

Conditional logistic analysis of case-control studies with complex sampling

BRYAN LANGHOLZ*

*Department of Preventive Medicine, University of Southern California, 1540 Alcazar
Street CHP-220, Los Angeles, CA 90089-9011, USA*
langholz@hsc.usc.edu

LARRY GOLDSTEIN

*Department of Mathematics, University of Southern California, 1042 West 36th Place, Los Angeles,
CA 90089-1113, USA*

SUMMARY

Methods for the analysis of unmatched case-control data based on a finite population sampling model are developed. Under this model, and the prospective logistic model for disease probabilities, a likelihood for case-control data that accommodates very general sampling of controls is derived. This likelihood has the form of a weighted conditional logistic likelihood. The flexibility of the methods is illustrated by providing a number of control sampling designs and a general scheme for their analyses. These include frequency matching, counter-matching, case-base, randomized recruitment, and quota sampling. A study of risk factors for childhood asthma illustrates an application of the counter-matching design. Some asymptotic efficiency results are presented and computational methods discussed. Further, it is shown that a ‘marginal’ likelihood provides a link to unconditional logistic methods. The methods are examined in a simulation study that compares frequency and counter-matching using conditional and unconditional logistic analyses and indicate that the conditional logistic likelihood has superior efficiency. Extensions that accommodate sampling of cases and multistage designs are presented. Finally, we discuss the analysis methods presented here to other approaches, compare counter-matching and two-stage designs, and suggest areas for further research.

Keywords: Binary data; Case-control studies; Counter-matching; Epidemiology; Logistic model; Multistage study designs.

*pl. check case-control
- ' ' or ' ' -*

1. INTRODUCTION

Case-control studies are the most commonly used study design in chronic disease epidemiology and have played a central role in identifying risk factors for disease (Breslow, 1996). Classical unmatched case-control studies consist of a random sample of the diseased subjects (cases) from a population and a random sample of the non-diseased (controls). The differences in exposure levels between cases and controls serves as a basis for estimation of the association of exposure with disease. Inference is typically based on Mantel–Haenszel methods, unconditional logistic likelihood, and, if the case-control sets are small, conditional logistic likelihood methods (e.g., Breslow and Day, 1980). The probabilistic paradigm typically used to develop the statistical theory is depicted in Figure 1. Case and control covariate values

*To whom correspondence should be addressed.

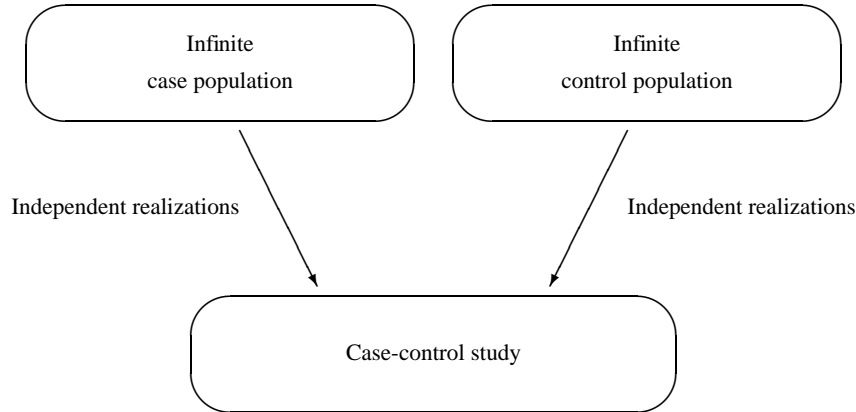


Fig. 1. Classical model for case-control data.

are assumed to be independent representatives from their respective (assumed ‘infinite’) populations. This approach is ‘retrospective’ in that case/control status is fixed and covariate values are random. However, the analysis methods typically treat the data as ‘prospective’, with case/control status as random and covariates fixed in the case-control set. The validity of this analysis approach for estimation of odds ratio parameters in the logistic model is based on ‘inverting’ a corresponding retrospective logistic model (e.g., Cornfield, 1951; Prentice and Pyke, 1979; Wild, 1991). While the retrospective model serves well when cases and controls are independently sampled, it cannot easily accommodate more complex sampling. In particular, this classical approach is not well suited for designs in which the sampling of controls depends on characteristics of the set of cases. An alternative approach, based on finite population sampling, is illustrated in Figure 2. Here, we are given a finite population, or ‘study base’, that consists of all cases and controls in a finite population. We then sample controls (and possibly cases) from this study base to obtain our case-control set. This approach is ‘prospective’ in that the conditioning is on the covariate values and disease status is random, and the case-control study is simply a sample from this larger population. Although this approach has been well developed for rate ratio estimation in nested case-control studies (e.g. Oakes, 1981; Borgan *et al.*, 1995; Langholz and Goldstein, 1996), and discussed in the unmatched case-control context in terms of control selection principles (e.g. Wacholder *et al.*, 1992; Rothman and Greenland, 1998), analytic methods have not been derived from this perspective. In this paper, we will show that the finite sampling approach leads to conditional logistic likelihood methods and can accommodate quite general sampling in a very natural way.

2. A STUDY OF RISK FACTORS FOR EARLY CHILDHOOD ASTHMA

The particular goals and available resources should always be considered in the design of an epidemiologic case-control study. Until recently, these design considerations have been limited to how much to stratify and how large a sample to take. Although the standard case-control design will continue to serve for ‘general purpose’ studies, it can be advantageous, in terms of validity and cost-efficiency, to tailor the study design to exploit particular features of the study setting. As an example which illustrates how study features can be used to advantage, we describe a study of risk factors for early childhood

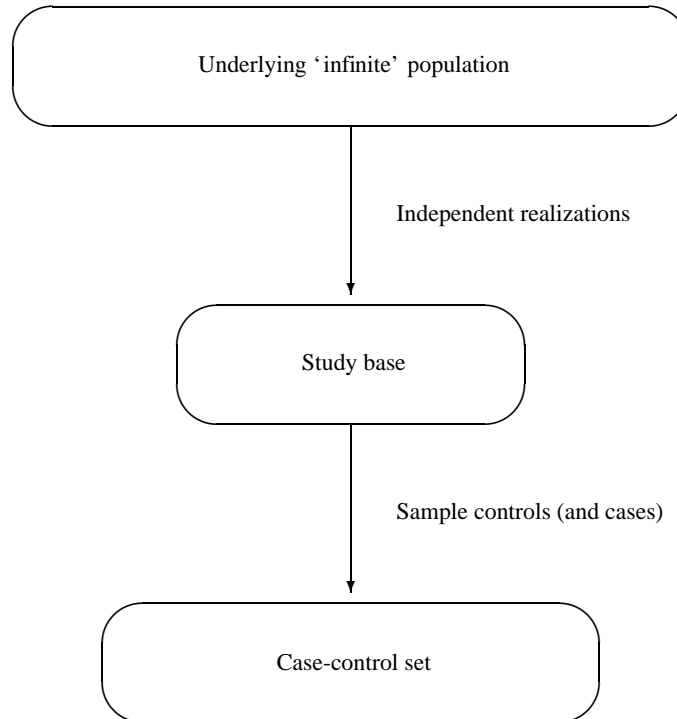


Fig. 2. Finite population model for case-control studies.

asthma. This study is part of the Children's Health Study currently underway at the University of Southern California Department of Preventive Medicine (Peters *et al.*, 1999). Children from 12 communities and three grade levels were enrolled to participate in a longitudinal study of childhood respiratory health. Information collected at enrolment to the study included whether the student had ever been diagnosed with asthma, exposed to tobacco smoke *in utero* and during childhood, and other factors that are potentially related to respiratory health. Using this baseline data, it was found that an asthma diagnosis age five or younger was associated with maternal smoking during pregnancy (*in utero* smoke exposure) but not with environmental tobacco smoke exposure in early childhood. The association with *in utero* smoke exposure and asthma has a plausible biologic explanation. Briefly, toxins in tobacco smoke are known to pass from pregnant woman who smoked to her fetus via the placenta and it has been established that these toxins can adversely affect the fetus (e.g. Werler, 1997). Thus, it is not unreasonable that smoke exposure may affect development of various organ systems, including respiratory, in the fetus.

Now, the adverse affects of tobacco smoke may be influenced by genetic differences in enzymatic detoxification systems. Detoxification of tobacco smoke is, to a large degree, controlled by GST-T1 and GST-M1 genes, with 'wild' genotype coding for the detoxification enzyme and the 'null' not. If this biologic model is correct then it would be expected that, among children whose mothers smoked during pregnancy, null genotype ('genotype positive') children would be at higher risk of asthma than wild genotype ('genotype negative') children. In order to test this hypothesis, we undertook the early asthma risk factors study (EARS) using a sample of 3465 'non-active' subjects from the main study which

included 249 cases with asthma diagnosis at age five years or younger. The goals of this case-control study are, first, to augment the smoking-during-pregnancy information (this was just a yes/no question in the baseline questionnaire) to assess dose response and, second, to ascertain the child's genotype status and to test the detoxification hypotheses. Additional smoking exposure information is obtained in a short interview and genotype status is obtained using standard PCR methods from buccal cells collected from subjects by swabbing the inside of the mouth.

Considerations in choosing the control sampling method were (1) we intended use all cases and sample controls from the 96 study bases determined by community, gender and grade, (2) the study bases are relatively small (13–52 subjects) with numbers of cases ranging from zero to seven and (3) we have yes/no information on *in utero* smoke exposure for all of these subjects. Simple frequency matching of controls within the study bases using a standard conditional logistic analysis would provide an appropriate and valid design and analysis given considerations (1) and (2). However, it has been shown that, with *in utero* smoke exposure a key variable in our analysis, it can be very efficient to use consideration (3) in a 'two-stage' sampling technique (Breslow and Cain, 1988; Zhao and Lipsitz, 1992). In this design, controls are sampled independently from the *in utero* smoke exposed and unexposed with a 'balanced' design (with equal numbers from each exposure category) being suggested as close to optimal. However, the analysis methods for this design are not valid when there are small numbers of cases per study base. Counter-matching is another design which takes advantage of exposure-related information in the sampling but has only been described for individually matched case-control studies (Langholz and Borgan, 1995). In Section 4.7, we present a natural extension of the counter-matching design for unmatched case-control data and the EARS study is of this design. The appropriate counter-matching conditional logistic likelihood is valid with small numbers of cases per study base and will be much more efficient than frequency matching for investigation of our research questions.

This study is on-going and we do not yet have data for analysis. Thus, in Section 9, we illustrate how the exposure-genotype data will be analysed using simulated genotype data.

3. NOTATION AND MODELS

Following Figure 2, consider a study base of n subjects and let $\mathcal{R} = \{1, \dots, n\}$ index the subjects. With D_i the disease indicator and Z_i covariates for subject i , the probability of disease is assumed to follow the proportional odds (logistic) model:

$$\text{pr}(D_i = 1|Z_i) = \frac{\lambda_0 r(Z_i; \beta_0)}{1 + \lambda_0 r(Z_i; \beta_0)} = \frac{\lambda_0 r_i}{1 + \lambda_0 r_i} \quad (1)$$

where λ_0 is the baseline odds and $r(Z_i, \beta_0)$ is the odds ratio associated with covariate value Z_i . We will always condition on the covariate value and so will suppress it in the following. Let \mathbf{D} be the (random) set of indices for disease subjects. Then

$$\begin{aligned} \text{pr}(\mathbf{D} = \mathbf{d}) &= \prod_{i \in \mathbf{d}} \frac{\lambda_0 r_i}{1 + \lambda_0 r_i} \prod_{j \in \mathcal{R} \setminus \mathbf{d}} \frac{1}{1 + \lambda_0 r_j} \\ &= \prod_{i \in \mathbf{d}} \lambda_0 r_i \prod_{j \in \mathcal{R}} \frac{1}{1 + \lambda_0 r_j} \\ &= \lambda_0^d r_{\mathbf{d}} Q_{\mathcal{R}} \end{aligned}$$

where, throughout, we use the italic letter for the number in the set, (for example, $s = |\mathbf{s}|$ is the number of subjects in \mathbf{s}), $r_{\mathbf{d}} = \prod_{i \in \mathbf{d}} r_i$ and $Q_{\mathcal{R}} = \prod_{j \in \mathcal{R}} \frac{1}{1 + \lambda_0 r_j}$. Replacing the true parameters by dummy

variables, the likelihood for estimation of λ_0 and β_0 is given by

$$L(\lambda, \beta) = \lambda^D r_{\mathbf{D}}(\beta) Q_{\mathcal{R}}(\lambda, \beta). \quad (2)$$

This likelihood, of course, requires that Z be known for all subjects in the study base and, if n is large, obtaining this information is usually neither practical nor cost-efficient. So methods for sampling the study base in which Z information is obtained only for the sampled subjects are very desirable. For now, we consider inclusion of all diseased ‘cases’ and sampling of non-diseased ‘controls’. Sampling of cases is considered in Section 10. In the particular (and common) situation where cases are randomly sampled with equal probability of inclusion, the methods are essentially the same as when all cases are used. Because we will only have complete covariate information on subjects in the case-control set, it is natural to base the likelihood on the probability of a set of cases given the case-control set. Let $\tilde{\mathcal{R}}$ be the sampled case-control set and $\pi(\mathbf{r}|\mathbf{s})$ be the probability that \mathbf{r} is the sampled case-control set given that the case set is \mathbf{s} . Implicit in the specification of the π s is the information available in the study base needed to do the sampling. Then, applying Baye’s theorem and the proportional odds model (1), we have that

$$\begin{aligned} \text{pr}(\mathbf{D} = \mathbf{d} | \tilde{\mathcal{R}} = \mathbf{r}) &= \frac{\text{pr}(\tilde{\mathcal{R}} = \mathbf{r} | \mathbf{D} = \mathbf{d}) \text{pr}(\mathbf{D} = \mathbf{d})}{\sum_{\mathbf{s} \subset \mathcal{R}} \text{pr}(\tilde{\mathcal{R}} = \mathbf{r} | \mathbf{D} = \mathbf{s}) \text{pr}(\mathbf{D} = \mathbf{s})} \\ &= \frac{\lambda_0^d r_{\mathbf{d}} Q_{\mathcal{R}} \pi(\mathbf{r}|\mathbf{d})}{\sum_{\mathbf{s} \subset \mathcal{R}} \lambda_0^s r_{\mathbf{s}} Q_{\mathcal{R}} \pi(\mathbf{r}|\mathbf{s})} \\ &= \frac{\lambda_0^d r_{\mathbf{d}} \pi(\mathbf{r}|\mathbf{d})}{\sum_{\mathbf{s} \subset \mathcal{R}} \lambda_0^s r_{\mathbf{s}} \pi(\mathbf{r}|\mathbf{s})}. \end{aligned}$$

Thus, the likelihood for the case-control set is given by

$$L(\lambda, \beta) = \frac{\lambda^D r_{\mathbf{D}}(\beta) \pi(\tilde{\mathcal{R}}|\mathbf{D})}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}} \lambda^s r_{\mathbf{s}}(\beta) \pi(\tilde{\mathcal{R}}|\mathbf{s})}. \quad (3)$$

This likelihood accommodates quite general sampling through specification of the case-control set selection probabilities $\pi(\mathbf{r}|\mathbf{s})$. In particular, the selection of controls can depend on the composition of cases in the study base. Because it is based on a ‘valid’ probability, (3) is a true likelihood and the usual ‘basic likelihood properties’ (expectation of the score equal to zero and the variance of the score equal to the expected information, at the true parameters) hold. As will be seen in the examples given in the next section, there is often cancellation of common factors in the $\pi(\tilde{\mathcal{R}}|\mathbf{s})$ so that the likelihood used to fit the data will have the form

$$L(\lambda, \beta) = \frac{\lambda^D r_{\mathbf{D}}(\beta) w_{\tilde{\mathcal{R}}}(\mathbf{D})}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}} \lambda^s r_{\mathbf{s}}(\beta) w_{\tilde{\mathcal{R}}}(\mathbf{s})} \quad (4)$$

where $w_{\tilde{\mathcal{R}}}(\mathbf{s})$ are convenient ‘risk weights’. Note that (4) is the natural generalization of the likelihood for ‘individually matched’ case-control sets (with one case per set) we derived earlier in the risk set sampling setting, with the risk set corresponding to the study base in the current context (Borgan *et al.*, 1995; Langholz and Goldstein, 1996). Consistent with that setting and with terminology often used to describe ‘prospectively’ collected data, we will refer to a collection of study bases as a ‘cohort’. The total likelihood from case-control sets sampled from a cohort is simply the product of likelihood contributions of the form of (4) from each of the study bases. Thus, in what follows, we consider a single study base.

4. EXAMPLES

The methods described in the last section accommodate very general sampling. In this section, we show how the methods are applied to a variety of interesting case-control designs. It is not our intention to evaluate these designs or to explore situations where they are potentially useful. But, by presenting them, we hope to show that there is much room for the development of creative study designs. We start off with some established designs and then present a few new ones.

4.1 *Frequency matching*

In this commonly used design, controls are selected in numbers proportional to the number of cases, usually within strata defined by age group or some other demographic variable. So suppose we sample md from the $n - d$ controls in the full set. Then

$$\pi(\mathbf{r}|\mathbf{d}) = \binom{n-d}{md}^{-1}$$

for each set \mathbf{r} of size $(m+1) \times d$ containing \mathbf{d} and the conditional likelihood is given by

$$\frac{\lambda^D r_{\mathbf{D}}(\beta) \binom{n-D}{mD}^{-1}}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}:s=D} \lambda^s r_{\mathbf{s}}(\beta) \binom{n-D}{mD}^{-1}} = \frac{r_{\mathbf{D}}(\beta)}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}:s=D} r_{\mathbf{s}}(\beta)}.$$

With risk weights identically equal to one, this is the usual conditional likelihood for ‘simple’ unmatched case-control data. From the full study base likelihood, it is easy to see that the sufficient statistic for λ is D , the number of cases. Because the frequency matching likelihood conditions on sets with size equal to the number of cases, the baseline odds parameter λ is ‘conditioned out’ and cannot be estimated. This reflects the requirements of the sampling. The study base does not need to be enumerated, it is only required that the case-control set controls are randomly sampled from all controls the study base. It is this feature that makes frequency matching (and some of the other designs discussed below) desirable for ‘population-based’ studies.

4.2 *Full study base*

It is not obvious that the study base likelihood (2) is a special case of the case-control likelihood (4). To see that it is, note that for the full study base $\pi(\mathbf{r}|\mathbf{d}) \equiv I(\mathbf{r} = \mathcal{R})$ for all \mathbf{d} . Thus,

$$\begin{aligned} L(\lambda, \beta) &= \frac{\lambda^D r_{\mathbf{D}}(\beta)}{\sum_{\mathbf{s} \subset \mathcal{R}} \lambda^s r_{\mathbf{s}}(\beta)} \\ &= \frac{\lambda^D r_{\mathbf{D}}(\beta) Q_{\mathcal{R}}(\lambda, \beta)}{\sum_{\mathbf{s} \subset \mathcal{R}} \lambda^s r_{\mathbf{s}}(\beta) Q_{\mathcal{R}}(\lambda, \beta)} \\ &= \lambda^D r_{\mathbf{D}}(\beta) Q_{\mathcal{R}}(\lambda, \beta) \end{aligned} \tag{5}$$

which is the study base likelihood (2). The denominator in (5) is one because the terms in the denominator define a probability distribution over all possible sets of cases for any λ and β .

Table 1. Asymptotic relative efficiencies under the null for frequency matching, case-base sampling, and counter-matching for a univariate exposure Z

1 : m frequency matching	Case-base sampling ^a	1 : 1 counter-matching ^b
Multiple cases in the case-control sets		
$\frac{m/(m+1)}{1-p}$	$\frac{m/(m+1)}{1-p\frac{m}{m+1}}$	$(\eta\gamma + (1-\eta)(1-\gamma))/(1-p)$ $- \frac{p}{1-p} \frac{p_Z q_Z}{p_{\tilde{Z}} q_{\tilde{Z}}} (\eta\gamma - (1-\eta)(1-\gamma))^2$
Individually matched case-control sets		
$m/(m+1)$	$m/(m+1)$	$\eta\gamma + (1-\eta)(1-\gamma)$

^a The m here refers to the number of subjects per case required to yield an expected m controls per case.

^b Sampling stratum variable \tilde{Z} is a surrogate for Z , both dichotomous. η and γ are the sensitivity and specificity of \tilde{Z} for Z , $p_{\tilde{Z}} = \text{pr}(\tilde{Z} = 1)$, $p_Z = \text{pr}(Z = 1)$, and p is the probability of disease.

4.3 Case-base sampling

In this design, md subjects are sampled from the n regardless of case-control status (Kupper *et al.*, 1975; Miettinen, 1982; Prentice, 1986). This control selection is characterized by

$$\pi(\mathbf{r}|\mathbf{d}) = \frac{\binom{md-(r-d)}{d}}{\binom{n}{md}}$$

for \mathbf{r} containing \mathbf{d} . Gathering terms that have a common risk weight, this yields the likelihood

$$L(\lambda, \beta) = \frac{\lambda^D r_{\mathbf{D}}(\beta)}{\sum_{k=|\tilde{\mathcal{R}}|}^{|\tilde{\mathcal{R}}|} \binom{mD-(|\tilde{\mathcal{R}}|-k)}{k} \sum_{\{s \subset \tilde{\mathcal{R}}: s=k\}} \lambda^s r_s(\beta)}$$

Unlike frequency matching, in which the baseline odds parameter is conditioned out, λ_0 is estimable in case-base likelihood. The information to estimate λ_0 comes from the number of cases that are selected as part of the md randomly sampled subjects. This provides an alternative approach to estimation of the baseline odds (or risk) from such data than previously suggested (Flanders *et al.*, 1990; Sato, 1992). As shown in Table 1, the ability to estimate λ_0 is obtained at some efficiency cost for the estimation of β_0 , relative to frequency matching. Like frequency matching, enumeration of the study base is not required, only that the study base is randomly sampled.

4.4 Independent Bernoulli trial sampling of controls: randomized recruitment

In the randomized recruitment design, controls (and cases) are included into the case-control sample according to the outcome of a Bernoulli trial where the probability of inclusion may depend on case-control status and covariate information available on the subject. This design was described in the context of a study of indoor radon exposure and lung cancer in which it was desired to over-sample non-smoking cases and smoking controls (Weinberg and Wacholder, 1990). Here, we consider only sampling of controls

but case sampling is easily incorporated using the methods described in Section 10. Let ρ_j be the probability of including j if j were a control in the case-control set. Then

$$\begin{aligned}\pi(\mathbf{r}|\mathbf{d}) &= \prod_{j \in \mathbf{r} \setminus \mathbf{d}} \rho_j \prod_{j \in \mathcal{R} \setminus \mathbf{r}} (1 - \rho_j) \\ &= \frac{1}{\rho_{\mathbf{d}}} [\rho_{\mathbf{r}} (1 - \rho)_{\mathcal{R} \setminus \mathbf{r}}].\end{aligned}$$

In the likelihood, the factors common to \mathbf{r} drop out and each of the ρ_j^{-1} can be incorporated into the corresponding ‘odds’ term as follows:

$$\begin{aligned}L(\lambda, \beta) &= \frac{[\prod_{i \in \mathbf{D}} \frac{1}{\rho_i}] \lambda^D r_{\mathbf{D}}(\beta)}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}} [\prod_{k \in \mathbf{s}} \frac{1}{\rho_k}] \lambda^s r_{\mathbf{s}}(\beta)} \\ &= \frac{\prod_{i \in \mathbf{D}} [\frac{1}{\rho_i} \lambda r_i(\beta)] Q_{\tilde{\mathcal{R}}}(\lambda, \beta)}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}} \prod_{k \in \mathbf{s}} [\frac{1}{\rho_k} \lambda r_k(\beta)] Q_{\tilde{\mathcal{R}}}(\lambda, \beta)}\end{aligned}\tag{6}$$

$$= \prod_{i \in \mathbf{D}} \left[\frac{1}{\rho_i} \lambda r_i(\beta) \right] Q_{\tilde{\mathcal{R}}}(\lambda, \beta)\tag{7}$$

where $Q_{\tilde{\mathcal{R}}} = \prod_{k \in \tilde{\mathcal{R}}} \{1 + \frac{1}{\rho_k} \lambda r_k(\beta)\}^{-1}$. The denominator in (6) equals one for the same reason as it does for the full set. This proves that the conditional logistic likelihood is exactly of an unconditional logistic form when sampling of controls is of independent Bernoulli trial type. Unlike the previous examples, randomized recruitment requires enumeration of the study base in order to independently sample subjects.

4.5 Quota sampling

This is a completely new design that generalizes the design for risk set sampled data described by us earlier Borgan *et al.* (1995). Let $C_i \in \{1, \dots, L\}$ be known for cases but unknown for controls in the study base and d_l be the number of cases with $C_i = l$. Associated with each level of C is a quota $m_l d \geq 0$. (Setting $m_l = 0$ corresponds to no quota.) Controls are selected sequentially and without replacement until all the quotas are met, i.e. there are $m_l d$ or more subjects (counting the cases) with $C = l$. Such a design should be useful if it is desired to assure a minimum level of absolute precision in the study. In particular, by setting a quota, this provides a way to assure that ‘rare cells’ are adequately represented in the case-control set. This may be especially important when the cohort consists of multiple study bases and the rarity of levels of C is highly variable. While this provides some intuition for the potential benefit of this design, further work needs to be done to determine if the design meets the expectations. The control selection probabilities associated with quota sampling are based on a negative multi-hypergeometric distribution (Schuster and Sype, 1987) with

$$\pi(\mathbf{r}|\mathbf{d}) = \frac{\sum_l (m_l d - d_l) I(|\mathbf{r}_l| = m_l d)}{\binom{n-d}{|\mathbf{r}|-d-1}} \frac{1}{n - (|\mathbf{r}| - d - 1)}$$

for sets with for sets with $|\mathbf{r}_l| \geq m_l d$ for all l and $|\mathbf{r}_l| = m_l d$ for at least one l . This yields the likelihood

$$L(\beta) = \frac{r_{\mathbf{D}}(\beta) [\sum_l (m_l D - D_l) I(|\tilde{\mathcal{R}}_l| = m_l D)]}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}: s=D} r_{\mathbf{s}}(\beta) [\sum_l (m_l D - s_l) I(|\tilde{\mathcal{R}}_l| = m_l D)]}$$

The sampling does not require enumeration of the study base as reflected in the weights, which only depend on the number of subjects sampled from each level of C . Thus, this design is appropriate for population based studies.

4.6 Independent Bernoulli sampling with stopping

It is not hard to imagine a ‘rare exposure’ situation where many controls will need to be selected to meet a quota in the quota sampling design. In such situations, it may be advantageous to sample controls sequentially but only include them based on the outcome of a Bernoulli trial. Or, from the point of view of the randomized recruitment design in Section 4.4, controls are examined sequentially and included depending upon the outcome of a Bernoulli trial until the quota in a ‘rare cell’ is met. This, in fact, is closer to the design described in Weinberg and Wacholder (1990) than pure Bernoulli trial sampling. Suppose that there is a ‘rare cell’ that is identified by $C = 1$ and suppose that all subjects in this cell are included with probability one, i.e. $C_i = 1 \Rightarrow \rho_i = 1$ and (non-diseased) subjects with $C_i \neq 1$ are sampled with $\rho_i < 1$. Then, with a quota of $m_1 d$ for subjects with $C = 1$,

$$\pi(\mathbf{r}|\mathbf{d}) = (m_1 d - d_1) \binom{\frac{1}{\rho}}{\mathbf{d}} K(\mathbf{r})$$

for sets \mathbf{r} with $|\mathbf{r}_1| = m_1 d$, where $K(\mathbf{r})$ is a common factor that depends only on the entire case-control set. This yields the likelihood

$$L(\beta) = \frac{\left[\frac{1}{\rho} r(\beta)\right]_{\mathbf{D}} (m_1 D - D_1)}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}: s=D} \left[\frac{1}{\rho} r(\beta)\right]_{\mathbf{s}} (m_1 D - s_1)}.$$

4.7 Counter-matching

Counter-matching for individually matched case-control studies has been described in earlier work (e.g., Langholz and Borgan, 1995; Langholz and Goldstein, 1996; Steenland and Deddens, 1997). These references discuss the efficiency benefits of counter-matching on exposure-related information. Here, we present the generalization of counter-matching to multiple cases per case-control set. It is assumed that the sampling strata variable $C \in \{1, \dots, L\}$ is known for all subjects in the study base. Case-control status and C define a $2 \times L$ table as shown in the top part of Figure 3. In counter-matching, the marginal total in the sampling stratum l is fixed to a value proportional to the number of cases, $m_l d$. Then $m_l d - d_l$ controls are randomly sampled without replacement from the $n_l - d_l$ total controls in stratum l . This is illustrated in the bottom half of Figure 3. Counter-matching control selection is characterized by

$$\pi(\mathbf{r}|\mathbf{d}) = \left[\prod_{l=1}^L \binom{n_l - d_l}{m_l d - d_l} \right]^{-1}$$

for \mathbf{r} with $|\mathbf{r}_l| = m_l d$, $l = 1, \dots, L$ with likelihood

$$L(\beta) = \frac{r_{\mathbf{D}}(\beta) \left[\prod_{l=1}^L \frac{n_l(n_l-1)\dots(n_l-D_l+1)}{m_l D(m_l D-1)\dots(m_l D-D_l+1)} \right]}{\sum_{\mathbf{s} \subset \tilde{\mathcal{R}}: s=D} r_{\mathbf{s}}(\beta) \left[\prod_{l=1}^L \frac{n_l(n_l-1)\dots(n_l-s_l+1)}{m_l D(m_l D-1)\dots(m_l D-s_l+1)} \right]}. \quad (8)$$

The sampling conditions on the number of cases, so that λ_0 is not estimable from the conditional likelihood. As shown below, asymptotic relative efficiency and simulation study results indicate that counter-matching for unmatched case-control studies will have efficiency benefits in the same situations as the individually matched version.

5. ASYMPTOTIC THEORY

It is straightforward to show that the general conditional logistic likelihood (4) has ‘basic likelihood properties’, i.e. that the expectation of the score equals zero and the variance of the score equals the

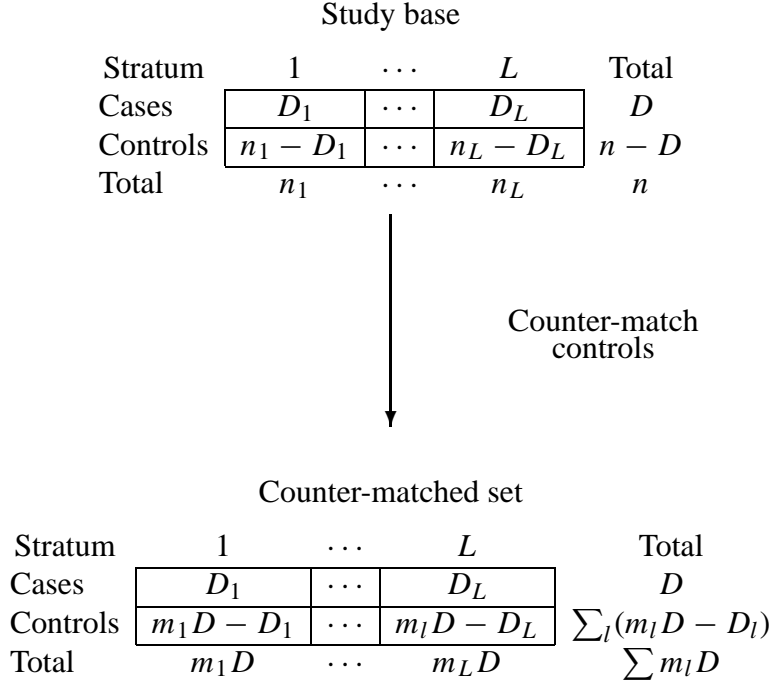


Fig. 3. Counter-matching for unmatched case-control studies.

expected information, at the true parameters. We conjecture that the estimators have the usual asymptotic properties of consistency and normality, and proving these rigorously is the subject of continuing research. The asymptotic theory is more difficult than in the usual independence setting due to the dependence induced between subjects in the sample by allowing control selection to depend on the characteristics of the case set. One design in which the sampling is independent is randomized recruitment (Section 4.4). In this ‘truly independent’ situation, the asymptotic theory is straightforward.

The information from (4) under the log-linear form is given by

$$\mathcal{I}(\beta) = \sum_{j \in \tilde{\mathcal{R}}} \mathbf{Z}_j^{\otimes 2} p_j (1 - p_j) + \sum_{j, k \in \tilde{\mathcal{R}}, j \neq k} \mathbf{Z}_j \mathbf{Z}_k (p_{j,k} - p_j p_k)$$

where $p_j = \text{pr}(j \in \mathbf{D} | \tilde{\mathcal{R}})$ and $p_{j,k} = \text{pr}(j, k \in \mathbf{D} | \tilde{\mathcal{R}})$. Technical difficulties arise in controlling the remainder terms away from the null, but we do not discuss these here. In the particular case where $\beta_0 = 0$, we have derived the asymptotic relative efficiencies for a number of designs. These are given in Table 1. Also given for comparison are the corresponding relative efficiencies for individually matched case-control sets in which the size of case-control sets stays (stochastically) the same but the number of case-control sets increase. Interestingly, although the designs are different, these latter relative efficiencies are the limit of the former as probability of disease $p \rightarrow 0$, i.e. when disease is rare in the study base. The finite sampling approach provides a $(1 - p)$ refinement to the well known $m/(m + 1)$ efficiency rule for frequency matching that accounts for the proportion diseased. Thus, in an extreme example, if $p = \frac{1}{2}$ and 1 : 1 frequency matching is used (so all study base subjects are included), the finite sampling relative

efficiency formula correctly predicts a relative efficiency of one, compared to the ‘rare disease’ efficiency of 1/2.

6. COMPUTATION

In general, the denominator of (4) can have many terms. For instance, the frequency matching denominator with 15 cases and 15 controls has over 150 million terms. Even if the number of terms is small, the risk weights in (4) are ‘set specific’ and cannot be accommodated using standard conditional logistic regression software. (This is in contrast to the individual matching situation where the risk weights can be incorporated as offsets in the model.) In this section, we consider three computational methods.

6.1 Exploit log-linear form

This approach is only applicable when the odds ratio term is log-linear. Then

$$\begin{aligned} \lambda^s r_s(\beta) w_{\tilde{\mathcal{R}}}(\mathbf{s}) &= \exp \left\{ \log(w_{\tilde{\mathcal{R}}}(\mathbf{s})) + \sum_{i \in \mathbf{s}} (\log(\lambda) + Z_{1i} \beta_1 + \cdots + Z_{pi} \beta_p) \right\} \\ &= \exp \left\{ \log(w_{\tilde{\mathcal{R}}}(\mathbf{s})) + s \log(\lambda) + \left(\sum_{i \in \mathbf{s}} Z_{1i} \right) \beta_1 + \cdots + \left(\sum_{i \in \mathbf{s}} Z_{pi} \right) \beta_p \right\}. \end{aligned}$$

The conditional likelihood (4) is formally equivalent to an individually matched conditional logistic likelihood treating each set as an ‘individual’ with covariate value equal to the sum of the set covariates. The ‘case’ in this pseudo case-control data set corresponds to the case set and each ‘control’ corresponds to a non-case-set term in the denominator of (4). The log risk weight associated with each set is added to each line of data and incorporated as an offset when fitting the model. The advantage of this approach is that it may be used for any control sampling scheme. The disadvantages are that it is restricted to the log-linear model and that the size of the pseudo case-control data set has records equal to the number of terms in the denominator (which will be extremely large if there are many cases per study base). The second of these is the most serious restriction but perhaps may be overcome by ‘sampling’ terms from the denominator along the lines of a method described for unweighted conditional likelihoods (Langholz and Goldstein, 1997).

6.2 Recursive algorithm

Currently, most software programs that fit conditional logistic regression models from the ‘unweighted’ likelihood use a recursive algorithm that is computationally very efficient compared to a term-by-term approach (Cox, 1972; Gail *et al.*, 1981). (An excellent description of the algorithm is given in the Epicure manual (Hirosoft International, Seattle, WA).) Let $a(\mathbf{r}, m) = \sum_{\{s \subset \mathbf{r}: s=m\}} r_s(\beta)$. The algorithm is based on the following fact that is a generalization of a well known identity for binomial coefficients:

$$a(\mathbf{r} \cup \{j\}, m) = a(\mathbf{r}, m) + r_j(\beta) a(\mathbf{r}, m - 1).$$

This algorithm can be used to fit other designs where the likelihood can be written as a risk weight times a function of the sums over sets of the same size. This is best seen using an example. For counter-matching, the denominator of likelihood (8) is:

$$\sum_{\{s \subset \tilde{\mathcal{R}}: s=D\}} r_s(\beta) w(s_1, \dots, s_L) = \sum_{k_1, \dots, k_L: \sum_l k_l = D} w(k_1, \dots, k_L) S(k_1, \dots, k_L)$$

where $S(k_1, \dots, k_L) = \sum_{\{s \subset \tilde{\mathcal{R}}: s_1=k_1, \dots, s_L=k_L\}} r_s(\beta)$. The key observation is that $S(k_1, \dots, k_L) = a(\tilde{\mathcal{R}}_1, k_1) \times \dots \times a(\tilde{\mathcal{R}}_L, k_L)$ so that the denominator equals

$$\sum_{k_1, \dots, k_L: \sum_l k_l = D} w(k_1, \dots, k_L) a(\tilde{\mathcal{R}}_1, k_1) \times \dots \times a(\tilde{\mathcal{R}}_L, k_L). \quad (9)$$

Thus, the algorithm can be run on each sampling stratum in the case-control set separately to compute the $a(\tilde{\mathcal{R}}_l, j)$, $j = 1, \dots, D$. These can then be combined in (9) with the appropriate weights. This procedure is linear in D and $|\tilde{\mathcal{R}}|$ rather than approximately exponential when using direct summing.

6.3 Unconditional logistic analysis

As discussed in the next section, there may be a marginal likelihood of the unconditional logistic form that will be valid for large samples and a useful estimation method.

7. UNCONDITIONAL LOGISTIC REGRESSION AS A ‘MARGINAL’ LIKELIHOOD

Unconditional logistic regression is a ‘standard’ analysis method for ‘standard’ case-control studies with large strata. As we showed in Section 4.4, the conditional finite sampling model is exactly of a weighted unconditional logistic likelihood form when control selection is by independent Bernoulli trials. Although this is still a subject in need of more research, we conjecture that there is always a ‘marginal’ likelihood that has the weighted unconditional logistic form and can be used to estimate parameters from the sampled data. This is based on the marginal probability that i is a case given that i is a member of the case-control set (e.g., Fears and Brown, 1986; Weinberg and Wacholder, 1993). Assuming all cases are used and applying Baye’s theorem, this probability is given by

$$\text{pr}(D_i = 1 | i \in \tilde{\mathcal{R}}) = \frac{\pi_i^{-1} \lambda_0 r_i}{1 + \pi_i^{-1} \lambda_0 r_i} \quad (10)$$

where

$$\begin{aligned} \pi_i &= \text{pr}(i \in \tilde{\mathcal{R}} | D_i = 0) \\ &= \sum_{s \subset \mathcal{R}: s \not\ni i} \left[\sum_{\mathbf{r} \subset \mathcal{R}: \mathbf{r} \ni i} \pi(\mathbf{r} | \mathbf{s}) \right] \text{pr}(\mathbf{D} = \mathbf{s} | D_i = 0) \\ &= E \left[\left(\sum_{\mathbf{r} \subset \mathcal{R}: \mathbf{r} \ni i} \pi(\mathbf{r} | \mathbf{D}) \right) \middle| D_i = 0 \right] \end{aligned}$$

where the expectation is with respect to case sets occurring in the study base. Now π_i cannot be computed because it requires covariate information for all study base members. Thus, the idea is to estimate π_i from the sample using a reasonably simple estimator. We propose estimation by replacing the expectation by the ‘observed’ value in a method of moments type estimator: $\hat{\pi}_i = \sum_{\mathbf{r} \subset \mathcal{R}: \mathbf{r} \ni i} \pi(\mathbf{r} | \mathbf{D})$. Note that this is only defined for controls in the case-control set. For cases, loosely, we envision using the control $\hat{\pi}$ that could replace the case and not ‘change the structure of the case-control set’, i.e. controls j such that $\pi(\mathbf{r} | \mathbf{d}) = \pi(\mathbf{r} | (\mathbf{d} \setminus \{i\}) \cup \{j\})$. This is illustrated in the counter-matching example below. Other methods could be used to estimate the π_i (Breslow and Holubkov, 1997; Robins *et al.*, 1994) but these are somewhat more complicated and may defeat the purpose of computational simplicity. The marginal

likelihood is formed as the product of contributions from (10) over case-control subjects,

$$L(\lambda, \beta) = \prod_{i \in \mathbf{D}} \left[\frac{1}{\hat{\pi}_i} \lambda r_i(\beta) \right] \prod_{k \in \tilde{\mathcal{R}}} \frac{1}{1 + \frac{1}{\hat{\pi}_k} \lambda r_k(\beta)}. \quad (11)$$

If the $\hat{\pi}$ are easy to compute, this likelihood can easily be fitted using standard unconditional logistic regression software by including $-\log(\hat{\pi}_k)$ as an offset in the model. Also, this approach provides a way to estimate λ_0 when it is ‘conditioned out’ in the conditional likelihood, such as in frequency and counter-matching. One difficulty is variance estimation. *Generally*, the expectation of score terms from (11) will not be zero, but the (normalized) *sum of the score terms* is consistent for zero. This implies that the inverse information will be conservative (too large) for estimation of the variance of the parameter estimates (Carroll *et al.*, 1995). Further work is needed to develop general variance estimation methods. Another topic for research is the relative efficiency of the conditional logistic likelihood estimator compared to the marginal likelihood. We explore these questions for the special cases of frequency and counter-matching.

7.1 Frequency matching

For 1 : m frequency matching,

$$\begin{aligned} \hat{\pi}_i &= \sum_{\mathbf{r} \in \tilde{\mathcal{R}}: \mathbf{r} \ni i} \pi(\mathbf{r} | \mathbf{D}) = \binom{n-D-1}{mD-1} \binom{n-D}{mD}^{-1} \\ &= \frac{mD}{n-D} \end{aligned}$$

for controls. Because switching case/control identity for any members of the case-control set does not change the control selection probabilities, this same weight is appropriate for the cases as well. In fact, because it is common to all subjects, this weight can be ignored as it will be ‘absorbed’ into the λ estimator. On the other hand, inclusion of this weight provides a way to estimate λ_0 from the frequency matched data when the number of subjects in the study base is known. As seen in the simulation studies below, frequency matching must be particularly simple (asymptotically equivalent to its Bernoulli trial counterpart) because no adjustment of the inverse information variance estimator is needed.

7.2 Counter-matching

With $m_l D$ subjects in stratum l (as in Figure 3),

$$\hat{\pi}_i = \sum_{\mathbf{r} \in \tilde{\mathcal{R}}: \mathbf{r} \ni i} \pi(\mathbf{r} | \mathbf{D}) = \frac{m_l D - D_l}{n_l - D_l} \quad (12)$$

for controls in stratum l . It is natural to assign the same weight for cases in stratum l . Noting that the sampling in Figure 3 has the same form as in the two-stage sampling design described by White (1982), and making the appropriate notational equivalencies, one finds that the marginal likelihood (12) is identical to the ‘conditional likelihood’ of Breslow and Cain (1988). This suggests the use of the Breslow–Cain estimator of the variance, which we explore in the simulation study in the next section. That the marginal likelihood is not appropriate when there are small numbers of cases is seen in the extreme situation of 1 : 1 counter-matching with one case in the study base. Then $\hat{\pi} = 0$ for the case’s sampling stratum, yielding an infinite weight for the case contribution.

8. SIMULATION STUDY

We compared estimators from cohort, frequency matched, and counter-matched data in a limited simulation study to address a number of questions. First, for counter-matching, we wanted to verify that the conditional logistic analysis, using the weights that we derived provided valid parameter and variance estimation. Second, we wanted to see if the efficiency of counter-matching with multiple cases per case-control set, relative to frequency matching, is similar to that found in the single case per set counter-matching. Third, we wished to see if the empirical relative efficiencies are as predicted by the derived formulae in Table 1. Finally, we wanted to assess the adequacy of the marginal likelihood for counter-matching and see whether the Breslow–Cain variance estimates the variance of the odds ratio estimate.

For each simulation, the structure of the cohort was defined by the number of study bases and the number of subjects within each study base. The joint distribution of dichotomous sampling strata variable \tilde{Z} and exposure variable Z was parametrized by the probability of exposure ($\text{pr}(Z = 1)$), and the sensitivity and specificity of \tilde{Z} for Z . Each subject's \tilde{Z} and Z values were randomly assigned based on this distribution. The case status of each subject in a given study base was independently and randomly determined given the odds ratio for Z . The cohort data set consisted of a line of data for each subject that included the study base number, case-control status indicator, \tilde{Z} , and Z . A 1 : 1 frequency-matched set was sampled from each study base and one \tilde{Z} -counter-matched set was also sampled with D subjects per sampling stratum (i.e. $m_0 = m_1 = 1$), the balanced 1 : 1 design. The odds ratio for Z was estimated from the cohort data using the logistic likelihood (2). For each frequency- and counter-matched data set, the odds ratio was estimated using the appropriate conditional and (marginal) unconditional logistic likelihoods. The counter-matching conditional likelihood was fitted using the strategy described in Section 6.1, exploiting the additional simplification possible for dichotomous covariates. For the marginal likelihoods, a baseline odds parameter was estimated for each study base and the Breslow–Cain variance computed. We also computed the score and information at the true parameters.

The results from many repetitions were then tabulated to include the average and empirical variance of the exposure (Z) log odds ratio estimate, the average of the estimated log odds ratio variance, as well as the score and information.

8.1 Results

Table 2 gives selected results from the simulation study of cohorts with 30 study bases and 100 subjects per study base. The prevalence of exposure ($Z = 1$) was taken to be 20%. The probability of disease in the unexposed was set so that the overall disease probability is about 10%, an average of 10 cases per study base. The sensitivity and specificity \tilde{Z} for Z were both set to 90%. Table 2 gives the results with exposure odds ratio equal to one and four. For the conditional logistic likelihood analyses, there is no evidence of bias for either frequency- or counter-matched designs and, comparing the second and third rows, the variance is well estimated by the inverse information. In this situation, with \tilde{Z} well correlated with Z , counter-matching offers a substantial efficiency advantage over frequency matching. In fact, under the null, about three times the number of frequency-matched controls are needed to obtain an efficiency comparable to the 1 : 1 counter-matched design. As seen in the ‘Efficiency’ rows, the empirical relative efficiencies are larger than the individually matched (single case) asymptotic relative efficiency formula would predict. The finite-sampling model formulae under the null (see Table 1) correct for the proportion diseased and appear to correctly predict the efficiency (within sampling error). For counter-matching, the marginal likelihood estimates shows evidence of some bias, although in this case, it is not severe. (This bias increases as the number of cases per study base gets smaller.) For frequency matching, the variance of $\hat{\beta}$ from the marginal likelihood is well estimated by the inverse information but,

Table 2. Results of simulation studies comparing cohort, frequency-matched, and counter-matched designs using conditional and marginal (unconditional) likelihood methods. Cohorts consisted of 30 study bases with 200 subjects each. The probability of exposure was $pr(Z = 1) = 20\%$ and the overall probability of disease was about 10%. The counter-matching sampling stratum variable \tilde{Z} had 90% sensitivity and specificity for the exposure variable Z . Based on 1000 trials

	Odds ratio = 1 (log odds = 0)					
	Cohort	Frequency matching		Counter matching		Balanced two-stage
		Conditional	Marginal	Conditional	Marginal	
$\hat{\beta}$	0.00	0.01	0.01	0.00	0.04	-0.01
$\text{Var}(\hat{\beta})$	0.024	0.040	0.044	0.027	0.033	0.031
I^{-1}	0.024	0.042	0.044	0.028	0.040	0.040
$\widehat{\text{Var}}\hat{\beta}^a$			0.044		0.030	0.030
Efficiency:						
Empirical ^b	1	59%	54%	86%	79%	78%
Single case ^c	1	50.0%		82.0%		
Multiple case ^c	1	55.6%		85.2%		
Odds ratio = 4 (log odds = 1.39)						
$\hat{\beta}$	1.40	1.39	1.47	1.38	1.42	1.43
$\text{Var}(\hat{\beta})$	0.016	0.034	0.038	0.019	0.021	0.023
I^{-1}	0.016	0.037	0.039	0.020	0.030	0.030
$\widehat{\text{Var}}\hat{\beta}^a$			0.042		0.021	0.023
Efficiency:						
Empirical ^b	1	46%	41%	83%	75%	68%
Single case ^d	1	36.8%		76.2%		

^a Breslow–Cain estimator.

^b $\text{Var}(\hat{\beta})$ for the cohort divided by that of the sampled.

^c See Table 1.

^d Using the formula given in the appendix of Langholz and Borgan (1995).

for counter-matching, it is extremely conservative. The Breslow–Cain estimator adequately captures the counter-matching variability. There is some efficiency loss in using the marginal likelihood compared to the conditional and the loss tends to be larger for counter-matching than for frequency matching. These results were consistent over the range of parameters we investigated in our simulation and indicate that there is some efficiency advantage of the conditional over the marginal likelihood for case-control studies with complex sampling.

9. DATA EXAMPLE

To illustrate the methods for the asthma study, we will consider only the *in utero* smoke exposure–genotype interaction analysis for a single gene, using the dichotomous smoke exposure. If more precise smoke information were collected, it would be used instead of the crude smoking variable in the analysis. The weights are determined by the crude smoke variable used in the sampling and do not

change throughout the analysis. Of course, we do not yet have genotype information so, for purposes of illustration, we randomly generated these data for subjects in the asthma data set assuming an interaction odds ratio of about two, independence of genotype status and *in utero* smoke exposure, and a genotype prevalence of 50%, about the prevalence of the genotypes in the general population. For simplicity we assume that the risk of asthma in children of non-smoking mothers would be expected to be about the same regardless of genotype status.

From these study data with simulated genotype information, controls were counter-matched on *in utero* smoke exposure with the number of subjects from each smoke exposure stratum equal to the number of cases in the study base ($m_0 = m_1 = 1$). Table 3 shows the data for four counter-matched study bases. The basic information, obtainable from the cohort, to do the counter-matched sampling includes study base identifier, case/control status, sampling stratum (*in utero* smoke exposure) indicator, number of sampling stratum subjects in the study base and number of sampling stratum subjects in the sample. The latter variables are needed to compute the risk weights in equation (8). The last column gives the genotype information which would be collected only on the sample. (In the actual study we would also include the detailed *in utero* smoke exposure information and genotype information for GST-M1 and GST-T1 genes.) The first study base has one case who is *in utero* smoke exposed. So a control was randomly sampled from the 29 non-exposed. The second study base has two cases, both *in utero* unexposed, so that two controls were randomly sampled from the eight exposed subjects. To analyze the counter-matched data, a pseudo case-control set was created as described in Section 6.1. Although this process involved some programming challenges, these were addressed in a straightforward manner using SAS Macro language. The resulting analysis data set contained a total of 13 793 records. The odds ratio parameters were estimated from this analysis data set using SAS PHREG, with the log weights into the model using the ‘offset’ option. Although the analysis data set is rather large, the required computing time for fitting models was a matter of seconds because there was only one ‘case’ (corresponding to the set of cases) per pseudo case-control set. The results of fitting the interaction model for the counter-matched sample are given in Table 4. Of course, the cohort data will not be available to us in the actual study, but cohort results may be given here because this is simulated data. The pattern of estimated odds ratios from the counter-matched data set is consistent with the cohort and, in this realization, the counter-matched study statistically detects the interaction.

Operationally, the selection of counter-matched study is an administrative and statistical process. Once the list of subjects in the study is produced, the data collection for the study subjects proceeds exactly as would a standard case-control study. With frequency matching, subject refusal is generally ignored in the analysis whereas for counter-matched data, subject refusal needs to be considered in terms of the appropriate weighting. This issue will be explored elsewhere, but we note that the methods developed in this paper (in particular, case-sampling and multistage sampling likelihoods described in Section 10) provide ways of addressing this problem.

Finally, we wish to stress that the genotype data is simulated so that the analyses presented in Table 4 are for the purpose of illustrating the statistical methods only.

10. OTHER CONDITIONAL LIKELIHOODS

In this section, we propose conditional likelihoods that naturally accommodates case sampling and another adapted for designs with multistage sampling of controls.

10.1 Case sampling

When disease is common, it may be advantageous to sample cases as well as controls. And, for the same reason as in control sampling, the sampling could depend on characteristics of the cases, such as exposure

Table 3. *Some data from the counter-matched asthma study (EARS) data set. The genotype data was simulated for illustration*

Study base number	Sampling stratum ^a	Case-control status	Total in sampling stratum	Sampled from sampling stratum	Genotype
1	0	0	29	1	1
1	1	1	7	1	1
2	0	1	32	2	1
2	0	1	32	2	1
2	1	0	8	2	1
2	1	0	8	2	0
3	0	1	35	1	1
3	1	0	7	1	1
4	0	1	26	4	1
4	0	1	26	4	0
4	0	0	26	4	1
4	0	0	26	4	0
4	1	1	4	4	1
4	1	1	4	4	1
4	1	0	4	4	0
4	1	0	4	4	0

^a *In-utero* smoke exposure with 1 = yes and 0 = no.

Table 4. *Estimated odds ratios (95% confidence intervals) in the analysis of in utero smoke-genotype interaction in the EARS data sets with simulated genotype*

Variable	Cohort	Counter-matched
<i>In-utero</i> smoke	0.9 (0.6–1.5)	0.8 (0.5–1.5)
Genotype	1.1 (0.8–1.5)	1.0 (0.5–1.7)
Interaction	2.2 (1.2–4.1)	2.6 (1.2–5.8)
χ^2 for interaction	6.4	5.5
P-value	0.01	0.02
Number of subjects	3216	495 ^a

^a A few sampling stratum had fewer subjects than the number of cases. For such sets, all subjects in the sampling stratum were included.

or exposure–correlate levels. The conditional probability $\text{pr}(\mathbf{D} = \mathbf{d} | \tilde{\mathcal{R}} = \mathbf{r})$ that we have used as the basis for control sampling does not accommodate case-sampling because, as it is easy to show, the probability will depend on the covariate values of the cases that were not sampled. Generalizing our notation (in a way which is not inconsistent with it), let \mathbf{D} denote cases that occur *inside* and $\bar{\mathbf{D}}$ those *outside* the case-control sample. We propose basing the likelihood on

$$\begin{aligned} \text{pr}(\mathbf{D} = \mathbf{d} | \tilde{\mathcal{R}} = \mathbf{r}, \bar{\mathbf{D}} = \bar{\mathbf{d}}) &= \frac{\lambda_0^d r_{\mathbf{d}} \text{pr}(\tilde{\mathcal{R}} = \mathbf{r}, \bar{\mathbf{D}} = \bar{\mathbf{d}} | \mathbf{D} = \mathbf{d})}{\sum_{\mathbf{s} \subset \mathbf{r}_m} \lambda_0^s r_{\mathbf{s}} \text{pr}(\tilde{\mathcal{R}} = \mathbf{r}, \bar{\mathbf{D}} = \bar{\mathbf{d}} | \mathbf{D} = \mathbf{s})} \\ &= \frac{\lambda_0^d r_{\mathbf{d}} \phi(\mathbf{d} | \mathbf{d} \cup \bar{\mathbf{d}}) \pi(\mathbf{r} | \bar{\mathbf{d}}, \mathbf{d})}{\sum_{\mathbf{s} \subset \mathbf{r}_m} \lambda_0^s r_{\mathbf{s}} \phi(\mathbf{s} | \mathbf{s} \cup \bar{\mathbf{d}}) \pi(\mathbf{r} | \bar{\mathbf{d}}, \mathbf{s})} \end{aligned}$$

where $\phi(\mathbf{s} | \mathbf{s} \cup \bar{\mathbf{d}})$ is the probability of sampling case set \mathbf{s} , given $\mathbf{s} \cup \bar{\mathbf{d}}$ is the set of all cases and $\pi(\mathbf{r} | \bar{\mathbf{d}}, \mathbf{d})$ is the probability of the case-control set given the case sampling outcome. In a simple example, it is easy to see that sampling cases (and controls) based on independent Bernoulli trials leads to the randomized recruitment likelihood previously derived (Weinberg and Wacholder, 1990).

10.2 Multistage sampling

In this class of designs, controls are sampled in m successive ‘stages’, using additional information collected for the controls remaining at each stage (Whittemore, 1997). Full covariate information is collected on the cases and last-stage controls. In this class, the sampling designs are described by π_l that are the probability of the l th stage sample ($\tilde{\mathcal{R}}_l$) given the previous stages and the set of cases. Then, an appropriate likelihood is based on the probability

$$\begin{aligned} \text{pr}(\mathbf{D} = \mathbf{d} | \tilde{\mathcal{R}}_1 = \mathbf{r}_1, \tilde{\mathcal{R}}_2 = \mathbf{r}_2 \cdots \tilde{\mathcal{R}}_m = \mathbf{r}_m) &= \frac{\lambda_0^d r_{\mathbf{d}} \pi_1(\mathbf{r}_1 | \mathbf{d}) \cdots \pi_m(\mathbf{r}_m | \mathbf{r}_{m-1}, \dots, \mathbf{r}_1, \mathbf{d})}{\sum_{\mathbf{s} \subset \mathbf{r}_m} \lambda_0^s r_{\mathbf{s}} \pi_1(\mathbf{r}_1 | \mathbf{s}) \cdots \pi_m(\mathbf{r}_m | \mathbf{r}_{m-1}, \dots, \mathbf{r}_1, \mathbf{s})}. \end{aligned}$$

Note that the likelihood will only depend on data available at each stage, with covariates to be used in the proportional odds model only required for cases and last-stage controls.

11. DISCUSSION

In this paper, we have derived a general conditional logistic likelihood for the analysis of unmatched case-control studies with complex sampling, presented some new case-control study designs, and sketched a number of aspects of this on-going research. The finite sampling model for case-control data that is at the heart of this methodology is ‘prospective’ in that the probability of disease for a subject in the study base is conditional on that subject’s covariate values. The base is then sampled, generally in a way that depends on the composition of the case set to obtain the case-control sample. We find this approach more natural than the ‘retrospective’ classical approach because it is true to the principle that cause (exposure) precedes effect (disease) and, in fact, better describes the reality of most case-control studies. We have presented a number of designs, some of them new and potentially useful in specialized situations, to illustrate both that there are many design options yet to be explored and that the conditional logistic likelihood can accommodate these types of complex sampling of controls. This likelihood has the advantages that it is a ‘true’ likelihood so that analysis is, in principle, straightforward and applies when there are a small, as

well as large, number of cases. Because of this property, it provides a unified approach to individually matched and unmatched case-control study analysis. Also, because the conditional logistic likelihood is the likelihood for the data given the information required to do the sampling, we conjecture that this likelihood is ‘efficient’ with respect to the logistic model for disease occurrence (1), the information needed to do the sampling (captured in the $\pi(\mathbf{r}|\mathbf{d})$), and the data collected in the sample.

Methods based on the classical model treat cases and controls as independent realizations from separate populations. It is interesting to compare designs and analysis methods developed under the two paradigms for two particular situations. In a first example, randomized recruitment was developed under the classical model (Weinberg and Wacholder, 1990). As we showed in Section 4.4, the finite-sampling conditional likelihood is the same as the weighted unconditional logistic likelihood developed under the classic because the sampling is truly independent over individuals. However, the actual implementation of the design in the lung cancer case-control study described by Weinberg and Wacholder (1990) is closer to the randomized recruitment with stopping design described in Section 4.6, a complication that is not readily accommodated by the classical approach. Whether accounting for the stopping makes an appreciable difference in the analysis is a topic for future research. A second example is the ‘two-stage’ sampling design described by White (1982) and Breslow and Cain (1988). The ‘first-stage sample’ has the structure as shown in Table 3, although the subjects in the first-stage sample may already be a case-control study with cases and controls (randomly) sampled from the study base. The design specifies that a fixed number of subjects are independently sampled from each of the case/control by exposure-level cells. Based on the classical model, Breslow and Cain develop a ‘conditional’ likelihood (identical to our marginal likelihood in Section 7.2) and other estimation methods have been suggested (Zhao and Lipsitz, 1992; Breslow and Holubkov, 1997; Lawless *et al.*, 1999). Because the structure of the case-control set identifies the number of cases in each exposure level, the conditional logistic likelihood we have developed ‘conditions out’ the main effects of exposure and these cannot be estimated from two-stage design data. Thus, the conditional likelihood is not a candidate analysis method for this design. The two-stage design is quite different from the counter-matching method we have described. First, rather than sampling independently from each cell, it is the sampling (exposure) stratum marginal totals that determine the counter-matching structure. This means that the number of cases in each sampling stratum determine the number of controls that will be sampled from that stratum. And, although cases can be sampled in conjunction with counter-matching of controls, it is not clear how best to do this—a topic for further work. In Section 8, we compared the counter-matching design analysed using conditional logistic and the marginal likelihoods in a limited simulation study. A further interesting comparison is to the ‘balanced’ two-stage design that includes all the D cases and a sample of $D/2$ controls from each of the two sampling strata (Breslow and Cain, 1988). Although the two-stage design has the same number of total controls (D), it is quite different from the ‘balanced’ counter-matching described in Section 8 with equal numbers of *total* subjects (cases plus sampled controls) from each sampling stratum. The performance the balanced two-stage design using the marginal likelihood was investigated in the simulation discussed in Section 8.1: the results are given in the last column of Table 2. Under the null, the two-stage design had efficiency similar to the counter-matched design using the marginal likelihood but for odds ratio equal to four, the two-stage design is somewhat less efficient than the counter-matched, using the marginal likelihood. The counter-matched study using the conditional logistic likelihood had superior efficiency to either design analysed using the marginal likelihood. Development of the asymptotic theory will help to further evaluate the efficiency of these designs and analysis methods. We note that the two-stage design is only valid for ‘large strata’, where there are a few study bases with many cases. Thus, we could not use this method for the EARS study, which has many study bases with few cases.

Although it may be a matter of semantics, we do not think of counter-matching as a two-stage design. Rather, the counter-matched sample is taken directly from the study base and is thus a first-stage sample. However, a simple two-stage variant is easy to describe. First ‘randomly sample’ (i.e. either equal-

probability Bernoulli trials or simple random sampling) controls from the study base. Exposure-related information would then be collected on this first-stage sample. A second-stage sample would then be obtained by counter-matching from within the first-stage sample. Using the multistage likelihood from Section 10.2, it is easy to see that this results in weights that are formally the same as counter-matching directly from the study base, replacing the study base sampling stratum totals (the n_i) by the corresponding first-stage sample totals.

The issues addressed by comparing counter-matching to the two-stage design are not unique to these designs. Each of the designs we have presented suggests a corresponding design based on the classical approach. So, for instance, in a 'classical' quota sampling design, the quotas would be set for each of the case/control by exposure-level cells separately. Along the same lines, there are other 'classes' of designs. For instance, each of the new designs we have presented has a case-base analogue. Another class is to sample in numbers proportional to size of the study base, rather than based on the number of cases. For each of these last two classes of designs, the baseline odds is estimable using the conditional logistic likelihood, but we conjecture that this is achieved only with some efficiency loss of the odds ratio parameters.

When parameters are conditioned out by the conditional logistic likelihood and appropriate study base information is available, we conjecture that an efficient general analysis strategy will be to use the conditional logistic likelihood for those parameters which are not conditioned out, and a marginal likelihood for estimation of parameters that are conditioned out, plugging in the estimated maximum conditional logistic likelihood estimates into the marginal likelihood. For instance, the baseline odds parameter is conditioned out in frequency- and counter-matching conditional logistic likelihood. In this situation, we propose estimating the odds ratio parameters using the conditional logistic likelihood and plugging these into a marginal likelihood (incorporating the total number of subjects in the study base) to estimate the baseline odds parameter. This strategy is consistent with the estimation of the baseline hazard for failure time data.

What are the implications of this and other work for epidemiologic research? There is a commonly held belief that control selection should not depend on exposure (e.g. Rothman and Greenland, 1998). This misconception has resulted in confirmatory case-control studies which simply replicate earlier ones, rather than build on them. Using creative study designs, confirmatory studies can specifically avoid the pitfalls that might have been present in the hypothesis generating study and use exposure related information to increase cost-efficiency. Thus, a flexible framework for development and a unified approach to the analysis of study designs with complex sampling can accelerate the pace of epidemiologic research and increase its validity. The 'random sampling of controls' principle needs to be replaced by the principle that 'the method of control selection must be incorporated into the analysis'.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Frank Gilliland for graciously offering the EARS study as the example for use in this paper. Peter Kraft is acknowledged for his work on the unconditional logistic likelihood for counter-matching. This work was supported by grants CA42949 of the US National Cancer Institute and 5P30ES07048 of the National Institute for Environmental Health Sciences.

REFERENCES

- BORGAN, Ø., GOLDSTEIN, L. AND LANGHOLZ, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Annals of Statistics* **23**, 1749–1778.
- BRESLOW, N. E. AND DAY, N. E. (1980). *Statistical Methods in Cancer Research. Volume I—The Design and*

- Analysis of Case-Control Studies*, Vol 32 of *IARC Scientific Publications*. Lyon: International Agency for Research on Cancer.
- BRESLOW, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Society* **91**, 14–28.
- BRESLOW, N. E. AND CAIN, K. C. (1988). Logistic regression for two stage case-control data. *Biometrika* **75**, 11–20.
- BRESLOW, N. E. AND HOLUBKOV, R. (1997). Weighted likelihood, pseudo-likelihood, and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* **16**, 103–116.
- CARROLL, R. J., WANG, S. AND WANG, C. Y. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association* **90**, 157–169.
- CORNFIELD, J. (1951). A method of estimating comparative rates from clinical data, applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* **11**, 1269–1275.
- COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B* **34**, 187–220.
- FEARS, T. R. AND BROWN, C. C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* **42**, 955–960.
- FLANDERS, W. D., DERSIMONIAN, R. AND RHODES, P. (1990). Estimation of risk ratios in case-base studies with competing risks. *Statistics in Medicine* **9**, 423–435.
- GAIL, M. H., LUBIN, J. AND RUBENSTEIN, L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* **68**, 703–707.
- KUPPER, L. L., MCMICHAEL, A. J. AND SPIRTAS, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk. *Journal of the American Statistical Association* **70**, 524–528.
- LANGHOLZ, B. AND BORGAN, Ø. (1995). Counter-matching: a stratified nested case-control sampling method. *Biometrika* **82**, 69–79.
- LANGHOLZ, B. AND GOLDSTEIN, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science* **11**, 35–53.
- LANGHOLZ, B. AND GOLDSTEIN, L. (1997). Fitting logistic models using conditional logistic regression when there are large strata. *Computing Science and Statistics* **29**, 551–555.
- LAWLESS, J. F., KALBFLEISCH, J. D. AND WILD, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society B* **61**, 413–438.
- MIETTINEN, O. S. (1982). Design options in epidemiology research: an update. *Scandinavian Journal of Work, Environment, and Health* **8**(Suppl. 1), 1295–1311.
- OAKES, D. (1981). Survival times: aspects of partial likelihood (with discussion). *International Statistical Review* **49**, 235–264.
- PETERS, J. M., AVOL, E., NAVIDI, W., LONDON, S. J., GAUDERMAN, W. J., LURMANN, F., LINN, W. S., MARGOLIS, H., RAPPAPORT, E., GONG, H. AND THOMAS, D. C. (1999). A study of twelve Southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity. *American Journal of Respiratory & Critical Care Medicine* **159**(3), 760–767.
- PRENTICE, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.
- PRENTICE, R. L. AND PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.
- ROBINS, J. M., ROTNITZKY, A. AND ZHOA, L. P. (1994). Estimation of regression coefficients when some regres-

- sors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.
- ROTHMAN, K. J. AND GREENLAND, S. (1998). *Modern Epidemiology*, 2nd edn. Philadelphia PA: Lippincott-Raven.
- SATO, T. (1992). Maximum likelihood estimation of the risk ratio in case-cohort studies. *Biometrics* **48**, 1215–1221.
- SCHUSTER, E. F. AND SYPE, W. R. (1987). On the negative hypergeometric distribution. *International Journal of Mathematical Education in Science and Technology* **18**, 453–459.
- STEENLAND, K. AND DEDDENS, J. A. (1997). Increased precision using counter-matching in nested case-control studies. *Epidemiology* **8**, 238–242.
- WACHOLDER, S., MCLAUGHLIN, J. K., SILVERMAN, D. R. AND MANDEL, J. S. (1992). Selection of controls in case-control studies: I. Principles. *American Journal of Epidemiology* **135**, 1019–1028.
- WEINBERG, C. R. AND WACHOLDER, S. (1990). The design and analysis of case-control studies with biased sampling. *Biometrics* **46**, 963–975.
- WEINBERG, C. R. AND WACHOLDER, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika* **80**, 461–465.
- WERLER, M. M. (1997). Teratogen update: smoking and reproductive outcomes. *Teratology* **55**, 382–388.
- WHITE, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115**, 119–128.
- WHITTEMORE, A. (1997). Multistage sampling designs and estimating equations. *Journal of the Royal Statistical Society B* **59**, 589–602.
- WILD, C. J. (1991). Fitting prospective regression models to case-control data. *Biometrika* **78**, 705–717.
- ZHAO, L. P. AND LIPSITZ, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine* **11**, 769–782.

[Received xx Month 2000 revised xx Month 2000]