University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2002 *Paper* 116

Case-Control Current Status Data

Nicholas P. Jewell* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, jewell@uclink.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley

Case-Control Current Status Data

Nicholas P. Jewell and Mark J. van der Laan

Abstract

Current status observation on survival times has recently been widely studied. An extreme form of interval censoring, this data structure refers to situations where the only available information on a survival random variable, T, is whether or not T exceeds a random independent monitoring time C, a binary random variable, Y. To date, nonparametric analyses of current status data have assumed the availability of i.i.d. random samples of the random variable (Y, C), or a similar random sample at each of a set of fixed monitoring times. In many situations, it is useful to consider a case-control sampling scheme. Here, cases refer to a random sample of observations on C from the sub-population where T is less than or equal to C. On the other hand, controls provide a random sample of observations from the sub-population where T is greater than C. In this paper, we examine the identifiability of the distribution function F of T from such case-control current status data, showing that F is identified up to a one parameter family of distribution functions. With supplementary information on the relative population frequency of cases/controls, a simple weighted version of the nonparametric maximum likelihood estimator for prospective current status data provides a natural estimate for case-control samples. Following the parametric results of Scott and Wild (1997), we show that this estimator is, in fact, nonparametric maximum likelihood.

1 Introduction

In some survival analysis applications, observation of the random variable T is restricted to knowledge of whether or not T exceeds a random monitoring time C. This structure is known as current status data. Nonparametric estimation of the survival function, and semi-parametric techniques for related regression models, based on current status data, have been much studied of late. See Jewell and van der Laan (1997) for a brief review, references, and some extensions. Usually data are assumed to arise from simple random samples from a population so that both the monitoring time and current status information are random, but available methods also directly apply to situations where monitoring times are fixed in advance.

Often, the failures of interest are rare in the population so that random samples will provide very few observations where failure has occurred at the observed monitoring time, whether the latter is random or fixed. In these contexts, it is natural to consider a case-control strategy where separate samples of individuals to whom an event has already occurred (cases) and those for whom the event has not yet occurred (controls) are obtained. This paper considers identifiability and estimation of the survival distribution based on case-control samples of current status data.

Let T be the survival random variable of interest, with associated distribution function F, and survival distribution $S = 1 - F$. For the remainder of the paper we focus on F, rather than S. For any individual, we assume that the monitoring time, C, is random and follows the distribution function G, independently of T. In standard current status data, with random C, an i.i.d. sample of n individuals is drawn from the joint distribution of

(T, C); however, only $\{(\Delta_i, C_i : i = 1, \ldots, n\}$ is observed where $\Delta = I(T \leq C)$. With case-control sampling, we obtain two separate samples, the first an i.i.d. random sample of size n_1 from individuals for whom $T \leq C$ (cases), the second an i.i.d random sample of size n_0 from those for whom $T > C$ (controls).

Section 2 describes some examples where case-control current status data naturally arises. In Section 3, we consider identifiability of F from such data. Following the elegant work of Scott and Wild (1997), nonparametric maximum likelihood estimation of F, based on case-control current status data supplemented by information on the population frequency of cases/controls, is developed in Section 4. In the absence of the population information, F can only be identified up to a one parameter family of distribution functions. Illustrative examples are presented in Section 5.

2 Motivating Examples

Two prime examples of current status data yield situations where case-control samples may often be desirable or unavoidable. The first case arises from partner studies of HIV infection (Jewell and Shiboski, 1990; Shiboski, 1998a) where HIV infection data is collected on both partners in a long-term sexual relationship. These partnerships are assumed to include a primary infected individual (index case) who has been infected via some external source, and a susceptible partner who has no other means of infection other than contact with the index case. Suppose T denotes the time (or number of infectious contacts) from infection of the index case to infection of the susceptible partner, and that the partnership is evaluated at a single time C after infection of the index case; then, the infection status of the susceptible partner provides current status data on T at time C.

Conventional sampling schemes for partner studies have largely been based on convenience samples that are assumed random in that the probability of selection of a partnership does not depend on T, and thus whether $T \leq C$ or not. Recently, case-control partner studies have been proposed as a practical alternative to these ad hoc sampling approaches. Further, the methods described here provide an approach to sensitivity analyses for previous partner studies when there is concern that selection probabilities may be associated with the infection status of the susceptible partner at monitoring. In Section 5, we apply our results to data from the California Partners Study (Padian et al., 1997).

The second common area of application is to estimation of the distribution of age at incidence of an occult non-fatal disease for which accurate diagnostic tests are available. If a cross-sectional sample of a given population receives such a diagnostic test, then the presence/absence of disease in an individual of age C yields current status information on the age, T, at disease incidence. Keiding (1991) describes the nonparametric maximum likelihood estimator of the distribution of the age at incidence of Hepatitis A infection, based on cross-sectional data obtained by K. Dietz.

For many diseases of low incidence, this approach to age incidence is only viable if a case-control sampling scheme is used. For example, with Alzheimer's disease, it is feasible to obtain a random sample of prevalent Alzheimer's patients, measuring their age at sampling. Population control samples are straightforward to obtain. Using the methods of this paper, the age incidence distribution of Alzheimer's disease in a given population is then easily obtained so long as the overall population prevalence is known. Our results are also relevant to genetic studies where the age of disease onset is a phenotype of interest in linkage analyses, and where individuals are ascertained according to their current disease status.

3 Identifiability

It will be helpful to introduce some additional notation at this point. We define the binary random variable Y to be 1 if $T \leq C$ and 0 if $T > C$. Thus, $E(Y|C = c) = P(T \leq C|C = c) = F(c)$, and so estimation of F can be viewed in terms of estimation of the conditional expectation of Y for all c. Further, we reparametrize the distribution function F by writing

$$F(t) = \frac{1}{1 + e^{-\phi(t)}} \tag{1}$$

for all t in the support of F. Note that, for Fs with infinite support, $\lim_{t \to\ 0} \phi(t) = -\infty$ and $\lim_{t \to\ \infty} \phi(t) = \infty$, with suitable modifications when F has finite support. Further, the function $\phi(t)$ is non-decreasing in t. Any such function ϕ that meets these conditions defines a distribution function via (1) and vice-versa. Expressing ϕ in terms of F is straightforward: $\phi(t) = \log\left[\frac{F(t)}{1-F(t)}\right]$. We refer to the function ϕ as the *log odds function* associated with F.

For any such function ϕ_0, define a one parameter family of distribution functions $\mathcal{F}_{\phi_0}$ as $\left\{F : F(t) = \frac{1}{1+e^{-\phi_0(t)+a}} \text{ for some constant } a\right\}$. We refer to this as the *proportional odds family associated with* ϕ_0 or *with* $F_0 = \left(1 + e^{-\phi_0}\right)^{-1}$, since all members of the family share identical odds ratios $\frac{F(t)/(1-F(t))}{F(s)/(1-F(s))}$ for any s, t in their support. The choice of ϕ_0 within $\mathcal{F}_{\phi_0}$ is arbitrary.

To address identifiability, we need to consider the relation between the function ϕ (or F) and the density functions of the observed data, that is $f_1(c) = P(C|Y = 1)$ and $f_0(c) = P(C|Y = 0)$. This is immediate since, by Bayes' rule,

$$\begin{aligned}
\phi(t) - \phi(s) &= \log\left[\frac{P(C = t|Y = 1)}{P(C = t|Y = 0)}\right] - \log\left[\frac{P(C = s|Y = 1)}{P(C = s|Y = 0)}\right] \\
&= \log\left[\frac{f_1(t)}{f_0(t)}\right] - \log\left[\frac{f_1(s)}{f_0(s)}\right]
\end{aligned}$$

for any s, t in the support of C. Since both f_0 and f_1 are identified from case-control current status data, it follows immediately that $\phi(t) - \phi(s)$ is also identified for any s, t in the support of C.

With prospective current status data, full identification of F requires the support of F to be contained in the support of G. Assuming the same condition here, the above observation then shows that the function ϕ (and thus F) is identifiable, up to a constant, from case-control current status data. In other words, the data points to the proportional odds family containing F, but cannot—without further information—identify the specific member of the family that generates the observations.

This non-identifiability is most easily demonstrated if F is known to belong to a parametric family. For example, if F is a member of the two parameter logistic distribution family, with $\log\left[\frac{F(t}{1-F(t)}\right] = \alpha + \beta t$, case-control current status data can only identify β and not α, absent other information. With a simple one parameter family, such as the Exponential, identifiability is theoretically possible; in this case, if the data is generated by a given Exponential, then identifying the proportional odds family that contains this distribution must, at the same time, identify the particular Exponential parameter since the proportional odds family of an Exponential distribution contains one and only one Exponential distribution. Of course, this identifiability is extraordinarily dependent on the particular parametric assumption invoked.

Ingeneral, we thus set the problem of estimation of F with case-control current status data within a broader problem where additional information is available. In particular, suppose that N individuals are sampled from the joint distribution of (Y, C). The numbers of individuals for whom $Y = i, (i = 0, 1)$, say N_0 and N_1, respectively, are observed.

However, no data on the random variable C is available at this point. To obtain such information, fixed samples of size $n_0(\leq N_0)$ and $n_1(\leq N_0)$ are selected, by simple random sampling, separately from the two groups, with $Y = 0$ and $Y = 1$, in the original sample of N. The random variable C is then measured for each of the $n_0 + n_1$ sampled individuals at this stage. In practice, the sampling rates, at this second stage, that is (n_0/N_0) and (n_1/N_1) will usually be quite different. Following Scott and Wild (1997), we refer to this structure as *case-control data supplemented by information on population totals.*

The supplemented data is thus $\{(y_{ij}, c_{ij}) : i = 0, 1; j = 0, \ldots, n_i; N_0, N_1\}$. We assume that, given N_0 and N_1, the sample sizes, n_0 and n_1, are random but non-informative, meaning that the conditional distribution of n_0 and n_1, given N_0 and N_1, has support on $n_i \leq N_i$, only depends on the original sample of N through N_0 and N_1, and does not depend on F or G. Note that we cannot assume n_i is fixed since N_i is random, but the latter assumption includes the possibility that each n_i is a fixed fraction of N_i.

For parametric estimation of F, the iterative method of Scott & Wild (1997) now applies directly by casting the problem in terms of estimating the binary regression model that links Y and the monitoring time C, namely $E(Y|C = c) = P(Y = 1|C = c)$, based on case-control data. Their approach is particularly straightforward when this regression model reduces to a standard generalized linear model, since each iteration involves only estimation of the (prospective) version of the latter that can be addressed using standard software. For example, if F is assumed to follow a Weibull distribution, with hazard $e^a b t^{b-1}$, then $\log - \log[E(Y|C = c)] = a + b \log c$, so that each iteration includes estimation of the parameters of a binary generalized linear model with complementary log-log link. In the next section, we adapt this approach to nonparametric estimation of F.

4 Nonparametric Maximum Likelihood Estimation

As at the end of the last section, we assume that the available case-control data is $\{(y_{ij}, c_{ij}) : i = 0, 1; j = 0, \ldots, n_i; N_0, N_1\}$, where N_0 and N_1 are known. With this population information in hand, a simple consistent estimator of F is immediately available by weighting observations inversely proportional to their probability of selection. In particular, the weights are (N_0/n_0) for controls and (N_1/n_1) for cases. Subsequently, the nonparametric maximum likelihood estimator for prospective current status data (Groeneboom & Wellner, 1980) can be directly applied to the weighted data to yield an estimator $\hat{F}$ of F. Computation of the estimator is achieved through use of the pool-adjacent-violators algorithm (Ayer et al., 1955, Barlow et al., 1972). We now show that this simple estimator is, in fact, the nonparametric maximum likelihood estimator based on case-control data supplemented by information on population totals.

First, it is straightforward to show that the likelihood function of the supplemented data is given by

$$L = \prod_{i=0}^{1} \left\{ \prod_{j=1}^{n_i} Pr(c_{ij}|Y = i) \right\} Pr(Y = i)^{N_i}; \tag{2}$$

see (4) in Scott and Wild (1997), or (2) in Wild (1991). This likelihood, (2) can be written in terms of F and G as follows:

$$\begin{aligned} L &= \prod_{i=0}^{1} \left\{ \prod_{j=1}^{n_i} Pr(y_{ij} = i|c_{ij}) Pr(c_{ij}) \right\} Pr(Y = i)^{N_i - n_i} \\ &= \prod_{j=1}^{n_0} (1 - F(c_{ij})) dG(c_{ij}) \prod_{j=1}^{n_1} F(c_{ij}) dG(c_{ij}) \\ &\quad \times \left\{ \int (1 - F(c)) dG(c) \right\}^{N_0 - n_0} \left\{ \int F(c) dG(c) \right\}^{N_1 - n_1}. \end{aligned} \tag{3}$$

Note that the first two terms in this product correspond to the likelihood from a prospective

sample of current status data where n_i observations have $Y_i = i$.

We wish to find the nonparametric maximum likelihood estimate (NPMLE) of F (and G) based on this likelihood, assuming N_i is known for $i = 0, 1$. For basic case-control data, where these population totals are not observed, our strategy is to assume specific values for N_0 and N_1, compute the NMPLE, and then allow the population totals to vary as a sensitivity parameter.

Following Scott and Wild (1997), we first profile the 'parameter' G out of the likelihood, that is we maximize (3) solely in terms of G, holding F fixed; the resulting estimate of G will, of course, be a function of F. As in Scott and Wild (1997), it is straightforward to see that this maximum likelihood estimate of G only places mass at observed values of C, namely $\{c_{ij} : i = 0, 1; j = 1, \ldots, n_i\}$. For notational simplicity, rewrite these observed monitoring times as a distinct set $x_k : k = 1, \ldots, K$, and set n_{+k} to be the number of $c_{ij} = x_k$. Then, if $\hat{\delta}_k$ is the maximum likelihood estimate of $dG(x_k)$, Scott and Wild (1987) show that

$$\hat{\delta}_k = \frac{n_{+k}}{N(\mu_0(1 - F(x_k)) + \mu_1 F(x_k))},$$

where each μ_i is implicitly defined through $\mu_i = \frac{n_i - \gamma_i}{N_i - \gamma_i}$, together with

$$\gamma_i = n_i - \sum_{k=1}^{K} n_{+k} \bar{F}_i(x_k) = n_i - \sum_{m=0}^{1} \sum_{j=1}^{n_m} \bar{F}_i(c_{mj}) \tag{4}$$

where

$$\bar{F}_0(c) = \frac{\mu_0(1 - F(c))}{\mu_0(1 - F(c)) + \mu_1 F(c)}, \tag{5}$$

and

$$\bar{F}_1(c) = 1 - \bar{F}_0(c).$$

Note that each μ_i (or γ_i) implicitly depends on F; also, $\gamma_0 + \gamma_1 = 0$.

By substituting $\hat{\delta}_k$, in terms of γ_i, for dG in (3), we can now write the profile log likelihood, up to a constant, as

$$\log L_P = \log L_P(F, \gamma_0(F), \gamma_1(F)) \simeq \sum_{i=0}^{1} \sum_{j=1}^{n_i} \log \bar{F}_i(c_{ij}) + \sum_{i=0}^{1} N_i \log(N_i - \gamma_i) - \sum_{i=0}^{1} n_i \log(n_i - \gamma_i). \tag{6}$$

It now remains to find the NPMLE, F_{ML}, that maximizes L_P over the space of distribution functions F. We describe F_{ML} through the appropriate score equation. This is calculated by considering the one dimensional model $F_\epsilon = (1 + \epsilon h) d\hat{F}_{ML}$, regarded as a function of ϵ, for $h \in L_0^2(F_{ML})$. Since F_{ML} is the NPMLE in the full model, it follows that

$$\frac{d}{d\epsilon} \log L_P(F_\epsilon, \gamma_0(F_\epsilon), \gamma_1(F_\epsilon)) \Big|_{\epsilon=0} = 0,$$

with the constraint that $\gamma_0(F_\epsilon) + \gamma_1(F_\epsilon) = 0$. This score equation is thus

$$\frac{\partial}{\partial \epsilon} \log L_P(F_\epsilon, \gamma_0, \gamma_1) + \left(\frac{\partial}{\partial \gamma_0} \log L_P \right) \left(\frac{\partial \gamma_0(F_\epsilon)}{\partial \epsilon} \right) - \left(\frac{\partial}{\partial \gamma_1} \log L_P \right) \left(\frac{\partial \gamma_0(F_\epsilon)}{\partial \epsilon} \right) \Big|_{\epsilon=0} = 0, \tag{7}$$

the last negative sign arising since $\frac{\partial \gamma_1}{\partial \gamma_0} = -1$.

First, note that

$$\frac{\partial}{\partial \gamma_i} \log \bar{F}_i(c_{ij}) = -a_i \left(1 - \bar{F}_i(c_{ij})\right)$$

where $i = 0, 1$, and $a_i = (n_i - \gamma_i)^{-1} - (N_i - \gamma_i)^{-1}$. Similarly, for $i \neq m$,

$$\frac{\partial}{\partial \gamma_i} \log \bar{F}_m(c_{mj}) = a_i \bar{F}_i(c_{mj}).$$

Thus,

$$\frac{\partial}{\partial \gamma_i} \log L_P \Big|_{\epsilon=0} = a_i \left[\gamma_i - \left(n_i - \sum_{m=0}^{1} \sum_{j=1}^{n_m} \left(\bar{F}_{ML} \right)_i (c_{mj}) \right) \right], \tag{8}$$

where it is implicit here that γ_i and thus a_i are evaluated at F_{ML}. It follows immediately from (4) that $\frac{\partial}{\partial \gamma_i} \log L_P \Big|_{\epsilon=0} = 0$ for $i = 0, 1$.

To evaluate the score equation, it remains, from (6), to consider $\frac{\partial}{\partial \epsilon} \log L_P(F_\epsilon, \gamma_0, \gamma_1) = \frac{\partial}{\partial \epsilon} \sum_{i=0}^{1} \sum_{j=1}^{n_i} \log \bar{F}_{\epsilon i}(c_{ij})$. Treating γ_0 and γ_1 as constants, we have

$$\begin{aligned} \frac{\partial}{\partial \epsilon} \log L_P(F_\epsilon)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} \sum_{i=0}^{1} \sum_{j=1}^{n_i} \left\{ y_{ij} \log \bar{F}_{\epsilon 1}(c_{ij}) + (1-y_{ij}) \log(1-\bar{F}_{\epsilon 0}(c_{ij}) \right\} \Big|_{\epsilon=0} \\ &= \sum_{i=0}^{1} \sum_{j=1}^{n_i} \frac{y_{ij}}{\bar{F}_{\epsilon 1}(c_{ij})} \frac{\partial}{\partial \epsilon} \bar{F}_{\epsilon 1}(c_{ij}) \Big|_{\epsilon=0} - \frac{(1-y_{ij})}{(1-\bar{F}_{\epsilon 0}(c_{ij}))} \frac{\partial}{\partial \epsilon} \bar{F}_{\epsilon 0}(c(_{ij}) \Big|_{\epsilon=0} \end{aligned}$$

$$= \sum_{i=0}^{1} \sum_{j=1}^{n_i} \left(\int_0^{c_{ij}} h dF_{ML} \right) (\mu_0 \mu_1) \left\{ \frac{y_{ij}}{(\bar{F}_{ML})_1(c_{ij})} - \frac{(1-y_{ij})}{(\bar{F}_{ML})_0(c_{ij}))} \right\} \{\mu_1 F_{ML}(c_{ij}) + \mu_0 (1-F_{ML})(c_{ij})\}^{-2}.$$

Since this holds for all $h \in L_0^2(F)$, we have, for every pair of support points (t_k, t_{k+1}) of F_{ML},

$$\sum_{i=0}^{1} \sum_{j=1}^{n_i} \frac{y_{ij}}{(\bar{F}_{ML})_1(c_{ij})} \{\mu_1 F_{ML}(c_{ij}) + \mu_0 (1-F_{ML})(c_{ij})\}^{-2} I\{c_{ij} \in (t_k, t_{k+1}]\} =$$

$$\sum_{i=0}^{1} \sum_{j=1}^{n_i} \frac{(1-y_{ij})}{(\bar{F}_{ML})_0(c_{ij}))} \{\mu_1 F_{ML}(c_{ij}) + \mu_0 (1-F_{ML})(c_{ij})\}^{-2} I\{c_{ij} \in (t_k, t_{k+1}]\}. \quad (9)$$

Since F_{ML} is piecewise constant, terms in $F_{ML}(c_{ij})$ can be factored out of the summations in (9) to yield the score equation

$$\sum_{i=0}^{1} \sum_{j=1}^{n_i} \frac{y_{ij}}{(\bar{F}_{ML})_1(c_{ij})} I\{c_{ij} \in (t_k, t_{k+1}]\} = \sum_{i=0}^{1} \sum_{j=1}^{n_i} \frac{(1-y_{ij})}{(\bar{F}_{ML})_0(c_{ij}))} I\{c_{ij} \in (t_k, t_{k+1}]\}. \quad (10)$$

When $\mu_0 = \mu_1$, $\bar{F}_1 = F$ and $\bar{F}_0 = 1 - F$, so that this is merely the score equation for standard prospective current status data. Suppose F^* yields the solution to this standard current status problem, then the score equation (10) is solved by the distribution function F given by

$$F(c) = \frac{\mu_0 F^*(c)}{\mu_1 (1 - F^*(c)) + \mu_0 F^*(c)}. \quad (11)$$

This, of course, can be seen directly by noting that maximization of $\sum_{i=0}^{1} \sum_{j=1}^{n_i} \log \bar{F}_i(c_{ij})$ over all distribution functions F (holding μ_0 and μ_1 constant) is equivalent to maximization of $\sum_{i=0}^{1} \sum_{j=1}^{n_i} \log \bar{F}_i(c_{ij})$, over all log odds functions ϕ, where now $\bar{F}_0(c_{ij}) =$

$\left(e^{a-\phi(c_{ij})}\right)\left(1+e^{a-\phi(c_{ij})}\right)^{-1}$ and $\bar{F}_1(c_{ij}) = \left(1+e^{a-\phi(c_{ij})}\right)^{-1}$, with $a = \log(\frac{\mu_0}{\mu_1})$. In turn, this is equivalent to maximization of $\sum_{i=0}^{1}\sum_{j=1}^{n_i}\log \bar{F}^*{}_i(c_{ij})$, over all log odds functions ϕ^*, where now $\bar{F}^*{}_0(c_{ij}) = \left(e^{-\phi^*(c_{ij})}\right)\left(1+e^{-\phi^*(c_{ij})}\right)^{-1}$ and $\bar{F}^*{}_1(c_{ij}) = \left(1+e^{-\phi^*(c_{ij})}\right)^{-1}$. The latter problem simply yields the prospective current status nonparametric maximum likelihood estimator $\hat{\phi}^*$; thus, the solution to the original maximization is the distribution function $\hat{F}$ given by $\hat{F} = \frac{\mu_0\hat{F}^*(c)}{\mu_1(1-\hat{F}^*(c))+\mu_0\hat{F}^*(c)}$, where $\hat{F}^* = \left(1+e^{-\phi^*}\right)^{-1}$. Note that the solution only depends on μ_0 and μ_1 through the ratio $(\frac{\mu_0}{\mu_1})$.

The solution of (10) depends on the particular values of μ_i, or γ_i, that, in turn, are evaluated at F_{ML}. However, from (5) and (11), it follows that

$$(\bar{F}_{ML})_0(c) = 1 - F^*(c).$$

Thus, at F_{ML},

$$\begin{aligned}\gamma_0 &= n_0 - \sum_{m=0}^{1}\sum_{j=1}^{n_m}(\bar{F}_{ML})_0(c_{mj}) \\ &= n_0 - \sum_{m=0}^{1}\sum_{j=1}^{n_m}(1-F^*)(c_{mj}).\end{aligned}$$

Since F^* is the prospective NPMLE of the data, it is a piecewise weighted average of the observed Y_{ij}'s so that $\sum_{m=0}^{1}\sum_{j=1}^{n_m}(1-F^*)(c_{mj}) = n_0$; this can also be seen directly from the score equation (10). Thus, $\gamma_0(F_{ML}) = 0$, or, equivalently, $\mu_0(F_{ML}) = n_0/N_0$. Similarly, $\gamma_1(F_{ML}) = 0$ and $\mu_1(F_{ML}) = n_1/N_1$.

We have thus shown that the NPMLE, F_{ML}, satisfies (10) with $\gamma_i = 0, i = 0, 1$. Hence, F_{ML} can be computed using the prospective current status NPMLE on the data, F^*, ignoring the design, and then using (11) with $\mu_0 = n_0/N_0$ and $\mu_1 = n_1/N_1$. This is, of course, just the weighted version of F^* with weights inversely proportional to the probability of selection.

The NPMLE assumes knowledge of the population totals N_0 and N_1 (in fact only the ratio N_1/N_0 need be known). Without such information, we can hypothesize a value for N_1/N_0, compute the NPMLE, and then vary the assumed N_1/N_0 as a sensitivity parameter over a range of plausible values. If N_1/N_0 is allowed to take on all values, the corresponding NPMLEs trace out the population odds family associated with any particular choice of N_1/N_0. This, of course, merely reinforces the identifiability findings of Section 3.

5 Example

We now apply the results of the previous section to a study of HIV transmission from males to females. The available data is based on 94 long-term heterosexual partnerships where the index case was male and data was available on both the time and number of sexual contacts between infection of the index case and the time when the infection status of the female partner was monitored. The range of these 94 times, and number of contacts, was 18 to 144 months, and 4 to 3334 contacts, respectively. At monitoring, 18 female partners were observed to be infected. Previously, this data has been analyzed assuming that the data was randomly sampled from a larger population, although it was recognized that this did not exactly reflect the operation of the study. n particular, there is a plausible concern that partnerships where transmission has already occurred (cases) might have been recruited at a higher rate than those where infection had not yet taken place (controls). For further details, see Padian et al (1997).

Figure 1 provides estimates of the distribution of the number of contacts between infection of the index case and the partner, based on this data, and assuming that either the selection probabilities are identical for cases and controls, or the selection rate is double or

fivefold greater for cases than controls.

FIGURE 1 ABOUT HERE

Figure 2 replots these estimates in terms of the associated log odds functions, which are parallel as discussed in Section 3, each possessing the exact same relative shape that for two different values s and t, give a identical odds ratio. Similar estimates, not shown here, are available for the distribution of the chronological time between infection of the index case and the partner.

FIGURE 2 ABOUT HERE

6 Discussion

We have approached the problem of nonparametric estimation of F by supplementing the data with the population information N_1/N_0. In many cases, different forms of population information might be known, such as the mean of F say. Here, the NPMLE with this mean can be found by selecting the unique distribution function in the population odds family identified by the data that shares this mean.

It would be of interest to extend the results of Huang and Wellner (1995) to smooth functionals of F, estimated from case-control current status data supplemented with the population case/control frequencies. Following the methods of Groeneboom and Wellner (1980) to establish convergence of the NPMLE, in this situation, (thereby determining the limit distribution), as n_0 and n_1 (and thus N_0 and N_1) tend to infinity appropriately, would also be of theoretical interest.

With regard to the example pertaining to age incidence, described in Section 2, future

work will address the common scenario where mortality may be increased by presence of the disease. Ignoring this possibility leads to biased estimates of the population age incidence distribution.

Finally, we note that the ideas discussed here apply more generally to estimation of a regression model linking T to a set of covariates $\mathbf{Z}$ using regression parameters β. In this situation, the latter model induces a binary regression model for Y, parametrized by β and the distribution function, F_0, of T at the baseline value of the covariates $\mathbf{Z} = \mathbf{0}$. With case-control data, the methods introduced in Section 4 are particularly easy if T follows the proportional odds regression model (Bennett, 1983) defined by

$$1 - S(t|\mathbf{Z} = \mathbf{z}) = \frac{1}{1 + e^{-\alpha(t) - \beta\mathbf{z}}},$$

where $S_0(t) = \frac{1}{1+e^{\alpha(t)}}$. Here, Y is associated with $\mathbf{Z}$ via the logit link:

$$\log \frac{p(\mathbf{z}|c)}{(1 - p(\mathbf{z}|c))} = \alpha(c) + \beta\mathbf{z}.$$

Here, the 'intercept' term, $\alpha(C) = \log \frac{(1-S_0(C))}{S_0(C)}$ is just the log odds function associated with F_0. If the baseline survival function S_0 is assumed to follow a particular parametric form, the corresponding binary regression model will often simplify to a familiar generalized linear model, so that again the techniques of Scott & Wild (1997) can be used to estimate both S_0 and the regression parameters β from case-control data. On the other hand, if S_0 is left arbitrary, a backfitting algorithm to compute estimates of β can be used along with the weighted NPMLE estimate of α studied in section 4. Shiboski (1998b) provides an excellent review of these methods for prospective current status data. With other regression models, including the proportional hazards and accelerated failure time models, the situation is more complex; presumably, an iterative technique, analogous to that used

by Scott and Wild (1997) will be needed here, iterating between estimation of β and the weighted NPMLE of the appropriate monotonic intercept term. These regression models can all be viewed as special cases of a generalized additive model (Hastie and Tibshirani, 1990), with a single isotonic component (in the random variable C), whose shape depends on F_0, 'added' to a regression model in $\mathbf{Z}$, with estimation based on case-control data.

References

Ayer, M, Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics* **26**, 641-647.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D. (1972) *Statistical Inference under Order Restrictions*. New York: Wiley.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273-277.

Groeneboom, P, Wellner, J.A. (1980) *Nonparametric Maximum Likelihood Estimators for Interval Censoring and Denconvolution*. Boston: Birkhäuser-Boston.

Hastie, T.J., Tibshirani, R.J. (1990) *Generalized Additive Models*. London: Chapman and Hall.

Huang, J., Wellner, J.A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I. *Statistica Neerlandica* **49**, 153-163.

Jewell, N.P., van der Laan. (1997). Singly and doubly censored current status data with extensions to multi-state counting processes. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Lecture Notes in Statistics #123, 171-184, Berlin: Springer.

Jewell, N.P., Shiboski, S. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics* **46**, 1133-1150.

KEIDING, N. (1997). Age-specific incidence and prevalence: a statistical perspective (with discussion). *Journal of the Royal Statistical Society A* **154**, 371-412.

PADIAN, N.S., SHIBOSKI, S.C., GLASS, S.O., VITTINGHOFF, E. (1997). Heterosexual transmission of HIV in northern California: results from a ten year study. *American Journal of Epidemiology* **146**, 350-357.

SCOTT, A.J., WILD, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57-71.

SHIBOSKI, S.C. (1998a). Partner Studies. In *The Encyclopedia of Biostatistics*, P. Armitage and T. Colton (eds.) 3270-3275.

SHIBOSKI, S.C. (1998b). Generalized additive models for current status data. *Lifetime Data Analysis* **4**, 29-50.

WILD, C.J. (1991). Fitting prospective regression models to case-control data. *Biometrika* **78**, 705-717.

Figure 1: Estimates of the Distribution of Number of Contacts Between Infection of the Male Index Case and the Female Partner, Assuming $\mu_1/\mu_0 =$ (i) 1, (o) (ii) 2 $(+)$, or (iii) 5 (x)

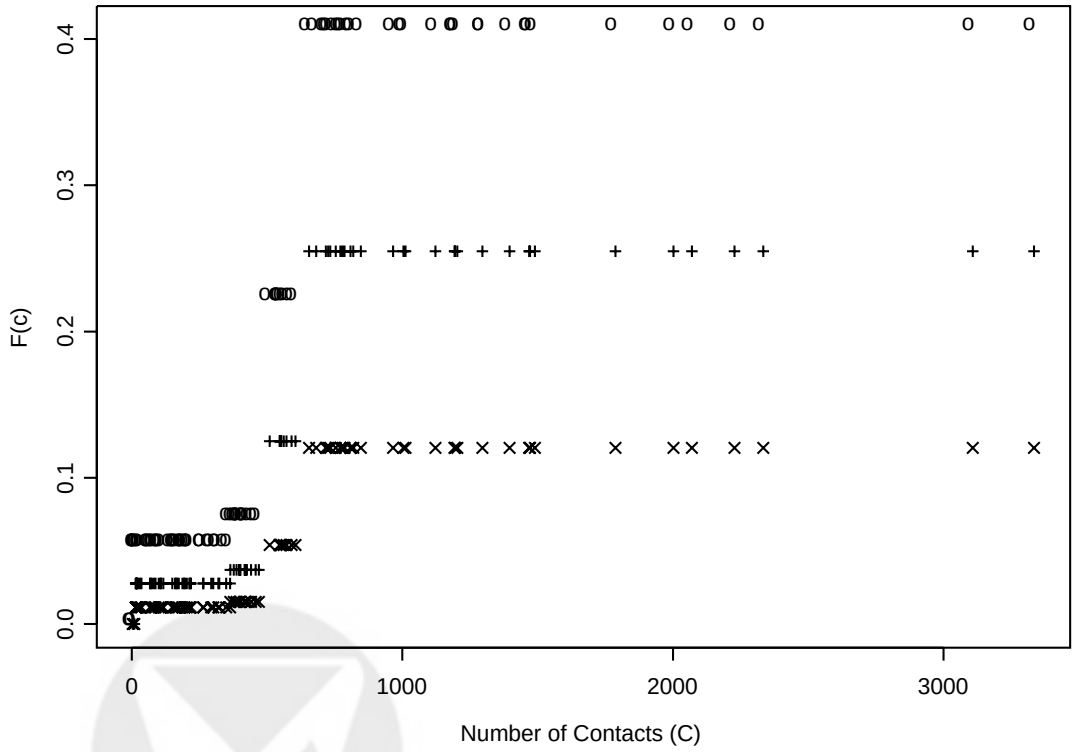

Figure 2: Estimates of the Log Odds Function Associated With the Distribution of Number of Contacts Between Infection of the Male Index Case and the Female Partner, Assuming $\mu_1/\mu_0 =$ (i) 1, (o) (ii) 2 $(+)$, or (iii) 5 (x)