Fixed Effects Methods for the Analysis of Non-Repeated Events

Paul D. Allison and Nicholas Christakis*

February 2000

Fixed-effects methods have become increasingly popular in the analysis of longitudinal data for one powerful reason: they make it possible to control for *all* stable characteristics of the individual, even if those characteristics cannot be measured. Fixed-effects methods are now readily available for 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 in which repeated events are observed for each individual (Chamberlain 1985, Yamaguchi 1986, Allison 1996). But fixed-effects Cox regression is not feasible when each individual experiences no more than one event.

In this paper, we explore fixed-effects methods for non-repeated events using conditional logistic regression on discrete-time data. There are several peculiar features of non-repeated event data that make a conventional fixed-effects approach problematic. As we shall see, no method works well for covariates that change monotonically with time (unless they are transformed into non-monotonic functions) For covariates that are not monotonic with time, one approach works well when covariates are uncorrelated with time, but is 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 find many applications.

* Paul D. Allison is Professor of Sociology at the University of Pennsylvania. Nicholas Christakis is Associate Professor of Medicine and Sociology at the University of Chicago. All communications should go to Allison at 3718 Locust Walk, Philadelphia, PA 19104, 215-898-6717, allison@ssc.upenn.edu. The latest version of this paper may be found on the Web at www.ssc.upenn.edu/~allison.

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 available through May 30, 1994. During that 17-month interval, there were 5,769 deaths of the husband and 1,918 deaths of the wife.

Given data like 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

$$\log h_i(t) = \alpha(t) + \beta W_i(t) + \delta X_i \qquad (1)$$

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 fixed covariates for couple i. We then estimate the model using 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 at the origin. Illness burden is scale based on medical records for the three years prior to the start of observation. The observed values range 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 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 be only last for 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.

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 w/in 30 days	--	--	1.47	.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.

THE CASE-CROSSOVER METHOD

In the absence of additional measured control variables, let's 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 design, which has been recently developed in the epidemiological literature (Maclure 1991, Marshall and Jackson 1993, Greenland 1996). In its basic form, the case-crossover design says to choose a sample of individuals who have experienced events, and record the values of their covariates at the time of the event. 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. While 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's 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

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

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_i's from the estimating equations. 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 can be 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:

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

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.

This likelihood function is 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. (For our analyses we used the SAS procedure PHREG.)

With a separate parameter for every day of observation, the model in equation (2) is too general for estimation. So we 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 nothing to the likelihood. 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 5,769 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.

RESULTS FOR COUPLE MORTALITY DATA

We first attempted to estimate a model in which W_{it} was coded 1 for wife dead on day t otherwise 0. However, this model did not converge. The reason can be seen in Table 2. If the

husband is dead (on the final day of the sequence), the wife is necessarily dead and there is 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.

Table 2. Cross-Classification of Husband Dead by Wife Dead, 39992 Person-Days

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

In general, convergence problems arise 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 of time can give useful information about the 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 are about double the odds if

the wife did not die during that interval. It's 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.

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

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 39,992 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 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, however, possible to include non-monotonic 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 doesn't 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.

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, the key innovation in this approach is the computational device of reversing the dependent and independent variables in the estimation of the conditional logit 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.[2] 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 an event (husband's death) on a given day and some appropriate representation of time, for example, a linear function. Again 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 method, it is critically important to include data from all individuals, both those who experienced the event and those who are censored. 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's a real boon 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, one cannot estimate a model with a completely arbitrary dependence on time, that is, with dummy variables for every point in time.

Of course, if the censored cases are available (as in our data set), more precise estimates can be obtained 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. 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 and uncensored cases. Because conditional logistic regression requires variation on the dependent variable for each conditioning stratum, we can eliminate couples whose wife did not die, with no loss of information. To avoid an unwieldy number of observations, we took a systematic sample of couple-days. All couple-days on which the husband died are included. For the remainder, we sampled the first couple-day (January 1, 1993) and every 30^{th} day thereafter, yielding a working

sample of 31,755 couple days. If the data are restricted to couples for whom both the husband and the wife died, the effective sample is reduced to 2,649 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

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

Estimation of the conditional logistic regression is a bit more complicated in the case-time-control 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. One approach is to use a program explicitly designed for conditional logistic regression with *m*:*k* matching (like Stata's clogit command). Alternatively, equivalent results may be obtained with the SAS procedure PHREG with its DISCRETE option for estimating a logit model with tied data.

Table 4 gives estimates for the full sample, and also for the subsample in which husbands died. For the reduced sample, the results are quite similar to those in Table 3, which used the case-crossover method on an equivalent sample. For the full sample, the odds ratios are a bit larger and the *p*-values are noticeably smaller. These smaller *p*-values are primarily due to the larger odds ratios, not to reduced sampling variability. The standard errors for the subsample are only 10-15% larger than those in the full sample. Again, the evidence suggests that the effects of wife's death are limited in time, with considerable fading after about two months.

Although our working dependent variable is wife's death, the odds ratios must be interpreted as the effect of wife's death on the odds of husband's death. That's 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 completely different data set that would sample couple-days prior to the wife's death, but not thereafter.

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 (1918 couples)	Odds-Ratio	2.37	2.41	1.72	1.28	1.13
	p-value	.008	<.0001	.007	.20	.52
Both Husband and Wife Died (126 couples)	Odds-Ratio	2.07	2.05	1.56	1.12	.90
	p-value	.05	.006	.05	.59	.60

SIMULATION RESULTS

Although the case-time-control method seems like the most promising approach for fixed-effects analysis, the method has seen few applications and is still somewhat controversial (Greenland 1996, Schneeweiss et al. 1997, Suissa 1998, Greenland 1999). To verify the appropriateness of this method for the kind of data considered here, we undertook a simulation study which investigated the possibility of large-sample bias under several scenarios. For each scenario, we constructed a sample of 10,000 "couples" who were followed for a maximum of 20 "months". Since our aim is only to investigate bias and not sampling variation, a single large sample is sufficient for each scenario. At each month, the husband could die or not die, with a probability determined by a logistic regression equation. Also at each time 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 non-existent effects when the covariate is correlated with time. The model used to generate the data had the following form:

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

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

where H_{it} is a dummy variable for husband's death in couple i at time t, T_{it} is 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 136,728 couple-months with 6481 husband deaths. The treatment dummy was equal to 1 in 45 percent of the couple-months. Conventional logistic regression of death on treatment and time yielded an odds ratio for treatment of 1.20 ($p<.0001$). Conditional logistic regression using the case-crossover method produced an odds ratio of 1.704 ($p<.0001$). Only the case-time-control method gave an appropriate answer. The odds ratio was .951, with a 95 percent confidence interval of .897 to 1.108. The estimated coefficient for time was .10, exactly what the model specified.

In other variations of this model, we set the coefficient for t to 0 in either the first equation or the 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 non-zero 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 102,987 couple-months with 8,851 husband deaths. Conventional logistic regression of death on treatment and time yielded an odds ratio for treatment of 2.39 ($p<.0001$). Conditional logistic regression using the case-crossover method produced an odds ratio of 3.13 ($p<.0001$). The case-time-control method estimated the odds ratio at 1.94, with a 95 percent confidence interval of 1.84 to 2.05.

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 111,415 couple-months with 8349 husband deaths. When we applied the case-time-control method without a control for *X*, the estimated odds ratio relating *T* and *H* was 2.75, well above the expected 2.0. When the model included *X*, the odds ratio for the relation between *T* and was 1.97 with a 95 percent confidence interval of 1.86 to 2.09. The estimated coefficient for *X* was .48, close to the .5 in the equation for *T*.

DISCUSSION AND CONCLUSION

Fundamental problems can arise when attempting to apply fixed-effects logistic regression to discrete-time event history data with non-repeated events (the case-crossover method). In particular, the conditional likelihood estimates will not converge if any monotonic function of time is included as a covariate. This would include linear, polynomial or logarithmic functions of time. It would also include any covariate, such as a dummy for spouse alive or dead, which 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 approaches.

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 of 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 a crucial component of this method—is unnecessary if multiple control times are available for those who do experience events. 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 variables by estimating a conditional multinomial logit model. At this point, however, we are unable 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.

REFERENCES

Allison, Paul D. (1982) "Discrete-Time Methods for the Analysis of Event Histories." Pp. 61-98 in *Sociological Methodology 1982*, edited by Samuel Leinhardt. Jossey-Bass.

Allison, Paul D. (1996) "Fixed-Effects Partial Likelihood for Repeated Events." *Sociological Methods & Research* 25: 207-222.

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

Chamberlain, Gary A. (1980) "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47: 225-238.

Chamberlain, Gary A. (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 University Press.

Greene, William T. (1990) *Econometric Analysis*. Macmillan.

Greenland, Sander (1996) "Confounding and Exposure Trends in Case-Crossover and Case-Time Control Designs." *Epidemiology* 7: 231-239.

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

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: 413-419.

Iwashyna T. J., J. Zhang , D. Lauderdale and N. A. Christakis (2000) "The Detection of Married Couples in Data from the Health Care Financing Administration and the Social Security Administration: Evidence of Differences." *Demography*. In press.

Maclure, Malcolm (1991) "The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events." *American Journal of Epidemiology* 133: 144-153.

Marshall, Roger J. and Rodney J. Jackson (1993) "Analysis of Case-Crossover Designs." *Statistics in Medicine* 12: 2333-2341.

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-S59.

Suissa, Samy (1995) "The Case-Time-Control Design." *Epidemiology* 6:248-253.

Suissa, Samy (1998) "The Case-Time-Control Design: Further Assumptions and Conditions." *Epidemiology* 9: 441-445.

Yamaguchi, Kazuo (1986) "Alternative Approaches to Unobserved Heterogeneity in the Analysis of Repeatable Events." Pp. 213-49 in *Sociological Methodology 1986*, edited by Nancy Brandon Tuma. American Sociological Association.

Zhang J., T. J. Iwashyna, N.A. Christakis (1999) "The Performance of Different Lookback Periods and Sources of Information for Charlson Comorbidity Adjustment in Medicare Claims." *Medical Care* 37: 1128-1139.

NOTES

[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 people were in marriages where both spouses were 65 or older. From this file, we identified husband/wife pairs using a data 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%. 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% of the total population. Of these couples, 3,247,729 are ones in which both members were older than 68. From this group, we

took a simple random sample of 50,000. We subsequently deleted 10 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).

[2] This symmetry is exact when the model is "saturated" in the control covariates but only approximate for unsaturated models.