FIXED-EFFECTS METHODS FOR THE ANALYSIS OF NONREPEATED EVENTS

*Paul D. Allison**
Nicholas A. Christakis†

For repeated events, fixed-effects regression methods—which control for all stable covariates—can be implemented by doing Cox regression with stratification on individuals. For nonrepeated events, we consider the use of conditional logistic regression to estimate fixed-effects models with discrete-time data. Known in the epidemiological literature as the case-crossover design, this method fails when any covariate is a monotonic function of time. Hence, no control for time itself can be included, leading to potentially spurious estimates. As an alternative, we consider the case-time-control method for estimating the effect of a dichotomous predictor. This method allows for the introduction of a control for time by reversing the role of the dependent and independent variables. In contrast to earlier work, we show that the method can be applied to data that contain only uncensored cases, and that it is possible to control for additional covariates, both categorical and quantitative. Simulation studies indicate that the case-time-control method is substantially superior to the case-crossover method and to conventional logistic regression. The methods are illustrated by estimating the effect of a wife's death on the hazard of death for her husband.

Direct correspondence to Paul D. Allison at Sociology Department, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104 (e-mail: allison@soc.upenn.edu).

*University of Pennsylvania
†Harvard University

1. INTRODUCTION

Fixed-effects methods have become increasingly popular in the analysis of longitudinal data for one compelling reason: They make it possible to control for *all* stable characteristics of the individual, even if those characteristics cannot be measured (Halaby 2004; Allison 2005). Using widely available software, fixed-effects methods can be applied to linear models (Greene 1990), logistic regression models (Chamberlain 1980), and Poisson regression models (Cameron and Trivedi 1998). For event-history analysis, a fixed-effects version of Cox regression (partial likelihood) is available for data with repeated events for each individual (Chamberlain 1985; Yamaguchi 1986; Allison 1996). But fixed-effects Cox regression is not feasible when no more than one event is observed for each individual.

In this paper, we explore fixed-effects methods for nonrepeated events using conditional logistic regression with discrete-time data. There are several peculiar features of nonrepeated event data that make a conventional fixed-effects approach problematic. As we shall see, none of the available methods works well for covariates that change monotonically with time (unless they are transformed into nonmonotonic functions). For covariates that are not monotonic with time, one approach works well when those covariates are uncorrelated with time but may be badly biased otherwise. Another method works well for covariates that are correlated with time, but only when the covariate is dichotomous, a situation that may still find many applications.

2. AN EXAMPLE

To make things concrete, we shall consider these issues in the context of an empirical example. Consider the following question: Does the death of a wife increase the hazard for the death of her husband? We have data on 49,990 married couples in which both spouses were alive and at least 68 years old on January 1, 1993.[1] Death dates for both spouses are

[1]To assemble a population-based sample of elderly couples, we linked Medicare claims data and other files at an individual level (using individual identifiers). We began with the 1993 Denominator File which includes 32,180,588 people 65 years of age or older. Based on Census data, we estimate that 13.2 million of these

available through May 30, 1994. During that 17-month interval, there were 5769 deaths of the husband and 1918 deaths of the wife.

Given data such as these, how can we answer our question? One straightforward approach is to do a Cox regression for husband's death with wife's vital status as a time-varying covariate. More specifically, let t_i be the husband's time of death for couple i, in days since the origin (January 1, 1993). If a death is not observed, then t_i is the censoring time (515 days). Let $W_i(t)$ be a time-varying covariate coded 1 if the wife is alive at time t and 0 otherwise. We postulate a proportional hazards model

$$\log h_i(t) = \alpha(t) + \beta W_i(t) + \delta X_i \tag{1}$$

where $h_i(t)$ is the hazard for husband's death at time t for couple i, $\alpha(t)$ is an unspecified function of time, and X_i is a vector of time-invariant covariates for couple i. This model may be estimated with standard partial likelihood software.

Estimates for one such model are shown in the first two columns of Table 1. "Black" is a dummy variable coded 1 for black race, otherwise 0. "Age" is the age in years on January 1, 1993. "Illness burden" is a scale based on medical records for the three years prior to the start of observation, with observed values ranging from 0 to 15. We see that the hazard of death for blacks is approximately 7 percent higher than for other races, but the effect is not statistically significant. On the other

people were in marriages where both spouses were 65 or older. From this file, we identified husband/wife pairs using a method described by Iwashyna et al. (1998, 2000). The method exploits Medicare's complex system of identification codes to find spousal pairs, and it has a sensitivity of up to 80 percent. While representing a majority of married people, these couples are somewhat more likely to be those in which the husband had been employed and the wife had either never earned money or earned less than her husband. However, in the current generation of elderly, this is the modal pattern. The application of this method resulted in the identification of 4,313,221 couples, 65 percent of the total population. Of this group, 3,247,729 are couples in which both members were older than 68. Out of this group, we took a simple random sample of 50,000. We subsequently deleted ten cases due to data inconsistencies, leaving 49,990 for analysis. For these couples, we have detailed hospitalization information for three years prior to 1993 and mortality and hospitalization information for both members of each couple until mid-1994. Using established methods of quantifying illness burden, we assigned each individual a morbidity burden based on their medical records for the three years preceding cohort inception (Zhang et al. 1999).

TABLE 1
Cox Regression Estimates for Models Predicting the Hazard of Husband's Death

Covariate	Hazard Ratio	*p*-Value	Hazard Ratio	*p*-Value
Black	1.07	.22	1.07	.23
Age	1.08	<.0001	1.08	<.0001
Illness burden	1.35	<.0001	1.35	<.0001
Wife dead	1.02	.86	—	—
Wife died within 30 days	—	—	1.47	.07

hand, there is a highly significant coefficient for age, with each year of age being associated with an 8 percent increase in the hazard. Each 1-point increase in illness burden is associated with a 35 percent increase in the hazard. There is, however, no evidence for an effect of wife's death on husband's hazard of death.

One possible explanation for the null effect of wife's death is that any such effect may last for only a limited period of time. To investigate this possibility, we estimated a second model in which the time-dependent covariate for wife's vital status was coded 1 if the wife had died within the previous 30 days, otherwise 0. Results in the last two columns of Table 1 offer modest support for this hypothesis. The hazard for husband's death is about 47 percent higher during the 30-day period after wife's death, with a *p*-value of .07.

Would we be justified in interpreting the hazard ratio for wife's death within 30 days as representing a causal relationship? An obvious objection is that these models omit many variables that are common to husbands and wives, or at least highly correlated, and which also have an impact on mortality. Possibilities include income, education, dietary habits, exercise patterns, smoking behavior, and drinking behavior. The omission of these variables could produce a spurious relationship between wife's death and husband's death. So it would be desirable to find a way to reduce or eliminate such biases. Putting additional appropriate variables into the model would be helpful, but such variables are not always available.

3. THE CASE-CROSSOVER METHOD

In the absence of additional measured control variables, we consider a fixed-effects approach in which each couple is compared with itself at different points in time, thereby controlling for all time-invariant

variables. One way of doing this is the case-crossover method, which has been widely used in the epidemiological literature (Maclure 1991; Marshall and Jackson 1993; Greenland 1996). According to the basic form of the case-crossover design, we must choose a sample of individuals who have experienced events and record the values of their covariates at the time of the event. We then choose some previous point in time when the event did not occur (a "control" period), and record the values of the covariates for the same individuals at that time. The data are analyzed by doing a matched-pair conditional logistic regression predicting whether or not the event occurred. A critical issue is how to choose the "control period" in order to minimize bias. More complicated forms of the design involve drawing more than one control period for each event. Although this can improve statistical efficiency, it is unclear how to do this in an optimal fashion (Mittleman, Maclure, and Robins 1995).

For our mortality data, we extend the case-crossover design by using information from *all* observed periods prior to the husband's death. Taking a discrete-time approach (Allison 1982), we treat each day as a distinct unit of analysis. Suppose that a husband died on day 78. We then ask the question: Given that he died, why did he die on this day and not on one of the preceding 77 days? Was there something different about those days compared with the day on which he died? As in the usual case-crossover design, we answer this question by way of conditional logistic regression.

Let p_{it} be the probability that the husband in couple i dies on day t, given that he is still alive at the beginning of that day. Let W_{it} be an indicator of wife's vital status on day t. For example, we could let W_{it} be 1 if the wife was dead on day t, otherwise 0. Alternatively, we could let W_{it} be 1 if her death occurred within, say, 60 days prior to day t, otherwise 0. We postulate the following logistic regression model

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \alpha_i + \gamma_t + \beta W_{it} \tag{2}$$

where γ_t represents an unspecified dependence on time and α_i represents the effects of all unmeasured variables that are specific to each couple but constant over time. Note that no time-invariant covariates are included in the model as their effects are absorbed into the α_i term.

We estimate the model by conditional maximum likelihood, thereby eliminating the α_is from the estimating equations. The mechanics are as follows. For couples in which the husband died, a separate

observation is created for each day that he is observed, from the origin until the day of death. For each day, the dependent variable Y_{it} is coded 0 if the husband remains alive on that day, and coded 1 if the husband died on that day. Thus, a man who died on June 1, 1993, would contribute 152 person-days; 151 of those would have a value of 0 on Y_{it}, while the last would have a value of 1. The wife's vital status is coded 1 if she was dead on the given day, otherwise 0. For a different representation of wife's vital status, the variable is coded 1 if her death occurred within, say, 60 days prior to the given day, otherwise 0.

All couples in which the husband did not die are effectively deleted from the sample. If the husband is alive on every day of observation, there is no within-couple variation on the dependent variable, and hence no information is contributed to the likelihood function. After deleting couples with no husband deaths, the likelihood function has the following form:

$$L = \prod_i \left(\frac{\exp(\gamma_T + \beta W_{iT})}{\sum_{t=1}^{T} \exp(\gamma_t + \beta W_{it})} \right) \tag{3}$$

In this equation, i runs over all couples whose husband died, and T represents the final day of observation—that is, the day on which the husband died. Notice that α_i has been factored out of likelihood.

Most comprehensive statistical packages have routines to maximize this likelihood, usually under the name "conditional logistic regression." The likelihood function is also identical in form to the stratified partial likelihood for a Cox proportional hazards model. Hence, the model may be estimated by any Cox regression program that allows for stratification.

With a separate parameter for every day of observation, the model in equation (2) is rather complex for estimation. We thus consider only models which impose some restrictions on γ_t. We begin by setting $\gamma_t = 0$—that is, no variation over time in the likelihood of a death. Because the observation period covers only 17 months, this is not an unreasonable assumption.

It so happens that couples who have no variation on the covariates over time can also be deleted from the sample because they contribute

TABLE 2
Cross-Classification of Husband Dead by Wife Dead, 39,942 Couple-Days

	Wife Alive	Wife Dead
Husband Dead	0	126
Husband Alive	19344	20472

nothing to the likelihood.[2] In our case of a single dichotomous covariate (wife's death), we delete any couple whose wife did not die before the husband. Of the 5769 couples in which the husband died, there were only 126 cases in which the wife's death preceded the husband's in this 17-month interval. So our usable set of couples declines from 49,990 to 126, a rather drastic reduction by any standard. These 126 couples contributed a total of 39,942 couple-days.

4. RESULTS FOR COUPLE MORTALITY DATA

We first attempted to estimate a conditional logistic regression model in which W_{it} was coded 1 for wife dead on day t, otherwise 0. However, this model did not converge. The reason is quasi-complete separation, which can be seen in Table 2. If the husband is dead (on the final day of the sequence), the wife is necessarily dead, yielding a 0 frequency count in one cell of the contingency table. (Remember that conditional likelihood necessarily restricts the sample to couples where the husband dies and the wife dies before the husband.) This will also be true in every couple-specific subtable. As is well known, the log-odds ratio for a 2×2 table is not defined when there is a zero in the any of the cells.

In general, quasi-complete separation arises whenever the time-varying covariate can only change monotonically with time. In our case, the dummy variable for wife dead can change from 0 to 1 over time but stays at 1 until the end of the series. The problem does not occur, however, if we estimate a model in which the covariate is an indicator of whether the wife died within, say, the previous 60 days. This covariate increases from 0 to 1 when the wife dies, but then goes back to 0 after 60 days (if the husband is still alive). Estimating the model with varying windows

[2]When $\gamma_t = 0$ and W_{it} constant for all t, the expression in parentheses in equation (3) is identically equal to 1.

TABLE 3
Odds Ratios for Predicting Husband's Death from Wife's Death Within Varying Intervals of Time, Case-Crossover Method

	Wife Died Within				
	15 Days	30 Days	60 Days	90 Days	120 Days
Odds-ratio	1.26	1.96	1.61	1.27	1.26
p-value	.54	.006	.03	.24	.25

of time can give useful information about how the effect of wife's death starts, peaks, and stops.

Table 3 gives estimated odds ratios for several different intervals of time, using conditional logistic regression. In all cases, the odds ratios exceed 1.0 and are statistically significant for the 60-day interval and the 30-day interval. For the latter, the odds of husband's death on a day in which the wife died during the previous 30 days is about double the odds if the wife did not die during that interval. It is worth keeping in mind, however, that in this data set there were only 22 couples in which the husband died within 30 days after the wife's death.

A major limitation of these analyses is that they assume no dependence on time itself, that is, $\gamma_t = 0$. Unfortunately, it has been shown that case-crossover designs can be extremely sensitive to violations of this assumption (Suissa 1995; Greenland 1996). For our example, if there is *any* tendency for the incidence of wife death to increase over the period of observation, this can produce a spurious relationship between wife's death (however coded) and husband's death. Intuitively, the reason is that husband's death always occurs at the end of the sequence of observations for each couple so any variable that tends to increase over time will appear to increase the hazard of husband's death.

Fortunately, there is little evidence for such a trend in these data. Going back to the original data set of 49,990 couples, a Weibull model for *wife's* death shows that the hazard of a death actually declines slightly with time. Similarly, in our sample of 39,942 person-days (from 126 couples) the correlation between wife's death within 30 days and time since the origin was $-.04$. So we seem to be in good shape for this analysis.

But what if there *were* a correlation between time and wife's death? How could the model be adapted to adjust for time dependence? A natural approach is to relax the assumption that $\gamma_t = 0$ and include

some function of time in the model. Unfortunately, this strategy will not generally work for this kind of data. If the covariates include any monotonic function of time with coefficients to be estimated, the maximum likelihood estimates for those coefficients do not exist and the model will not converge. Again, the problem is that any covariate that may increase with time but never decrease (or that may decrease but never increase) will be a "perfect" predictor of husband's death because a death always occurs at the last point in time.

It is possible, however, to include nonmonotonic functions of time. For example, to allow for cyclic annual variation in the hazard of husband's death, we fit a conditional logistic regression model with three covariates: wife death within 30 days, $\sin(2\pi t/365)$, and $\cos(2\pi t/365)$ where t is the number of days since the origin. All three covariates were highly significant, and the odds ratio for wife's death remained at about 2.0.

While such a model provides useful information, it still does not solve our problem of needing to control for monotonic functions of time. As one possible solution, we estimated models with increasing functions of time in which the coefficients of time were fixed rather than estimated. These models converged, and the estimated hazard ratios were similar to those in Table 3. Since the results could depend on the fixed values of the coefficients, we performed a sensitivity analysis in which the time coefficients were systematically varied over a range of plausible values. Although the empirical application seemed to work well, results of simulation studies (not shown) convinced us that this approach is not valid. In particular, the coefficient for wife's death was badly biased unless the coefficients for time were ridiculously large, and there was no apparent way to determine the correct values for the time coefficients.

5. THE CASE-TIME-CONTROL METHOD

We now consider an alternative fixed-effects method that appears to solve the problems that arise when the distribution of the covariate is not in fact stable over time. Introduced by Suissa (1995), who called it the "case-time-control" design, this approach uses the computational device of reversing the dependent and independent variables in the estimation of the conditional logistic regression model. This makes it possible to

introduce a control for time, something that cannot be done with the case-crossover method.

As is well known, when both the dependent and independent variables are dichotomous, the odds-ratio is symmetric—reversing the dependent and independent variables yields the same result, even when there are other covariates in the model.[3] In the case-time-control method, the working dependent variable is the dichotomous covariate—in our case whether or not the wife died during the preceding specified number of days. Independent variables are the dummy variable for the occurrence of the event (husband's death) on a given day and some appropriate representation of time—for example, a linear function. As in the case-crossover method, a conditional logistic regression is estimated with each couple treated as a separate stratum. Under this formulation, there is no problem including time as a covariate because the working dependent variable is not a monotonic function of time.

In Suissa's formulation of the case-time-control method, it is essential to include data from all individuals, both those who experienced the event and those who did not (the censored cases). However, his model was developed for data with only two points in time for each individual, an event period and a control period. In that scenario, the covariate effect and the time effect are perfectly confounded if the sample is restricted to those who experienced events. On the other hand, censored individuals provide information about the dependence of the covariate on time, information that is not confounded with the occurrence of the event.

By contrast, our data set (and presumably many others) has multiple "controls" at different points in time for each individual. That eliminates the complete confounding of time with the occurrence of the event (husband death), making it possible to apply the case-time-control method to uncensored cases only. That is a real advantage in situations where it is difficult or impossible to get information for those who did not experience the event. The only restriction is that when the model is estimated without the censored cases, we cannot estimate a model with a completely unrestricted dependence on time—that is, with dummy variables for every point in time.

[3]This symmetry is exact when the model is "saturated" in the control covariates but only approximate for unsaturated models (Breslow and Powers 1978).

Of course, if the censored cases are available (as in our data set), it may be possible to get more precise estimates by including them. But even if censored cases are available, there is a potential advantage to limiting the analysis to those who experienced the event. Suissa's version of the case-time-control method has been criticized for assuming that the dependence of the covariate on time is the same among those who did and did not experience the event (Greenland 1996). This criticism has no force if the data are limited to those with events.

For our mortality data, the working data set can be constructed as before, with one record for each day of observation, from the origin until the time of husband's death or censoring. Unlike the case-crossover analysis, we now include both censored cases (couples in which the husband did not die) and uncensored cases. However, because conditional logistic regression requires variation on the dependent variable for each conditioning stratum, we can eliminate couples whose wife did not die before the husband, with no loss of information. This restriction gives us 1743 couples who contributed a total of 872,697 couple-days.

We estimated the following model. Let H_{it} be a dummy variable for the death of husband i on day t, and let P_{it} be the probability that wife's death occurred within a specified number of days prior to day t. Our working logistic regression model is

$$\log\left(\frac{P_{it}}{1-P_{it}}\right) = \alpha_i + \beta H_{it} + \gamma t. \tag{4}$$

Again, we estimate the model by conditional logistic regression with each couple as a stratum.[4]

Table 4 gives estimates for the 1743 couples in which the wife died, and for the more restricted sample of 126 couples in which both the husband died and the wife died before the husband. The estimates and p-values for the two subsamples are very close and also quite similar

[4]We used SAS PROC LOGISTIC with the STRATA statement. Estimation of the conditional logistic regression could also be done by way of a Cox regression program, but that would be more complicated in the case-time-control method than in the case-crossover method because a couple may have more than one day on which wife had died within the preceding specified number of days. Consequently, a conventional Cox partial likelihood is not appropriate. However, a Cox regression program that can estimate a discrete model for tied data (available in SAS or Stata) can produce the correct likelihood function.

TABLE 4
Odds Ratios for Predicting Husband's Death from Wife's Death Within Varying Intervals of Time, Case-Time-Control Method

		Wife Died Within				
		15 Days	30 Days	60 Days	90 Days	120 Days
Wife died (1743 couples)	Odds-ratio	1.36	2.03	1.51	1.09	.97
	p-value	.41	.004	.05	.69	.88
Both died (126 couples)	Odds-ratio	1.30	1.99	1.50	1.01	.80
	p-value	.48	.005	.06	.95	.27

to those in Table 3 for the case-crossover method. Again, the evidence suggests that the effects of wife's death are limited in time, with considerable fading after about two months. The standard errors (not shown) are virtually identical for the sample of 1743 and the sample of 126, so little was gained here by including the censored cases.

Although our working dependent variable is wife's death, the odds ratios should be interpreted as the effect of wife's death on the odds of husband's death. That is because of the time ordering of the observations—wife's death always precedes husband's death. If our goal was to estimate the effect of husband's death on wife's mortality, we would have to construct a different data set that would sample couple-days prior to the wife's death, but not thereafter.

5. SIMULATION RESULTS

Although the case-time-control method seems like a promising approach for fixed-effects analysis, the method has seen only a few applications in the epidemiological literature and is still considered somewhat experimental (Greenland 1996; Schneeweiss et al. 1997; Suissa 1998; Greenland 1999; Donnan and Wang 2001; Hernandez-Diaz et al. 2003; Schneider et al. 2005). To verify the appropriateness of this method for the kind of data considered here, we undertook a simulation study that investigated the performance of the estimators under several scenarios. For each scenario, we constructed 100 samples, each with 500 "couples" who were followed for a maximum of 20 "months." At each month, the husband could die or not die, with a probability determined by a logistic

regression equation. Also at each month, a "treatment" variable could take on a value of 1 or 0, again with probability determined by a logistic regression equation.

Model 1. We first tested to see whether the case-time-control method avoids the key flaw of the case-crossover method: a tendency to detect nonexistent effects when the treatment is correlated with time. The model used to generate the data had the form

$$\text{Logit}[\Pr(H_{it} = 1)] = -4 + .10t + .50u_i$$
$$\text{Logit}[\Pr(T_{it} = 1)] = -1 + .10t + .50u_i,$$

where H_{it} is a dummy variable for husband's death in couple i at time t, T_{it} is a dummy variable for treatment for couple i at time t, and u_i is a random draw from a standard normal distribution that is specific to couple i but which does not vary over time. Thus, the model does not allow for an effect of treatment on death but does assume substantial effects of time on both treatment and death (approximately a 10 percent increase in the odds at each succeeding month). Furthermore, there is substantial unmeasured heterogeneity (u_i) that is common to both death and treatment. Application of this model produced samples that averaged 6868 couple-months and 323 husband deaths. The treatment dummy was equal to 1 in 45 percent of the couple-months.

Table 5 shows the results for three different estimation methods. For each method, the table gives the true parameter value (for the effect of treatment on the log-odds of death), the mean of the 100 parameter estimates, the mean of the 100 estimated standard errors, the standard deviation of the 100 parameter estimates (which, ideally, should be the same as the mean of the standard errors), and the proportion of nominal 95 percent confidence intervals that actually include the true value ("coverage"). The case-time-control method performed about as well as could be hoped for—the mean parameter estimate is close to 0, the two estimates of the standard error are identical, and 94 percent of the nominal 95 percent confidence intervals contain the parameter value.

By contrast, the case-crossover method did poorly. The average coefficient estimate was .549 (corresponding to an odds ratio of 1.7) and only 1 percent of the confidence intervals included the true value.

TABLE 5
Estimates from Simulated Data Using Three Methods

Model	Method[a]	Parameter[b]	Estimate[c]	Average SE[d]	Standard Deviation[e]	Coverage[f]
1	CTC	.00	−.025	.132	.132	.94
1	CC	.00	.549	.126	.125	.01
1	LR	.00	.192	.118	.116	.67
2	CTC	.69	.721	.168	.168	.96
2	CC	.69	1.311	.164	.160	.00
2	LR	.69	.930	.151	.166	.48
3	CTC	.69	.687	.166	.165	.96
3	CTC(-X)	.69	1.018	.163	.166	.47
3	CC	.69	1.253	.163	.159	.06
3	LR	.69	.918	.150	.145	.49

[a]CTC = case-time-control, CC = case-crossover, LR = conventional logistic regression, CTC(-X) = case-time-control without covariate X.
[b]True value of the coefficient in the model producing the data.
[c]Mean of 100 parameter estimates.
[d]Mean of 100 standard error estimates.
[e]Standard deviation of 100 parameter estimates.
[f]Percentage of nominal 95 percent confidence intervals that include the true value.

Conventional logistic regression did a little better but was still unsatisfactory. The average coefficient estimate was .192, and only 67 percent of the confidence intervals contained the true value.

In other variations of this model (not shown), we set the coefficient for t to 0 in either the first or second equation. The case-time-control method performed well in either variation. As expected, the case-crossover method did well when there was no effect of time on treatment, but not otherwise.

Model 2. The second model modified the equation for death to allow for a nonzero effect of treatment. The equation for T was the same as before. The equation for H was

$$\text{Logit}[\Pr(H_{it} = 1)] = -3.5 + .10t + .69T_{it} + .50u_i.$$

The coefficient of .69 corresponds to an odds ratio of 2.0. This model produced samples that averaged 7987 couple-months and 217 husband deaths.

Again, as seen in Table 5, the case-time-control method does well, with a mean coefficient estimate of .721 (corresponding to an odds ratio

of 2.06), with 96 percent of the confidence intervals containing the true value. By contrast, the case-crossover method greatly overestimates the coefficient, and not a single confidence interval contains the true value. As before, conventional logistic regression gives intermediate results with a 30 percent overestimate of the coefficient and nominal 95 percent confidence intervals that contain the true value in only 48 percent of the samples.

Model 3. To our knowledge, the case-time-control method has never been considered as a method to control for other time-varying covariates. Model 3 introduces a covariate that varies with time and affects both treatment and death. The equations are

$$\text{Logit}[\Pr(H_{it} = 1)] = -3 + .10t + .69T_{it} + .8X_{it} + .50u_i$$
$$\text{Logit}[\Pr(T_{it} = 1)] = -1 + .10t + .5X_{it} + .50u_i.$$

Since X and T are correlated, we expect that omitting X from the estimated model will bias the estimated coefficient of T in the equation for husband's death. To control for X in the case-time-control method, we shall include it as a covariate in the conditional logistic regression predicting T. The model produced samples that averaged 7409 couple-months and 279 husband deaths.

As shown in Table 5, the case-time-control method does just as well here as with the previous scenarios. However, when the model is estimated without the covariate X, the parameter estimate is much too high (odds ratio of 2.77 rather than 2) and only 47 percent of the confidence intervals contained the true value. Even with the inclusion of X, the case-crossover method does poorly, with an odds ratio of 3.5 and only 6 percent coverage. As before, conventional logistic regression (with X included) produces estimates that are a bit too high and coverage of only 49 percent.

6. DISCUSSION AND CONCLUSION

Fundamental problems can arise when attempting to apply fixed-effects logistic regression to discrete-time event history data with nonrepeated events (the case-crossover method). In particular, the conditional likelihood estimates will not converge if the model includes any covariate

that is a monotonic function of time. This includes linear, polynomial, or logarithmic functions of time itself, as well as any covariate, such as a dummy for spouse alive or dead, that can only change in one direction with time. Since time dependence cannot be controlled, the method can also produce highly spurious estimates of the effects of any covariates that happen to be correlated with time. Of course conventional Cox models could still be estimated, but that would lose the advantage of the fixed-effects approach.

The case-time-control method provides a solution to the inability to control for time. This method also relies on conditional logistic regression, but reverses the role of the dichotomous event and a dichotomous covariate. Simulations suggest that the case-time-control method produces approximately unbiased estimates of the odds ratio of interest, even in cases where both the event hazard and the dichotomous covariate are strongly dependent on time. We have extended this method in two ways. First, we argue that the inclusion of individuals who did not experience events—previously thought to be an essential feature of this method—is unnecessary if multiple control times are available for those who do experience events and the dependence on time is not left unrestricted. Second, our simulation results suggest that additional time-varying covariates can be included as controls in the regression model.

Application of both the case-crossover method and the case-time-control method to mortality data of elderly couples provides evidence that there is indeed an effect of wife's death on husband's odds of death, even when all stable covariates are controlled, but that the effect is of limited duration.

At this point, the case-time-control method is still restricted to situations in which the aim is to estimate the effect of a dichotomous covariate on an outcome event, while controlling for other covariates, either dichotomous or continuous. In principle, one ought to be able to estimate effects of multiple dichotomous covariates by estimating a separate model for each covariate as the "dependent" variable. It may also be possible to handle polytomous covariates by estimating a conditional multinomial logit model (although commercial software for estimating such models is not widely available at present). At this point, however, we are unable to use the case-time-control approach to estimate the effect of a continuous covariate. And there is little hope for estimating the effects of covariates that are monotonic with time. Still, as we saw

here, many such variables can be reformulated in ways that eliminate the monotonicity.

The methods described here would be appropriate for events like deaths or loss of virginity that are, in principle, not repeatable. They may also be appropriate for events like arrests or promotions which, although repeatable in principle, may be sufficiently rare that they are observed no more than once for any individual in the sample. However, in the case of rare but repeatable events, the case-time-control method should be necessary only if observation ceases at the occurrence of the first observed event. When observation continues and the covariates continue to be measured after the occurrence of the event, we can use a conventional conditional logistic regression predicting the event, with a control for time or any other covariate that increases monotonically with time. That is because, in that observational setting, the event does not always occur at the end of the sequence of observations, and hence there is no problem of quasi-complete separation.

REFERENCES

Allison, Paul D. 1982. "Discrete-Time Methods for the Analysis of Event Histories." Pp. 61–98 in *Sociological Methodology*, vol. 13, edited by Samuel Leinhardt. San Francisco: Jossey-Bass.

———. 1996. "Fixed Effects Partial Likelihood for Repeated Events." *Sociological Methods & Research* 25:207–22.

———. 2005. *Fixed Effects Regression Methods for Longitudinal Data Using SAS*. Cary, NC: SAS Institute.

Breslow, N., and W. Powers. 1978. "Are There Two Logistic Regressions for Retrospective Studies?" *Biometrics* 34:100–105.

Cameron, A. Colin, and Pravin K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge, England: Cambridge University Press.

Chamberlain, Gary A. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47:225–38.

———. 1985. "Heterogeneity, Omitted Variable Bias, and Duration Dependence." Pp. 3–38 in *Longitudinal Analysis of Labor Market Data*, edited by James J. Heckman and Burton Singer. Cambridge, England: Cambridge University Press.

Donnan, Peter T., and Jixian Wang. 2001. "The Case-Crossover and Case-Time-Control Designs in Pharmacoepidemiology." *Pharmacoepidemiology and Drug Safety* 10:259–62.

Greene, William T. 1990. *Econometric Analysis*. New York: Macmillan.

Greenland, Sander. 1996. "Confounding and Exposure Trends in Case-Crossover and Case-Time-Control Designs." *Epidemiology* 7:231–39.

———. 1999. "A Unified Approach to the Analysis of Case-Distribution (Case Only) Studies." *Statistics in Medicine* 18:1–15.

Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." *Annual Review of Sociology* 30:507–44.

Hernandez-Diaz, Sonia, Miguel A. Hernan, Katie Meyer, Martha M. Werler, and Allen A. Mitchell. 2003. "Case-Crossover and Case-Time-Control Designs in Birth Defects Epidemiology." *American Journal of Epidemiology* 158:385–91.

Iwashyna, T. J., J. Zhang, D. Lauderdale, and N. A. Christakis. 1998. "A Methodology for Identifying Married Couples in Medicare Data: Mortality, Morbidity, and Health Care Use Among the Married Elderly." *Demography* 35:413–19.

———. 2000. "Refinements of a Methodology for Detecting Married Couples in the Medicare Data." *Demography* 37(2):251–52.

Maclure, Malcolm. 1991. "The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events." *American Journal of Epidemiology* 133:144–53.

Marshall, Roger J., and Rodney J. Jackson. 1993. "Analysis of Case-Crossover Designs." *Statistics in Medicine* 12:2333–41.

Mittleman, Murray A., Malcolm Maclure, and James M. Robins. 1995. "Control Sampling Strategies for Case-Crossover Studies: An Assessment of Relative Efficiency." *American Journal of Epidemiology* 142:91–98.

Schneeweiss, Sebastian, Til Sturmer, and Malcom Maclure. 1997. "Case-Crossover and Case-Time-Control Designs as Alternatives in Pharmacoepidemiologic Research." *Pharmacoepidemiology and Drug Safety* 6 suppl. 3:S51–59.

Schneider, M. F., S. J. Gange, J. B. Margolick, R. Detels, J. S. Chmiel, C. Rinaldo, and H. K. Armenian. 2005. "Application of Case-Crossover and Case-Time-Control Study Designs in Analyses of Time-Varying Predictors of T-Cell Homeostasis Failure." *Annals of Epidemiology* 15:137–44.

Suissa, Samy. 1995. "The Case-Time-Control Design." *Epidemiology* 6:248–53.

———. 1998. "The Case-Time-Control Design: Further Assumptions and Conditions." *Epidemiology* 9:441–45.

Yamaguchi, Kazuo. 1986. "Alternative Approaches to Unobserved Heterogeneity in the Analysis of Repeatable Events." Pp. 213–49 in *Sociological Methodology*, vol. 16, edited by Nancy Brandon Tuma. Washington, DC: American Sociological Association.

Zhang, J., T. J. Iwashyna, and N. A. Christakis. 1999. "The Performance of Different Lookback Periods and Sources of Information for Charlson Comorbidity Adjustment in Medicare Claims." *Medical Care* 37:1128–39.