its standard error derived from the observed information matrix).

The crude confidence interval ignoring C should be reported if and only if there is no CI confounding. Below we detail situations in which, given data on C, no point estimate confounding exists, and show which of these also display no CI confounding. Mathematical verifications of the following propositions are quite straightforward.

In a case control study without prior knowledge and with sOR as the parameter of interest if one but not both of $\tilde{u}OR_{EC|\bar{D}}$ and $\tilde{u}OR_{DC|\bar{E}}$ equal one in the data, we will have no point estimate confounding (since $\tilde{c}OR = \tilde{u}sOR$), but we will have CI confounding and the observed crude confidence interval will generally be falsely narrow as in Example F. If the odds ratio is known *a priori* to be constant over strata, point estimate nonconfounding generally requires the odds ratio to be constant over strata in the data.

Likewise, in a follow-up study without prior information and with sMR as the parameter of interest, if $\tilde{u}OR_{DC|\bar{E}} = 1$ (C does not appear to be a risk factor in the data), but an E–C association exists in the data, there is no point estimate confounding, but there is CI confounding. The observed crude confidence interval is, in general, falsely narrow.

In a case-control study given $OR_{EC|\bar{D}} = 1$ *a priori* or in a case control or follow-up study given $OR_{DC|\bar{E}} = 1$, we have no CI confounding as shown in Examples A and D for all data outcomes. This holds whether or not the stratum specific odds ratios in case-control studies or risk ratios in follow-up studies are known to be constant across strata.

In a follow-up study with sMR as parameter of interest, given no E–C association in the data, we have no point estimate confounding, but unless we know $OR_{DC|\bar{E}} = 1$ we have CI confounding with a falsely wide observed crude confidence interval. In general, this remains true in the presence of priori knowledge of no $E - C$ association in the target population if standard errors are calculated (as they should be) conditional on the approximate ancillary $\tilde{B}$ (as shown in Example C).

5(P). *Efficient Estimators Other Than the Maximum Likelihood Estimator*

Given the data from Example A with $OR_{DC|\bar{E}} = 1$ known *a priori*, data analysts using efficient methods of estimation other than maximum likelihood (e.g., weighted least squares or minimum Chi-square methods) would report a restricted estimate different than $\tilde{c}RR = \tilde{r}sMR$. All three methods give identical crude estimates and unrestricted estimates since all three methods are Fisher consistent*. We still call C a point estimate nonconfounder, since $\tilde{c}RR$ does equal some, although not all, appropriate estimators (i.e. $\tilde{r}sMR = \tilde{c}RR$) in the data at hand. Furthermore, whenever $\tilde{r}sMR \neq \tilde{c}RR$ in the observed data, the restricted estimate calculated under the other methods will also in general differ from $\tilde{c}RR$. Therefore the restriction of our previous discussion to maximum likelihood estimation is adequate.

Bayes' estimators, although efficient, are unlikely to be accepted by epidemiologists as intuitively appropriate estimators for the evaluation of confounding. For example, in the absence of true prior knowledge, a Bayes estimator of the sMR will differ from $\tilde{r}sMR$, the intuitively appropriate estimator, usually in the direction of one's prior beliefs about the magnitude of the sMR. Intuitive rejection of Bayes' estimators as appropriate implies that, to epidemiologists, appropriate estimators are not only efficient, but also Fisher consistent.

The Mantel–Haenszel estimator of a common odds ratio is Fisher consistent, but inefficient. Although by our definitions, it is technically confounded, it is so nearly efficient [21] that the magnitude of confounding (i.e. the magnitude of the asymptotic bias conditional on an ancillary statistic that measures the degree to which the OR is not constant over strata in the data) is usually negligible.

5(Q). *Implications for Design and Data Analysis*

If C is a confounder for the parameter of interest in the absence of data on C, data on C must be collected for valid inference.

If, given data on C, C is a CI nonconfounder for all possible data outcomes, then data on C, even if collected, would not be useful for the analysis.

If C is a nonconfounder absent data on C, but a confounder given data on C (e.g. a cumulative incidence follow-up study with $OR_{EC} = 1$), then although data on C are not necessary, if they are collected, inference on the sMR or a common RR will usually be more precise in large samples.

Once data on C have been collected, inference must be performed conditional on the observed E–C association even if by chance a large E–C association in the data produces conditional

confidence intervals based on r̃RR (assuming a constant RR across strata) or r̃sMR (not assuming constancy) of greater length than the unconditional crude confidence interval.

5(R). Sparse Follow-up Data—the Case of Many Covariates

If in an unmatched follow-up study, data on p rather than just two potential risk factors (each at two levels) is obtained, the data will consist of 2^p rather than two 2×2 tables of E by D. For a fixed sample size as data are collected on more and more covariates (i.e. $p \to \infty$), even though E is unassociated with the covariates in the population, (e.g. a randomized trial), it is almost certain that none of the 2^p tables will contain simultaneously an exposed and unexposed case, and r̃RR (we assume here the risk ratio is constant over strata) will be undefined (i.e. its variance conditional on the observed exposure-covariate association can be regarded as infinite). Therefore, in direct contrast to the case where the sample size to covariate ratio is large, inference on the RR based on the unconditional crude confidence interval (which is appropriate if the covariate data had not been obtained) would almost certainly be more precise than inference based on conditional confidence intervals centered on r̃RR. This result is not simply a reflection of the failure of the normal approximation to the conditional distribution of r̃RR since identical results obtain if the outcome, exposure, and all covariates are jointly multivariate normally distributed with exposure independent of all covariates. The exact conditional confidence interval for the partial regression coefficient of the outcome on exposure controlling for all p measured covariates will be wider than the crude unmatched confidence interval for the regression coefficient of outcome on exposure with probability approaching 1 as $p \to \infty$. Rather the reason why the width of the exact conditional confidence interval exceeds that of the crude confidence interval when data on many covariates have been obtained is that the exact conditional confidence interval must cover the true coefficient for the exposure effect 95% of the time along every possible exposure–covariate margin while the crude confidence interval need only cover the exposure coefficient 95% of the time unconditionally. Thus, for exposure–covariate margins that occur with small probability, the coverage rate of the crude confidence interval is unconstrained and may be either much less or much greater than 95%.

Therefore, if Fisher's argument for conditioning on ancillary statistics is accepted, frequentist inference naturally leads, in sparse data, to conditional confidence intervals that are so wide as to be useless for inferential purposes. Additional problems for a conditional frequentist approach to inference in sparse data arise if, following Buehler [4], we require adherence to a continuity principle of inference which requires that continuous small perturbations in a model (i.e. in our assumed prior knowledge) give continuous small perturbations in the inferences drawn and actions chosen based on a given data set (assuming a continuous action space).

Together the conditionally and continuity principles require that in Example C, rather than performing unconditional inference when no exact ancillary exists, one must condition on an approximate ancillary such as $\tilde{B}$. Otherwise, in conflict with the continuity principle, one would abruptly jump from reporting the variance of r̃sMR based on the observed information (the conditional variance) when the risk ratio was known to be exactly constant over strata (and thus an exact ancillary exists) to reporting the variance based on the expected information (the unconditional variance) when the risk ratio was known to be nearly but not exactly constant over strata (and no exact ancillary exists). See Section 8(D).

Although we have acted as if corresponding to the prior knowledge that $\mathrm{OR}_{EC} = 1$ in Example C, a unique approximate ancillary (i.e. $\tilde{B}$) existed, in actuality, other statistics (e.g. the likelihood ratio statistic for testing $\mathrm{OR}_{EC} = 1$) are also approximate ancillaries [5]. Nonetheless in large samples, for almost all data outcomes, the observed value of any of the various approximate ancillaries would be nearly equal. Thus, in large samples, no difficulty arises since approximate ancillary statistics are essentially unique.

But when data on a large number of covariates have been obtained (i.e. sparse data) the distribution of any approximate ancillary, rather than being $N(0, 1)$, can depend strongly on the parameter of interest, and furthermore, the observed values of various approximate ancillary statistics may differ markedly from one another. Therefore in sparse data we cannot uniquely determine which approximate ancillary to condition on. Thus we cannot uniquely determine the set of hypothetical repetitions on which to base our inferences.

These difficulties with a conditional frequentist approach to inference strongly suggest that one consider a Baysian approach to statistics.

5(S). Summary of Section 5 Results

If our prior knowledge was such that C was a CI confounder for all possible data outcomes, it would be fruitless to collect data on C. In such a case, we will call C a design nonconfounder.

For all unmatched study designs and for all effect parameters considered in this section (i.e. sMR, sOR, IDR), if C is known *a priori* not to be a risk factor for disease in the unexposed, then C is a design nonconfounder except as discussed in Section 5(J) on "Failures of Principle 4". In case-control studies incorporating matching on C, C would remain a point estimate and CI confounder.

For the effect parameters considered in this section, if E and C are known to be unassociated among those at risk in an incidence density case-control study within a stable population; or if (a) E and C are known to be unassociated at start of follow-up in a case-control study within a fixed cohort; (b) at each joint level of E and C the cumulative incidence of disease (as well as the cumulative loss due to non-disease related reasons) is small over the follow-up period and (c) the control sampling fraction is small; then C is a design nonconfounder (irrespective of whether C was matched on in the design). This requirement that the disease be rare in a fixed cohort holds even for those case-control study designs in which the rare disease assumption is unnecessary for valid estimation of the parameter of interest (as for example in a time-matched case-control study within a cohort with the parameter of interest being the incidence density ratio). Technically C is a design nonconfounder in case control studies in fixed cohorts given $OR_{EC} = 1$ at start of follow-up only in the limit as the cumulative incidence of disease at each joint level of E and C and the control sampling fraction both approach zero. If the cumulative incidence is small but nonzero, C will be a design nonconfounder to an excellent approximation.

The last two paragraphs correct and make precise Principle 1 in the Introduction. Principles 2 and 3 remain as given. Principle 4 is generally false under our definition of confounding.

6. AN EXACT THEORY OF CONFOUNDING GIVEN DATA ON C

In the Introduction, C was defined to be an (exact) nonconfounder for a particular parameter of interest given data on C if and only if, for every possible outcome of the study, inference on the parameter of interest does not depend on the realized value of C. Since our interest is in inference rather than in particular decision problems: for a Bayesian, we will require that the marginal posterior distribution of the parameter of interest does not depend on C; for a pure likelihoodist, we will require that the maximized relative likelihood ratio [23, pp. 18, 19] for any two values of the parameter of interest does not depend on C; for a frequentist, we will require that the appropriate sample space does not depend on the data through C. Other approaches to frequentist inference, e.g. large sample maximum likelihood inference and exact uniformly most accurate unbiased confidence intervals will also be considered.

We restrict our attention to parameters than can be consistently estimated from data on E, C and D.

An alternative definition of nonconfounding given C would be to define C to be a *data-based* nonconfounder given data on C if inference on the parameter of interest does not depend on the data through C for the particular data outcome obtained. An example will be given in Section 6(E).

6(A). Cuts, S-*Sufficiency and* S-*Ancillarity*

In this section we review some concepts that will be necessary for further development.

Consider a statistical model with minimal sufficient statistics (T, Q) and parameters $p = (p_1, p_2)$ where T, Q, p_1, p_2 are possibly vector-valued with the following property:

$$f(T, Q; p_1, p_2) = f(T|Q; p_1) f(Q; p_2), \tag{5}$$

where f either is a density function or a probability function. If any value of p_2 can arise with any value of p_1, Q is called a cut [6]. The following properties are trivial to show;

(1) in a Bayesian framework, if p_1 and p_2 have independent priors, then the posterior distribution of p_2 depends on the data only through Q;

(2) the maximum likelihood estimator of p_2 and its estimated asymptotic variance derived from either the observed information matrix or expected information evaluated at the maximum likelihood estimate depend on the data only through Q;
(3) pure likelihood inference based on maximized relative likelihoods [23, pp. 18, 19] for p_2 depend on the data only through Q.

Furthermore if Q is a cut, by definition, Q is said to be S-sufficient for p_2 and S-ancillary for p_1 [6]. By extended principles of conditionality and sufficiency, frequentist inference on p_2 is from the marginal distribution of Q, and on p_1 is from the conditional distribution of T given Q [6]. The sample space for inference on p_2 will be uniquely defined if Q is boundedly complete for p_2 given p_1 known [24]. The standard motivating example for the extended principles of sufficiency and conditionality identifies T with a measurement on a rod of unknown length p_1, and Q with the outcome of a coin toss with unknown probability p_2 of heads [20, Chap. 2]. If heads is obtained, the measurement is performed with a precise instrument of known variance, and otherwise with an imprecise instrument also of known variance. It seems imperative that inference on p_2 should depend only on the outcome of the coin toss, while inference on p_1 should depend only on T and the known precision of the instrument actually used.

6(B). Cross-overs

Consider that $p_1 - p_2 = 0$ was known *a priori*. Then the priors for p_1 and p_2 are no longer independent, and thus the posterior distribution will depend on the data through T; the restricted maximum likelihood estimator of p_2 given $p_1 = p_2$ and maximized relative likelihoods for p_2 will depend on the data through T; every value of p_2 cannot arise with every value of p_1, so Q is by definition no longer S-sufficient for p_2 or S-ancillary for p_1. Information on p_1 from $f(T|Q; p_1)$ now supplies information on p_2. A prior restriction that functionally depends on both p_1 and p_2 will be called a cross-over since it crosses over the cut between p_1 and p_2. On the other hand a restriction such as $p_1 = 5$ does not cross the cut, and properties 1–3 above continue to hold.

If p_1 and p_2 are of high dimensionality while the data set available is of only moderate size, then for frequentist or pure likelihoodist inference, we must further model the probability mechanism generating the data since priors with which to downweight "unlikely parameter values" are not allowed. Of course due to technical limitations and to the problems of prior specification, this is also in fact true for a Bayesian. For example, if the conditional distribution T given Q and the marginal distribution of Q were completely unrestricted, then equation (5), although true, would often be unsuitable for inference concerning p_2. Since further modelling represents additional prior restrictions on the class of allowed probability distributions, it is important to recognize that certain models will themselves introduce cross-over restrictions. If in the Bayesian sense an individual's prior for p_1 and p_2 are not independent, then that individual's best models, in the sense of being the dogmatic priors closest to his nondogmatic priors, would probably introduce a cross-over restriction.

Extensions of the ideas in this section to partial likelihoods are discussed in Appendix A of Robins [25].

6(C). IDR Case-control Study in a Steady State Population

Throughout the remainder of the paper, let x_{ijk} indicate the number of individuals with status (E_i, D_j, C_k), $i, j, k \in (1, 2)$. As before (E_1, D_1, C_1) is (E, D, C) but, in a change of notation, $(\bar{E}, \bar{D}, \bar{C})$ is henceforth to be written as (E_2, D_2, C_2) rather than as (E_0, D_0, C_0). Let

$$\text{sOR} = \frac{p(E|D)}{p(\bar{E}, C|D)\dfrac{p(E|\bar{D}, C)}{p(\bar{E}|\bar{D}, C)} + p(\bar{E}, \bar{C}|D)\dfrac{p(E|\bar{D}, \bar{C})}{p(\bar{E}|\bar{D}, \bar{C})}},$$

$$\text{cOR} = \frac{p(E|D)[1 - p(E|\bar{D})]}{p(E|\bar{D})[1 - p(E|D)]},$$

and

$$\tilde{c}OR = \left(\frac{x_{11+}}{x_{+1+} - x_{11+}}\right) \Big/ \left(\frac{x_{12+}}{x_{+2+} - x_{12+}}\right).$$

Suppose x_{+1+} incident cases and x_{+2+} individuals at risk for disease are randomly sampled over a fixed case-ascertainment period from an infinite, dynamic, steady state population. (The population of all individuals of age 40–45 in the United States would serve as an excellent approximation to such a population over time periods of a few years.)

Let I_{ij} be the constant steady state incidence of disease in the population among individuals of status (E_i, C_j). Define the standardized incidence density ratio,

$$\text{sIDR} = \frac{I_{1+}}{p(C \mid E, \bar{D}) I_{21} + [1 - p(C \mid E, \bar{D})] I_{22}},$$

to be the parameter of interest. Here $\bar{D}$ denotes individuals at risk. If $I_{11}/I_{21} = I_{12}/I_{22} = \text{IDR}$, then sIDR = IDR. As noted in Section 4, sIDR = sOR without any rare disease assumption [13]. Furthermore, C is a nonconfounder absent data on C if sOR = cOR where the exposure covariate condition for nonconfounding, i.e. $\text{OR}_{EC|\bar{D}} = 1$, implies E and C are unassociated in those at risk without reference to any rare disease assumption.

Assume the investigator is interested in cOR rather than sOR perhaps because C is known to be an intermediate variable. Intuitively, in the absence of prior knowledge, inference on cOR should not depend on the data through C. This intuition can be formalized by noting the product multinomial likelihood conditional on x_{+1+}, x_{+2+} can be written $(S_1)(S_2)$ where

$$S_1 = p\left[x_{11+}, x_{12+}; \quad \frac{\text{OR}_{ED|C}}{\text{OR}_{ED|\bar{C}}}, \text{OR}_{DC|\bar{E}}, \text{OR}_{EC|\bar{D}}, p(C \mid \bar{E}, \bar{D})\right] \tag{6}$$

$$S_2 = p[x_{11+}, x_{12+}; \quad \text{cOR}, p(\bar{E} \mid \bar{D})], \tag{7}$$

where $x = (x_{111} \cdots x_{222})$. Thus the crude data (x_{11+}, x_{12+}) are a cut. If priors for the crude parameters are independent of the other 4 parameters; then for all three schools of inference, inference on cOR does not depend on the data through C. Now consider sOR as the parameter of interest where C is known to be a nonconfounder absent data on C either because $\text{OR}_{DC|\bar{E}} = 1$ or $\text{OR}_{EC|\bar{D}} = 1$. Even if due to sampling variability an E–C association exists among the controls and a D–C association exists among the unexposed, nonetheless for all three schools, inference on the sOR still depends only on (x_{11+}, x_{12+}) since both of the above restrictions are noncrossovers. Furthermore if, e.g. $\text{OR}_{EC|\bar{D}} = 1$ is known *a priori*, further noncrossover prior knowledge that

$$\text{OR}_{DC|\bar{E}} = 4, p(C \mid \bar{E}, \bar{D}) = p(C \mid \bar{D}) = \frac{1}{3} \quad \text{or} \quad \frac{\text{OR}_{ED|C}}{\text{OR}_{ED|\bar{C}}} = 1$$

supplies no additional information for inference on sOR. On the other hand additional prior knowledge that the effect of E and C on the incidence of disease was additive, i.e. $I_{11} = I_{21} + I_{12} - I_{22}$, is a cross-over restriction. In fact it is easy to show that the restricted maximum likelihood estimator of sOR, given both $\text{OR}_{EC|\bar{D}} = 1$ and the above additive restriction, depends on the data through C, and has strictly smaller asymptotic variance than $\tilde{c}OR$, which is the restricted MLE given only $\text{OR}_{DC|\bar{E}} = 1$.

This result could in theory have implications for study design. If $\text{OR}_{EC|\bar{D}} = 1$ was known *a priori*, data on C, even if collected, would not be used in the analysis. But if the investigator, a non-Bayesian, thought it possible that in years to come biological theories might prove E and C act additively but not multiplicatively on incidence, he might collect data on C anyway. Then if, in the future, additivity is proved, a reanalysis incorporating data on C will allow increased precision in the estimation of sOR.

We now examine commonly practiced forms of frequentist inference that fail to explicitly recognize the principle of S-sufficiency. As mentioned above, large sample inference for the sOR based on the restricted maximum likelihood estimator and its observed or estimated expected information evaluated at the restricted maximum likelihood estimate will depend on the data only through (x_{11+}, x_{12+}) if $\text{OR}_{EC|\bar{D}} = 1$ or $\text{OR}_{DC|\bar{E}} = 1$ is known *a priori*. In fact in Appendix D we show

that the restricted maximum likelihood estimator based on the crude data will have a strictly smaller asymptotic variance than the unrestricted maximum likelihood estimator unless both $OR_{EC|\bar{D}} = 1$ and $OR_{DC|\bar{E}} = 1$. If the odds ratio is known *a priori* to be constant over levels of C, then in the absence of prior knowledge of $OR_{DC|\bar{E}}$ or $OR_{EC|\bar{D}}$, frequentists may calculate uniformly most accurate unbiased confidence intervals or conditional maximum likelihood estimates by conditioning on the margins $\{x_{i+k}, x_{+jk}\}$. Given $OR_{DC|\bar{E}} = 1$ *a priori*, the complete minimal sufficient statistic is $(x_{11+}, x_{12+}, x_{1+1}, x_{2+1})$ and thus the uniformly most accurate unbiased confidence interval and conditional restricted maximum likelihood estimator would be based on $p(x_{11+}|x_{1++}, x_{1+1}, x_{2+1}; \mathrm{OR})$. Likewise given $OR_{EC|\bar{D}} = 1$ *a priori*, the complete minimal sufficient statistic is $(x_{11+}, x_{12+}, x_{+11}, x_{+21})$, and thus $p(x_{11+}|x_{1++}, x_{+11}, x_{+21}; OR)$ is the relevant conditional distribution. But

$$p(x_{11+}|x_{1++}, x_{1+1}, x_{2+1}; \mathrm{OR}) = p(x_{11+}|x_{1++}, x_{+11}, x_{+21}; \mathrm{OR}) = p(x_{11+}|x_{1++}; \mathrm{OR}),$$

and thus the optimal frequentist exact intervals do not depend on the data through C.

Efficient estimators and large sample optimal confidence intervals for the common OR based on weighted least squares, using the prior knowledge that either $OR_{EC|\bar{D}} = 1$ or $OR_{DC|\bar{E}} = 1$, will not be functions of the minimal sufficient statistic, and will depend on the data through C. Therefore first order efficient large sample inference can be performed utilizing data on C even when C is a nonconfounder given C.

6(D). *Cumulative Incidence Follow-up Study*

Suppose, as described in Section 4, x_{1++} exposed, and x_{2++} nonexposed individuals are randomly sampled from a near-infinite nondiseased population that is followed for a specified time period. Individuals developing disease are labelled D or D_1. Define $p_{ij} = p(D|E_i, C_j)$.

Then

$$\mathrm{sMR} = \frac{p_{11}p(C|E) + p_{12}p(\bar{C}|E)}{p_{21}p(C|E) + p_{22}p(\bar{C}|E)} \quad \text{and} \quad \check{\mathrm{c}}\mathrm{RR} = (x_{11+}/x_{1++})/(x_{21+}/x_{2++})$$

and $\mathrm{sMD} = (p_{11} - p_{21})p(C|E) + (p_{12} - p_{22})p(\bar{C}|E)$. In follow-up studies, except as discussed in Section 6(E), all the following results for the sMR apply equally to sMD.

The likelihood for this follow-up study is $(S_1')(S_3)$ where $S_1' = S_1$ from equation (6) with x_{12+} replaced by x_{21+} and $S_3 = p[x_{11+}, x_{21+}|\mathrm{cRR}, p(D|\bar{E})]$. If cRR is the parameter of interest, and the priors for the crude parameters are independent of the other four, then all three schools of inference would ignore data on C. Furthermore, with sMR as parameter of interest, if $OR_{DC|\bar{E}} = 1$ is known *a priori*, then inference on $\mathrm{sMR} = \mathrm{cRR}$ is independent of data on C; but if $OR_{EC} = 1$ is known *a priori*, then inference on sMR for all schools of inference depends on the data through C since $OR_{EC} = 1$ is a cross-over restriction.

In fact, if the risk ratio is constant over levels of C, then the likelihood is

$$p(x|x_{1+1}, x_{2+1}; \quad \mathrm{RR}_{ED|\bar{C}}, \mathrm{RR}_{DC|\bar{E}}, p_{++})\, p(x_{1+1}, x_{2+1}; \quad p(C|\bar{E}), OR_{EC}). \tag{8}$$

Therefore (x_{1+1}, x_{2+1}) is S-ancillary for $\mathrm{RR}_{ED|\bar{C}} = \mathrm{sMR}$. Thus inference on the common RR will be independent of any prior knowledge concerning OR_{EC} for all three school of inference provided priors for $[p(C|\bar{E}), OR_{EC}]$ are independent of those for the other three parameters.

6(E). *Cumulative Incidence Type Case-control Study*

Consider an unmatched case-control study of a near-infinite superpopulation, in which x_{+1+} diseased and x_{+2+} nondiseased are randomly sampled at the end of follow-up. As discussed in Section 4, given data on C, sOR but not sMR can be consistently estimated [10]. If the disease is rare at each level of E and C, i.e. $p_{++} = p_{11} + p_{12} + p_{21} + p_{22} \approx 0$, then $\mathrm{sOR} \approx \mathrm{sMR}$. In the absence of data on C, consistent estimation of sOR requires that cOR = sOR which holds if and only if $OR_{EC|\bar{D}} = 1$ or $OR_{DC|\bar{E}} = 1$ [11, Chap. 13].

Since in a case-control study, sMR, the parameter of interest, is consistently estimable only as $p_{++} \to 0$, we extend our definition of confounding absent data on C (given in Section 4) as follows: C is a nonconfounder absent C in a cumulative incidence case-control study under the rare disease assumption if and only if the parameter of interest equals a crude parameter in the limit as $p_{++} \to 0$.

Given $\mathrm{OR}_{EC} = 1$ and $p_{++} \to 0$, then $\mathrm{OR}_{EC|\bar{D}} \to 1$. Thus, as noted in Section 4, as $p_{++} \to 0$ the conditions for nonconfounding absent data on C for the SMR are the same in both a cumulative incidence follow-up study and an unmatched case-control study. Note that the rare disease assumption is formally $p_{++} \to 0$ and not that $p(D) \to 0$.

The likelihood for an unmatched cumulative incidence case-control study is $S_1 S_2$ as defined in equation (6) and (7). The likelihood is independent of p_{++}. Therefore in analogy with the above definition of nonconfounding absent C under the rare disease assumption, C is defined to be a nonconfounder given (data on) C in a cumulative incidence case-control study if in the limit as $p_{++} \to 0$, inference on the sMR does not depend on the data through C.

Given $\mathrm{OR}_{EC} = 1$ known *a priori*, $\mathrm{OR}_{EC|\bar{D}} = (1 - p_{11})(1 - p_{22})/[(1 - p_{12})(1 - p_{21})]$. With x_{+1+}, x_{+2+} fixed by the sampling scheme, a Taylor series expansion of the likelihood $S_1 S_2$ around $\mathrm{OR}_{EC|\bar{D}} = 1$ shows that as $p_{++} \to 0$ the likelihood is just $S_1 S_2$ evaluated at $\mathrm{OR}_{EC|\bar{D}} = 1$ multiplied by a factor of $[1 + O(p_{++})]$. Therefore, given $\mathrm{OR}_{EC} = 1$, as $p_{++} \to 0$ since both sOR and sMR equal cOR $[1 + O(p_{++})]$, inference on the cOR, sOR or sMR will not depend on the data through C for all three schools of inference given the appropriate prior independences. In the Bayesian case, we have used the fact that as the prior probability mass of p_{++} converges to 0, the posterior distributions for cOR and sMR become identical.

Note that it is only by means of the above Taylor series expansions of the likelihood that Miettinen and Cook's Principle 1, described in the Introduction, can be justified in cumulative incidence case-control studies under the rare disease assumption.

If instead $\mathrm{OR}_{DC|\bar{E}} = 1$ is known *a priori*, C will also be a nonconfounder given C since the prior restriction remains a noncrossover.

We now compare the inference that would be drawn from prospective follow-up of a population in which the disease is rare and $\mathrm{OR}_{EC} = 1$. To do so, since the prospective likelihood $S_1 S_3$ depends on p_{++}, we consider sequences of follow-up studies in which $p_{++} \to 0$, $x_{1++}, x_{2++} \to \infty$ in such a way that x_{1++}/x_{2++} and $E(x_{i+j})p_{ij}$ are constant. The results given in the remainder of this section do not apply to the sMD. We first consider the case where $\mathrm{RR}_{ED|C} = \mathrm{RR}_{ED|\bar{C}} = \mathrm{RR}$. By expression (8), given appropriate prior independences, inference on the RR is independent of the prior knowledge $\mathrm{OR}_{EC} = 1$. Nevertheless since $\widetilde{\mathrm{OR}}_{EC} \equiv (x_{1+1})(x_{2+2})/(x_{2+1})(x_{1+2})$ converges to 1 in probability as $x_{1++}, x_{2++} \to \infty$, it is of interest to examine the limiting case in which $\widetilde{\mathrm{OR}}_{EC} = 1$. Consider the approximation

$$(x_{ilj} | x_{i+j}) \sim \text{Poisson}\,(p_{ij} x_{i+j}). \tag{9}$$

For data outcomes satisfying

$$\frac{x_{i+j}}{x_{i++}} \geqslant \alpha > 0 \text{ for } \textit{some fixed } \alpha, \text{ and } (x_{ilk}) < (x_{i+k})^{1/2}, \tag{10}$$

as $x_{i++} \to \infty, p_{ij} \to 0$

$$P(x, \Theta) = B(x, \Theta)[1 + o(1)], \tag{11}$$

where $B(x, \Theta)$ is the true product binomial likelihood conditional on $\{x_{i+k}\}$, and $P(x, \Theta)$ is the product over i, j of Poisson likelihoods based on expression (9).

Equation (11) justifies calculating likelihood ratios based on $P(x, \Theta)$ in the limit as $p_{ij} \to 0$, $x_{i++} \to \infty$ for data outcomes satisfying expression (10). Furthermore since outcomes satisfying expression (10) occur with probability 1 as $x_{i++} \to \infty$, frequentist inference, conditional on both $\{x_{i+k}\}$ and on such outcomes, can be based on $P(x, \Theta)$ as $x_{i++} \to \infty$.

$$P(x, \Theta) = F_1 F_2,$$

where

$$F_1 \equiv \prod_{k=1}^{2} pr\,[x_{11k} | x_{+1k}, x_{1+k}, x_{2+k}, \mathrm{RR}] \tag{12}$$

$$F_2 \equiv pr\,[x_{+1+}, x_{+11} | x_{1+1}, x_{1+2}, P_1, P_2], \tag{13}$$

where

$$(x_{11k} | x_{+1k}, x_{1+k}, x_{2+k}) \sim \text{Binomial}\,[x_{+1k}, x_{1+k}\mathrm{RR}/(x_{1+k}\mathrm{RR} + x_{2+k})]$$

$$P_k = p_{2k}[\mathrm{RR} + (x_{2+k}/x_{1+k})].$$

If $\widetilde{OR}_{EC} = 1$, then $x_{2+k}/x_{1+k} = m$ where $m \equiv x_{2++}/x_{1++}$ is fixed by the sampling scheme and $P_k = p_{2k}[RR + m]$. Therefore

$$F_1 = pr(x_{111}, x_{211} | x_{+1+}, x_{1+1}, x_{1+2}) pr(x_{11+} | x_{+1+}; RR, m], \tag{14}$$

where

$$(x_{11+} | x_{+1+}) \sim \text{Binomial}[x_{+1+}, RR/(RR + m)]. \tag{15}$$

If we regard RR, P_1, P_2 as the model "parameters" by expressions (13) and (14), we have for outcomes satisfying expression (10) that as $p_{ij} \to 0$, $x_{1++} \to \infty$, C is a *data-based* nonconfounder given C whenever $\widetilde{OR}_{EC} = 1$ for all 3 schools of inference provided every value of RR can occur with every value of (P_1, P_2) and priors for RR and $(P_1, P_2, P_{C|\bar{E}}, OR_{EC})$ are independent. From a frequentist perspective, the justification for calling C a data-based nonconfounder given C when $\widetilde{OR}_{EC} = 1$ is by appeal to a pointwise version of the extended principle of conditionality as in Barndorff-Nielsen [26] that would require inference for the RR based on $P(x, \Theta)$ to depend only on expression (15) when $\widetilde{OR}_{EC} = 1$. In the Bayesian case, justification depends on equation (11) together with the fact that the prior probability masses for the p_{ij} are converging to 0. For the pure likelihoodist, justification depends only on the fact that data outcomes satisfy expression (10) and $x_{i++} \to \infty$ without regard for restrictions on the p_{ij}.

$0 \leqslant p_{ij} \leqslant 1$ and $0 \leqslant RR \leqslant \infty$ together imply $\max(P_k) = RR + m$ if $RR < 1$ and $\max(P_k) = 1 + m/RR$ if $RR > 1$. Therefore it is necessary that we restrict the P_k such that

$$P_k < \min(1, m) \tag{16}$$

in order that every value of RR and (P_1, P_2) can occur together. For fixed m, expression (16) will certainly hold in the limit as $p_{ij} \to 0$. If RR and (P_1, P_2) are independent at a given value of m, then they will be dependent for other values of m. Thus it would be unusual for an individual to actually hold priors with RR and (P_1, P_2) independent.

We now consider other forms of frequentist inference for the RR based on $P(x, \Theta)$ when $\widetilde{OR}_{EC} = 1$. The maximum likelihood estimator and its estimated asymptotic variance, as well as exact confidence intervals and the conditional maximum likelihood estimator based on expression (15), all depend on the data only through (x_{11+}, x_{21+}).

In practice $(x_{i1k} | x_{i+k})$ will not be quite Poisson, and $\widetilde{OR}_{EC}$ will not be exactly one. Nevertheless, for data sets in which x_{i1k}/x_{i+k} is small and $\widetilde{OR}_{EC} \approx 1$, the crude inference based on expression (15) will serve as an excellent approximation to inference based on $B(x, \Theta)$ provided that in the Bayesian case small values of p_{ij} are accorded significant prior probability, and the necessary prior independences hold. In this paper we restrict our attention to the theoretical limiting relationships and do not discuss the accuracy of the approximation.

If the risk ratio had not been constant over levels of C, $\{x_{i+k}\}$ are not S-ancillary for sMR, and no factorization similar to equation (14) appears to exist. Nonetheless, straightforward calculation demonstrates that if $\widetilde{OR}_{EC} = 1$, the maximum likelihood estimator of sMR is $\tilde{c}RR$. Furthermore, in the limit as $x_{i++} \to \infty$, for data outcomes satisfying expression (10) both the estimated asymptotic variance of the maximum likelihood estimator based on either the observed or expected information and maximized relative likelihood ratios for the sMR depend on the crude data alone. These results remain true if $OR_{EC} = 1$ is known *a priori*.

The above results for follow-up studies under the rare disease assumption imply that, given $OR_{EC} = 1$ is known, the asymmetry between cumulative incidence case-control studies and follow-up studies as expressed for example in Principles 1 and 3 of the Introduction would, for most (but not all) data outcomes, be of little consequence since in a follow-up study, as $x_{i++} \to \infty$, and $p_{++} \to 0$, $\widetilde{OR}_{EC}$ converges to 1 in probability.

6(F). Cumulative Incidence Case-control Study Within a Cohort

Consider a sample of size N randomly selected at start of follow-up from a near-infinite superpopulation. x_{+1+} cases develop during follow-up. All x_{+1+} cases and a further x_{+2+} controls sampled at random from the $(N - x_{+1+})$ non-cases have their E and C statuses determined. sMR is again the parameter of interest. As noted in Section 4, since $p(D)$ can be estimated as x_{+1+}/N, cRR can now be estimated from the crude data even when $p_{++} \not\approx 0$ [12]. Therefore, sRR can be

consistently estimated absent data on C whenever sRR = cRR, i.e. $\mathrm{OR}_{EC} = 1$ or $\mathrm{OR}_{DC|E} = 1$.

Consider further this case-control study within a cohort with sMR as parameter of interest. Let $\mathbf{N} = \{N_{i+j}\}$ be the vector of the unobserved number of individuals of status E_i, C_j in the cohort. Assume prior knowledge is available that $\mathrm{OR}_{EC} = 1$ in the infinite super population. The frequency functions for the observed data can be written

$$p(x) = \sum_{N_{i+j} > x_{i+j}} p(x, \mathbf{N}). \tag{17}$$

Straightforward calculation shows that the restricted maximum likelihood estimator given $\mathrm{OR}_{EC} = 1$ as well as the relative maximized likelihood ratio depend on the data through C.

Nonetheless since a case-control study within a cohort design is usually chosen in the study of a rare disease, it is of interest to examine inference on the sMR under the rare disease assumption. Since the likelihood depends on p_{++} we must consider sequences of studies in which $p_{++} \to 0$ as $N \to \infty$ in such a way that $E(N_{i+j})p_{ij}$ is constant. Let a be the fraction of the $(N - x_{+1+})$ nondiseased individuals selected as controls. If $a = 1$, we have the equivalent of the full follow-up study considered in Section 6(E).

At the other extreme if we consider the case in which as $N \to \infty$, aN is constant, we can show that (x_{11+}, x_{12+}) form a cut for sMR as $N \to \infty$. Consider data outcomes in which as $N \to \infty$, $x_{i1k} < N^{1/2}$ and $x_{i2k} > bN$ for some fixed $b > 0$. Such outcomes will occur with probability 1 as $N \to \infty$ provided $p_{i+j} = p(E_i, C_j)$ is bounded away from 0 by ϵ, for some $\epsilon > 0$. Then as $N \to \infty$, by a Taylor series expansion

$$p(x) = UV[1 + o(1)],$$

where

$$U = p[x \mid x_{+1+}, x_{+2+}, E(\mathbf{N})]$$

$$V = p[x_{+1+}, x_{+2+} \mid E(\mathbf{N})]\alpha P^{x_{+1+}}(1 - P)^{N - x_{+1+}}$$

$$P = \sum_{i,k} p_{ik} p_{i+k}.$$

As $N \to \infty$, U is just $S_1 S_2$ as defined in equations (6) and (7). Therefore x_{+1+}, x_{+2+} are S-ancillary for the 6 parameters of S_1, S_2. Furthermore if $\mathrm{OR}_{EC} = 1$ is known *a priori*, as $N \to \infty$ by a Taylor series expansion, U equals $S_1 S_2$ evaluated at $\mathrm{OR}_{EC|\bar{D}} = 1$, multiplied by a factor $[1 + o(1)]$. Thus regarding x_{+1+}, x_{+2+} as fixed and given $\mathrm{OR}_{EC} = 1$, as $p_{ij} \to 0$ and $N \to \infty$, (x_{11+}, x_{12+}) become S-sufficient for sMR conditional on data outcomes with $x_{i1k} < N^{1/2}$, $x_{i2k} > bN$. Miettinen and Cook [1] claim that in a cumulative incidence case control study within a cohort, the crude analysis is appropriate. The above Taylor series expansion defines conditions under which their claim is true.

Therefore, given $\mathrm{OR}_{EC} = 1$ known *a priori*, for actual data sets with $a = x_{+2+}/(N - x_{+1+}) \approx 0$ and $(x_{i1k}/x_{i2k})a \approx 0$, inference based on the crude data will be an excellent approximation to inference based on equation (17) irrespective of associations observed in the data, provided that in the Bayesian case the necessary prior independences hold and small values of p_{ij} are accorded significant prior probability. Again we do not consider the accuracy of the approximation.

7. EXTENSION TO ARBITRARY RANDOM VARIABLES

Let (E, C, D) represent any random variables, continuous or discrete. Define reference levels (E_o, C_o, D_o). In epidemiologic studies D_o signifies the nondiseased state and E_o a nonexposed state whenever such categorizations are epidemiologically meaningful.

7(A). Confounding Absent Data on C*—General Case*

Consider an unmatched case-control study either of cumulative incidence type or of incidence density type within a stable population. Define:

$$g(C, E, D) \equiv \frac{p(C, E \mid D)}{p(C, E \mid D_o)} \frac{p(C_o, E \mid D_o)}{p(C_o, E \mid D)}$$

$$h(C, E, D) \equiv \frac{g(C, E, D)}{g(C, E_o, D_o)}$$

$$q(E, D) \equiv \frac{p(C_o, E \mid D)}{p(C_o, E \mid D_o)} \frac{p(C_o, E_o \mid D_o)}{p(C_o, E_o \mid D)}$$

where $p(\bullet, \bullet \mid \bullet)$ represents a density or frequency function.

$g(C, E_0, D), h(C, E, D), q(E, D)$ generalize $\mathrm{OR}_{DC|\bar{E}}, \dfrac{\mathrm{OR}_{ED|C}}{\mathrm{OR}_{ED|\bar{C}}}, \mathrm{OR}_{ED|\bar{C}}$, respectively.

$$\mathrm{cOR}(E, D) \equiv \frac{p(E \mid D)}{p(E_o \mid D)} \frac{p(E_o \mid D_o)}{p(E \mid D_o)} = \frac{A}{B},$$

where

$$A \equiv \int h(C, E, D)\, g(C, E_o, D) q(E, D) p(C \mid E, D_o)\, \mathrm{d}C$$

$$B \equiv \int g(C, E_o, D) p(C \mid E_o, D_o)\, \mathrm{d}C$$

$$\mathrm{sOR}(E, D) \equiv \frac{A}{F},$$

where

$$F \equiv \int g(C, D, E_o) p(C \mid E_o, D_o)\, \mathrm{d}C$$

$$\mathrm{sOR}(E, D) = \mathrm{cOR}(E, D) \Leftrightarrow B = F \tag{18}$$

Expression (18) extends to arbitrary random variables results of Whittemore [27] and Shapiro [28] for contingency tables. Since $\mathrm{cOR}(E, D)$ can be consistently estimated from the crude data in an unmatched case-control study, expression (18) gives conditions under which C is a non-confounder absent data on C in an unmatched case-control study.

If $h(C, E, D) = 1$, then $g(C, E, D) = g(E, D)$ and $\mathrm{sOR}(E, D) = q(E, D)$. We then say that E and C interact multiplicatively. This generalizes the notion of constancy of the odds ratio over levels of C from the contingency table context.

Furthermore, in a cumulative incidence study, so that $p(D \mid E, C)$ is defined, we find, by use of the identity

$$p(E, C \mid D) = \Lambda(E, C, D) p(E, C \mid D_o) \frac{p(D_o)}{p(D)},$$

where

$$\Lambda(E, C, D) = \frac{p(D \mid E, C)}{p(D_o \mid C, C)}$$

that if for all $E, C, p(D_o \mid E, C) \approx 1$ (which defines the rare disease assumption), then

$$\mathrm{cOR}(E, D) = \frac{\int \Lambda(E, C, D) p(C \mid E, D_o)\, \mathrm{d}C}{\int \Lambda(E_o, C, D) p(C \mid E_o, D_o)\, \mathrm{d}C} \approx$$

$$\frac{\int p(D \mid E, C) p(C \mid E)\, \mathrm{d}C}{\int p(D \mid E_o, C) p(C \mid E_o)\, \mathrm{d}C} = \frac{p(D \mid E)}{p(D \mid E_o)} \equiv \mathrm{cRR}(E, D)$$

$$\mathrm{sOR}(E,D)=\frac{\int \Lambda(E,C,D)p(C\mid E,D_o)\,dC}{\int \Lambda(E_o,C,D)p(C\mid E,D_o)\,dC}\approx$$

$$\frac{p(D\mid E)}{\int p(D\mid E_o,C)p(C\mid E)\,dC}\equiv \mathrm{sMR}(E,D).$$

Also, without any rare disease assumption, $\mathrm{sMR}(E,D)=\mathrm{cRR}(E,D)\Leftrightarrow$

$$\int p(D\mid E_o,C)p(C\mid E_o)\,dC=\int p(D\mid E_o,C)p(C\mid E)\,dC, \tag{19}$$

which is a generalization of results of Shapiro [28] for contingency tables. Equation (19) provides conditions under which C is a nonconfounder absent C in a cumulative incidence follow-up study.

7(B). Confounding Given Data on C*—Case-control Studies*

Consider an unmatched case-control study likelihood with N_k individuals randomly sampled from D_k, $k\in(1\cdots K)$ with $\mathrm{sOR}(E,D)$ as parameter of interest. For control j from disease level k

$$p(E_{jk},C_{jk}\mid D_k)=L_{1jk}L_{2jk},$$

where

$$L_{1jk}\equiv\frac{p(E_{jk},C_{jk}\mid D_k)}{p(E_{jk}\mid D_k)}=\frac{h(C_{jk},E_{jk},D_k)g(C_{jk},E_o,D_k)p(C_{jk}\mid E_{jk},D_k)}{\int h(C,E_{jk},D_k)\,g(C,E_o,D_k)p(C\mid E_{jk},D_k)\,dC}$$

and

$$p(E_{jk}\mid D_k)\equiv L_{2jk},$$

where we note $h(C_{jo},E_{jo},D_o)=g(C_{jo},E_o,D_o)=1$. Thus the likelihood for the full case-control study is $L_{1+}L_{2+}$ where

$$L_{1+}\equiv\left(\prod_{k=1}^{K}\prod_{j=1}^{N_k}L_{1jk}\right)L_{2+}\equiv\left(\prod_{k=1}^{K}\prod_{j=1}^{N_k}L_{2jk}\right).$$

L_{1+} is parameterized by $h(C,E,D),g(C,E_o,D),p(C\mid E,D_o)$. L_{2+} depends on the data only through $\{(E_{jk},D_k)\}$ and is parameterized by $\mathrm{cOR}(E,D)$, $p(E\mid D_o)$. Since

$$p(E_{jk}\mid D_k)=\frac{\mathrm{cOR}(E_{jk},D_k)p(E_{jk}\mid D_o)}{\int \mathrm{cOR}(E,D_k)p(E\mid D_o)\,dE},$$

$\{(E_{jk},D_k)\}$ is a cut and will be complete for $[p(E,D_o),\mathrm{cOR(E,D)}]$ at fixed values of the other parameters in the class of all probability distributions. Furthermore, if $\mathrm{cOR}(E,D)=\mathrm{sOR}(E,D)$ then $B=F$ from expression (18). $B=F$ is a noncrossover restriction. Thus, if sOR is the parameter of interest, C is a nonconfounder given C if it is a nonconfounder absent C.

7(C). Confounding Given Data on C*—Follow-up Study*

Let sMR be the parameter of interest. We intercharge E with D in our previous notation, i.e. E is now effect, and D is dose, to allow reuse of formulae in Section 7(B).

L_{ijk}, L_{jk}, L_{1+}, and L_{2+} are as defined previously and the product $L_{1+}L_{2+}$ is the full follow-up study likelihood. Since $p(E_{jk}\mid D_k)=cRR(D_k,E_{jk})p(E_{jk}\mid D_o)$, $\{(E_{jk},D_k)\}$ remains a cut, is S-sufficient for $\mathrm{cRR}(D,E),p(E\mid D_o)$, and S-ancillary for $h(C,E,D),g(C,E_o,D),p(C\mid E,D_o)$.

If $\mathrm{sMR}(D,E)=\mathrm{cRR}(D,E)$ then after interchanging E with D, formula (19) implies

$$1=\int p(C\mid E,D_o)\frac{p(C\mid D)}{p(C\mid D_o)}\,dC.$$

However,

$$\frac{p(C\mid D)}{p(C\mid D_o)}=\frac{\int L_1 \mathrm{cRR}(D,E)\,p(E\mid D_o)\,\mathrm{d}E}{\int p(C\mid E,D_o)\,p(E\mid D_o)\,\mathrm{d}E},$$

where L_1 is L_{1jk} without subscripts j, k.

Thus formula (19) represents a complex cross-over restriction unless the noncross-over restriction $p(C\mid E,D_o)=p(C\mid D_o)\Rightarrow p(E\mid D_o,C)=p(E\mid D_o)$ holds, i.e. unless C is not a risk factor for disease in the unexposed. If data on another covariate C' is available and prior knowledge exists concerning the joint distribution of (E, C, D) conditional on C', the results of Sections 7(A)–(C) hold, provided each expression is interpreted as being conditional on C'. This remains true even if C' was matched on in the design, e.g. in a case-control study controls were sampled from individuals at the same level of C' as a case.

7(D). Model Restrictions

In applications various statistical models, e.g. logistic models, are commonly employed. Since models represent additional prior restrictions on the class of allowed probability distributions, it is important to recognize that certain models will themselves introduce cross-over restrictions. For example, in Section 6(C) the additional prior knowledge that $\mathrm{OR}_{EC\mid\bar{D}}=1$ in an additive incidence model represented a cross-over restriction, while in a multiplicative incidence model, i.e. a logistic model without an interaction term, it did not. We must assume for our theoretical results that model restrictions are known *a priori*, and thus are not subject to revision in light of the data.

8. A STATISTICAL INTERPRETATION OF THE EPIDEMIOLOGIST'S CONCEPT OF CONFOUNDING

Our perspective that confounding given data on C depends on certain likelihood factorizations is far from the intuitive understanding of confounding that exists in the epidemiologic literature. In Section 5 we informally described a coherent statistical interpretation of Miettinen and Cook's [1] intuitive concept of confounding. In this section, we formalize the results given in Section 5.

8(A). Appropriate Estimators are Unconditionally Efficient

Miettinen and Cook define C to be a nonconfounder given data on C if and only if the "crude estimate of effect equals an appropriate adjusted estimate in the data at hand". They fail to define the class of appropriate adjusted estimators. We show appropriate adjusted estimators must be asymptotically efficient if Principles 1 and 2 in the Introduction are to hold.

Given a model $F(x,\Theta)$, $\Theta=(\Theta_1\cdots\Theta_k)$, let $\tilde{u}\Theta$ be the vector of unrestricted maximum likelihood estimators of Θ, and $\tilde{r}\Theta$ be the restricted maximum likelihood estimators of those components of Θ that are not known *a priori*.

Consider the cumulative incidence follow-up study of Section 6(D) in which $\mathrm{OR}_{DC\mid\bar{E}}=1$ is known *a priori*, but $\tilde{c}\mathrm{RR}=\tilde{r}\mathrm{sMR}\neq\tilde{u}\mathrm{sMR}$ in the data at hand, since $\mathrm{OR}_{DC\mid\bar{E}}\neq\tilde{u}\mathrm{OR}_{DC\mid\bar{E}}$. Principle 1 suggests $\tilde{r}\mathrm{sMR}$ be reported in lieu of $\tilde{u}\mathrm{sMR}$. Similarly by Principle 2, if $\mathrm{OR}_{DC\mid\bar{E}}=2$ say, is known *a priori*, and $\tilde{u}\mathrm{sMR}=\tilde{c}\mathrm{RR}\neq\tilde{r}\mathrm{sMR}$ in the data at hand because $\tilde{u}\mathrm{OR}_{DC\mid\bar{E}}=1$, then $\tilde{c}\mathrm{RR}$ must be adjusted to $\tilde{r}\mathrm{sMR}$ by symmetry with the previous example. Thus in both instances the efficient $\tilde{r}\mathrm{sMR}$ is "appropriate", while the asymptotically unbiased but inefficient estimator $\tilde{u}\mathrm{sMR}$ is not.

8(B). Appropriate Estimators are Conditionally Asymptotically Unbiased

Epidemiologists in general assume that a crude estimate is confounded by C if it fails to equal any intuitively unbiased estimator in the observed data. The epidemiologists' "intuitive bias" is not exact bias since $\tilde{c}\mathrm{RR}$ is intuitively unbiased for cRR but has an infinite expectation. Rather estimators considered to be "intuitively unbiased" appear to be estimators that are locally uniformly asymptotically unbiased; see Section 8(E). Until Section 8(E), we restrict our attention to maximum likelihood estimation, since implicit in Miettinen and Cook's examples and in discussions of confounding in the epidemiologic literature in general is concern with inference based

only on the method of maximum likelihood as demonstrated by the choice of $\check{c}RR = \check{r}sMR$ rather than, e.g. a minimum Chi-square estimator, as the "appropriate estimator" given $OR_{DC|\bar{E}} = 1$ *a priori*.

Epidemiologists who view $\tilde{u}sMR$ in the above examples with $OR_{DC|\bar{E}}$ known as "intuitively biased" are *implicitly* conditioning on an approximate ancillary statistic such as

$$\tilde{A} = \frac{(\tilde{u}OR_{DC|\bar{E}} - OR_{DC|\bar{E}})}{\tilde{V}ar(\tilde{u}OR_{DC|\bar{E}})_{|\tilde{r}\Theta}},$$

that measures the degree to which the observed $OR_{DC|\bar{E}}$ association differs due to sampling variability from its known population value.

We show that $\tilde{u}sMR$ is conditionally asymptotically biased. The level of argument is rather heuristic in the spirit of Cox and Hinckley [20, Chap. 9].

(i) Let $x_{i_1} \ldots x_{i_{N_i}}$ be N_i independent identically distributed random vectors with distribution $F_i(x, \Theta)$ for $i = (1 \ldots L)$ where $x \in R^p$, $\Theta = (\Theta_1 \ldots \Theta_k) \in \Lambda c R^k$. Each F_i may depend on only a subset of Θ. Define $P = (P_1 \ldots P_L)$, where $P_i \equiv N_i/N$ and P_i is presumed known and constant for all

$$N = \sum_{i=1}^{L} N_i \quad \text{as} \quad N \to \infty.$$

(ii) Assume $(\tilde{u}\Theta_1 \ldots \tilde{u}\Theta_k)$ are minimal sufficient for Θ.

(iii) Assume $\tilde{u}\Theta \sim \mathbf{N}[\Theta, \mathbf{I}^{-1}]$ as $N \to \infty$ where $\mathbf{N}[\Theta, \mathbf{I}^{-1}]$ is the multivariate normal distribution with mean Θ and covariance matrix $\mathbf{I}^{-1}$; $\mathbf{I} = E(I)$; I is the matrix

$$\left[\frac{-\partial^2 L(\Theta, x)}{\partial\Theta_i \partial\Theta_j}\right] \quad \text{for} \quad i,j \in \{1 \ldots k\}; \quad L(\Theta, x) = \sum_{i=1}^{L} \sum_{j=1}^{N_i} \log f_i(x_{ij}, \Theta); \quad f_i(\bullet, \bullet)$$

is the density or frequency function associated with $F_i(\bullet, \bullet)$.

(iv) Furthermore, we assume that $\tilde{u}\Theta$ is locally uniformly asymptotically normal as defined in Section 8(E) over the interior of the parameter space and that the elements of $\mathbf{I}$ are smoothly differentiable functions of Θ.

If $\Theta_k = \Theta_{k_o}$ is known *a priori*, then

$$\tilde{A} = \frac{\tilde{u}\Theta_k - \Theta_{k_o}}{\sqrt{Var(\tilde{u}\Theta_k)_{|\tilde{r}\Theta}}} \sim N(0, 1) \quad \text{as} \quad N \to \infty.$$

Now we show that a reparameterization $(\Theta_1 \ldots \Theta_{k-1}, D)$ exists such that $\tilde{u}D = (\tilde{A}/\sqrt{N})$. Since $\tilde{u}\Theta$ is a minimal sufficient, $\tilde{r}\Theta = g_1(\tilde{u}\Theta)$ for some function g_1. Therefore, $[Var(\tilde{u}\Theta_k)_{|\tilde{r}\Theta} N]^{1/2} = g_2(\tilde{u}\Theta, P, \Theta_{k_o})$ for some function g_2. Thus, $\tilde{A}/\sqrt{N} = g_3(\tilde{u}\Theta, \Theta_{k_o}, P)$ for some g_3. Let $D = g_3(\Theta, \Theta_{k_o}, P)$. Then $\tilde{A} = \sqrt{N}\tilde{u}D$ and $\Theta_k = \Theta_{k_o} \Rightarrow D = 0$. Thus, following Cox and Hinckley [20, Chap. 9] by condition (iii)

$$(\tilde{u}\Theta_1 | \tilde{A}) \sim N\left(\Theta_1 - \frac{Cov(\tilde{u}\Theta_1, \tilde{u}D)}{Var(\tilde{u}D)}\tilde{u}D, \sigma^2\right) \quad \text{as} \quad N \to \infty,$$

where

$$\sigma^2 = Var\,\tilde{u}\Theta_1 - \frac{Cov^2(\tilde{u}\Theta_1, \tilde{u}D)}{Var(\tilde{u}D)}$$

and Var, Cov are asymptotic variances and covariances determined by the appropriate row and column of $\mathbf{I}^{-1}$.

Thus, asymptotic bias $(\tilde{u}\Theta_1 | \tilde{A}) = [Cov(\tilde{u}\Theta_1, \tilde{u}D)/Var(\tilde{u}D)]\tilde{A}$. Cox and Hinckley point out that an asymptotically efficient estimator of Θ_1 conditional on $\tilde{A}$ is

$$\tilde{Q} = \tilde{u}\Theta_1 + \left[\frac{Cov(\tilde{u}\Theta_1, \tilde{u}D)}{Var(\tilde{u}D)}\right](\tilde{u}D)$$

and furthermore that a first order expansion of the log-likelihood gives $\tilde{r}\Theta_1 = \tilde{Q} + O_p(1/N)$. Therefore, $\tilde{r}\Theta_1$ is to $O_p(1/N)$ the conditional as well as unconditional maximum likelihood

estimator of Θ_1. Therefore, asymptotic bias $(\tilde{r}\Theta_1|\tilde{A}) = 0$, and asymptotic bias $(\tilde{u}\Theta_1|\tilde{A}) = 0 \Leftrightarrow \mathrm{Cov}(\tilde{u}\Theta_1, \tilde{u}D) = 0 \Leftrightarrow \mathrm{Cov}(\tilde{u}\Theta_1, \tilde{u}\Theta_k) = 0 \Leftrightarrow \mathrm{Var}(\tilde{r}\Theta_1) = \mathrm{Var}(\tilde{u}\Theta_1)$ which links conditional asymptotic bias to the unconditional efficiency of $\tilde{u}\Theta_1$. Since $\mathrm{Var}(\tilde{u}\Theta_1|\tilde{A}) = \mathrm{Var}(\tilde{r}\Theta_1|\tilde{A}) = \sigma^2$, conditionally it is *only asymptotic bias* that distinguishes the two estimators. Furthermore to $O(1)$ the asymptotic bias of $\tilde{u}\Theta_1$ will not depend on the parameterization of the restriction.

8(C). Conditional Inference to $O(N^{-3/2})$ in Variance

Efron and Hinckley originally pursued the idea that inference should be performed conditional on approximate ancillaries. They give both theoretical and Monte Carlo justification in one parameter problems for preferring in moderate sized samples $I^{-1}_{\tilde{r}\Theta}$ to $\mathbf{I}^{-1}_{\tilde{r}\Theta}$ as an estimator of the conditional variance for use in conditional Wald-type confidence intervals. Their preference for $I^{-1}_{\tilde{r}\Theta}$ to $\mathbf{I}^{-1}_{\tilde{r}\Theta}$ depended on the fact that $[\mathrm{Var}(\tilde{r}\Theta) - \mathrm{Var}(\tilde{r}\Theta|\tilde{A})]$ is $O(N^{-3/2})$. Thus the standard large sample results of Section 8(B), accurate only to $O(1/N)$ in variance, could not be invoked. Barndorff-Nielsen [29] and Amari [30] extended these results to the multiparameter situation. They considered independent and identically distributed observations from continuous curved exponential families with the unrestricted parameter space an open subset of finite dimensional Euclidian space.

Given $\Theta_k = \Theta_{k_o}$ redefine Θ as $\Theta = (\Theta_1 \ldots \Theta_{k-1})$. They essentially proved that if I is a function of $\tilde{A}/\sqrt{N}$ then:

(1) $N\,\mathrm{Var}(\Theta_1|\tilde{A})$ and $N\,\mathrm{Var}(\tilde{\Theta}_1)$ differ conditional on $\tilde{A}$ by $O(1/\sqrt{N})$.
(2) $N[I^{-1}_{\tilde{r}\Theta}]_{1,1}$ is to at least $O_P(1/N)$ the conditional maximum likelihood estimator of $N\,\mathrm{Var}(r\Theta_1|\tilde{A})$, and is thus conditionally efficient.
(3) $N[\mathbf{I}^{-1}_{\tilde{r}\Theta}]_{1,1}$ is conditional on $\tilde{A}$, consistent but asymptotically biased for $N\,\mathrm{Var}(\tilde{r}\Theta|\tilde{A})$.
(4) $N[\mathbf{I}^{-1}_{\tilde{r}\Theta}]_{1,1}$ is unconditionally efficient for $N\mathrm{Var}(\tilde{r}\Theta)$. $N[I^{-1}_{\tilde{r}\Theta}]_{1,1}$ will be unconditionally, inefficient but asymptotically unbiased for $\mathrm{Var}N(\tilde{r}\Theta_1)$.

The extension of their results to our multiple sample situation should not be difficult. The proper extension of their proofs to discrete random variables will not be considered although at a minimum we must bound the parameter space of multinomial probabilities away from 0 or 1 by ϵ for some ϵ to obtain asymptotic normality even to $O(1/N)$ in variance uniformly over the entire parameter space. Rather we show below that if $\mathrm{Var}(\tilde{r}\Theta_1|\tilde{A})$ is formally defined to be $[E(I(\Theta, x)|\tilde{A})]^{-1}_{1,1}$, then conditions (1)–(4) can be derived in a heuristic manner by a Taylor series expansion to $O(N^{-3/2})$. We offer this as a heuristic demonstration of the reasonableness of conditions (1)–(4) in the discrete case even if as in our examples the moments of $\tilde{r}\Theta_1$ are infinite. Practically in curved models we would expect as in Efron and Hinckley [5] that approximate Wald-type confidence intervals based on $I^{-1}_{\tilde{r}\Theta}$ more closely approach the nominal confidence coefficient than intervals based on $\mathbf{I}^{-1}_{\tilde{r}\Theta}$ in repeat Monte Carlo trials conditional on $\tilde{A}$.

If $(\tilde{u}\Theta_1 \ldots \tilde{u}\Theta_k)$ is mapped 1–1 onto $(\tilde{r}\Theta, \tilde{u}\Theta_k - \Theta_{k_o})$ and thus onto $(\tilde{r}\Theta, \tilde{A}/\sqrt{N})$, then by a Taylor series expansion around $(\tilde{r}\Theta - \Theta)$ we have

$$N\,\mathrm{Var}(\tilde{r}\Theta_1|\tilde{A}) \equiv N\left\{E\left[I\left(\Theta, \tilde{r}\Theta, \frac{\tilde{A}}{\sqrt{N}}\middle|\tilde{A}\right)\right]\right\}^{-1}_{1,1} = N\left[I^{-1}\left(\Theta, \Theta, \frac{\tilde{A}}{\sqrt{N}}\right)\right]_{1,1} + O\left(\frac{1}{N}\right). \quad (20)$$

Since $E[\tilde{r}\Theta - \Theta|\tilde{A}]$ and $\mathrm{Var}(\tilde{r}\Theta|\tilde{A})$ are $O(1/N)$. The final $O(1/N)$ term in expression (20) is $O(1/N)$ conditional on $\tilde{A}$ and $O_p(1/N)$ unconditionally since $\tilde{A}$ is $O_p(1)$. Similarly by a Taylor expansion

$$N[\mathbf{I}^{-1}]_{1,1} \equiv N\,\mathrm{Var}(\tilde{r}\Theta_1) \equiv N[E(I)]^{-1}_{1,1} = N[I^{-1}(\Theta, \Theta, 0)] + O\left(\frac{1}{N}\right), \quad (21)$$

since

$$E(\tilde{r}\Theta - \Theta), \quad \mathrm{Var}(\tilde{r}\Theta), \quad \mathrm{Var}\left(\frac{\tilde{A}}{\sqrt{N}}\right), \quad E\left(\frac{\tilde{A}}{\sqrt{N}}\right) \quad \text{are} \quad O\left(\frac{1}{N}\right).$$

Therefore from expression (20) and (21), respectively

$$N \operatorname{Var}(\tilde{r}\Theta_1 \mid \tilde{A})_{\mid \tilde{r}\Theta} = N[I^{-1}_{\tilde{r}\Theta}]_{1,1} + O_p\left(\frac{1}{N}\right) \tag{22}$$

$$N \operatorname{Var}(\tilde{r}\Theta_1)_{\mid \tilde{r}\Theta} = N[I^{-1}(\tilde{r}\Theta, \tilde{r}\Theta, 0)] + O_p\left(\frac{1}{N}\right). \tag{23}$$

Furthermore if

$$I\left(\Theta, \tilde{r}\Theta, \frac{\tilde{A}}{\sqrt{N}}\right) \text{ is a function of } \left(\frac{\tilde{A}}{\sqrt{N}}\right),$$

then by a further Taylor expansion,

$$N\left[I^{-1}\left(\Theta, \Theta, \frac{\tilde{A}}{\sqrt{N}}\right)\right]_{1,1} = N[I^{-1}(\Theta, \Theta, 0)]_{1,1} + O\left(\frac{1}{\sqrt{N}}\right), \text{ conditional on } \tilde{A} \tag{24}$$

$$\text{since } \tilde{A} \text{ and } N\frac{\partial I^{-1}(\Theta, \Theta, x)}{\partial x}\bigg|_{x=0} \text{ are } O(1).$$

Likewise,

$$N\left[I^{-1}\left(\tilde{r}\Theta, \tilde{r}\Theta, \frac{\tilde{A}}{\sqrt{N}}\right)\right]_{1,1} = N[I^{-1}(\tilde{r}\Theta, \tilde{r}\Theta, 0)]_{1,1} + O_p\left(\frac{1}{\sqrt{N}}\right) \text{ unconditionally.} \tag{25}$$

Furthermore, by equations (21) and (23),

$$[E(N\mathbf{I}^{-1}_{\tilde{r}\Theta} \mid \tilde{A})]_{1,1} = N[I^{-1}(\Theta, \Theta, 0)]_{1,1} + O\left(\frac{1}{N}\right). \tag{26}$$

Finally expressions (20), (21) and (24) imply condition (1); equation (22) implies condition (2); expressions (20), (24) and (26) imply condition (3); equation (23) and (25) imply condition (4). Conditions (1)–(4) remain true if $\Theta_k = \Theta_{k_o}$ represents p restrictions on a p-vector of parameters provided $\tilde{A}$ is then the corresponding p-vector of approximate ancillaries $(\tilde{A}_1 \ldots \tilde{A}_p)$.

As stressed by Sprott [31], if

$$\tilde{r}\Theta_1 z_{\alpha/2} \pm \sqrt{[I^{-1}_{\tilde{r}\Theta}]_{1,1}}$$

are to be used to form large sample $(1-\alpha)$ conditional confidence intervals, Θ_1 should be parameterized such that at the observed sample size the maximized relative likelihood for Θ_1 is nearly normal in shape. $z\alpha/2$ is the upper $\alpha/2$ percentage point of a normal distribution with mean 0 and variance 1.

If it is accepted that $I^{-1}_{\tilde{r}\Theta}$ is preferred over $\mathbf{I}^{-1}_{\tilde{r}\Theta}$ because $N\mathbf{I}^{-1}_{\tilde{r}\Theta}$ is conditionally asymptotically biased for $N \operatorname{Var}(\tilde{r}\Theta \mid \tilde{A})$, then efficient rather than inefficient but consistent estimators of $N \operatorname{Var}(\tilde{r}\Theta_1)$ would be preferred for Wald-type confidence intervals even when I is functionally independent of $\tilde{A}$. For example

$$N[I^{-1}_{\tilde{u}\Theta}]_{1,1} \quad \text{and} \quad N[\mathbf{I}^{-1}_{\tilde{u}\Theta}]_{1,1}$$

are unconditionally consistent but inefficient estimator of $N \operatorname{Var}(\tilde{r}\Theta_1)$ if $\operatorname{Cov}(\tilde{u}\Theta_1, \tilde{u}\Theta_k) \neq 0$. Therefore

$$N[I^{-1}_{\tilde{u}\Theta}]_{1,1}$$

will be conditionally asymptotically biased for $N \operatorname{Var}(\tilde{r}\Theta \mid \tilde{A})$.

8(D). A Justification for Conditioning on Approximate Ancillaries

In this section we illustrate the necessity of conditioning on approximate ancillaries in curved models for large sample inference to $O(N^{-3/2})$ in variance if the conditionality principle for exact ancillaries and a continuity principle of inference, as outlined by Buehler [4], are both accepted. Such a continuity principle states that if $p, p_1, p_2 \ldots$ are specifications such that $\lim p_n = p$ then if $s, s_1, s_2 \ldots$ are their solutions then $\lim s_n = s$. We take s_i to be Wald-type large sample confidence

intervals accurate to $O(N^{-3/2})$ in variance with standard errors derived from efficient estimators of the appropriate expected information matrix. If we do not condition on approximate ancillaries, by equation (23) estimated standard errors can be based on either $\mathbf{I}^{-1}_{\tilde{r}\Theta}$ or $I^{-1}(\tilde{r}\Theta, \tilde{r}\Theta, 0)$ for inference accurate to $O(N^{-3/2})$ in variance. In fact for the class of generalized linear models [32] $\mathbf{I}^{-1}_{\tilde{r}\Theta}$ and $I^{-1}(\tilde{r}\Theta, \tilde{r}\Theta, 0)$ are exactly equal.

For inference conditional on the vector of approximate ancillaries $\tilde{A}$, by equation (22) estimated standard errors can be based on either

$$I^{-1}_{\tilde{r}\Theta} \quad \text{or} \quad [E(I|\tilde{A})]^{-1}_{\tilde{r}\Theta}.$$

We choose $I^{-1}_{\tilde{r}\Theta}$ for reasons of computational convenience.

If an S-ancillary statistic exists all expectations are conditional on the S-ancillary statistic irrespective of whether or not they are taken conditional on any approximate ancillaries as well.

Consider as an illustrative example the unmatched cumulative incidence follow-up study of Section 6(D). With $\mathrm{OR}_{EC} = 1$ known *a priori*, c̃RR and r̃sMR are unconditionally asymptotically unbiased for sMR. In Appendix C we show that, in general, $\mathrm{Var}(\tilde{r}\mathrm{sMR}) < \mathrm{Var}(\tilde{c}\mathrm{RR})$ to $O(N^{-1})$ unless c̃RR is asymptotically unbiased conditional on $\tilde{A}$. Conditional on $\tilde{A}$, the approximate ancillary, we show in Appendix C that in general $E(\tilde{c}\mathrm{RR} - \mathrm{sMR}|\tilde{A})$ is $O(N^{-1/2})$ and thus c̃RR is conditionally asymptotically biased. Furthermore, $\mathrm{Var}(\tilde{r}\mathrm{sMR}|\tilde{A}) \neq \mathrm{Var}(\tilde{r}\mathrm{sMR})$ to $O(N^{-3/2})$ since I^{-1} is a function of $\tilde{A}$.

Now given $\alpha = \alpha_0$ is known *a priori* as well, where $\mathrm{RR}_{ED|C} = \mathrm{RR}_{ED|\bar{C}} + \alpha$, if $\alpha_o > 0$ no exact ancillary exists. But when $\alpha = 0$, (x_{1+1}, x_{2+1}) are S-ancillary for sMR. Let $\tilde{A}_1$, $\tilde{A}_2$ be the approximate ancillaries associated with prior knowledge that $\mathrm{OR}_{EC} = 1$ and $\alpha = \alpha_o$ respectively. Since $\tilde{A}_1 \neq 0$ implies

$$[\mathbf{I}^{-1}_{\tilde{r}\Theta}]_{1,1} - [E(I|x_{1+1}, x_{2+1})_{|\tilde{r}\Theta}]^{-1}_{1,1} \neq 0,$$

without an extension of the conditionality principle to include approximate ancillaries, unconditional confidence intervals based on $[\mathbf{I}^{-1}_{\tilde{r}\Theta}]_{1,1}$ for $\alpha_o \neq 0$ known *a priori* would abruptly change to conditional confidence intervals based on

$$[E(I|x_{1+1}, x_{2+1})_{|\tilde{r}\Theta}]^{-1}_{1,1} \quad \text{at} \quad \alpha_o = 0.$$

On the other hand, if inference is conditional on $(\tilde{A}_1, \tilde{A}_2)$, confidence intervals for all α_o would be based on $(I^{-1}_{\tilde{r}\Theta})_{1,1}$, a continuous function of α_o.

Thus if we plan to condition on exact S-ancillary statistics, a continuity principle of inference (that requires continuous perturbations in model specification give continuous perturbations in inferences drawn from a given data set) necessitates conditioning on approximate ancillaries.

Furthermore, Wald-type confidence intervals accurate to $O(N^{-3/2})$ in variance are obtained by conditioning on approximate ancillaries $\tilde{A}$ whenever I is a function of $\tilde{A}$, and performing unconditional inference otherwise, even in the presence of exact ancillaries since $(I^{-1}_{\tilde{r}\Theta})_{1,1}$ is unaffected by conditioning on exact ancillaries.

In our example of Section 6(A) since the restriction $\mathrm{OR}_{DC|\bar{E}} = 1$ was on a natural parameter of the exponential model, no curvature was introduced, and thus $\tilde{A}$ was not a function of the minimal sufficient statistic. Efron and Hinkley [5], Barndorff-Nielsen [29] and Amari [30] assumed reduction by sufficiency in the restricted model prior to conditioning on approximate ancillaries. Epidemiologists in viewing ũsMR as biased conditional on $\tilde{A}$ are conditioning prior to reducing by sufficiency. But when $\mathrm{OR}_{DC|\bar{E}}$ is known, $I^{-1}_{\tilde{r}\Theta} = \mathbf{I}^{-1}_{\tilde{r}\Theta}$. Therefore conditional and unconditional Wald-type confidence intervals centered on $\tilde{r}\Theta$ will be identical. Thus for large sample inference to $O(N^{-3/2})$ in variance, in an uncurved model it appears to be of only philosophical rather than practical concern whether ũsMR is viewed as conditionally biased or unconditionally inefficient.

Thus inference conditional on approximate ancillaries offers a parsimonious, intuitive, and unified view of Wald-type confidence intervals accurate to $O(N^{-3/2})$ in variance, whether or not exact ancillaries exist, and whether or not the approximate ancillary is a function of the minimal sufficient statistic. Inappropriate estimators are then in accord with the intuition of epidemiologists "biased" estimators, albeit asymptotically biased. If on the other hand the sufficiency principle is preserved and as a consequence the conditionality principle is extended to approximate ancillaries only in curved models, then the previous unified summary must be modified such that in uncurved

models inappropriate estimators are inefficient estimators and in curved models they are asymptotically biased estimators.

Approximate confidence intervals based on the acceptance regions of likelihood ratio tests give large sample conditional confidence intervals accurate to $O(N^{-3/2})$ without the need to explicitly consider the existence of approximate ancillaries in forming the confidence interval [5]. Confidence intervals based on the acceptance region of a score test would require that in curved models, expectations be taken conditional on approximate ancillary statistics.

8(E). Preliminary Test Estimators and Uniform Asymptotic Unbiasedness

We have shown that nonconfounding given data on C requires in general correct prior knowledge concerning nature. It may be argued that we never have correct prior knowledge, but only prior beliefs, i.e. extreme data results will always make us revise our prior opinions, no matter how firmly held.

Frequentist statisticians, when faced with this difficulty, have often suggested that an investigator with interest in a parameter Θ_1, who in addition has a strong prior belief but not true prior knowledge that $\Theta_k = \Theta_{k_o}$, perform a preliminary test of the hypothesis $\Theta_k = \Theta_{k_o}$ at some predetermined significance level. If the test rejects, the estimate $\tilde{u}\Theta_1$ is reported and $\tilde{r}\Theta_1$ otherwise. This so-called preliminary test estimator, $\tilde{p}\Theta_1$, is asymptotically normal with asymptotic expectation Θ_1 and asymptotic variance equal to that of $\tilde{u}\Theta_1$, provided $\Theta_k \neq \Theta_{k_o}$ [33]. It remains asymptotically unbiased but not asymptotically normal if $\Theta_k = \Theta_{k_o}$ [33]. Furthermore, in moderate-sized samples $\tilde{p}\Theta_1$ may have better performance in terms of mean square error than $\tilde{u}\Theta_1$ if $\Theta_k = \Theta_{k_o}$ or $\Theta_k \approx \Theta_{k_o}$. Thus $\tilde{p}\Theta_1$ is often recommended by statisticians because it is asymptotically efficient and yet has better moderate sample performance than $\tilde{u}\Theta_1$ provided the investigator's prior beliefs are correct or nearly correct. Nevertheless in the epidemiologic literature preliminary tests to determine whether C is a confounder, e.g. testing whether $\mathrm{OR}_{DC|\bar{E}} = 1$ in a cumulative incidence follow-up study, are uniformly disparaged, for example see Kleinbaum, Kupper and Morgenstern [11, p. 254]. The preliminary test estimator is viewed as intuitively biased by epidemiologists. We suggest that epidemiologists regard as intuitively biased estimators that systematically deviate from the true population value to such an extent that Wald-type confidence intervals centered on the estimator fail to give correct coverage properties even in large samples. A formal statistical concept that corresponds to this intuitive notion of unbiasedness is local uniform asymptotic unbiasedness.

Let $\tilde{\Theta}_{1n}$ be an estimator of Θ_1 based on n observations. Let $\partial\Lambda$ be the topological boundary of the parameter space in R^k. The boundary may or may not be included in the parameter space. Let $\Lambda^* = \Lambda - \partial\Lambda$. We define an estimator to be locally uniformly asymptotically unbiased if

$$\begin{gathered} \forall \boldsymbol{\Theta}_o \in \Lambda^*, \forall \textit{ sequences } k_n = \boldsymbol{\Theta}_o + \gamma/\sqrt{n}, \gamma \in R^k \\ \textit{under } k_n,\ E[n^{1/2}(\tilde{\Theta}_{1n} - \Theta_{1n})] = 0 \\ \textit{where } \Theta_{1n} \textit{ is the first component of } k_n. \end{gathered} \tag{27}$$

We are restricting attention to estimators $\tilde{\Theta}_{1n}$ for which $\forall \boldsymbol{\Theta}_o \in \Lambda^*$, $\sqrt{n}\,(\tilde{\Theta}_{1n} - \Theta_1)$ has under any sequence k_n a nondegenerate, although not necessarily normal, limiting distribution with finite mean and variance. Our definition follows closely from Sen [33, Sections 4 and 5]. It follows immediately from Sen's Section 5 that $\tilde{p}\Theta_1$ is locally uniformly asymptotically biased if $\mathrm{Cov}(\tilde{u}\Theta_1, \tilde{u}\Theta_k) \neq 0$.

A slightly different but more easily interpretable condition is as follows. An estimator $\tilde{\Theta}_{1n}$ is locally uniformly asymptotically median unbiased [34] if

$$\begin{gathered} \forall \Theta_o \in \Lambda^*, \quad \text{there exists} \quad \sigma > 0 \text{ s.t.} \\ \boldsymbol{\Theta} \in J \equiv \left\{ \boldsymbol{\Theta} \,\middle|\, |\boldsymbol{\Theta} - \boldsymbol{\Theta}_o| < \sigma \right\}, \sup_{\boldsymbol{\Theta} \in J} \{ |p[\tilde{\Theta}_{1n} < \Theta_1 | \boldsymbol{\Theta}] - \tfrac{1}{2}| \} \to 0 \quad \text{as} \quad n \to \infty. \end{gathered} \tag{28}$$

For estimators that are locally uniformly asymptotically normal and unbiased, expressions (27) and (28) coincide; where by definition an estimator is locally uniformly asymptotically normal and

unbiased if

$$\forall \boldsymbol{\Theta}_o \in \Lambda^*, \quad \text{there exist} \quad \sigma > 0 \quad \text{and } \textit{functions} \quad \sigma_n(\boldsymbol{\Theta}) \text{ s.t.} \tag{29}$$

$$\sup_{\boldsymbol{\Theta} \in J} \{|p[(\tilde{\Theta}_{1n} - \Theta_1)/\sigma_n(\boldsymbol{\Theta}) \leqslant z_\alpha \Big| \boldsymbol{\Theta}] - \alpha|\} \to 0 \quad \text{as} \quad n \to \infty \text{ [34]},$$

where J is as in expression (28). In fact expression (29) implies expressions (28) and (27).

We now consider the relationship between equation (27) and Wald-type confidence intervals. By definition $\tilde{\Theta}_{1n}$ can center a Wald type confidence interval for $\Theta \in \Lambda$ only if there exists $\sigma_n(\boldsymbol{\Theta})$ such that

$$\sup_{\boldsymbol{\Theta} \in \Lambda} \{|p[(\tilde{\Theta}_{1n} - \Theta_1)/\sigma_n(\boldsymbol{\Theta}) \leqslant z_a \Big| \boldsymbol{\Theta}] - \alpha|\} \to 0 \quad \text{as} \quad n \to \infty. \tag{30}$$

It is obvious that if equation (27) is false, expression (30) is false. Even if expression (29) holds, if for some $\Theta \in \partial\Lambda$, $\sqrt{n}(\tilde{\Theta}_{1n} - \Theta_1)$ has a degenerate limiting distribution, expression (30) is false. For example, consider the unbiased estimator, x/n, of p when either $p = 0$ or $p = 1$ where $x \sim$ *Binomial* (n, p), $p \in [0, 1]$. In this example, even if the parameter space is restricted to Λ^*, i.e. (0,1), and expression (29) holds, expression (30) is false because

$$\lim_{p \to 1} [n \text{ Var } (x/n)] = 0.$$

Thus the intuitive concept of unbiasedness, i.e. equations (27) or (28), are necessary but not sufficient for $\tilde{\Theta}_{1n}$ to center Wald-type confidence intervals. If expression (29) holds, then $\tilde{\Theta}_{1n}$ can center Wald intervals on any compact subspace of Λ^* [34].

REFERENCES

1. O. S. Miettinen and E. F. Cook, Confounding: essense and detection. *Am. J. Epidemiol.* 114, 593–603 (1981).
2. N. E. Day, D. P. Byar and S. B. Green, Overadjustment in case-control studies. *Am. J. Epidemiol.* **112,** 696–706 (1980).
3. N. E. Breslow and N. E. Day, Statistical methods in cancer research. *Int. Agency on Cancer Publ.*, Lyon, France (1980).
4. R. J. Buehler, *Rejoinder. J. Am statist. Ass.* **77,** 593–4 (1982).
5. B. Efron and D. V. Hinkley, Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65,** 657–87 (1978).
6. O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory.* Wiley, Chichester (1978).
7. S. Greenland and J. M. Robins, Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* **15,** 413–419 (1986).
8. B. Definetti, *Probability, Induction, and Statistics.* Wiley, New York (1972).
9. J. B. Copaz, Randomization models for the matched and unmatched 2 × 2 tables. *Biometrika* **60,** 467–476 (1973).
10. O. S. Miettinen, Components of the crude risk ratio. *Am. J. Epidemiol.* **96,** 168–172 (1972).
11. D. G. Kleinbaum, L. L. Kupper and H. Morgenstern, *Epidemiologic Research.* Belmont. Lifetime Learning Publications, California (1982). Chap. 13
12. R. R. Neutra and M. E. Drolette, Estimating exposure-specific disease rates from case-control studies using Bayes Theorem. *Am. J. Epidemiol.* **108,** 214–22 (1978).
13. O. S. Miettinen, Estimability and estimation in case-referent studies. *Am. J. Epidemiol.* **103,** 226–35 (1976).
14. R. L. Prentice and N. E. Breslow, Retrospective studies and failure time models. *Biometrika* **65,** 153–158 (1978).
15. S. Greenland and R. Neutra, Control of confounding in the assessment of medical technology. *Int. J. Epidemiol.* **9,** 361–367 (1980).
16. K. J. Rothman, Epidemiologic methods in clinical trials. *Cancer* **39,** 1771–1775 (1977).
17. J. M. Robins and S. Greenland, The role of model selection in causal inference from nonexperimental data. *Am. J. Epidemiol.* **123,** 392–402 (1986).
18. R. J. Baker and J. A. Neider, *The GLIM System: Release 3.* Numerical Algorithms Group, Oxford (1978).
19. R. A. Fisher, Two new properties of mathematical likelihood. *Proc. R. Soc. A.* **144,** 285–307 (1934).
20. D. R. Cox and D. V. Hinkley, *Theoretical Statistic*, Vol. 38. Chapman and Hall, London (1974).
21. N. E. Breslow, Odds ratio estimators when the data are sparse. *Biometrika* **68,** 73–84 (1981).
22. B. Efron, Bootstrap methods: another look at the jacknife. *Am. Statist.* **7,** 1–26 (1979).
23. A. W. F. Edwards, *Likelihood.* Cambridge University Press (1972).
24. A. P. Dawid, On the concepts of sufficiency and ancillarity in the presence of nuisance parameters. *J. R. stat. Soc. B* **37,** 248–258 (1975).
25. J. M. Robins, Another approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathl Modelling* **7,** 1393–1512 (1986).
26. O. Barndorff-Nielsen, Nonformation. *Biometrika* **63,** 567–71 (1976).
27. A. S. Whittemore, Collapsibility of multidimensional contingency tables. *Jl R. stat Soc B* **40,** 328–40 (1978).
28. S. H. Shapiro Collapsing contingency tables—a geometric approach. *Am. Statist.* **36,** 43–46 (1981).
29. O. Barndorff-Nielsen, Conditionality resolutions. *Biometrika* **67,** 239–310 (1980).
30. S. Amari, Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika* **69,** 1–19 (1982).
31. D. A. Sprott, Comments on a paper by B. Efron and D. V. Hinkley. *Biometrika* **65,** 485–486 (1981).

32. J. Nelder and R. Wedderburn, Generalized linear models. *Jl R. stat. Soc. A* **135,** 370–384 (1972).
33. P. K. Sen, Asymptotic properties of maximum likelihood estimators based on conditional specification. *A. Statist.* **7,** 1019–1033 (1979).
34. B. Lindsay, Nuisance parameters, mixture models and the efficiency of partial likelihood estimators. *Phil. Trans. R. Soc. A* **296,** 639–65 (1980).

APPENDIX A

In this Appendix m_1, m_0, $\tilde{m}_1$, $\tilde{m}_0$, $\tilde{c}RD = \tilde{m}_1 - \tilde{m}_0$, P_E, $P_{\bar{E}}$, N_E, $N_{\bar{E}}$, N and β are as defined in Section 3(**B**). N_1, N_2, N_3 and N_4 are as in Section 2(**B**).

Theorem A.1

Under hypothetical rerandomizations $N\,\mathrm{Var}^A(\tilde{c}RD)$ is given by equation (1).

Proof. It follows from the equations on page 471 of Copas [9] (s_i/n_i in Copas is $\tilde{m}_1$ in our notation) that $E(\tilde{c}RD) = m_1 - m_0$ and

$$\mathrm{Var}(\tilde{c}RD) = \frac{1}{N-1}\left[\frac{N_{\bar{E}}}{N_E} m_1(1-m_1) + \frac{N_E}{N_{\bar{E}}} m_0(1-m_0) + m_0 + m_1 - 2m_0m_1 - \beta\right], \tag{A.1}$$

where $|m_1 - m_0| \leqslant \beta \leqslant \min(m_0 + m_1, 2 - m_0 - m_1)$. One can show that $\tilde{c}RD$ is asymptotically normal. Equation (A.1) implies that $N\,\mathrm{Var}^A(\tilde{c}RD)$ is given by equation (1), since equation (1) is derived from equation (A.1) by replacing $N-1$ by N and rearranging terms.

Proof of Lemma 3.1

It is straightforward to show that $R = 0$ if $\beta = 0$ and $m_1 = m_0$. Since R is a strictly decreasing function of β, we have that $R < 0$ if $m_1 = m_0$ and $\beta \neq 0$.

Next, without loss of generality, assume $m_1 - m_0 > 0$. Then, from the bounds on the range of β, we have, upon substituting $m_1 - m_0$ for β in equation (2), that $R \leqslant (2m_0 - m_1)(1 - m_1) - m_0(1 - m_0)$. Write $m_1 = m_0 + \alpha$. Then $\partial R/\partial\alpha = -(2m_0 - m_1) - (1 - m_1) = 2(m_1 - m_0) - 1 = 2\alpha - 1$. Therefore, for $0 < \alpha < \frac{1}{2}$, R is a strictly decreasing function of α, and thus it follows that $R < 0$ from the results given in the preceeding paragraph for the case $m_1 = m_0$.

It only remains to show that for $\frac{1}{2} < \alpha \leqslant 1$, $R < 0$. Since for $\frac{1}{2} < \alpha < 1$, R is strictly increasing in α, we need only evaluate R at $\max(\alpha) = 1 - m_0$, i.e., at $m_1 = 1$. But at $m_1 = 1$, $R = -m_0(1 - m_0) < 0$ assuming $m_0 > 0$.

Theorem A.2

$$\tilde{c}RD \pm 1.96\sqrt{\frac{\tilde{m}_0(1-\tilde{m}_0)N}{(N_E N_{\bar{E}})}} \tag{A.2}$$

is a 95% large sample prediction interval for $(O - \mathrm{EX})/N_E \equiv [(N_{1E} + N_{2E}) - (N_{1E} + N_{3E})]/N$.

Proof. Since $N_{1E} + N_{3E} = N_1 + N_3 - N_{1\bar{E}} + N_{3\bar{E}}$, we have

$$(O - \mathrm{EX})/N_E = \left[\tilde{m}_1 + \tilde{m}_0\frac{N_{\bar{E}}}{N_E}\right] - \frac{m_0 N}{N_E}.$$

Now $\tilde{m}_1 + \tilde{m}_0(N_{\bar{E}}/N_E)$ is an observed random variable and $m_0 N/N_E$ is a population parameter. Furthermore, since

$$E(\tilde{m}_0) = m_0 \quad \text{and} \quad \mathrm{Var}(\tilde{m}_0) = \frac{N_E m_0(1-m_0)}{(N-1)N_{\bar{E}}}$$

[9] a 95% large sample confidence interval for $(N_E/N)m_0$ is

$$\frac{N}{N_E}\tilde{m}_0 \pm 1.96\frac{N}{N_E}\sqrt{\frac{N_E\tilde{m}_0(1-\tilde{m}_0)}{NN_{\bar{E}}}} = \frac{N}{N_E}\tilde{m}_0 \pm 1.96\sqrt{\frac{\tilde{m}_0(1-\tilde{m}_0)N}{(N_E N_{\bar{E}})}}.$$

This implies that a 95% prediction interval for $(O - \mathrm{EX})/N_E$ is

$$\tilde{m}_1 + \tilde{m}_0\frac{N_{\bar{E}}}{N_E} + \frac{N}{N_E}\tilde{m}_0 \pm 1.96\sqrt{\frac{\tilde{m}_0(1-\tilde{m}_0)N}{(N_E N_{\bar{E}})}}$$

which simplifies to equation (A.2).

APPENDIX B

1. An *estimator* is a rule (function) that takes any observed data set and produces a numerical estimate of a population parameter. "Estimates" are thus the observed value of an "estimator" in the data at hand. Two estimators are identical if they give identical estimates for all possible data outcomes.
2. A *confidence interval* is a rule that gives for any data outcome an interval (the observed value of the confidence interval in the data at hand). A 95% confidence interval must by definition cover the X parameter of interest 95% of the time in hypothetical repetitions for all possible values of the model parameters.
3. *Local uniform asymptotic unbiasedness* is the statistical concept that corresponds to the epidemiologist's intuitive concept of "bias" [see Section 8(C)–(E)]. Estimators that are not locally uniformly asymptotically unbiased (i.e. intuitively biased estimators) systematically deviate from the true population parameter in hypothetical repetitions of the study to such an extent that large-sample confidence intervals based on the estimate ± 1.96 standard errors fail to give 95% probability of coverage of the true parameter (since such intervals are not centered on the true parameter on average). The usual statistical definition of exact small sample bias has no relation to the intuitive concept of bias. For example, the ratio of observed proportions, the intuitively unbiased estimator of the risk ratio, has an exact bias of infinity. Consistent estimators and even asmpotically unibased estimators that are not locally uniformly asymptotically unbiased are intuitively biased, and are statistically unsuitable for centering large-sample confidence intervals. For example the

backward elimination (i.e. preliminary test) estimator described in Section 5(B) on prior knowledge versus prior belief is consistent and asymptotically unbiased but locally uniformly asymptotically biased [see Section 8(E)].

4. *Efficient estimators* have minimum variance (asymptotic) among all (locally uniformly asymptotically unbiased) estimators. Practically, this means that 95% confidence intervals based on an efficient estimator ± 1.96 times an estimated standard error are narrower on the average than those based on inefficient estimators. In this paper we consider any estimator that is not efficient to be inefficient. Therefore, inefficient estimators include all biased estimators and all nonoptimal unbiased estimators.
5. *Inference in large samples and sparse data.* Large sample (asymptotic) inference generally refers to inference based on tests and confidence intervals which attain their nominal significance levels and coverage probabilities only in samples of sufficient size that normal approximations are adequate. In this paper, we use the term "large-sample inference" specifically to refer to situations in which the number of parameters remains fixed as the sample size increases. A more appropriate choice of terms might have been "large stratum inference." We use the term "sparse data" to refer to situations in which the number of parameters would increase with sample size as, for example, in a matched pair case control study.
6. *Consistency.* An estimator of a population parameter is consistent if the estimate obtained from a sample consisting of the entire near-infinite population equals the population parameter.
7. *Fisher consistency.* A method of estimation is Fisher consistent if, whenever the proportions in the observed data equal true population proportions, the estimate of any parameter equals its true population value.
8. *Maximum likelihood estimators and the centering of confidence intervals.* In sufficiently large samples the maximum likelihood estimator of any parameter ± 1.96 standard errors determine a 95% confidence interval. The sample size required will be much less if the maximum likelihood estimator is symmetrically distributed even in moderate sized samples. Thus confidence intervals for a given parameter of interest (e.g. the odds ratio or risk ratio) would often be constructed by exponentiating a confidence interval for the ln OR or ln RR since the distribution of the maximum likelihood estimator of the logarithm is more nearly symmetric in moderate sized samples.
9. The *minimal sufficient statistic* is the smallest summary of the data that contains all the information concerning the model parameters. For example, for inference on the probability of success based on 50 binomial trials, the total number of successes is the minimal sufficient statistic. The observed order of successes and failures contains no additional information.

APPENDIX C

Theorem C.1

In the follow-up study of Section 6(D) with simple random sampling conditional on exposure statistics, if $\mathrm{OR}_{EC} = 1$, then, in the notation defined in Sections 6(B)–(D),

$$E(\check{\mathrm{c}}\mathrm{RR}\,|\,\tilde{b}) = \mathrm{sMR} + \frac{K\tilde{b}}{2\sqrt{N}}\left[\frac{(p_{11}-p_{12})}{p_{21}p_{1+1}+p_{22}(1-p_{1+1})} + \frac{(p_{21}-p_{22})[p_{11}p_{1+1}+p_{12}(1-p_{1+1})]}{[p_{21}p_{1+1}+p_{22}(1-p_{1+1})]^2}\right] + o\left(\frac{1}{\sqrt{N}}\right),$$

where without loss of generality, we have assumed the special case $x_{1++} = x_{2++} = N$ for notational simplicity, and we use as our approximate ancillary

$$\tilde{b} = \frac{\tilde{\mathrm{u}}\mathrm{RD}_{EC} - 0}{\tilde{\mathrm{V}}\mathrm{ar}(\tilde{\mathrm{u}}\mathrm{RD}_{EC})} = \frac{(x_{1+1}-x_{2+1})/N}{\sqrt{\frac{2}{N}\left(\frac{x_{++1}}{2N}\right)\left(\frac{1-x_{++1}}{2N}\right)}}$$

$$p_{1+1} \equiv p(C\,|\,E), \text{ and } K \equiv \sqrt{\frac{2}{N}(p_{1+1})(1-p_{1+1})}.$$

Proof. Use the fact that

$$E\left(\frac{x_{1+1}}{N}\,\middle|\,\tilde{b}\right) = p_{1+1} + \frac{K\tilde{b}}{2\sqrt{N}} + o(N^{-1/2})$$

to expand conditional on $\tilde{b}$,

$$E(\check{\mathrm{c}}\mathrm{RR}) = E\left[\frac{x_{111}}{x_{1+1}}\frac{x_{1+1}}{N} + \frac{x_{112}}{x_{1+2}}\frac{x_{1+2}}{N}\right] \Big/ E\left[\frac{x_{211}}{x_{2+1}}\frac{x_{2+1}}{N} + \frac{x_{212}}{x_{2+2}}\frac{x_{2+2}}{N}\right]$$

in a Taylor series.

It follows that a sufficient condition for $\check{\mathrm{c}}\mathrm{RR}$ to be asymptotically biased for the sMR conditional on $\tilde{b}$ is that $\tilde{b} \neq 0$ and that the effect of C on disease be in the same direction in the exposed and unexposed.

Theorem C.2

In the follow-up study of Section 6(D) with simple random sampling with conditional on exposure and $\mathrm{OR}_{EC} = 1$,

$$\mathrm{Var}^A(\check{\mathrm{c}}\mathrm{RR}) \geqslant \mathrm{Var}^A(\check{\mathrm{r}}\mathrm{sMR}) = \mathrm{Var}^A(\check{\mathrm{c}}\mathrm{RR}\,|\,\tilde{b}=0) = \mathrm{Var}^A(\check{\mathrm{r}}\mathrm{sMR}\,|\,\tilde{b}=0)$$

with equality only if, $\forall \tilde{b}$ asymptotic bias $(\check{\mathrm{c}}\mathrm{RR}\,|\,\tilde{b}) = 0$ for the sMR where Var^A are to $O(N^{-1})$.

Proof. We first prove the following lemma.

Lemma I

In the follow-up study of Section 6(D) with simple random sampling with $\mathrm{OR}_{EC} = 1$, to $O(N^{-1})$

$$E[\mathrm{Var}^A(\check{\mathrm{c}}\mathrm{RR}\,|\,\tilde{b})] = E[\mathrm{Var}^A(\mathrm{rsMR}\,|\,\tilde{b})] = \mathrm{Var}^A(\check{\mathrm{c}}\mathrm{RR}\,|\,\tilde{b}=0) = \mathrm{Var}^A(\check{\mathrm{r}}\mathrm{sMR}\,|\,\tilde{b}=0)$$

Proof. $E[\mathrm{Var}^A(\check{\mathrm{c}}\mathrm{RR}\,|\,\tilde{b})] = \mathrm{Var}^A[\check{\mathrm{c}}\mathrm{RR}\,|\,\tilde{b} = E(\tilde{b}) = 0]$ to $O(N^{-1})$ by the usual properties of large sample expectations. A similar result holds with $\check{\mathrm{r}}\mathrm{sMR}$ in place of $\check{\mathrm{c}}\mathrm{RR}$. But if $\tilde{b} = 0$, then $\check{\mathrm{c}}\mathrm{RR} = \check{\mathrm{r}}\mathrm{sMR}$, proving the lemma.

Proof of Theorem C.2. Since $E(\tilde{r}sMR - sMR|\tilde{b}) = o(N^{-1/2})$, $Var^A[E(\tilde{r}sMR|\tilde{b})] = 0$ to $O(N^{-1})$. Thus we have equality in Theorem C.2 if and only if $Var^A[E(\tilde{c}RR|\tilde{b})] = 0$ to $O(N^{-1})$ which will occur if and only if $E[(\tilde{c}RR - sMR|\tilde{b})] = o(N^{-1/2})$ Q.E.D.

APPENDIX D

Theorem D.1

In the unmatched case-control study of Section 6(E) if $OR_{DC|\bar{E}} = 1$ so that $\tilde{c}OR = \tilde{r}sOR$, $Var^A(\tilde{u}sOR) \leqslant Var^A(\tilde{c}OR)$ with equality if and only if $OR_{EC|\bar{D}} = 1$ as well where Var^A are to $O(N^{-1})$.

Proof. From the results of Section 6, since $\tilde{c}OR$ is $\tilde{r}sOR$, the theorem will be true if

$$OR_{EC|\bar{D}} = 1 \quad \text{implies} \quad E(\tilde{u}sOR|\tilde{A}) = sOR \quad \text{to} \quad O(N^{-1/2}) \tag{D.1}$$

when $\tilde{A}$ is an approximate ancillary corresponding to $OR_{DC|\bar{E}} = 1$. Equation (D.1) will hold if and only if

$$E(\tilde{u}sOR|\tilde{A}) = E(\tilde{c}OR|\tilde{A}) \quad \text{to} \quad O(N^{-1/2}). \tag{D.2}$$

Equation (D.2) will hold if and only if to $O(N^{-1/2})$ conditional on $\tilde{A}$,

$$E(XYF + CBF) = E\left[\frac{(XF + CB)(YF + EB)}{(F + B)}\right] \Leftrightarrow 0 = E(C - X)E(BF)E(Y - E),$$

where

$$F = x_{221}\, XF = x_{211}\, YF = x_{121}\, QXYF = x_{111},\ B = x_{222},\ CB = x_{212},\ EB = x_{122}\, UCBE = x_{112}.$$

But $E(X - C|\tilde{A})$ is $O(N^{-1/2})$ since $X/C = \tilde{u}OR_{DC|\bar{E}}$. Therefore equation (D.2) will hold if and only if $E(Y - E|\tilde{A})$ is $O(N^{-1/2}) \Leftrightarrow OR_{EC|\bar{D}} = 1$ since Y/E is $\tilde{u}OR_{EC|\bar{D}}$.

By essentially the same proof, we have.

Corollary D.1

The theorem is true with the roles of $OR_{DC|\bar{E}}$ and $OR_{EC|\bar{D}}$ reversed.

Corollary D.2

In the unmatched follow-up study of Section 6(D), if $OR_{DC|\bar{E}} = 1$, $Var^A(\tilde{u}sMR) \leqslant Var^A(\tilde{c}RR)$ to $O(N^{-1})$ with equality only if $OR_{EC} = 1$ as well.

Computers Math. Applic. Vol. 18, No. 5, pp. 475, 1989
Printed in Great Britain.

0097-4943/89 $3.00 + 0.00

ERRATA TO "THE FOUNDATIONS OF CONFOUNDING IN EPIDEMIOLOGY"

Computers Math. Applic. **14**(9–12), 869–916 (1987)

J. M. ROBINS
Occupational Health Program, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, U.S.A.

H. MORGENSTERN
Division of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90024, U.S.A.

Page/Paragraph/Line	Uncorrected	Corrected
877/3rd to last/3rd to last	is not given	may not be given
880/2nd to last/equation	$OR_{EC \mid \bar{D}} = 1$	$OR_{EC \mid \bar{D}} = 1$
882/2nd to last/3rd to last	replacing "exposed"	interchanged with "exposed"
890/last/2nd to last	unexposed sample,	unexposed sample
890/last/2nd to last	in the data differs	differs
891/last/last	values	value
900/3/equation (6)	$S_1 = p[x_{11+}$	$S_1 = p[x \mid x_{11+}$
905/last/denominator of $\Lambda(E, C, D)$	$p(D_0 \mid C, C)$	$p(D_0 \mid E, C)$
909/last/4	Since	since
909/last/4	The final	, the final
910/2/2	$\hat{r}\theta_1 z_{\alpha/2\pm}$	$\hat{r}\theta_1 \pm z_{\alpha/2}$
910/2/5	$z\alpha/2$	$z_{\alpha/2}$
913/Ref. [9]	Copaz	Copas
913/Ref. [25]	Another	A new
914/Proof of Lemma 3.1/2	$\beta \neq 0$	$\beta \neq 0$ [9].
915/Theorem C2/1	sampling with conditional	sampling conditional