Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2006
Methodological Issues in Measuring Population Health

The Analysis of Population-based Case Control Studies[1]

Alastair Scott[2]

Abstract

We discuss methods for the analysis of case-control studies in which the controls are drawn using a complex sample survey. The most straightforward method is the standard survey approach based on weighted versions of population estimating equations. We also look at more efficient methods and compare their robustness to model mis-specification in simple cases. Case-control family studies, where the within-cluster structure is of interest in its own right, are also discussed briefly.

KEY WORDS: Case-control studies; Response-selective sampling; Retrospective sampling; Weighting.

1. Introduction

The case-control study, in which separate samples are drawn from 'cases' (people with a disease of interest, say) and from 'controls' (people without the disease), is one of the most common designs in health research. In fact, Breslow (1996) has described such studies as "the backbone of epidemiology". We shall concentrate on biostatistical applications, but the basic design is an efficient sampling strategy whenever cases are rare and examples are common in many other fields as well (business, social science, ecology, market research, for example). In particular, there has been a parallel development of much of the theory in the econometric literature on choice-based sampling (see Manski and McFadden 1981, Cosslett 1981 for example).

There are two fundamentally different types of case-control study: (set-)matched studies, in which each case is matched with one or more controls, and unmatched studies, in which the case and control samples are drawn independently, although there may be loose "frequency matching", with the control sample allocated across strata defined by basic demographic variables in such a way that the distribution of these variables in the control sample is similar to their expected distribution in the case sample. We are only concerned with unmatched studies here and, more specifically, only with the restricted class of population-based studies in which the controls (and occasionally the cases as well) are selected using standard survey sampling techniques.

An excellent introduction to the strengths and potential pitfalls of case-control sampling is given by Breslow (1996, 2004). One of the most important and difficult challenges confronting anyone designing such a study is to ensure that controls really are drawn from the same population, using the same protocols, as the cases. In the words of Miettinen (1985), cases and controls "should be representative of the same base experience". Failure to ensure this adequately in some early examples led to case-control sampling being regarded with some suspicion by many researchers. A comprehensive discussion on the principles that should govern the selection of controls is given in Wacholder, McLaughlin, Silverman and Mandel (1991). Since the essence of survey sampling lies in methods for drawing representative samples from a target population, it became natural at some stage to think about using survey methods for obtaining controls. Increasingly over the last 25 years or so, the controls (and occasionally the cases as well) are being drawn using complex stratified multi-stage designs. A good history of this development can be found in Chapter 9 of Korn and Graubard (1999).

The analysis of such studies is a particularly appropriate topic for this paper since Joe Waksberg himself was one of the principal drivers behind the adoption of survey methods (and random digit dialing, in particular) for obtaining controls (see, for example, Waksberg 1998 and DiGaetano and Waksberg 2002).

---

[1] This paper originally appeared in the December 2006 issue of Survey Methodology (Volume 32, No. 2, pp. 123-132). It is reprinted here with the permission of the editors.

[2]Alastair Scott, Department of Statistics, University of Auckland, Auckland 1, New Zealand.
E-mail: a.scott@auckland.ac.nz.

2. Examples

We start with two examples to illustrate the sort of problem that we want to handle. The first example is typical of the large scale studies conducted by the National Cancer Institute whose personnel have been responsible for much of the development of the area. Joe Waksberg, along with his colleagues at Westat, had a strong influence on the sampling methods used for these studies (see Hartge, Brinton, Rosenthal, Cahill, Hoover and Waksberg 1984, who also gives a description of a number of other similar studies) so it is a natural place to start.

Example 1.

In 1977 – 78, the National Cancer Institute and the US Environmental Protection Agency conducted a population-based case-control study to examine the effects of ultraviolet radiation on non-melanoma skin cancer over a one-year period (Hartge, Brinton, Rosenthal, Cahill, Hoover and Waksberg 1984, Fears and Gail 2000)**.** The study was conducted at eight geographic locations with varying solar ultraviolet intensities. Samples of non-melanoma skin cancer patients aged 20 to 74 and samples of general population controls from each region were interviewed by telephone to obtain information on risk factors. At each location, a simple random sample of 450 patients and an additional sample of 50 patients in the 20 – 49 age group were selected for contact. For the controls, 500 households were sampled at each location using Mitofsky-Waksberg random-digit dialing (Waksberg 1978). An attempt was made to interview all adults aged 65 – 74 as well as a randomly selected individual of each sex aged 20 to 64. In addition, a second Mitofsky-Waksberg sample of between 500 to 2,100 households was taken and information gathered on all adults aged 65 to 74. This resulted in samples of approximately 3,000 cases and 8,000 controls, with the sampling rate for cases being roughly 300 times the rate for controls, depending on age.

The second example is important to me personally since it first introduced Chris Wild and myself to the area.

Example 2.

The Auckland Meningitis Study was commissioned by the NZ Ministry of Health and Health Research Council to study risk factors for meningitis in young children which was reaching epidemic proportions in Auckland at that time (see Baker, McNicholas, Garrett, Jones, Stewart, Koberstein and Lennon 2000). The target population was all children under the age of nine in the Auckland region in 1997 – 2000.

All cases of meningitis in the target age group over the three year duration of the study were included in the study, resulting in about 250 cases. A similar number of controls was drawn from the remaining children in the study population using a complex multi-stage design. At the first stage of sampling, 300 census mesh blocks (each containing roughly 70 households) were drawn with probabilities proportional to the number of houses in the block. At the second stage, a systematic sample of 20 households was selected from each chosen mesh block and children from these households were selected for the study with varying probabilities that depended on age and ethnicity and were chosen to match the expected frequencies among the cases. Selection probabilities are shown in the table below: (PI means Pacific Islander) Cluster sample sizes varied from one to six and a total of approximately 250 controls was achieved. This corresponds to a sampling fraction of about 1 in 400 on average, so that cases are sampled at a rate that is 400 times that for controls here.

These two studies are fairly typical of the sort of study that we want to discuss. They also illustrate the two main sampling methods used, namely random digit dialing and area sampling. A lively discussion of the relative merits of these two strategies are given in Brogan, Denniston, Liff, Flagg, Coates and Brinton (2001) and DiGaetano and Waksberg (2002).

Table 1
Selection Probabilities

AGE	MAORI	PACIFIC ISLANDER	OTHER
≤ 1 year	0.29	0.70	0.10
≤ 3 years	0.15	0.50	0.07
≤ 5 years	0.15	0.31	0.04
≤ 8 years	0.15	0.17	0.04

3. General Set-Up

Suppose that we have a binary response variable, Y, with $Y = 1$ denoting a case and $Y = 0$ denoting a control, and a vector of potential explanatory variables, $\mathbf{x}$. We assume that the value of Y is known for all N units in some target population but that at least some components of $\mathbf{x}$ are unknown. We stratify the population into cases and controls, draw a sample from each stratum based on the variables that we know for all units, and measure the values of the missing covariates for the sampled units (in practice, the control sample is often drawn from the whole population, rather than the units with $Y = 0$. If the proportion of cases is small, the difference will be negligible. Otherwise it is simple to adapt the results below to this variant – for a rigorous development, see Lee, Scott and Wild 2006). Typically, we then want to use the sample data to fit a binary regression model for the marginal probability of a being a case as a function of the covariates. The model used is almost always logistic with

$$\begin{aligned} \text{logit}\,\{P(Y = 1 \mid \mathbf{x})\} &= \log\left(\frac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})}\right) \\ &= \beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1 \qquad (1) \end{aligned}$$

say, where β_0 and $\boldsymbol{\beta}_1$ are unknown parameters, and we shall assume model (1) throughout the paper. Extensions to more general regression models are straightforward in principle (see Scott and Wild 2001b) but the resulting expressions are somewhat clumsier than those for the logistic model.

How should we go about fitting the model (1) given sample data? Efficient methods are straightforward with simple or stratified random sampling, but we are interested in more complex sampling procedures here. Very often the complex sampling is simply ignored. Potentially, this could lead to all the usual problems that arise from ignoring sampling design structure. Varying selection probabilities can distort the mean structure and estimates produced by standard programs may be inconsistent. Intra-cluster correlation can reduce the effective sample size so that routinely-produced standard errors are too small, confidence intervals are too short, $p-$ values too low, and so on. A simple strategy that has been adopted by some researchers to minimize the effect is to keep the numbers of subjects in each cluster small (see Graubard, Fears and Gail 1989, for example). This reduces the design effect and hence the impact of clustering, but it can be a very expensive remedy. We look at some possible ways of coping with standard, more cost-effective, sampling schemes in the next few sections.

4. Survey Weighted Approach

One obvious possibility is to use the standard weighted estimating equation approach embodied in most modern packages for analyzing survey data (see Binder 1983). Suppose first that we had data from the whole finite population. If we assume this finite population is drawn from a superpopulation in which the conditional logistic model **(**1**)** holds, then we could estimate $\boldsymbol{\beta}$ by solving the whole-population or census estimating equations

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{1}^{N} \mathbf{x}_i(y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta})) = 0, \qquad (2)$$

where $p_1(\mathbf{x}; \boldsymbol{\beta}) = e^{\beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1}/(1 + e^{\beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1})$. (These are the likelihood equations if population units are assumed to be sampled independently from a superpopulation but the resulting estimators are consistent under much more realistic population structures as long as model (1) holds marginally – see Rao, Scott and Skinner 1998 for more discussion.)

Now, for any fixed value of $\boldsymbol{\beta}$, $\mathbf{S}(\boldsymbol{\beta})$ in equation (2) is just a vector of population totals. This means that we can estimate it from the sample, say by

$$\hat{\mathbf{S}}(\boldsymbol{\beta}) = \sum_{\text{sample}} w_i\mathbf{x}_i(y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta})), \qquad (3)$$

where w_i is the inverse of the selection probability, perhaps adjusted for non-response and post-stratification. Setting $\hat{\mathbf{S}}(\boldsymbol{\beta})$ equal to $\mathbf{0}$ gives us our estimator, $\hat{\boldsymbol{\beta}}$. We could use linearization or the jackknife directly on $\hat{\boldsymbol{\beta}}$ to get standard errors. Alternatively, we can expand $\hat{\mathbf{S}}(\hat{\boldsymbol{\beta}})$ about the true value, $\boldsymbol{\beta}$, and obtain as our estimated covariance matrix the "sandwich" estimator

$$\hat{\mathrm{C}}\mathrm{ov}\{\hat{\boldsymbol{\beta}}\} \approx \mathbf{J}(\hat{\boldsymbol{\beta}})^{-1}\,\hat{\mathrm{C}}\mathrm{ov}\{\hat{S}(\hat{\boldsymbol{\beta}})\}\,\mathbf{J}(\hat{\boldsymbol{\beta}})^{-1}, \qquad (4)$$

where $\mathbf{J}(\boldsymbol{\beta}) = -\,\partial\hat{\mathbf{S}}/\partial\boldsymbol{\beta}^T = \sum_{\text{sample}} w_i\, p_1(\mathbf{x}_i;\boldsymbol{\beta})\, p_0(\mathbf{x}_i;\boldsymbol{\beta})\, \mathbf{x}_i\, \mathbf{x}_i^T$ with $p_0 = 1 - p_1$. Since $\hat{\mathbf{S}}(\boldsymbol{\beta})$ is a vector of totals, $\hat{\mathrm{C}}\mathrm{ov}\{\hat{\mathbf{S}}(\boldsymbol{\beta})\}$ should be available as a matter of course for any standard design. Most major statistical packages (for example, SAS (PROC SURVEYLOGISTIC), SPSS (CSLOGISTIC), STATA (SVY:LOGIT), SUDAAN (LOGISTIC)) can handle logistic regression with complex sampling and weighting routinely these days. Thus producing weighted estimates and making associated inferences is reasonably straightforward.

Strictly speaking, the selection probabilities will themselves often be random variables in our model-based framework, based on a finite population that we assume is generated from the model. We can account for this by using the results in Rao (1973), but the correction is of order $1/N$ and can be ignored in most large studies.

The downside of weighting in general is that it tends to be inefficient when the weights are highly variable. **(**A rule-of-thumb sometimes suggested is that $w_{\max}/w_{\min}$ should be no more than 10.**)** In case-control studies, the variation in weights is about as extreme as it can get. For instance, the ratio of $w_{\max}$ to $w_{\min}$ is approximately 300:1 in Example 1 and approximately 1,000:1 in Example 2. Even more extreme ratios are not uncommon. No experienced survey sampler would be surprised to find that weighting is not very efficient under these circumstances.

Can we do something more efficient? The answer is certainly "Yes" in some special cases. Fully efficient likelihood methods have been developed in situations where both cases and controls are drawn using simple or stratified random sampling and these can be very much more efficient than weighted methods. We review these results in the next section.

5. Review: Simple Case

We start with the very simplest case where cases and controls are selected by simple random sampling and we have no population information about any of the covariates at the design stage. Here fully efficient semi-parametric maximum-likelihood procedures are well-developed. Moreover, these methods are very simple to implement using standard software (Prentice and Pyke 1979). (The methods are *semi-parametric* because the full likelihood depends on the unknown distribution of the covariates and we do not want to model this in general.)

It turns out that all we have to do is fit model (1) using a standard logistic regression program without any weighting at all. More specifically, solving the unweighted equation

$$\sum_{\text{sample}} \mathbf{x}_i\,(y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta})) = \mathbf{0}, \qquad (5)$$

produces efficient estimates of all the coefficients except the intercept. Perhaps more importantly, all the standard errors and resulting inferences that we get from the standard program are also valid, again with the exception of anything involving the intercept. It is simple enough to correct inferences involving the intercept provided that we know the ratio of the two sampling fractions but we are often only interested in the other coefficients anyway.

The results extend directly to stratified random sampling, provided that separate intercepts for each stratum are included in the model. Again efficient semi-parametric estimators of all coefficients except the stratum intercepts can be obtained simply by running the data through an ordinary (unweighted) logistic regression program. Again, the estimated standard errors and associated inferences are also valid. As with simple random sampling, we can correct the results for the stratum intercepts provided that we know the stratum sampling fractions but, again, these are usually of minor interest.

Thus in these simple situations, maximum likelihood estimates are simpler to compute than the weighted estimates, as well as being more efficient. How much more efficient are they? This depends on the number of covariates, the magnitude of their coefficients and the ratio of the sampling fractions, but the difference is often substantial. (The weighted estimates are about 50% efficient in Example 2 of the introduction, for example, and less than 20% efficient in the brain cancer

example we look at in Section 8. Lawless, Kalbfleisch and Wild 1999 discuss situations where the efficiency is even lower than this.)

Finally, we note that the maximum likelihood estimates have yet another advantage over weighted estimates: they tend to have much better small sample performance, especially in situations where the efficiency of the weighted estimates is low. Essentially, weighting results in a reduction in the effective sample size and it is this effective sample size that governs when the asymptotic theory starts to give a good approximation. (See Scott and Wild 2001a for more details.) Clearly we can pay a very heavy price for a rigid adherence to population weights.

6. More Complex Sampling

In both the examples in Section 2, the controls were obtained from a complex multi-stage survey rather than a simple random sample. As we noted in the introduction, this is increasingly common in large scale case-control studies. (Occasionally, as in Example 1, the cases are also selected using a complex sampling scheme.) It is possible to derive semi-parametric efficient estimators for stratified multistage sampling, assuming that primary sampling units are selected independently within strata (which is the assumption that all the computer packages are making with the survey-weighted approach anyway), but this requires us to build multivariate models for the vector of responses within a primary sampling unit. Details can be found in Neuhaus, Scott and Wild (2002, 2006). Unless we are interested in the within-cluster structure in its own right (as in the family case-control studies considered in Section 9, for example), this requires far too much effort for it to be practicable, certainly for routine analysis.

Can we do something simpler without losing too much efficiency? Weighted estimates are always available, of course. However, they are just as inefficient with complex designs as they are in the simple case considered in the previous section. It turns out that we can do considerably better without too much extra complication.

Return for a moment to the situation of the previous section where we have a simple random sample of size n_1 from the case stratum and an independent simple random sample of size n_0 from the control stratum. Here all units in Stratum ℓ have weight $w_i \propto W_\ell / n_\ell$, where W_ℓ denotes the proportion of the population in the stratum, for $\ell = 0,1$. If we divide throughout by N and set $p_0(\mathbf{x};\boldsymbol{\beta}) = 1 - p_1(\mathbf{x};\boldsymbol{\beta})$, then we can re-write equation (3) for the weighted estimator in the form

$$W_1 \frac{\sum_{\text{cases}} \mathbf{x}_i \, p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - W_0 \frac{\sum_{\text{controls}} \mathbf{x}_i \, p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}. \quad (6)$$

Similarly, we can write equation (5) for the efficient maximum likelihood estimator in the form

$$\omega_1 \frac{\sum_{\text{cases}} \mathbf{x}_i p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - \omega_0 \frac{\sum_{\text{controls}} \mathbf{x}_i p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}, \quad (7)$$

where $\omega_\ell = n_\ell / (n_0 + n_1)$, for $\ell = 0,1$. Both these are special cases of the general set of estimating equations

$$\lambda_1 \frac{\sum_{\text{cases}} \mathbf{x}_i p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - \lambda_0 \frac{\sum_{\text{controls}} \mathbf{x}_i p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}. \quad (8)$$

As $n_0, n_1 \to \infty$, under mild conditions on the way that the finite population is generated from the superpopulation the solution of (8) converges almost surely to the solution $\boldsymbol{\beta}^*$ of

$$\lambda_1 E_1\{\mathbf{X} \, p_0(\mathbf{X};\boldsymbol{\beta}^*)\} - \lambda_0 E_0\{\mathbf{X} \, p_1(\mathbf{X};\boldsymbol{\beta}^*)\} = \mathbf{0}, \quad (9)$$

where $E_\ell\{\cdot\}$ denotes the conditional expectation given that $Y = \ell$ for $\ell = 0,1$. If model (1) is true, then equation (8) has solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ and $\beta_0^* = \beta_0 + b_\lambda$ with $b_\lambda = \log(\lambda_1 W_0 / \lambda_0 W_1)$ for any positive λ_0, λ_1 (see Scott and Wild 1986 for details of the proof). Thus the solution to equation (8) produces consistent estimators of all the regression coefficients apart

from the constant term for any $\lambda_\ell > 0\,(\ell = 0,1)$. As in the simple case, it is easy to correct the inferences about the constant term, provided that we know the proportion of cases in the population.

Now turn to more complex sampling schemes. Since the left hand side of equation (9) just involves two subpopulation means, we can still estimate these means for any standard survey design. This suggests an estimator, $\hat{\boldsymbol{\beta}}_\lambda$ say, for general sampling schemes satisfying

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \lambda_1 \hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta}) - \lambda_0 \hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta}) = \mathbf{0}, \qquad (10)$$

where $\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ is the sample estimator of the subpopulation mean $E_\ell\{\mathbf{X}(1 - p_\ell(\mathbf{X};\, \boldsymbol{\beta}))\}(\ell = 0,1)$. The covariance matrix of $\hat{\boldsymbol{\beta}}_\lambda$ can then be obtained by standard linearization arguments. This leads to an estimated ('sandwich') covariance matrix

$$\hat{\mathrm{C}}\mathrm{ov}\{\hat{\boldsymbol{\beta}}_\lambda\} \approx \mathbf{J}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1}\, \hat{\mathrm{C}}\mathrm{ov}\{\hat{\mathbf{S}}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)\}\, \mathbf{J}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1}, \quad (11)$$

with $\mathbf{J}_\lambda(\boldsymbol{\beta}) = (-\partial\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T)$ and $\hat{\mathrm{C}}\mathrm{ov}\{\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta})\} = \lambda_1^2\, \hat{\mathrm{C}}\mathrm{ov}\{\hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta})\} + \lambda_0^2\, \hat{\mathrm{C}}\mathrm{ov}\{\hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta})\}$. Here, $\hat{\mathrm{C}}\mathrm{ov}\{\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})\}$ denotes the usual survey estimate which should be available routinely for any standard survey design since $\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ is just an estimated mean.

All of this can also be carried out straightforwardly in any package that can handle logistic regression for complex survey designs simply by specifying the appropriate vector of weights. More specifically, suppose that

$$\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta}) = \frac{\sum_{i \in S_\ell} w_i\, \mathbf{x}_i\, (y_i - p_1(\mathbf{x}_i;\, \boldsymbol{\beta}))}{\sum_{i \in S_\ell} w_i}, \qquad (12)$$

where S_1 denotes the case subpopulation (*i.e.*, the set of all units with $Y = 1$) and S_0 denotes the control subpopulation (the set of all units with $Y = 0$). Then the estimating equation (9) can be written in the form

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \sum_{\text{sample}} w_i^*\, \mathbf{x}_i (y_i - p_1(\mathbf{x}_i;\, \boldsymbol{\beta})) = \mathbf{0}, \qquad (13)$$

with $w_i^* \propto \lambda_\ell\, w_i / \sum_{i \in S_\ell} w_i$ for units in $S_\ell(\ell = 0,1)$. In other words, we simply have to scale the case weights and control weights separately so that the sum of the case weights is proportional to λ_1 and the sum of the control weights is proportional to λ_0 and put them, along with the usual specification of the design structure (strata, primary sampling units), into our program of choice. Note that the choice of proportionality constant does not affect the result.

We still have to decide on good values for λ_1 and λ_0. We can often get substantial gains using sample weights $(\lambda_\ell = n_\ell / n)$ compared with using population weights $(\lambda_\ell = W_\ell)$. Scott and Wild (2002) report efficiency gains of 50% or more in Example 2 and in simulations based on that population. The gains became larger as the strength of the relationship increased, and as the effect of clustering increased. Moreover the coverage of confidence intervals was closer to the nominal value for sample weighting in the simulations.

Using sample weights is the most efficient possible strategy when we have simple random samples of cases and controls but for more complex schemes using the sample weights will no longer be fully efficient. We might expect weights based on some form of equivalent sample sizes to perform better. This does indeed produce some gain in efficiency in some limited simulations reported in Scott and Wild (2001a). However, the gains are relatively small, at least when the control sample design effect is less than 2, since $\mathrm{Cov}\{\hat{\boldsymbol{\beta}}_\lambda\}$ is very flat as a function of $\boldsymbol{\lambda}$ near its minimum. Considerations of robustness that we discuss in Section 8 are possibly more important in the choice of $\boldsymbol{\lambda}$.

The gains from sample weighting may depend very much on the particular problem under examination. Korn and Graubard (1999, page 327) comment that, in their experience, the sample weighting strategy rarely produces big gains in efficiency. Obviously more work, both empirical and theoretical, is needed here. In any event, it would seem prudent to fit the model using both sample weights and population weights routinely. If the coefficient estimates are similar, then we can make a judgement based on the estimated standard errors. However, significant differences in the coefficient estimates indicate that the model has been mis-specified. If we are unable to fix up the deficiencies in the model, then we need to think very carefully about just what it is that we are trying to estimate. We look at this again in Section 8.

7. Stratified Sampling

The compromise suggested in the previous section (*i.e.*, use standard survey weighting within the subpopulations defined by case/control status but combine the sub-populations using sample proportions) seems to work reasonably well in practice but it is all completely *ad hoc*. Could we do better with a more systematic approach?

In the special case of stratified random sampling, where independent case-control samples are taken within each stratum, fully efficient procedures are well-developed and easy to implement. In particular, if our model includes a separate intercept for each stratum, then ordinary unweighted logistic regression (with a simple adjustment for the stratum intercepts if they are wanted) is the efficient semi-parametric maximum likelihood procedure (Prentice and Pyke 1979). It is reasonably straightforward to extend this to more general stratified designs. Our model is now

$$\text{logit}\{P(Y = 1 \mid \mathbf{x}, \text{ Stratum h})\} = \beta_{0h} + \mathbf{x}^T\beta_1, \qquad (14)$$

and the stratified equivalent of the estimating equation (7) is

$$\sum_h \left(\lambda_{1h} \frac{\sum_{\text{cases}} \mathbf{x}_i p_{0h}(\mathbf{x}_i; \boldsymbol{\beta})}{n_{1h}} - \lambda_{0h} \frac{\sum_{\text{controls}} \mathbf{x}_i p_{1h}(\mathbf{x}_i; \boldsymbol{\beta})}{n_{0h}} \right) = \mathbf{0}. \quad (15)$$

As $n_{0h},\ n_{1h} \to \infty,$ the solution of (7) converges almost surely to the solution of

$$\sum_h (\lambda_{1h} E_{1h}\{\mathbf{X}\, p_{0h}(\mathbf{X};\ \boldsymbol{\beta})\} - \lambda_{0h} E_0\ \{\mathbf{X}\, p_{1h}(\mathbf{X};\ \boldsymbol{\beta})\}) = \mathbf{0}, \qquad (16)$$

with the obvious extension of the notation from the unstratified case. If model (13) is true, then equation (8) has solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ and $\beta_{0h}^* = \beta_{0h} + b_{\lambda h}$ with $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$. Since equation (14) only involves stratum means, we can estimate them easily using the data coming from any reasonable survey design, for example by

$$\hat{\boldsymbol{\mu}}_{\ell h}(\boldsymbol{\beta}) = \frac{\sum_{i \in S_{\ell h}} w_{ih}\, \mathbf{x}_{ih}(y_{ih} - p_1(\mathbf{x}_{ih};\ \boldsymbol{\beta}))}{\sum_{i \in S_{\ell h}} w_{ih}}.$$

Substituting these estimators in place of the sample means in equation (14) leads to the estimating equation

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \sum_h \sum_{i \in S_h} w_{ih}^*\, \mathbf{x}_i (y_i - p_{1h}(\mathbf{x}_i;\ \boldsymbol{\beta})) = \mathbf{0}, \qquad (17)$$

with $w_{ih}^* \propto \lambda_{\ell h}\, w_{ih} / \sum_{i \in S_{\ell h}} w_{ih}$ for units in $S_{\ell h} (\ell = 0,1;\ \ h = 1, \ldots, H)$. This can be fitted in any standard survey program by including these weights and the appropriate design information. Note that we need to be careful about how we include the so-called 'strata' in the design specification. If primary sampling units are nested within the 'strata', as with the geographical locations in Example 1, there is no problem and the strata should be included in the standard way. However, if the primary sampling units cut across the 'strata', as with age in Example 1 and age and ethnicity in Example 2, then these are not strata in the usual survey sampling sense. They should not be included in the design specifications but simply handled through the weights.

Sometimes we want to model the contribution of the stratum variables using some smooth parametric curve rather than including them through dummy variables. For example, we might well want to include a linear function of age in our model in both Examples 1 and 2. The survey weighted method and the compromise weighting suggested in Section 6 both apply directly and no new theory is needed. More efficient methods are not nearly so simple, however. Fully efficient methods have been developed in the case where simple random samples of cases and controls are drawn within each of the strata (see

Scott and Wild 1997, and Breslow and Holubkov 1997) but the resulting estimating equations are not linear combinations of stratum means and there is no obvious way of generalizing them to more complex sampling schemes. There is a slightly less efficient way that does extend easily, however. If we modify model (14) by including $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$ as an offset, *i.e.*, we set

$$\text{logit}\{P^*(Y = 1 \mid \mathbf{x}, \text{ Stratum h})\} = b_{\lambda h} + \beta_{0h} + \mathbf{x}^T\beta_1, \quad (18)$$

then equation (15) produces consistent, fully efficient, estimates of all the coefficients including $\beta_{0h} (h = 1, \ldots, H)$. Including the same offsets in models where there is no β_{0h} term and the $\mathbf{x}$ vector includes functions of the stratifying variable produces consistent estimators of all the coefficients with typically high (although not full) efficiency (see Fears and Brown 1986, and Breslow and Cain 1988). This generalizes to arbitrary designs immediately. We just use equation (16) with p_{1h} replaced by p^*_{1h} defined by setting $\text{logit}(p^*_{1h}) = b_{\lambda h} + \mathbf{x}^T\beta$. Then any survey program that caters for offsets can be used to fit the model and provide estimated standard error, *etc*.

How much extra efficiency do we get in this case? We have carried out a number of simulations, some of which are reported in Scott and Wild (2002). Most of the scenarios are based on the meningitis study in Example 2 and we set the ratio of the largest to smallest stratum sampling fraction in the control sample at about 10:1. Without any clustering, the gain in efficiency from using the offset method (which is full maximum likelihood in this case) compared to the *ad hoc* procedure was never more than 10%. The relative efficiencies stayed about the same as clustering that cut across strata was introduced. When clustering nested within strata was introduced, the gains disappeared progressively as the design effect increased and the *ad hoc* procedure actually became more efficient than the offset method when the design effect reached about 1.5.

As we stated earlier, it is possible to produce fully efficient semi-parametric estimators if we are willing to model the dependence structure within primary sampling units. We have begun to carry out some simulation. The early results suggest that the extra work involved in the modeling will almost never be worth the effort if we are only interested in the parameters of the marginal model (1). Our tentative conclusion is that, the *ad hoc* partially weighted procedures (with sample weights) are simple to use and work well enough for most practical purposes in the range covered by our experience but this is another area where more empirical work is needed yet. We note, however, that there are some problems, like the case-control family design discussed in Section 9, where the within-cluster behaviour is of interest in its own right. These require more sophisticated methods.

8. Robustness

There must be a catch somewhere. What if the model is not correct? What price do we pay for efficiency then?

By its construction, the population-weighted estimator is always estimating the linear logistic approximation that we would get if we had data from the whole population. By contrast, what the more efficient sample-weighted estimator is estimating depends on the particular sample sizes used. Some people would regard this alone as a strong enough reason for using the population weighted estimator and I suspect that very few people would regard it as completely satisfactory to have the target of their inference depend on the arbitrary choice of sample size.

Our general estimator $\hat{\boldsymbol{\beta}}_\lambda$ satisfying (10) converges to the solution of equation (9), $\mathbf{B}_\gamma$ say, with $\gamma = \lambda_0 / (\lambda_0 + \lambda_1)$, which depends on the true model and distribution of the covariates, as well as on γ. In Scott and Wild (2002), we looked at what happens to $\mathbf{B}_\gamma$ under mild deviations from the assumed model. (We are interested in small deviations since large ones should be picked up by routine model-checking procedures and the model then improved.) For simplicity, suppose that we fit a linear model with a single explanatory variable for the log odds ratio but that the true model is quadratic, say

$$\text{logit}\{P(Y = 1 \mid x)\} = \beta_0 + \beta_1 x + \delta x^2 \qquad (19)$$

with δ small.

Obviously, the actual slope on the logit scale, $\beta_1 + 2\delta x$, changes as we move along the curve. For any $0 < \gamma < 1$, $\mathbf{B}_{\gamma 1}$ is equal to the actual slope at some point along the curve. Denote this value by $x = x_\gamma$. Let x_0 be the expected value of x in the control population and let x_1 the expected value of x in the case population. We shall assume that $\beta_1 > 0$ so that $x_0 < x_1$. It turns out that x_γ always lies between x_0 and x_1 .and that x_γ increases as γ increases from 0 to 1. Recall that survey weighting corresponds to $\gamma = W_0$ and sample weighting to $\gamma = \omega_0 = n_0 / n$. Typically, W_0 is much larger than

ω_0 so that survey weighting gives an estimate of the slope at larger values of x, where the probability of a case is higher, while the slope estimated from sample weighting is closer to the average value of x in the population. Figure 1, adapted from Scott and Wild (2002), illustrates the position in two scenarios, one with positive curvature and one with negative, based roughly on Example 2. The value of δ is chosen so that it would be detected with a standard likelihood ratio test about 50% of the time if we took simple random samples of $n_0 = n_1 = 200$ from the population.

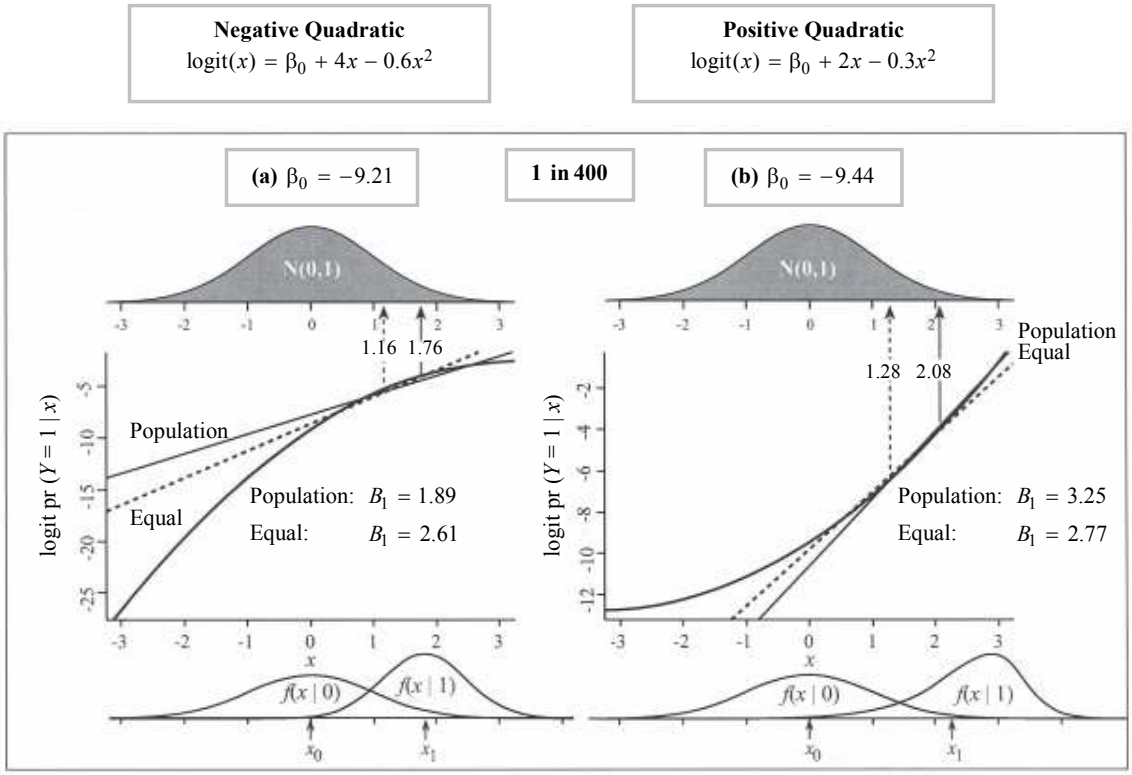

Figure 1. Comparison of population and equal weights.

In both scenarios, the value of β_0 is set so that the proportion of cases in the population is 1 in 400, *i.e.*, so that $W_0 = 0.9975$. The overall density of x is shown at the top of the graph and the conditional densities for cases and controls are shown at the bottom. Values of x_γ and $\mathbf{B}_{\gamma 1}$ are shown for $\gamma = W_0$ (labeled "population") and $\gamma = 0.5$ (labeled "equal"). The latter value corresponds to sample weighting if we draw equal numbers of cases and controls. Clearly, survey weighting is estimating the appropriate slope for values of x further out in the upper tail of the distribution (*i.e.*, for individuals at higher risk) than equal weighting in both scenarios.

Note that if we took simple random samples of $n_0 = n_1 = 200$ from the population in Figure 1 (a), the relative efficiency of survey weighting is only about 16%, and the small sample bias is 0.24. In this case, even if we take the population value as our target, the survey weighting leads to a larger mean squared error than sample weighting.

More results are given in Scott and Wild (2002) where we also look at the effect of omitted covariates. This turns out to have a similar, but somewhat smaller, effect to omitting a quadratic term.

Which is the right value of γ to use? That clearly depends on what we want to use the resulting model for. If our primary interest is in using the model for estimating odds ratios at values of x where the probability of a case is higher, and the sample is large enough so that variance and small sample bias are less important, we might use population weights. For smaller sample sizes, or if we are interested in values of x closer to the population mean, sample weights would be better. A value intermediate between population weighting and sample weighting might sometimes be a sensible compromise. For example trimming the weights to 10:1 (*i.e.*, setting $\gamma \approx 0.91$) in the example, instead of 1:1 (sample weighting) or 400:1 (population weighting), leads to an efficiency of 70% and a small sample bias of 0.04. The corresponding values for population weighting were 16% and 0.24. The value of $x_{0.91}$ lies almost exactly half way between $x_{0.5}$ and $x_{0.9975}$.

9. Case-Control Family Studies

If we are primarily interested in the parameters of the marginal model (1), then the methods that we have discussed in previous sections are simple to implement and reasonably efficient. Fully efficient methods require building parametric models for the within-cluster dependence and the extra effort that this would entail is rarely worthwhile. However, there are situations where the dependence structure is of interest in its own right. In particular, it has become increasingly common for genetic epidemiologists to augment data from a standard case-control study with response and covariate information from family members, in an attempt to gain information on the role of genetics and environment. This can be regarded as a stratified cluster sample, with families as clusters, and the intra-cluster structure is of the primary focus of attention here. The following example is fairly typical.

Example 3.

Wrensch, Lee, Miike, Newman, Barger, Davis, Wiencke and Neuhaus (1997) conducted a population-based case-control study of glioma, the most common type of malignant brain tumor, in the San Francisco Bay Area. They collected information on all cases of glioma that were diagnosed in a specified time interval and on a comparable sample of controls obtained through random digit dialing. They also collected brain tumor status and covariate information from family members of the participants in the original case-control sample. There were 476 brain cancer case families and 462 control families in the study.

We could use the methods that we have been discussing to fit a marginal model for the probability of becoming a glioma victim but a major interest of the researchers was the estimation of within-family characteristics. One way of approaching this would be to fit a mixed logistic model with one or more random family effects.

Note that, strictly speaking, the original sampling scheme in Example 3 is not included in this case-control set-up. The stratification here is related to the response variable but not completely determined by it. Stratum 1 contains the 476 families with a case diagnosed in a particular small time interval while Stratum 2 contains the remaining 1,942,490 families, some of which contain brain cancer victims.

In Neuhaus *et al*. (2006) we develop efficient semi-parametric methods for stratified multi-stage sampling in situations where the stratification depends on the response, possibly in an unspecified way that has to be modeled, and observations within a primary sampling unit are related through some parametric model. The estimates require the solution of $p + 1$ estimating equations, where p is the dimension of the parameter vector. The covariance matrix can also be estimated in a

straightforward way using an analogue of the inverse observed information matrix. The whole procedure can be implemented using any reasonably general maximization routine but this still requires some computing expertise.

We could also fit the same models using survey weighted estimators, which has the big advantage of requiring no specialist software. In our example, case families would have weight 1 and control families would have weight $1,942,490/462 \approx 4,200$. With such a huge disparity, we might expect the weighted estimates to be very inefficient indeed. Unfortunately it turned out to be almost impossible to fit an interesting model for which the weighted estimates converged. One problem is that the weighted estimates are based almost entirely on the control sample and there is very little information about family effects in the control families. (Another problem is that we did not have information on age for family members and any model without age was grossly mis-specified!) For this reason, we had to resort to simulation which is far from complete at this stage. It seems, however, that the efficiency of weighted estimates is less than 10% of the efficient semiparametric estimates here. More details are given in Neuhaus *et al.* (2002, 2006).

Although our simulations are at a very early stage, it is possible to draw a few tentative conclusions. The main one is that within-family quantities are very poorly estimated, even using fully-efficient procedures. Case-control family designs, where the information on family members is obtained as an add-on to a standard case-control design, simply do not contain enough information to estimate the parameters of interest to genetic epidemiologists unless the associations are extremely (even unrealistically) strong. (I should note that not all genetic epidemiologists would agree with this.) More efficient variants are possible, however. For example, if we can identify families containing more than one case, then it is possible to get much greater efficiency by heavily over-sampling such families. In essence, we would be taking the family as the sampling unit, defining a 'case family' as one containing multiple individual cases and then taking a case-control sample of families. This is an important area where a lot of work still needs to be done.

10. Conclusion

The population-based case-control study is one of those subjects where practice has forged ahead of theory. As far as I know, the only book that discusses the topic in any depth is Korn and Graubard (1999, Chapter 9). One aspect that has received a reasonable amount of theoretical attention in the literature is stratification. Efficient procedures for incorporating stratifying variables in the analysis have been developed by Scott and Wild (1997), Breslow and Holubkov (1997), and Lawless *et al.* (1999), among others, when the variables can take only a finite set of values. Breslow and Chatterjee (1999) have considered how best to use such information at the design stage. The extension of all this (both analysis and design) to situations where we have information on continuous variables such as age for all members of the population is an area that still needs work. Much less has been written on the effect of clustering, even though multi-stage sampling is in common use. Exceptions are Graubard *et al.* (1989), Fears and Gail (2000) and Scott and Wild (2001a). Perhaps this paper might stimulate more work on an important topic. In particular, since the essence of the problem boils down to estimating two population means (see equation (8)), it should be possible to transfer a lot of the expertise about efficient survey design across to this problem.

Acknowledgements

I would like to thank the referees and Barry Graubard and Graham Kalton, whose thoughtful discussion of an early version of this paper helped my understanding of the subject considerably. Finally, I would to give special thanks to my long term collaborators Chris Wild, with whom almost all the work underlying this paper was done, and Jon Rao, with whom I learnt essentially everything that I know about the analysis of survey data.

References

Baker, M., McNicholas, A., Garrett, N., Jones, N., Stewart, J., Koberstein, V. and Lennon, D. (2000). Household crowding: A major risk factor for epidemic meningococcal disease in Auckland children. *Pediatric Infectious Disease Journal*, 19, 983-990

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Breslow, N.E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91, 14-28.

Breslow, N.E. (2004). Case-control studies. In *Handbook of Epidemiology*. (Eds. W. Aherns and I. Pigeot). New York: Springer. 287-319.

Breslow, N.E., and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11-20.

Breslow, N.E., and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Applied Statistics*, 48, 457-468.

Breslow, N.E., and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society*, B, 59, 447-461.

Brogan, D.J., Denniston, M.M., Liff, J.M., Flagg, E.W., Coates, R.J. and Brinton, L.A. (2001). Comparison of telephone sampling and area sampling: Response rates and within-household coverage. *American Journal of Epidemiology*, 153, 1119-1127.

Cosslett, S.R. (1981). Maximum likelihood estimation for choice-based samples. *Econometrica*, 49, 1289-1316.

DiGaetano, R., and Waksberg, J. (2002). Trade-offs in the development of a sample design for case-control studies. *American Journal of Epidemiology*, 155, 771-775.

Fears, T.R., and Brown, C.C. (1986). Logistic regression models for retrospective case-control studies using complex sampling procedures. *Biometrics*, 42, 955-960.

Fears, T.R., and Gail, M.H. (2000). Analysis of a two-stage case-control study with cluster sampling of controls: Application to nonmelanoma skin cancer. *Biometrics*, 56, 190-198.

Graubard, B.I., Fears, T.R. and Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control sampling. *Biometrics*, 45, 1053-1071.

Hartge, P., Brinton, L.A., Rosenthal, J.F., Cahill, J.I., Hoover, R.N. and Waksberg, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.

Hartge, P., Brinton, L.A., Cahill, J.I., West, D., Hauk, M., Austin, D., Silverman, D. and Hoover, R.N. (1984). Design and methods in a multi-center case-control interview study. *American Journal of Public Health*, 74, 52-56.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

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, 413-38.

Lee, A.J., Scott, A.J. and Wild, C.J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 95 (to appear).

Manski, C.F., and McFadden, D. (Eds) (1981). *Structural Analysis of Discrete Data with Econometric Applications*. New York: John Wiley & Sons, Inc.

Miettinen, O.S. (1985). The case-control study: Valid selection of subjects. *American Journal of Epidemiology*, 135, 1042-1050.

Prentice, R.L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

Neuhaus, J., Scott, A.J. and Wild, C.J. (2002). The analysis of retrospective family studies. *Biometrika*, 89, 23-37.

Neuhaus, J., Scott, A.J. and Wild, C.J. (2006). Family-specific approaches to the analysis of retrospective family data. *Biometrics*, 62, in press.

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

Rao, J.N.K., Scott, A.J. and Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.

Scott, A.J., and Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society*, B, 48, 170-182.

Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 83, 57-72.

Scott, A.J., and Wild, C.J. (2001a). The analysis of clustered case-control studies. *Applied Statistics*, 50, 57-71.

Scott, A.J., and Wild, C.J. (2001b). Fitting regression models to case-control data by maximum likelihood. *Journal of Statistical Planning and Inference*, 96, 3-27.

Scott, A.J., and Wild, C.J. (2002). On the robustness of weighted methods for fitting model to case-control data by maximum likelihood. *Journal of the Royal Statistical Society*, B, 64, 207-220.

Wacholder, S., McLaughlin, J.K., Silverman, D.T. and Mandel, J.S. (1991). Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology*, 135, 1019-1028.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Waksberg, J. (1998). Random digit dialing sampling for case-control studies. In *Encyclopedia of Biostatistics.* (Eds. P.Armitage and T. Colton). New York: John Wiley & Sons, Inc., 3678-3682.

Wrensch, M., Lee, M., Miike, R., Newman, B., Barger, G., Davis, R., Wiencke, J. and Neuhaus, J. (1997). Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls. *American Journal of Epidemiology*, 145, 581-93.