Modeling the Days of Our Lives: Using Survival Analysis When Designing and Analyzing Longitudinal Studies of Duration and the Timing of Events

Judith D. Singer and John B. Willett
Graduate School of Education, Harvard University

Psychologists studying whether and when events occur face unique design and analytic difficulties. The fundamental problem is how to handle censored observations, the people for whom the target event does not occur before data collection ends. The methods of survival analysis overcome these difficulties and allow researchers to describe patterns of occurrence, compare these patterns among groups, and build statistical models of the risk of occurrence over time. This article presents a unified description of survival analysis that focuses on 2 topics: study design and data analysis. In the process, we show how psychologists have used the methods during the past decade and identify new directions for future application. The presentation is based on our own experience with the methods in modeling employee turnover and examples drawn from research on mental health, addiction, social interaction, and the life course.

Psychologists often study whether and, if so, when specific events occur. Researchers investigating the course of affective illness, for example, examined the age at first onset (Rice et al., 1987), the length of initial illness spells (Zito, Craig, Wanderling, & Siegel, 1987), how long treated patients remain free of symptoms (Lavori, Keller, & Klerman, 1984), whether individuals with a history of affective disorders have another episode (Amenson & Lewinsohn, 1981), how long successfully treated individuals remain well before relapse (Prien et al., 1984), and how long second and subsequent illness spells last (Shapiro, Quitkin, & Fleiss, 1989). Similar questions about the timing of events arise in studies of the age of attainment of developmental milestones (first word, first step, first date, marriage, first child), relapse after the cessation of undesirable behaviors (criminal activity, smoking, drug use, alcohol abuse), and pauses and shifts in social interactions (gaze, attention, conversation).

Research questions about time pose unique design and analytic difficulties. No matter when data collection begins and no matter how long any subsequent follow-up period, some people may not experience the target event before data collection ends; some people may not develop an affective disorder, some people in therapy may not stabilize, and some of those stabilized may not relapse. Should the researcher assume that none of these people will ever experience the event? All the researcher knows is that by the end of data collection, usually an arbitrary point in time, the event had not yet occurred. Statisticians say that such observations are censored.

The prospect of censoring complicates research design; the presence of censoring complicates statistical analysis. Many researchers respond to these complications with ad hoc strategies, none entirely satisfactory. Some try design solutions by restricting data collection to uncensored observations (Taber & Proch, 1987). Others try analytic solutions by categorizing the outcome and placing the censored observations in a single group (Condiotte & Lichtenstein, 1981), deleting the censored observations (Litman, Eiser, & Taylor, 1979), and using the censored outcome as a categorical predictor of another outcome that varies over time (Coelho, 1984). Appropriately shunning these strategies, others sidestep the "when" question entirely and ask only the "whether" question: Does the event occur by a particular point in time (Grey, Osborn, & Reznikoff, 1986) or by each of several successive points in time (Benfari & Eaker, 1984).

For years, psychologists recognized the severe limitations of these strategies, most notably their sensitivity to the length of data collection (e.g., Brownell, Marlatt, Lichtenstein, & Wilson, 1986; Furby, Weinrott, & Blackshaw, 1989; Hunt, Barnett, & Branch, 1971; McFall, 1978; Nathan & Lansky, 1978; Sutton, 1979; Wainer, 1977). Until recently, however, few alternatives were available. New developments in statistical theory accompanied by new developments in statistical computing have changed how researchers can study time. The new methods—known as survival analysis, event history analysis, or hazards modeling—were developed by biostatisticians modeling human lifetimes (Cox, 1972; Cox & Oakes, 1984; Kalbfleisch & Prentice, 1980; Miller, 1981) and have been extended by economists and sociologists studying social transitions (Allison, 1984; Blossfeld, Hamerle, & Mayer, 1989; Heckman & Singer, 1985; Tuma & Hannan, 1984). Differences in labels aside, these techniques use similar mathematical roots to address similar research goals: to help researchers simultaneously explore

The order of the authors was determined by randomization. We thank Richard J. Murnane for first raising the substantive questions that led to our study of survival analysis. Judith D. Singer also thanks the National Academy of Education for its support during her Spencer postdoctoral fellowship. We thank Victor J. Stevens for kindly providing the summary data used in Figure 1.

Correspondence concerning this article should be addressed to Judith D. Singer or John B. Willett, Harvard University, Graduate School of Education, Cambridge, Massachusetts 02138.

Psychological Bulletin, 1991, Vol. 110, No. 2, 268-290

whether events occur—do people develop an affective disorder, smoke a cigarette, interrupt a conversation—and if so, when. Using specific techniques within the broad class of methods, researchers can describe patterns of occurrence, compare these patterns among groups, and build statistical models of the risk of occurrence over time.

Owing to its genesis in modeling human lifetimes, where the target event is death, survival analysis is shrouded in dark, foreboding terms. But beyond the terminology lies a powerful methodology that appropriately uses data from all observations, uncensored and censored cases alike. Data collection can be prospective or retrospective, experimental or observational. Time can be measured continuously or discretely. The only requirements are that (a) at every time point of interest, each individual be classified into one of two or more mutually exclusive and exhaustive states, and (b) the researcher know, for at least some of these individuals, when the transition from one state to the next occurs.

Because the methods of survival analysis adapt easily to psychological phenomena, they now appear more often in the research literature. A search of the *Psychological Abstracts* and *Educational Resources Information Center* databases for articles published between 1980 and 1990,[1] for example, led to over 100 citations. Substantive fields allied with those in which the methods emerged (medicine, economics, and sociology) are at the forefront of application: research in mental health, organizational behavior, social psychology, and the life course. Several recent articles also described how the methods can be used to explore specific topics including social interaction (Allison & Liker, 1982; Gardner & Griffin, 1989; Griffin & Gardner, 1989), organizational behavior (Fichman, 1988; Morita, Lee, & Mowday, 1989), clinical trials (Greenhouse, Stangl, & Bromberg, 1989), and the life course (Johnson, 1988; Teachman, 1982).

Despite the growing use of survival analysis, a unified presentation of the methods written for research psychologists has yet to appear. In the present article, we begin to fill this void. After developing the basic concepts underlying survival analysis, we focus on two topics: study design and data analysis. For each, we outline issues researchers face and provide guidelines for making informed decisions about them. In the process, we review how psychologists used the methods to date, and identify new directions for future application. We base our presentation on our own experience with the methods in modeling teacher turnover (Murnane, Singer, & Willett, 1988, 1989; Murnane, Singer, Willett, Kemple, & Olsen, 1991; Singer, in press; Singer & Willett, 1988, in press; Willett & Singer, 1988, 1989, 1991, in press) and examples drawn from four areas of inquiry: (a) mental health research on the onset and duration of mental illness and relapse after treatment; (b) addiction research on the onset and duration of addictive behavior and relapse after treatment; (c) social interaction research on whether and when individuals interrupt a conversation, stop paying attention during class, or avert a gaze; and (d) life course research on the onset and duration of life events and developmental milestones.

Concepts Underlying Survival Analysis

The concepts underlying survival analysis differ markedly from the familiar means, standard deviations, and correlations of traditional statistics. We develop these concepts here using data reported by Stevens and Hollis (1989), who evaluated the efficacy of supplementing a smoking-cessation program with follow-up support sessions designed to help ex-smokers cope with abstinence. The researchers randomly assigned 587 adults who successfully completed a 4-day program to one of three conditions: (a) 3 weeks of coping-skills training; (b) 3 weeks of support sessions without skills training; or (c) no supplemental sessions. For 50 weeks after quitting, participants returned a biweekly postcard noting their smoking status. Defining abstinence as smoking no more than five cigarettes per month, Stevens and Hollis asked whether the follow-up support helped people remain abstinent and if it did not, when people were most likely to relapse.

Survivor Function

Survival analysis begins with the survivor function. When studying the efficacy of a smoking-cessation program, as in this example, the population survivor function represents the probability that a randomly selected ex-smoker will remain abstinent versus time. Researchers with a representative sample from a target population can compute the sample survivor function, which estimates the population probability that a randomly selected person will remain abstinent longer than each time assessed—in this example, 10 weeks, 20 weeks, and so on—until everyone relapses or data collection ends (whichever happens first).

The first panel of Figure 1 presents the sample survivor function for the 198 people in Stevens and Hollis's (1989) control group.[2] At the beginning of the study (the beginning of "time"), the survival probability is 1.00. As time passes and people relapse, the survivor function drops toward 0. In this study, 82% abstain ("survive") more than 4 weeks, 66% abstain more than 8 weeks, and 60% abstain more than 12 weeks. By 50 weeks, when data collection ends, 38% remain abstinent. These individuals have censored relapse times either because they never relapse or they do so after data collection ends. Because of censoring, sample survivor functions rarely reach 0. Allison (1984) and Kalbfleisch and Prentice (1980) presented formal definitions of the survivor function.

The sample survivor function provides information for answering the descriptive question: How many weeks pass before the average smoker relapses? When the sample survivor func-

[1] We used the following key words for the search: survival analysis, event history, hazard(s) model, proportional hazard(s), Cox('s) regression. For articles published between 1980 and 1983, we used three additional key words: recidivism rate(s), duration (near) longitudinal, duration (near) analysis.

[2] We estimated the sample survivor function in Figure 1 using data kindly provided by Victor J. Stevens (Stevens & Hollis, 1989) using the Kaplan-Meier product-limit method (Kalbfleisch & Prentice, 1980). We then smoothed the obtained discrete estimates using a spline function (after the recommendation of Miller, 1981). The same method was used to create all subsequent displays in the article. Our intentions were strictly pedagogic: We wished to use continuous-time survivor and hazard functions to introduce the concepts of survival analysis before discussing differences between continuous-time and discrete-time methods.

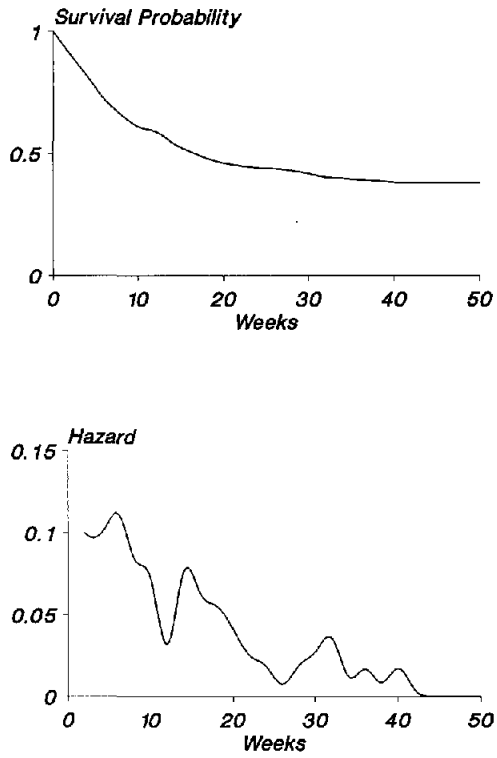

Figure 1. Sample survivor (top) and hazard (bottom) functions for 198 ex-smokers based on data reported by Stevens and Hollis (1989).

tion reaches .50, half the ex-smokers have relapsed and half have not. The estimated median lifetime identifies this midpoint, which indicates how much time passes before half the sample experiences the target event. As shown in Figure 1, among ex-smokers without follow-up support, the answer is 16 weeks. The statistic incorporates data from both the 123 uncensored individuals who relapsed within the 50 weeks of data collection and the 75 censored individuals who did not.

All survivor functions have a shape similar to that displayed in Figure 1: a negatively accelerating extinction curve, a monotonically nonincreasing function of time. This generalization was noted by Hunt and colleagues well before the advent of modern survival analysis (Hunt et al., 1971; Hunt & Bespalec, 1974a, 1974b; Hunt & General, 1973; Hunt & Matarazzo, 1970). After finding similar patterns in nearly 100 studies of smoking, heroin, and alcohol cessation, Hunt et al. (1971) presaged the utility of survivor functions, writing that they "hoped to use the differences in slope between individual curves as a differential criterion to evaluate various treatment techniques" (p. 455).

Hazard Function

If a large proportion of successful abstainers suddenly relapse in a given month, the survivor function drops sharply, as happens in Figure 1, during each of the first few months of the study. When the slope of the survivor function is increasingly negative, ex-smokers are at greater risk of relapse. Isolating time periods with steep slope changes is one way to identify risky time periods. But a better way to assess risk is to examine the hazard function, a related mathematical function that registers these changes in the slope of the (log) survivor function.

Mathematical definitions of hazard differ depending on whether time is measured discretely or continuously. If time is measured discretely, as in this example, hazard is the conditional probability that an ex-smoker will relapse in a particular time interval, given that the person has not relapsed by the beginning of the interval. As the interval length decreases, the probability that an event will occur during any given interval decreases as well. At the limit, when time is measured continuously, we must modify the definition of hazard because the probability that an event occurs at any instant approaches zero. So, in continuous time, hazard is the instantaneous rate of relapse, given uninterrupted abstinence until that time. Allison (1984) and Kalbfleicsh and Prentice (1980) provided formal definitions of hazard in both discrete and continuous time.

Like the survivor function, the hazard function can be plotted versus time, yielding a profile of the risk of relapsing each week, given uninterrupted abstinence until that week. The magnitude of each week's hazard indicates the risk of relapsing in that week: The higher the hazard, the greater the risk. Each interval's hazard is calculated using data on only those individuals still eligible to experience the event during the interval (the risk set); individuals who have already relapsed are not included.

The lower panel of Figure 1 contains the sample hazard function corresponding to the sample survivor function in the top panel. The risk of relapse is high in each of the first few weeks of the study and then declines over time. Ex-smokers are at greatest risk of relapse immediately after they quit; those who successfully abstain for several months are likely to abstain for at least a year.

Use of the hazard function in psychological research was proposed well before the arrival of modern survival methods, but because the associated statistical models were not yet available, much information in the function remained unexploited. For example, McFall (1978), Sutton (1979) and Litman et al. (1979) suggested that researchers examine relapse on a period-by-period basis, as the hazard function does, and identify who relapses and when. These authors appropriately dismissed the survivor function as too crude a summary because of its consistent shape regardless of the distribution of risk.

The hazard function, in contrast, effectively captures the distribution of risk across time. Figure 2 contains four hazard functions, each portraying a different risk profile. Because peaks indicate periods of elevated risk, they pinpoint when the target event is most likely to occur.

The hazard function in panel A is flat; risk is unrelated to time, and the event occurs at random. Because age, period, and cohort effects influence human behavior (Baltes & Nesselroade, 1972; Featherman & Lerner, 1985; Hogan, 1984; Schaie, 1965), flat hazard functions are uncommon in psychological research. Nevertheless, duration-independent behavior has been found in studies of time to marital breakdown after the birth of a child (Fergusson, Horwood, & Shannon, 1984), time to shifts in attention in the classroom (Felmlee & Eder, 1983; Felmlee, Eder, & Tsui, 1985), and time to changing a

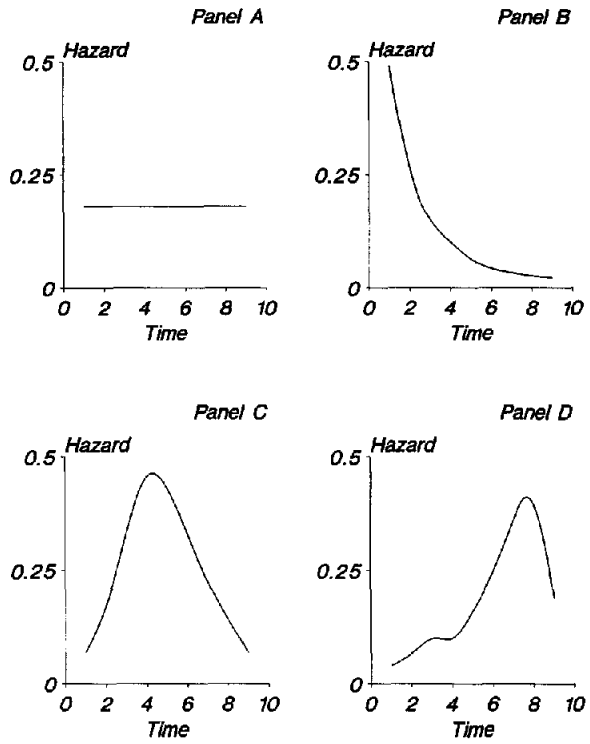

Figure 2. Four prototypical hazard functions: panel A, flat; panel B, early peak; panel C, middle peak; and panel D, late peak.

decision in the face of irrelevant information from a low-status partner (Hembroff & Myers, 1984).

Hazard functions with a definitive peak are more common. If the target event most likely occurs immediately after "the clock" starts, the hazard function peaks early (panel B). As shown in Figure 1, such hazard functions arise in addiction relapse studies. The early peak indicates the high risk of relapse immediately after cessation; the later level segment indicates the safe period when abstinent individuals rarely relapse. Hazard functions with early peaks have been found in studies of the relapse of mental illness (Lavori et al., 1984), recurrence of spousal violence (Berk & Sherman, 1985, 1988; Sherman & Berk, 1984), recidivism of sexual offenders (Soothill & Gibbens, 1978), marital dissolution (Morgan, Lye, & Condran, 1988; Tuma, Hannan, & Groeneveld, 1977), and employee turnover (Fichman, 1988; Murnane et al., 1989, 1991; Sorensen & Tuma, 1981).

Other hazard functions peak at intermediate (panel C) or late (panel D) points in time. If the risk of an event increases over time, the hazard function will peak later. Hazard functions with late peaks have been found in diverse substantive areas, including studies of human lifetimes (Gross & Clark, 1975), retirement (Campbell, Mutran, & Parker, 1987), and the time until the settlement of labor strikes (Kiefer, 1988).

Middle peaks arise primarily in studies with long data collection periods. This design dependency arises for a simple reason: In short studies, the particular time producing the middle peak appears to be late on the time axis simply because no data are collected afterward. As a result, retrospective studies, which easily cover extended periods of time, often reveal hazard functions with middle peaks. For example, when Diekmann and Mitter (1983) retrospectively asked young adults when they first shoplifted (if they ever did), the answers ranged from ages 4 to 16, with a peak during early adolescence—ages 12 to 14. Also, in a 12-month retrospective study of when Thai mothers stop exclusively breastfeeding their infants, Tognetti (1990) found a peak at 8 months.

Incidence and Prevalence: An Analogy for Hazard and Survival

Because hazard and survival functions may be unfamiliar concepts, we offer an epidemiological analogy to concepts that some readers may find more familiar: incidence and prevalence. Incidence measures the number of new events occurring during a time period (expressed as a proportion of the number of individuals at risk), whereas prevalence cumulates these risks to the total number of events that have occurred by a given time (also as a proportion; see, e.g., Kleinbaum, Kupper, & Morgenstern, 1982; Lilienfeld & Lilienfeld, 1980). Incidence and prevalence correspond directly to hazard and survival: Hazard represents incidence, and survival represents cumulative prevalence.

This analogy reinforces the importance of examining both the survivor and hazard functions. Epidemiologists have long recognized that although prevalence assesses the extent of a problem at a particular point in time, incidence is the key to disease etiology (Mausner & Bahn, 1974). Why? Because prevalence confounds incidence with duration. Conditions with longer durations may be more prevalent even if they have equal or lower incidence rates. To determine when people are at risk, epidemiologists study incidence, and when they study incidence, they are actually studying hazard.

Design: Collecting Survival Data

Survival analysis requires data summarizing the behavior of a sample of individuals over time. Data can be collected prospectively at several points in time or retrospectively at a single point in time, with probes that permit event-history reconstruction. The best studies tailor the time frame to the target event. When studying social interaction, for example, a series of 10-min segments might suffice, but when studying marital dissolution, even a 10-year window might not. In the following sections, we discuss nine issues that must be considered when the collection of survival data is being designed: deciding who to study, defining the possible states, identifying the beginning of time, selecting the length of data collection, choosing intervals for prospective data collection, reconstructing event histories in retrospective data collection, minimizing attrition, handling repeated events, and determining how many people to study. We discuss them in the order in which they usually arise during research planning.

Deciding Who to Study

As with any statistical method, the full advantages of survival analysis require a representative sample of individuals selected from an appropriate target population. Although data collected

from convenience samples can be used, probabilistic statements, population generalizations of sample summary statistics, and statistical inferences may be incorrect.[3] Because many psychologists using survival methods have worked with sociologists accustomed to stringent sampling, our literature search identified many articles whose findings were based on representative samples (e.g., Fergusson, Horwood, & Dimond, 1985; Teachman & Polonko, 1984; Yamaguchi & Kandel, 1987). We hope this standard will persist as the methods find their way into substantive areas unrelated to sociology.

A more problematic issue concerns the need to define carefully the target population from which the sample will be selected. Subtle variations in population definitions can inadvertently distort the distribution of time, the very quantity of interest. Consider the tempting strategy of eliminating censoring altogether by restricting the target population to only those individuals with known event times. When studying how long it took social service agencies in Alabama to discharge children from foster care, for example, Milner (1987) defined his target population as 222 children who were released from care during 1984 and 1985 (thus disregarding those who were not discharged). He then selected a sample of 75 children and found that of these 37% entered care within 5 months of discharge, 29% entered care within 6 to 11 months of discharge, 14% entered care within 12 to 24 months of discharge, and the remaining 20% entered care over 25 months before discharge.

The estimated median time to discharge in Milner's (1987) sample was 6 to 11 months. Should we conclude that the average child in foster care stays less than 1 year? Although Milner chose a probability sample from a well-defined target population, we do not know the answer to this question because his target population is unsuitable for answering it. Milner knew about discharge times only among children already discharged; he ignored those who remain in care. Children in foster care for long periods of time are likely to have been excluded from his study. Determining how long the average child stays in foster care requires a random sample of all children in foster care. Milner's sampling strategy leads to an underestimate of the average duration in the full population.

Some target-population definitions create more subtle biases. Especially problematic are retrospective studies of life-threatening events. A population of individuals still living obviously excludes those who have already succumbed. In their retrospective study of age at first suicide ideation among college students, for example, Bolger, Downey, Walker, and Steininger (1989) lamented the limited generalizability of their study because it necessarily excluded "those who have committed suicide or who are not attending college because of suicidal behavior" (p. 187).

When a sample excludes individuals who have already experienced the event of interest before data collection began, statisticians say that the sample is left truncated. Left truncation has received very little attention in the methodological literature, perhaps because the nature of the problem—the omission of any information—makes its difficult to evaluate the extent or impact of the truncation. As Hutchison (1988a, 1988b) noted, many methodologists ignore left truncation entirely or incorrectly fail to distinguish it from another methodological difficulty discussed later: left censoring. To avoid the complications arising from left truncation, we offer some design advice: Whenever possible, define the target population using delimiters unrelated to time, and if this is impossible, fully explore the potential biases created by whatever definition you use.

Defining the Possible States and the Target Event

At every time point of interest, each individual under study must occupy one, and only one, of two or more states. The states must be mutually exclusive (nonoverlapping) and exhaustive (of all possible states). Each individual is either ill or well, smoking or abstinent, breastfed or eating supplemental foods. The target event occurs when an individual moves from one state to another.

States must be defined precisely, with clear guidelines indicating the specific behaviors, responses, or scores constituting each state. Individual variation and measurement difficulties complicate this task because states may blend or have fuzzy boundaries. In their review of recidivism among sex offenders, for example, Furby et al. (1989) compared several definitions of recidivism: recommission of the same crime, of any sex crime, or of any crime. They also asked whether recidivism requires a conviction or whether commission of a crime is sufficient for classification.

Researchers developing guidelines for defining states can learn much from the addiction-relapse literature. Substance abuse—of alcohol, food, drugs, or tobacco—can be assessed in many ways using biochemical assays, clinical judgment, and self-reports to name a few. Whereas treatment programs may proffer the goal of total abstinence (e.g., Alcoholics Anonymous), researchers generally avoid such restrictive definitions because they lead to underestimates of the time to relapse. By defining relapse as a single sip of alcohol, more people will relapse sooner. Less limited definitions have the opposite effect: They bias estimates toward late relapse. Seeking the best of both worlds, Brownell et al. (1986) argued for using at least two definitions—lapse (a temporary slip that may or may not lead to relapse) and relapse—because different factors may predict the two events.

Variation in the definition of states may help explain the variation in relapse rates reported in the literature. In a prospective study of unaided smoking cessation, for example, Marlatt, Curry, and Gordon (1988) found that 1 month after quitting, 23% of the sample never actually quit (they smoked again within 24 h), 36% quit for at least 24 h but subsequently relapsed within the month, 16% were primarily abstinent but smoked one or two cigarettes, and only 25% were successfully abstinent. Consider the many different relapse rates that could be calculated from these figures: by setting aside individuals who never really quit, by pooling the primarily abstinent individuals with the relapsers, or by pooling them with the successfully abstinent individuals.

[3] When conducting a randomized experiment, of course, the decreased generalizability associated with convenience samples may be offset by the resultant increase in internal validity (Light, Singer, & Willett, 1990). Nevertheless, we strongly encourage researchers to select representative samples from well-defined target populations to increase the validity of their inferences.

We cannot review here all the measurement considerations necessary for deriving reliable and valid definitions of event states. Instead we offer more modest advice: Collect data with as much precision as possible so that you can appropriately code transitions from one state to the next.

Identifying the Beginning of Time

The problem of starting the clock is more complicated than it may appear. Because birth is both handy and meaningful, many researchers use it when studying developmental sequences and milestones. In a community survey of selected mental disorders, for example, Burke, Burke, Regier, and Rae (1990) appropriately used chronological age (time since birth) to examine when respondents first reported diagnoses.

But the simplest start time may not always be the best. Even developmental studies may appropriately count time from another beginning. When modeling outcomes among premature infants, for example, gestational age might be more appropriate. Time can also begin after related events occur or at a moment convenient to the researcher. This issue is more than academic. *Imprecise start times lead to imprecise event times.*

Some researchers start the clock when the individual experiences a precipitating event, such as leaving school (Marini, 1987), getting arrested (Zatz, 1985), getting pregnant (Yamaguchi & Kandel, 1987), having an abortion (Gibb & Millard, 1981), getting married (Teachman, 1982), having a child (Fergusson et al., 1984), taking a job (Mobley, Griffeth, Hand, & Meglino, 1979), or quitting smoking (Brownell et al., 1986). Use of these start times is appropriate because an individual is at risk of the target event only after experiencing the precipitating event. This is true even if the precipitating event only crudely approximates the beginning of time. Researchers modeling affective disorders, for example, typically use the date of psychological evaluation, diagnosis, or completion of therapy to start the clock, even though they might like to use the actual time of disease onset (Gallagher-Thompson, Hanley-Peterson, & Thompson, 1990; Simons, Murphy, Levine, & Wetzel, 1986; Zis & Goodwin, 1979).

Although the best start date is relevant to the research agenda, some researchers use an arbitrary start date if no other compelling argument can be marshalled. Researchers conducting experiments typically use the date of randomization (Peto, et al., 1976) or the date of intervention (Berk & Sherman, 1988; Greenhouse et al., 1989) for just this reason. When studying ongoing social interactions, the clock might as well start at a convenient moment, because there is little hope of identifying a substantively meaningful start time in a long-term continuing process. Using this argument, in their study of continuities and breaks in gazes between a husband and wife, Gardner and Griffin (1989) used a single arbitrarily selected 20-min segment of interaction in which the couple was asked to discuss a question about their marriage.

What happens if the start date is unknown for some individuals under study? Statisticians say that such observations are left censored (to distinguish them from right-censored observations in which the event times are unknown). Fichman (1989) encountered left censoring when studying absenteeism among 465 coal miners during a single calendar year. Fichman's interest was in the length of attendence spells, which begin when a coal miner returns to work after an absence and stop when the coal miner is absent another day. During the year, each coal miner generated several attendence spells, but the first spell was always left censored because the immediately prior absence occurred during the previous calendar year and was thus unknown to Fichman.

Methods for analyzing left-censored data remain in their infancy. Although Turnbull (1974, 1976) offered some basic descriptive approaches and Flinn and Heckman (1982) and Cox and Oakes (1984) offered some guidelines for developing models, most methodologists dismiss the problem soon after introducing the terminology (see, e.g., Blossfeld et al., 1989, p. 29; Tuma & Hannan, 1984, p. 135). The most common advice (Allison, 1984; Tuma & Hannan, 1984) is that researchers set the left-censored spells aside from analysis, the strategy eventually adopted by Fichman (1989) when faced with left-censored attendence spells.

We augment this analytic advice with related design advice: When possible, select a target population and a start time that eliminates or minimizes the possibility of left censoring. Track the employment durations of newly hired workers from their date of hire. Follow the breastfeeding patterns of infants immediately after they are born. The goal of these design strategies is to eliminate left censoring before it occurs, so that a researcher does not expend time and effort collecting data that will later be discarded or whose omission will distort the distribution of the very quantity of interest: time.

Selecting the Length of Data Collection

Once the clock starts, it must eventually stop. Clocks in retrospective studies stop on the date of interview; clocks in prospective studies can, in theory at least, continue indefinitely. As a practical matter, though, most prospective studies follow a sample for a finite, preselected period of time. The length of data collection determines the amount of right censoring (hereafter referred to as censoring). Because longer data collection periods yield fewer censored observations, the simple maxim is, the longer, the better. But beware: Longer data collection periods have their own disadvantages, including higher costs and increased attrition.

When deciding on the length of follow-up, remember that to determine when the event is likely to occur, it must actually occur for enough people under study. If the target event never occurs during data collection, all observations are censored. The researcher has little information, knowing only that it generally takes longer than this period for the event to occur.

There is no universally appropriate length of follow-up. The answer depends on the event under study. We try to make this decision by using information about the anticipated shape of hazard function and the probable median lifetime to apply a simple rule of thumb: The follow-up period should be long enough for at least half the sample to experience the target event during data collection. This ensures sufficient information for estimating a median lifetime and, as we show later in the section on determining sample size, it ensures reasonable statistical power. Using nonstatistical arguments, McFall (1978) suggested that smoking-relapse studies use a 6- to 12-month follow-

up. In our review of smoking-relapse studies published during the 1980s, we found that this guideline is widely accepted; the modal follow-up period was 1 year, and this period yielded an average censoring rate below 50%.

Events with legal ramifications may be easier to follow for longer periods of time. Hunt and Bespalec (1974b) made this point when they found that heroin-addiction studies had longer follow-up periods than smoking-cessation studies. Nathan and Lansky (1978) suggested that alcoholism- and drug-relapse studies use a 2-year follow-up. Prospective studies of recidivism among criminals often have even longer follow-ups (Blumstein & Cohen, 1987; Farrington, Ohlin, & Wilson, 1986; Schmidt & Witte, 1988).

Regardless of the length of follow-up, researchers must explicitly state its length. Relapse rates are meaningless unless linked to specific time periods. In a study of reinstitutionalization among mentally retarded adults, for example, Seltzer, Seltzer, and Sherwood (1982) reported that 65% of the subjects were not reinstitutionalized, but the researchers fail to specify in what time frame. Without this information, how can we know whether this percentage is low or high or know how to compare this rate to others found elsewhere? Even well-documented longitudinal studies using sophisticated analytic techniques occasionally omit this important piece of information (Zatz, 1985). The length of data collection is key to understanding the ultimate course of survival.

Choosing Intervals for Data Collection in Prospective Studies

Once the data collection window is set, the researcher must decide how often to collect follow-up data within this time period. Researchers studying social interactions can collect data in continuous time because they generally use videotapes of short duration that commence at randomly selected times (Bloxom, 1984, 1985; Drass, 1986; Felmlee & Eder, 1983; Gardner & Griffin, 1989; Hembroff & Myers, 1984).

Researchers studying other psychological phenomena generally observe a sample of individuals for longer periods of time; a further complication is that the clock generally starts at birth or after a precipitating event. Under these circumstances, logistical and financial constraints dictate data collection at discrete intervals. Although some researchers gather sufficient information using only one follow-up (with probes that permit retrospective reconstruction of event histories), we believe that systematic collection of data at regular intervals is far better. Even then, probes must be used to retrospectively reconstruct events transpiring between interviews.

The use of prespecified discrete data collection points adds measurement imprecision. If transitions occur in continuous time but data are collected in discrete time, a researcher will never know an individual's mental state at the crucial transition moment. This imprecision has serious consequences if information about this moment is key for predicting the timing of events, as in addiction relapse, where the coping skills of the ex-smoker, drinker, eater, or drug user may determine whether the person succumbs to temptation. Shiffman (1982) used an innovative design to overcome this restriction; he interviewed 183 ex-smokers who called a smoking-cessation hotline because they were in crisis. This design may be useful in other substantive areas requiring data describing the precise moment of transition.

Several strategies can be used to reconstruct event histories in prospective studies. Bounded-recall probes can help improve the quality of data describing behavior between interviews. At the beginning of the second and subsequent interviews, Neter and Waksberg (1964) suggested that interviewers ask about behavior only after reminding respondents of their responses during the previous interview. S. Cohen and Lichtenstein (1990) simply used multiple definitions of events. After each interview, they labeled individuals who said that they were not currently smoking and had not smoked even a puff during the last week as *point abstinent;* they labeled individuals who were point abstinent at all interviews up to the assessment point and had not smoked more than 3 days since quitting as *continuously abstinent.*

At what specific time points should limited data collection resources be targeted? Although the regular collection of data at equally spaced intervals is the most systematic approach, this strategy may omit information about the periods of greatest research interest. To maximize information about transitions from one state to the next, collect the most data when events are the most likely to occur.

We find it helpful to use the shape of the hazard function to inform this decision: Collect data more frequently when hazard is high and less frequently when hazard is low. This allocation strategy was used effectively by Hall, Rugg, Tunstall, and Jones (1984), who, in their 1-year prospective study of smoking cessation after behavioral skills training, placed their four data collection periods at 3, 6, 26, and 52 weeks after treatment. If they had spaced data collection episodes equally and waited until Week 13 to first collect follow-up data, they would have been unable to determine that the risk of relapse was highest in the few weeks immediately after cessation.

Reconstructing Event Histories in Retrospective Data Collection

In 1837, William Farr wrote, "Is your study to be retrospective or prospective? If the former, the replies will be general, vague, and I fear of little value" (cited in Lilienfeld & Lilienfeld, 1980). His words remain true today. Researchers studying the timing of events are well advised to collect data prospectively. But when studying infrequent events, prospective data collection may be infeasible. With few alternatives available, researchers choose the second-best approach: Interview people retrospectively and ask, "Has the event ever occurred? If so, when did it first occur?" Retrospective data collection has been used successfully by researchers studying age at first date (Dornbusch et al., 1981), first shoplift (Diekman & Mitter, 1983), first use of alcohol, tobacco, and drugs (Adler & Kandel, 1983), first suicide ideation (Bolger et al., 1989), and cessation of breastfeeding (Diamond, McDonald, & Shah, 1986).

Retrospective data are imperfect at best. Although rare events—marriage, childbirth, hospitalization—may be remembered indefinitely and highly salient events—major accidents or illnesses—may be remembered for 2 or 3 years, habitual events—activities of daily living—are forgotten almost im-

mediately (Bradburn, 1983; Sudman & Bradburn, 1982). The longer the time period, the greater the error. In addition, as we noted earlier, if the event of interest is death (as in the case of suicide), the collection of retrospective data from a given cohort ensures that sampling will be biased by the omission of those who have already experienced the event of interest.

Three errors are common: (a) memory failures, in which respondents forget events entirely; (b) telescoping, in which events are remembered as having occurred more recently than they actually did; and (c) rounding, in which respondents drop fractions and report even numbers or numbers ending in 0 or 5. These errors create different biases: Memory failures lead to underreporting, telescoping to overreporting, and rounding to both.

Supplemental aids and records can help reduce errors. Records control overreporting that is due to telescoping but have no effect on omission; aided recall, where the subject is explicitly presented with the possible options and is asked directly whether any particular event happened, reduces the number of omissions but may increase telescoping (Sudman & Bradburn, 1974). Researchers developing items for retrospective recall would do well to consult strategies described in the ongoing series, *Cognition and Survey Measurement,* published by the National Center for Health Statistics (see, e.g., Lessler, Tourangeau, & Salter, 1989; Means, Nigam, Zarrow, Loftus, & Donaldson, 1989).

If retrospective recall is the only alternative, is it worth the effort? We believe it is. In their retrospective study of suicide ideation, Bolger et al. (1989) successfully used several approaches to improve recall. Although studying a threatening event, they couched the study in less threatening terms, about the development of the concept of death and suicide. They never asked about respondents' mental health or suicidal behavior, only about thoughts and knowledge about others. Questionnaires were anonymous and self-administered in a group setting. Respondents were college students, close enough in age to the time period of interest (adolescence) but old enough to be removed.

Another word of caution is in order. Retrospective data collection is especially problematic if informants, not the individuals themselves, provide the data. Collecting retrospective data from some informants may not even be worth the effort. In a retrospective study of the familial transmission of affective disorders, for example, Stancer, Persad, Wagener, and Jorma (1987) ultimately had to set aside from analysis all parents and relatives of probands who were not personally interviewed.

Minimizing Attrition

Collecting prospective longitudinal data is difficult and expensive. Researchers most successful at minimizing attrition have used some of the following strategies: Explain to respondents why you need to follow them; ask them to contact you if they move; visit their homes and ask neighbors for information about them; pay them for participation in each interview; have them pay you an "earnest deposit" refundable at the end of the last interview; offer lottery prizes for those who successfully complete all required interviews (S. Cohen & Lichtenstein, 1990, used a videotape recorder); mail a newsletter at regular intervals; record the names and addresses of several relatives or friends not living with them; convene reunion meetings; maintain contact at regular intervals even if you are not recording data as frequently; and consult official records (jail, hospital, welfare, driver registration). Crider, Willits, and Bealer (1971, 1973), Farrington, Gallagher, Morley, St. Ledger, and West (1990), and Murphy (1990) offered many helpful strategies for minimizing attrition.

Despite diligent effort, most researchers lose some individuals to follow-up. Researchers attempting to improve their study by using a long follow-up period face a further conundrum: The longer the follow-up, the greater the attrition. Individuals lost to follow-up have censored event times. However, this type of censoring is not the noninformative censoring for which survival methods were developed. Individuals lost to follow-up can differ substantially from individuals who continue to participate. In the well-controlled Ontario Exercise–Heart Collaborative Study, for example, Oldridge et al. (1983) found that smokers and blue-collar workers were more likely to drop out.

What should a researcher do with the data on individuals lost to follow-up? Although multiple imputation methods offer much promise for handling these observations (Little & Rubin, 1987), two simple strategies can sometimes suffice. One is to assign each case a censored event time equal to the length of time the person was observed (without the event occurring). If an individual participated for the first 6 months of a 12-month study before dropping out, censor the event time at 6 months. The other approach is to use a worst-case scenario: Assume that the event actually occurred when the case was lost to follow-up. Under this strategy, the event time is not censored.

The appropriateness of these alternatives depends, in part, on the target behavior under study. As Tuma and Hannan (1984) pointed out, assuming that the event occurred when the observation is actually censored is tantamount to recoding a nonevent as an event. However, in some substantive fields, such as addiction relapse, this recoding may have much substantive basis. In fact, addiction researchers usually assume that individuals lost to follow-up have relapsed. They argue that these individuals are notoriously unfaithful subjects, and if they were "clean" they would keep in touch. Within 12 weeks after beginning a study of 221 treated alcoholics, opiate users, and cigarette smokers, for example, Hall, Havassy, and Wasserman (1990) lost 73 people (one third of their sample) to follow-up despite valiant attempts to minimize attrition. To ascertain the impact of attrition on analysis, the researchers conducted extensive sensitivity analyses, including coding relapse as occurring the week after the last interview completed and setting aside these cases from analysis. All analytic findings were similar in sign and magnitude, although the standard errors of parameter estimates were higher because of a loss of statistical power. The use of multiple strategies to analyze attrited cases increases confidence in the analytic results.

Handling Repeated Events

Many events are irreversible: first word, first step, puberty, high school graduation, and death to name a few. Once they occur, they cannot occur again. Other events—depression, in-

carceration, crimes, abortion, childbirth, marriage—can occur again and again. When studying the timing of potentially repeatable events, researchers must note the spell number under study, because the natural course of a first spell may differ from the natural course of second and subsequent spells.

Zatz (1985) recognized the ramifications of multiple spells. Studying the arrest histories of 257 boys referred to the Department of Corrections during an 11-year period, she found a total of 1,916 arrests. Rather than analyze each arrest separately (first arrest, second arrest, and so on), she combined the arrests together and used the number of priors as a predictor, finding that prior arrests increased the rate of commitment and decreased the rate of probation. We discuss this strategy further in the Analysis: Examining Survival Data section.

The presence of multiple spells may help explain many puzzles in relapse research. Klerman (1978) and Lavori et al. (1984) suggested that variation in relapse rates may be attributable to researchers' failure to note how many prior episodes of depression each subject had. Amenson and Lewinsohn (1981) suggested that multiple spells may explain the higher prevalence of depression among women. Although they agreed that the higher prevalence may, in fact, be attributable to an increased risk of depression or episodes of longer duration, they also suggested that relapse may hold the key. Previously depressed women may be at greater risk of recurrence than previously depressed men. Researchers studying addiction relapse have noted renewed abstinence on the part of formerly abstinent people who relapsed early after quitting. Previous treatment, even unsuccessful treatment, may increase the probability of success of subsequent treatments.

Determining How Many People to Study

Having specified in broad outline the design of a study, the final step is to determine how many people to study. Statisticians determine the minimum number of people a researcher should study by conducting a statistical power analysis (e.g., J. Cohen, 1990; Kraemer & Thiemann, 1988). Before conducting a power analysis, a researcher must specify the particular hypothesis to be tested, the desired Type I and Type II error rates, and the minimum effect size considered important; for survival analysis, the researcher must also specify the distribution of the hazard function and the length of follow-up.

Biostatisticians have derived methods for determining sample size with survival data, each applicable under somewhat different circumstances. Donner (1984) and Lachin (1981) reviewed the literature; Freedman (1982) provided tables for two group comparisons; Makuch and Simon (1982) provided formulas for multiple-group comparisons; Schoenfeld and Richter (1982) provided nomograms for the same purpose; Bernstein and Lagakos (1978) and Dupont and Plummer (1990) described computer programs that perform these and other calculations for several designs; and Rubinstein, Gail, and Santner (1981), Moussa (1988), and Lachin and Foulkes (1986) provided formulas for complex designs with stratification, covariate information, or allowances for loss of individuals to follow-up. In the presentation that follows, we have computed minimum sample sizes using the computer program of Dupont and Plummer.

No single table or formula can cover all possible design configurations. Here we provide ballpark estimates of sample size similar to those we provided elsewhere for more familiar statistical analyses (Light et al., 1990). Table 1 contains the minimum total sample sizes necessary to achieve a power of .80 for a simple two-group comparison at the .05 level (two-tailed). The rows of the table indicate minimum detectable effect sizes (R); the columns indicate the length of follow-up (F); the cell entries indicate the minimum total sample size used in the analysis (N). Researchers should inflate these sample-size estimates appropriately to adjust for cases lost to follow-up. These calculations were made assuming a flat hazard function, a restrictive assumption indeed, but the simplest, and the one researchers generally assume in the absence of more detailed information.

To use the table, the researcher must first specify the smallest effect size deemed important for detection. Although biostatisticians developed several measures of effect size, perhaps the simplest is the ratio of median lifetimes in the two groups, denoted by R. Letting m_1 be the median lifetime in one group and m_2 the median lifetime in the other, $R = m_1/m_2$. When $R = 1.25$, the median lifetime of one group is 25% longer than the median lifetime of the other; when $R = 1.50$, the median lifetime of one group is 50% longer; when $R = 2.00$, the median lifetime of one group is twice as long (100%) as the other group.[4]

How does a researcher specify the minimum detectable effect size in advance of data collection? One way is to use prior research. Consider a two-group experiment that might follow from Stevens and Hollis's (1989) smoking study. The median survival time in the control group of this experiment was 4 months ($m_2 = 4$). If the median survival time in a new experimental group is expected to be as high as 8 months ($m_1 = 8$), the new study can be designed to detect an R of 2.00; if the median survival time in the new experimental group is expected to be only 6 months ($m_1 = 6$), the study should be designed to detect an R of 1.50. In the absence of such prior information, Schoenfeld and Richter (1982) suggested that $R = 1.50$ be used because a 50% increase in survival is "clinically important and biologically feasible" (p. 163).

After specifying the minimum detectable effect size, the researcher must specify the length of follow-up. Because the length of follow-up can vary greatly across substantive domains, we need a standardized measure applicable to a variety of settings and metrics. We achieve this goal by dividing the length of follow-up by the average anticipated median lifetime in the two groups. More precisely, letting $A = (m_1 + m_2)/2$ be the average median lifetime in the two groups and T be the total length of follow-up, our standardized measure of follow-up, F, is T/A. If a study follows individuals to only half the average median lifetime, $F = 0.5$; if a study follows individuals to the average median lifetime, $F = 1.0$; if a study follows individuals for twice as long as the average median lifetime, $F = 2.0$.

By using a standardized measure of the length of follow-up, the table can be used with studies of widely varying length. It is

[4] For readers who prefer to think in terms of ultimate percentage surviving, an R of 1.50 corresponds to an improvement in survival from 50% to 63%, from 25% to 40%, or from 10% to 22%. An R of 2.00 corresponds to an improvement from 50% to 71%, from 25% to 50%, or from 10% to 32% (Freedman, 1982).

Table 1
Minimum Total Number of Individuals Needed to Detect Differences in Survival Between Two Groups

Effect size	Follow-up period				
	0.5	1.0	1.5	2.0	2.5
1.25	>2,162	1,260	976	840	766
1.50	654	382	296	254	232
1.75	344	200	156	134	122
2.00	224	130	102	88	80

Note. We have assumed a two-tailed test at the 0.05 level, power of 0.80, exponentially distributed survival times, and all individuals followed for the same period of time.

equally applicable if the average median lifetime is 6 min, 6 days, 6 months, or 6 years. If the average median lifetime (A) is 6 (in any of these units), a follow-up (T) of 3 yields an F of 0.5, a follow-up of 6 yields an F of 1.0, a follow-up of 9 yields an F of 1.5, and a follow-up of 12 yields an F of 2.0. The particular time units cancel each other out in the standardization.

We now examine the minimum sample sizes presented in Table 1, focusing first on differences in effect size displayed across the rows. Small effects ($R = 1.25$) are difficult to detect. Regardless of the length of follow-up, a study must include many hundreds or well over 1,000 individuals to have a reasonable chance of detecting such effects. Medium-sized effects ($R = 1.50$–1.75) can be detected with moderate-sized samples; somewhere between 200 to 400 individuals will generally suffice, depending on the length of follow-up. Large effects ($R = 2.00$) are relatively easy to detect, even using small samples. If the median lifetime in one group is twice as long as the median lifetime in the other, the researcher has an 80% chance of detecting this difference using only 100 to 200 individuals.

Table 1 can also be used for another purpose: to decide on the length of data collection. Reexamine the table, focusing now on the variation in sample sizes across the columns, corresponding to follow-ups of widely differing lengths. The great variation in minimum sample sizes for a given effect size emphasizes the importance of following individuals under study for as long as possible.

Consider, for example, how the minimum sample size needed to detect an R of 1.50 depends on the length of follow-up. If a researcher follows a sample only halfway to the average median lifetime, $F = .50$; such a study would require 654 people to detect the 50% difference in median lifetimes. If the researcher follows people for longer periods of time, however, fewer people are needed. If the follow-up extends to the average median lifetime ($F = 1.00$), the same power of .80 can be achieved with almost half as many individuals ($N = 382$). If the follow-up is extended further to twice the average median lifetime ($F = 2.00$), the same power can be achieved with only a third as many individuals ($N = 254$).

The message for research design is clear. Much statistical power can be gained by following people for longer periods of time. Researchers would do well to follow people for at least as long as the average median lifetime ($F = 1.00$). By doubling the length of follow-up, the researcher can achieve the same statistical power with approximately 33% fewer individuals. If the length of follow-up is less than the average median lifetime, only studies of many hundreds of individuals will have adequate statistical power.

Analysis: Examining Survival Data

Most researchers begin data analysis with exploratory and descriptive approaches; they move on to fitting statistical models and testing hypotheses only after a full exploration of the data (Ehrenberg, 1982; Mosteller & Tukey, 1977). In the following sections, we present a broad array of strategies for analyzing survival data, beginning with descriptive approaches and moving on to model building.

Describing Survival Data

Analysis of survival data typically begins with an examination of the sample survivor and hazard profiles and the comparison of these profiles computed separately for subsamples of individuals sharing characteristics of substantive interest. We illustrate these approaches using data reported by Telles and Spreat (1985), who studied discharge patterns among 404 mentally retarded residents of a short-term rehabilitation facility. The authors asked when residents were referred for discharge and whether time to referral differed by the residents' level of retardation. After computing an estimated median time to discharge of 3.2 years, the authors examined variation in referral time by displaying estimated hazard rates according to the residents' level of retardation. The bottom panel of Figure 3 contains hazard profiles for two of the resident groups: those with mild and severe retardation. The top panel contains the corresponding sample survival profiles we reconstructed from Telles and Spreat's data.

Sample survivor and hazard profiles contain a great deal of information. Examining the sample survivor profiles by retardation level shows that mildly retarded residents have better long-term cumulative prospects for referral than do severely retarded residents. About half the mildly retarded residents are referred for discharge within 2 to 3 years of admission; severely retarded residents wait a year longer on average.

The subsample hazard profiles disentangle these referral patterns year by year and provide a more sensitive magnifying glass for identifying when clients are likely to be referred.[5] Immediately after moving into the facility, the risk of referral rises as clients improve. After a few years, however, the risk of referral declines. In every year, the hazard for mildly retarded residents is higher than that for severely retarded residents, indicating that the former group is more likely to be referred for discharge at all times.

When we compare hazard profiles for the two groups of people, we implicitly treat level of retardation as a predictor of the entire hazard profile. The profile comparison shows how

[5] It may seem inappropriate to use the foreboding term *risk* to describe referral for discharge rather than the less emotive term *chance* used by the study's authors. We use risk to avoid possible confusions over the meaning of chance and to be consistent with usage in the current article and the research literature.

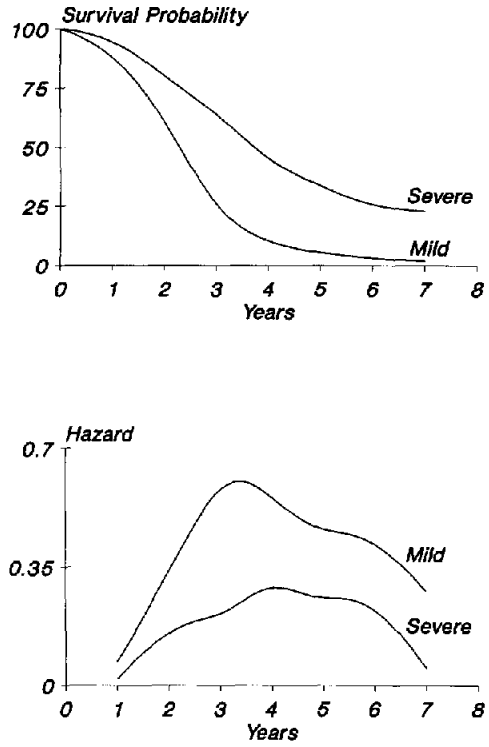

Figure 3. Sample survivor (top) and hazard (bottom) functions for mildly and severely retarded residents based on data reported by Telles and Spreat (1985).

risk of referral is related to retardation. We could divide the sample in other ways and treat these divisions as predictors of hazard as well. Telles and Spreat (1985), for instance, tabulated hazard by year of admission and asked whether the time to referral depends on entry cohort.

Within-group survivor and hazard profiles provide simple persuasive descriptions of when events occur and how timing patterns vary across groups (Lawless, 1982; Lee, 1980). But graphical displays cannot answer complex research questions. Continuous predictors would yield a cumbersome collection of profiles: one per predictor value. These methods are ill-suited for exploring the effects of several predictors simultaneously, for evaluating the influence of interactions among predictors, and for making inferences about the population from which the sample was drawn.

Ryan and Dent (1984) illustrated these limitations in their study of the relationship between the duration of infant breast-feeding and two dichotomous predictors: maternal employment (employed vs. unemployed) and high or low socioeconomic status (SES). After displaying survival profiles separately by employment status and SES, the authors explored both variables simultaneously by presenting four survival profiles (unemployed–high SES, unemployed–low SES, employed–high SES, employed–low SES). Were Ryan and Dent to add other predictors, or if these predictors were measured more precisely, the number of subsamples would rise multiplicatively, and their ability to estimate survivor profiles would decline as the number of subjects in each subgroup plummeted. Clearly, a more comprehensive approach is needed, and building statistical models of hazard offers one such strategy.

Building Statistical Models of Hazard

We represent the relationship between the hazard profile and predictors statistically in much the same way as we represent the relationship between noncensored outcomes and predictors in ordinary regression models. But because the outcome of interest is an entire continuous function, not just a conditional mean, hazard models are somewhat more complex conceptually. To illustrate the construction of a hazard model, we proceed heuristically by examining the two sample hazard profiles displayed in the bottom panel of Figure 3 and developing a population model that captures the relationship between the predictor (RETARDATION) and the outcome (the entire hazard profile).

We begin developing the model by replotting the two sample hazard profiles using a logarithmic transformation (Figure 4). In our model, we use a logarithmic transformation of hazard because untransformed hazard is bounded (it takes on only nonnegative values). To build a statistical model using a weighted linear combination of predictors (see Equation 1 later), an outcome's range should be unbounded (Mosteller & Tukey, 1977). When time is measured discretely, a logit transformation is used for the same reason.

Now we seek a representation of the relationship between the entire log-hazard profile and the predictor. Ignoring minor differences in shape, we see that in the sample the predictor RETARDATION essentially displaces the two risk profiles vertically relative to each other. When RETARDATION = 0 (mild), the log-hazard function is consistently higher relative to its location when RETARDATION = 1 (severe), indicating that, at every possible time among those residents who have not yet been referred, those who are mildly retarded have a greater risk of referral. Letting $h(t)$ represent the entire population hazard profile, we express this vertical displacement by relating the logarithmic

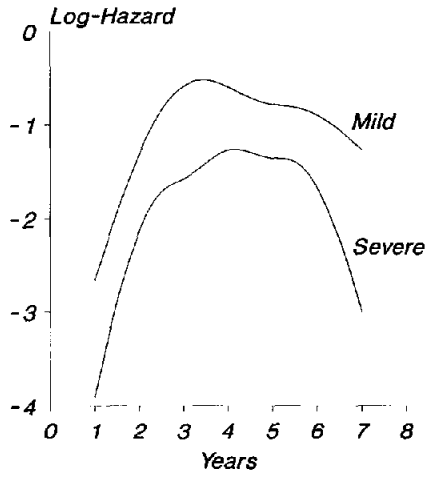

Figure 4. Sample log-hazard functions for mildly and severely retarded residents based on data reported by Telles and Spreat (1985).

transformation of the hazard profile to the predictor RETARDATION as follows:

$$\log h(t) = \beta_0(t) + \beta_1 \text{RETARDATION}. \quad (1)$$

$\beta_0(t)$ is called the *baseline log-hazard profile.* It represents the value of the outcome (the entire log-hazard function) in the population when the predictor (RETARDATION) is 0. We write the baseline as $\beta_0(t)$, a function of time, and not as β_0, a single term unrelated to time (as in regression analysis), because the outcome itself, $\log h(t)$, is an entire temporal profile. The model specifies that differences in the value of RETARDATION shift this time profile of log-hazard up or down. The slope parameter, β_1, captures the magnitude of this shift; it represents the vertical shift in log-hazard attributable to a one-unit difference in RETARDATION. Because RETARDATION is a dichotomy coded 0 and 1, β_1 captures the difference in the risk of referral between the mildly and severely retarded groups; because severely retarded residents have a lower risk of referral, in this example, β_1 will be negative.

Hazard models such as Equation 1 closely resemble familiar regression models. Several predictors can be included by adding other variables expressed as linear functions of additional unknown slope parameters on the right side of the equation. This model expansion allows researchers to examine one predictor's effect while controlling statistically for others'. Inclusion of cross-product terms enables examination of statistical interactions between predictors. Had Ryan and Dent (1984) built models of the hazard of breastfeeding cessation, for instance, they would have used three predictors: maternal employment, SES, and the interaction between the two.

Hazard models provide a powerful, flexible, and sensitive approach to survival analysis, subsuming the exploratory graphical approaches described earlier. The goodness of fit of a hypothesized population model can be evaluated with data, allowing inferences about population relationships between hazard and predictors. And as we show later, reconstructed survivor and hazard functions and estimated median lifetimes can depict the effects of predictors, providing answers to research questions in the original metric of interest: time.

Proportional Versus Nonproportional Hazards

Hazard models like Equation 1 implicitly assume that all the log-hazard profiles corresponding to successive values of a predictor differ only by their relative elevation (described here by β_1). Under such models, but in the antilogged world of raw hazard, all the hazard profiles are simply magnifications or diminutions of each other: They are proportional. Under this proportional hazards assumption, implicit in Equation 1, the entire family of log-hazard profiles represented by all possible values of the predictors share a common shape and are mutually parallel. Willett and Singer (1988) drew an analogy between this assumption and that of homogeneity of regression slopes in the analysis of covariance.

Proportional hazards models have become popular in psychology in part because many statistical packages now provide programs (PHGLM in SAS, BMDP 1L) for estimating their parameters using a method developed by Cox (1972). This ingenious strategy allows estimation of parameters like β_1 without consideration of the shape of the baseline hazard function, $\beta_0(t)$. For this reason, analogous to nonparametric methods (which make no underlying distributional assumptions), Cox regression is called semiparametric. Proportional hazards modeling is the most frequently used type of survival analysis in psychological research, having been applied to topics such as smoking relapse (Stevens & Hollis, 1989), affective disorders (Shapiro et al. 1989), childhood family breakdown (Fergusson, Horwood, & Dimond, 1985; Fergusson, Horwood, & Shannon, 1984), interruptions in conversation (Drass, 1986), employee turnover (Morita et al., 1989), and employee absences (Harrison & Hulin, 1989).

The tremendous boon of the semiparametric method—arising from its ability to evaluate effects of predictors independent of the shape of baseline hazard profile—leads to a marked disadvantage, however. The method is so general that it works for an unspecified baseline hazard profile of any shape. Without needing to explore the baseline hazard, many investigators examine effects of predictors without exploring overall levels of risk (see, e.g., Heckert & Teachman, 1985; Moen, Dempster-McClain, & Williams, 1989). Because the baseline hazard function can be easily ignored, researchers may fail to recognize substantively and statistically important information contained only in the shape of the baseline hazard function.

What kinds of information can be found? The baseline hazard function and, under the proportional hazards assumption, its magnified and diminished cousins describe the pattern of risk over time; it tells us when the target event is most likely to occur. The hazard profiles in Figure 3, for example, show that residents are at greatest risk of referral for discharge in the 3rd and 4th years after admission. All the predictor does is magnify or diminish this pattern.

The ease with which hazard functions themselves can be ignored has had further ill consequence: It has promoted unthinking and dubious acceptance of the proportional hazards assumption. It is all too easy to examine effects of predictors without examining the tenability of the underlying proportional hazards assumption. Notice, for example, that the sample log-hazard profiles in Figure 4 are neither identical in shape nor parallel, suggesting that the proportional hazards assumption might be untenable.

Researchers should be especially circumspect about the tenability of the proportional hazards assumption; estimated effects of predictors may be wrong if the adopted model incorrectly constrains the log-hazard profiles to be parallel with identical shapes. Ignoring such underlying failures can lead to incorrect substantive conclusions. Indeed, in our own empirical research on employee turnover, we have found that violations of the proportional hazards assumption are the rule, not the exception (see, e.g., Murnane et al., 1989, 1991; Singer, in press; Singer & Willett, in press). Other researchers documented similar violations. In a study of child mortality, for example, Trussel and Hammerslough (1983) documented differences in interpretation that arise when the proportional hazards assumption is injudiciously assumed tenable (compare their Tables 3 and 4, particularly the effects of gender, birth order, and age of mother at birth). Violations of the proportional hazards assumption have been detected in other substantive areas as well, including the age of onset of suicide ideation (Bolger et al., 1989) and the

length of time to a change of opinion (Hembroff & Myers, 1984).

So pervasive is the widespread acceptance of the proportional hazards assumption that we now begin our own data analyses with the entirely opposite view: Along with unicorns and normal distributions (Micceri, 1989), we regard the proportional hazards assumption as mythical in any set of data until proved otherwise. Before adopting a proportional hazards model, researchers should, at a minimum, subdivide the sample by values of predictors and compare the shapes of the hazard profiles within groups. Arjas (1988), Kalbfleisch and Prentice (1980), Harrell and Lee (1990), and Willett and Singer (1988) provided methods for exploring the tenability of the assumption. And as we discuss later, researchers can adopt a broader analytic approach, one that facilitates the statistical testing of the proportional hazards assumption and permits the fitting of nonproportional hazard models if necessary.

Different Types of Predictors That Can Be Included in Hazard Models

Hazard models can simultaneously include either or both time-invariant and time-varying predictors. As befits their label, time-invariant predictors describe immutable characteristics of individuals; the values of time-varying predictors, in contrast, may fluctuate with time. When building hazard models of the risk of premarital pregnancy, for example, Yamaguchi and Kandel (1985a) included time-invariant predictors (e.g., race, father's education, prior history of school absences) and time-varying predictors (e.g., illicit drug use, current school attendance) and found that "women who currently use drugs . . . are about twice as likely as women who never used these drugs to experience a premarital pregnancy" (p. 262). These authors also used hazard models to study the links between time-varying drug consumption and the risk of job turnover (Kandel & Yamaguchi, 1987).

The hazard model in Equation 1 includes a single time-invariant predictor: RETARDATION. The information captured by this predictor—mild or severe retardation—remains constant over time. β_1 quantifies the time-invariant effect of this time-invariant predictor on the risk of referral. Hazard models like Equation 1 can be extended to include time-varying predictors. Such extensions can be particularly helpful to psychological researchers studying predictors that vary naturally over time.

Hazard models with time-varying predictors closely resemble the model in Equation 1. In Yamaguchi and Kandel's (1985b) study of the risk of premarital pregnancy, for example, one possible population hazard model might include time-invariant RACE (authors' coding: 0 = not Black, 1 = Black) and time-varying DRUG USE (0 = not currently using, 1 = currently using) as follows:

$$\log h(t) = \beta_0(t) + \beta_1 \text{RACE} + \beta_2 \text{DRUG USE}(t). \qquad (2)$$

The parenthetical t in the predictor DRUG USE(t) indicates that the values of this predictor may vary over time. Unit differences in DRUG USE correspond to shifts in the log-hazard profile of β_2. Although the values of the predictor DRUG USE may differ over time, each one-unit difference anywhere produces the same shift of β_2 in the appropriate part of the log-hazard profile. So although the model includes a time-varying predictor, the per-unit effect of that predictor on log-hazard is constant over time.

Another way to understand the effects of time-varying predictors is to conceptually regard the outcome in Equation 2—the log-hazard profile—as a temporally sequenced list (a vector) of premarital pregnancy risks. The predictors also can be viewed as an ordered list of values that for each woman describe the values of RACE and DRUG USE over time. Each element in the hazard list corresponds to an element in each predictor's list. For a time-invariant predictor such as RACE, all elements in each woman's predictor list are identical: 1 for every person who is Black, 0 for every person who is not. For a time-varying predictor such as DRUG USE, in contrast, the values in the predictor list may differ across occasions. If a woman does not use drugs initially, the early elements in the DRUG USE vector are 0; when drug use begins, the values change to 1. If drug use persists, the values stay at 1; if it ends, the values revert to 0. Each woman has her own drug-use pattern: The number of patterns across women is limited only by the number of possible states and occasions. The hazard model simply relates the values in one list (the hazard vector) to the values in the other (the predictor vector) regardless of whether the elements in the latter list are identical.

Time itself is the fundamental time-varying predictor. So conceptually at least one might argue that it too should be included as a time-varying predictor in Equation 2, mapping intrinsic changes in the risk of pregnancy over time. Although intuitively appealing, this approach produces complete redundancy in the model because this time-varying effect is already captured by the baseline log-hazard function, $\beta_0(t)$. $\beta_0(t)$ describes the chronological pattern of baseline risk: The differences in log-hazard attributable solely to time. Estimation of the baseline hazard function is tantamount to estimation of the main effect of time. This analogy reinforces the need to examine the shape of the baseline hazard because it provides information about the effects of the fundamental time-varying predictor: time itself.

Interactions Between Predictors and Time

Not only can predictors themselves be time invariant or time varying, but their effect on hazard can be constant or vary over time as well. By including a main effect of the time-varying predictor DRUG USE in Equation 2, we assume that the vertical displacement associated with drug use is the same at ages 16 and 24 (and equal to β_2).

The assumption of temporally immutable effects may not hold in reality: The effects of some predictors will vary over time. The impact of drug use on the risk of pregnancy might decline as women mature and become less susceptible to peer pressure. If so, the distance between the log-hazard profiles associated with different values of the predictor DRUG USE would narrow as time passes (and women mature).

When the effect of one predictor on an outcome differs by levels of another predictor, statisticians say that the two predictors interact. If the effect of a predictor like DRUG USE differs across time, we say that the predictor interacts with time. Predictors that interact with time have important substantive interpretations, allowing researchers to build complex models of the

relationship between predictors and risk. If a predictor primarily affects early risks, the hazard profiles will be widely separated in the beginning of time and converge as time passes. If a predictor primarily affects late hazards, it will have little effect at the beginning of time, but will widen the distance between hazard functions on each subsequent occasion.

Much information about human behavior can be learned by exploring whether the effects of predictors are constant or variable over time. In their study of the age at first suicide ideation, Bolger et al. (1989) detected interactions between two predictors and time. Dividing time into two broad periods—adolescence and preadolescence—they found that the effects of both respondent race and parental absence in childhood differed across these periods. With regard to race, they found that during preadolescence Whites were less likely to consider suicide than non-Whites, but during adolescence they were more likely to do so; with regard to parental absence, they found that during preadolescence children who experienced a parental absence were more likely to consider suicide than those who did not experience such absence, but during adolescence parental absence had little impact on the risk of suicidal thought. Including interactions between predictors and time allows researchers to more accurately characterize the predictors of risk.

If a predictor interacts with time, the proportional hazards assumption is violated, and proportional hazards models do not represent reality. Although few researchers to date have explored the possibility of such interactions, those who do often find that they are pervasive. When evaluating the decision-making process in informal task groups, for example, Hembroff and Myers (1984) found an interaction between time and the status characteristics of participants. They demonstrated the extent of the violation by displaying the dramatically nonproportional fitted hazard functions obtained in each of four treatment groups (see their Figure 3, p. 345).

The proportional hazards assumption is easily tested by adding to the model an interaction with time and assessing the effect of this new predictor. If the proportional hazards assumption is not violated, the interaction term can be removed. If a violation is detected, the interaction with time remains in the model to ensure the appropriate estimation of predictor effects (see Singer & Willett, in press). We recommend that researchers routinely examine such interactions in hazard models, just as they would routinely examine interactions among other predictors in traditional linear models.

Competing Risks

All these examples assume that each individual can occupy only one of two possible states. Survival methods can also be extended to situations where each individual can occupy one of several states. This methodology, known as competing-risks survival analysis, can handle an unlimited number of states. The only requirement is that the multiple states be mutually exclusive and exhaustive.

We illustrate the competing-risks methodology by referring back to Zatz's (1985) study of the final case dispositions of juveniles arrested in Phoenix, Arizona. After arrest, each juvenile's case can be disposed of in one, and only one, of five ways: (a) dismissal; (b) informal processing; (c) probation; (d) commitment to the Department of Corrections (DOC); or (e) remand to adult court. Initially, each juvenile is at risk of every mode of disposition; after disposition by any particular mode, the juvenile is no longer at risk of any of the other modes (for this arrest episode). Dismissed offenders cannot be put on probation; those remanded to an adult court cannot be informally processed as juveniles. The five modes of disposition, then, are unique events that compete with each other to terminate the time between arrest and case disposition. Juveniles whose cases are disposed of by one method are protected against the others.

Many psychological phenomena lend themselves to a competing-risks formulation. This methodology is appropriate whenever the individuals under study are at risk of several mutually exclusive events. Lives end by one of many competing diseases or by accidents (Cox & Oakes, 1984). Employees leave their jobs voluntarily or involuntarily (Hachen, 1988; Singer & Willett, 1988; Sorenson, 1977). Adults not working can be unemployed or out of the labor force (Flinn & Heckman, 1982). Students leave school by graduating or by dropping out (Willett & Singer, in press). Patients completing psychotherapy can stay well, have a relapse, or have a new episode (Gallagher-Thompson et al., 1990). Even study attrition can be analyzed through a competing-risks formulation (Tuma & Hannan, 1979).

In competing-risks survival analysis, the researcher models the risk of each event separately. In this way, the predictors of risk can differ depending on which of the several competing events eventually occurs. Zatz (1985), for example, argued that the risk profiles for each method of case disposition might differ depending on the offender's age at arrest: First offenders and young offenders would be handled leniently, whereas older offenders would be remanded to adult court.

After modeling the risk profiles for each event separately, the researcher then assembles a global profile. Although several computer programs (e.g., RATE; Tuma, 1986) can conduct the calculations simultaneously, these models also can be estimated with standard survival analysis software. The idea is to conduct as many separate analyses as there are states; when studying the five possible modes of case disposition, we would conduct five concurrent analyses. However, the separate analyses are not conducted in separate subsamples—one per event type—as might be anticipated. Instead, all cases are included in every analysis, with modified definitions of censoring accounting for the competing risks.

We outline the strategy briefly using Zatz's (1985) example. Begin with juveniles given probation. When studying probation, their time to probation is recorded appropriately. When studying the other four modes of disposition—dismissal, informal processing, commitment to DOC, and remand to adult court—their time to disposition is censored at the time of probation. Why? Because after being given probation, the juvenile is not at risk of these four competing events. Redefining censoring similarly for each of the other four events, we estimate five sets of hazard models. After identifying predictors of hazard for each event separately, we recombine the component-risk profiles to create the overall risk profile for all the events taken together. Allison (1984) and Hachen (1988) provided further details.

Repeated Spells

As described earlier, some events can recur many times during an individual's lifetime. When analyzing data that may include repeated spells, researchers must take special care because subsequent spells may not be independent replications of the initial spell. Zatz (1985), for example, found that the disposition of first offenders differed from that of repeat offenders. Similarly, Lavori et al. (1984) found that the risk of relapse into depression was greater in cohorts with more prior episodes of depression.

Repeated spell data can often be analyzed spell by spell. This is easiest in prospective studies that follow a sample of individuals over time, allowing observation of the target event whenever it occurs. Each individual's lifetime can be divided into the separate spells—the first, the second, and so on—and hazard models of each spell built separately (see, e.g., Murnane et al., 1989).

Spell-by-spell analyses are not always feasible, however, For example, although Zatz (1985), in her study of juvenile offenders, wanted to analyze each arrest spell separately—all first arrests, all second arrests, and so on—she could not do so because too few juveniles experienced these later spells. Instead, she pooled the repeated-spell data together in a single analysis and used number of prior spells as a continuous predictor of risk. (Alternatively, she could have represented spell number using a system of dummy variables.) Judicious inclusion of terms capturing interactions between the spell variables and both time and other predictors can account for differences in the baseline hazard profile and variation in predictor effects from spell to spell.

Although attractive in its simplicity, we believe that such a combined-spells approach is inherently flawed. Why? Because the models fitted do not explicitly link each person's repeated spells. This omission may exaggerate degrees of freedom and lead to underestimates of the standard errors of the parameter estimates. Recognizing this limitation, Fichman (1988, 1989) and Harrison and Hulin (1989) went to the other extreme; they randomly selected a single spell for each individual when modeling employee absences. This strategy is also flawed, however, because it sets aside large amounts of information, thereby reducing statistical power.

For these reasons, we recommend caution when interpreting analyses based on repeated-spell data. This limitation may have severe consequences for researchers who use survival methods to study social interactions where each individual or dyad generates many spells of information, and this information is pooled together in a single analysis (Felmlee & Eder, 1983; Gardner & Griffin, 1989; Griffin & Gardner, 1989). Researchers who collect data that include large numbers of repeated spells may find it more appropriate to refocus their research questions on the probability of transitions between states so that the power of Markov modeling can be brought to bear (see Wickens, 1982).

Discrete-Time Survival Analysis

The models posited previously assume that time can take on any nonnegative value and represent the baseline hazard as a continuous function of time, $\beta_0(t)$. But many researchers collect data in discrete time, either because the events only occur or are only measured at specific times: every week, month, semester, or year. For example, when McLanahan (1988) used the *Panel Study of Income Dynamics* to examine whether and when adolescent girls moved out of their parents' house and became heads of households in their own right, she could not precisely determine the transition dates. She knew each girl's status (household head or not) only on a year-by-year basis.

Proportional and nonproportional hazards models can be used to analyze such discrete-time data using a modification of logistic regression known as *discrete-time survival analysis.* This methodology is developed and described elsewhere (Allison, 1982; Efron, 1988; Laird & Olivier, 1981); Guilkey and Rindfuss (1987) and McLanahan (1988) illustrated its application. Discrete-time survival analysis is easy to apply, facilitates the estimation of the baseline hazard function, encourages the testing of the proportional hazards assumption, and enables researchers to fit hazard models using procedures available in most statistical computer packages. For all these reasons, we encourage its wider application to studying questions about time.

Before using logistic regression to conduct a discrete-time survival analysis, the researcher alters the data structure, transforming the standard one-person, one-record data set (the *person* data set) into a one-person, multiple-period data set (the person–period data set). Figure 5 illustrates this conversion in a hypothetical data set based on McLanahan's (1988) study; it presents data for three adolescent girls; data collection began when the girls were 16 and continued until they either turned 26 or became a head of household or when data collection ended.

The original person data set records each girl's data in a single line. The two variables, "HOHAGE" and "CENSOR," describe whether and if so, when the girl became a head of household. For Subjects 1 and 3, who are not censored (CENSOR = 0), HOHAGE contains the age at which the girl actually became a head of household; for Subject 2, who is censored (CENSOR = 1), HOHAGE contains the age at which data collection ended. The time-invariant predictor BLACK records her race (0 = not Black, 1 = Black); the time-varying predictor SES describes her parent's SES (1 = low SES, 4 = high SES). Time-varying predictors like SES, which may take on different values in each data collection period, are represented by a series of variables, one per year of data collection (SES_{16} through SES_{26}). For data collection periods after the transition to household head, or after the end of data collection, the annual SES descriptors are missing (denoted by ".").

The records in the reconstructed person–period data set note what happened to each girl during each discrete-time period when the event of interest could have occurred, until it did occur, or until data collection ended (whichever comes first). In this example, this yields one record per year per person. If the event occurred during data collection, the girl became a head of household at the age recorded, and the case is not censored. Subject 1, who became a head of household at age 18, has records for ages 16, 17, and 18. Subject 3, who became a head of household at age 23 has eight records, one for each year between age 16 and age 23. Censored cases did not become a head of household during the data collection period. Subject 2, who

Original Person Data Set

ID	HOHAGE	CENSOR	BLACK	SES_{16}	SES_{17}	SES_{18}	SES_{19}	SES_{20}	SES_{21}	SES_{22}	SES_{23}	SES_{24}	SES_{25}	SES_{26}
01	18	0	1	3	3	3	.	.	.	.	.	.	.	.
02	21	1	1	1	1	2	2	2	2	.	.	.	.	.
03	23	0	0	3	3	3	3	4	4	4	4	.	.	.

Converted Person-Period Data Set

ID	D_{16}	D_{17}	D_{18}	D_{19}	D_{20}	D_{21}	D_{22}	D_{23}	D_{24}	D_{25}	D_{26}	BLACK	SES	HOUSEHEAD
01	1	0	0	0	0	0	0	0	0	0	0	1	3	0
01	0	1	0	0	0	0	0	0	0	0	0	1	3	0
01	0	0	1	0	0	0	0	0	0	0	0	1	3	1
02	1	0	0	0	0	0	0	0	0	0	0	1	1	0
02	0	1	0	0	0	0	0	0	0	0	0	1	1	0
02	0	0	1	0	0	0	0	0	0	0	0	1	2	0
02	0	0	0	1	0	0	0	0	0	0	0	1	2	0
02	0	0	0	0	1	0	0	0	0	0	0	1	2	0
02	0	0	0	0	0	1	0	0	0	0	0	1	2	0
03	1	0	0	0	0	0	0	0	0	0	0	0	3	0
03	0	1	0	0	0	0	0	0	0	0	0	0	3	0
03	0	0	1	0	0	0	0	0	0	0	0	0	3	0
03	0	0	0	1	0	0	0	0	0	0	0	0	3	0
03	0	0	0	0	1	0	0	0	0	0	0	0	4	0
03	0	0	0	0	0	1	0	0	0	0	0	0	4	0
03	0	0	0	0	0	0	1	0	0	0	0	0	4	0
03	0	0	0	0	0	0	0	1	0	0	0	0	4	1

Figure 5. Transforming a person data set into a person–period data set. (HOHAGE = age at which girl became head of household or when data collection ended; CENSOR = whether or not girl was censored, 0 = not censored, 1 = censored; BLACK = girl's race, 0 = not Black, 1 = Black; SES = socioeconomic status of girl from age 16 to age 26; D = dummy variable for discrete-time intervals from 16 to 26; HOUSEHEAD = whether the girl became head of household in each particular year, 0 = no, 1 = yes.)

was 21 when lost to follow-up, still was not a head of household; she therefore has six records, one for each of the ages 16, 17, 18, 19, 20, and 21.

Each person–period record contains period-specific values of three different types of predictors: (a) the age indicators, dummy variables D_{16} through D_{26}, specifying the discrete-time interval to which the record refers; (b) the time-invariant predictor, BLACK, whose values are constant across records for each person; and (c) the time-varying predictor, SES, whose values may fluctuate from year to year.

The person–period data set also includes a new outcome variable—here called HOUSE HEAD—that indicates whether the young woman became a head of household in that particular year. If she did not, HOUSE HEAD = 0; if she did, HOUSE HEAD = 1. Girls who never became household heads during the study are censored, and so for them HOUSE HEAD = 0 in every record, even the last (as for Subject 2). Girls who became household heads during the study are not censored, and so for them, HOUSE HEAD = 1, but only in the year that the transition occurred: age 18 for Subject 1, age 23 for Subject 3.

In discrete-time survival analysis, a researcher uses the person–period data set to model the relationship between the occurrence of the event of interest (becoming head of household) and the selected predictors. Because the outcome—HOUSE HEAD—is dichotomous, logistic regression is used to model the log-odds of becoming a head of household. Such estimation of discrete-time hazard models eliminates the need for dedicated software. Allison (1982) described this process using a worked example (see also Laird & Olivier, 1981). Willett and Singer (1991; Singer & Willett, in press) presented computer code (in SAS) for transforming the data set, fitting models, and reconstructing fitted hazard and survivor plots from parameter estimates.

A discrete-time population hazard model expressing the risk of becoming a head of household in terms of the main effect of BLACK is the following:

$$\text{logit } h(t) = [\delta_{16} + \delta_{17}D_{17} + \delta_{18}D_{18} + \cdots + \delta_{26}D_{26}] + \beta_1 \text{BLACK}. \quad (3)$$

The baseline hazard, once a continuous function of time, $\beta_0(t)$, is now a step function of time, a weighted linear combination of the age indicators $[\delta_{16} + \delta_{17}D_{17} + \delta_{18}D_{18} + \cdots + \delta_{26}D_{26}]$. Omitting one age indicator (here D_{16}) prevents complete linear redundancy; the logistic regression parameters δ_{17} through δ_{26} measure deviations of the baseline logit-hazard from an initial value of δ_{16}.

Logistic regression parameter estimates, standard errors, and goodness-of-fit statistics generated by predicting the dichotomous outcome HOUSE HEAD using the time indicators and predictors are exactly those required for testing hypotheses about hazard. Allison (1982) demonstrated that these estimates are "consistent, asymptotically efficient, and asymptotically normally distributed" and that, despite the apparent inflation of sample size on creation of the person–period data set, the estimated standard errors are consistent estimators of the true standard errors (p. 82; see also Singer & Willett, in press). So, as with continuous-time models, the estimate of β_1 quantifies the relationship between the girl's race and her risk profile. Estimates of the δ's lead to fitted hazard probabilities for each discrete time period and allow reconstruction of fitted hazard and survivor plots. In an extensive investigation of the convergence of discrete- and continuous-time survival methods, Efron (1988) showed that the estimated survival profiles recovered from logistic regression estimates approach the well-known Kaplan-Meier estimates as the overall time interval is more finely discretized. He demonstrated, in addition, that the information loss on discretization is inversely related to the square of the number of discrete intervals and therefore declines rapidly to zero with increasing discretization.

Interactions among predictors, and between predictors and the time indicators, are included by forming cross-products in the person–period data set and using them as predictors (Singer & Willett, in press). Adding these interactions facilitates easy testing of the proportional hazards assumption and, if the assumption is violated, retention of the interactions in the fitted model ensures the appropriate estimation of effects.

Interpreting Fitted Models

Statistical models are of little use unless a researcher can interpret and present them clearly and persuasively. Interpretation includes at least three components: identification of statistically significant effects, computation of numerical summaries of effect size, and graphic display of the magnitude and direction of effects. These three components have direct analogues in traditional methods. When conducting an analysis of variance, for example, researchers might first determine whether the difference in average outcome between two groups is statistically significant. If it is, they might then express one group's advantage in standard deviation units, and provide data plots comparing the distribution of the outcome across groups.

All three components play an important role in data analysis. But because hazard models may be difficult to grasp, relating as they do to variations in entire hazard profiles, we believe that graphical techniques may provide the best medium for understanding and reporting findings. As the figures in this article demonstrate, graphics can help communicate complex and unfamiliar ideas about whether an event occurs and if so, when. Yet even the most effective graphical displays must be supported by documentation of parameter estimates and associated standard errors. So we begin our discussion of interpretation with the computer output commonly generated by statistical packages.

Computer output documenting the results of fitting hazard models closely resembles output documenting the results of other statistical techniques. Most programs output estimates of slope parameters, standard errors of these estimates, ratios of each parameter estimate to its standard error (a t statistic), and a p value based on the t statistic for testing the null hypothesis that the corresponding parameter is zero in the population (given that the other predictors are in the model). Some programs output a χ^2 statistic in lieu of a t statistic; the accompanying p value assesses the improvement in fit resulting from adding the predictor to a reduced model containing all the other predictors.

Researchers seeking predictors having a statistically significant relationship with hazard often focus primarily on p values. Adopting a suitable α level (perhaps adjusting for the multiple tests performed), they tick off those predictors whose p values beat (i.e., are smaller than) the target value. Although simple and straightforward, we discourage this strategy: It provides no information on the relative sizes and directions of effects, and it fails to address the more important question, "What is the relationship between the predictor and risk?"

Because hazard models represent relationships between transformations of entire hazard profiles and predictors, answering this question is complex. But after learning how to interpret these complex outcomes (log-hazard profile and logit-hazard profile), parameter estimates associated with predictors can be interpreted similarly to regression coefficients. Depending on whether a continuous or discrete-time model has been fit, the parameter estimates represent the difference in elevation of the log- or logit-hazard profiles corresponding to predictor values one unit apart. We find it helpful to imagine the profile on a log- or logit-hazard plot moving up (or down if the estimate is negative) for a one-unit difference in the predictor (see Figure 4). Predictors with larger parameter estimates produce larger elevation differences per unit difference in the predictor.

Even after considerable experience with hazard models, visualizations in transformed hazard remain difficult. An alternative intuitive approach is to transform the outcome back into the more familiar metric of risk, antilogging parameter estimates as necessary. Of course, a researcher must use different transformations and interpretations depending on whether continuous- or discrete-time models have been fitted.

We illustrate these ideas beginning with the continuous-time hazard model in Equation 1. Antilogging both sides, we have the following:

$$h(t) = e^{\beta_0(t)} e^{\beta_1 \text{RETARDATION}}.$$

Because RETARDATION = 0 for mildly retarded residents and 1 for severely retarded residents, the hazard functions corre-

sponding to these two groups are $h(t\text{:mild}) = e^{\beta_0(t)}$ and $h(t\text{:severe}) = e^{\beta_0(t)}e^{\beta_1}$. The risk profile in the severely retarded group is simply the profile in the mildly retarded group multiplied by e^{β_1}. This multiplicative rule applies to both categorical and continuous predictors. So in continuous-time hazard models, antilogged parameter estimates yield numerical multipliers of risk per unit difference in the predictor.

This transformation strategy enabled Shapiro et al. (1989) to document the vast superiority of lithium over imipramine in reducing the recurrence of manic episodes in a group of patients with a history of affective disorders. After adjusting for several other predictors, including patient gender and psychological history, the authors obtained a parameter estimate of 2.378 for a dummy variable representing drug therapy, SE = 0.68, $t = 3.52$, $p < .0005$ (their Table 2, p. 403). They interpreted the antilog of this estimate ($e^{2.378} = 10.8$) by writing that "patients with a manic index episode taking imipramine were at almost 11 times the risk for recurrence of those taking lithium" (p. 403).

Another way to interpret this scaling factor is in terms of percentage difference in risk. Doubling the baseline risk (multiplying by a factor of 2) is equal to a 100% increase in risk; tripling the baseline risk (multiplying by a factor of 3) is equal to a 200% increase. So in the previously cited drug study, multiplying the baseline hazard by 10.8 corresponds to a 980% increase in the risk of relapse for those taking imipramine over those taking lithium. The general rule is simple: The percentage difference in risk per unit difference in the predictor is $100(e^{\beta} - 1)$ (see Allison, 1984; Tuma & Hannan, 1984). Some researchers add these estimates of e^{β}, or $100(e^{\beta} - 1)$, to tables reporting parameter estimates, standard errors, t statistics, and p values (see, for instance, Bolger et al., 1989, Table 1; Hagan & Zatz, 1985, Table 1; Yamaguchi & Kandel, 1987, Table 2).

Similar but modified interpretations can be made after fitting discrete-time hazard models. Because discrete-time hazard is the conditional probability that an event will occur in a particular time interval given that it has not yet occurred before the interval, the model in Equation 3, which uses logit-hazard as the outcome, expresses the relationship between predictors and the log odds of occurrence. Estimates of e^{β} or $100(e^{\beta} - 1)$ are therefore multipliers of, or percentage of increases in, the odds of an event.

After finding a statistically significant relationship between the risk of becoming a head of household and whether the girl had lived in a mother-only family when she was between 12 and 16 years of age, McLanahan (1988) used this transformation strategy to interpret her findings. In the discrete-time hazard model, McClanahan found that the estimated coefficient of the single-mother family predictor was 0.87 ($p < .05$, Table 2, p. 9). Because $e^{0.87} = 2.39$, the estimated odds that an adolescent would become a head of household in her own right were 2.39 times greater for girls from single-mother homes than for girls from two-parent families. This represents a 139% increase in the odds of becoming a head of household.

As these illustrations document, numeric and algebraic strategies are not the last word in communicating the findings of survival analysis. Apart from being arithmetically convoluted, they have at least two other flaws. First, they ignore the shape of the baseline hazard function; they indicate only the extent to which one risk profile is a magnification or diminution of another. As argued earlier, the shape of the hazard profile—the temporal placement of its peaks and valleys—tells us much about the survival process under investigation. Second, algebraic interpretations are useful only if the proportional hazards assumption is met. If the effect of predictors differs over time, the risk profiles are no longer parallel in log- or logit-space, and so it makes little sense to talk about one profile being rescaled to generate the other. If the shapes of the risk profiles differ dramatically, algebraic interpretions may not only oversimplify findings; they may even misrepresent them completely.

Presenting fitted hazard plots, fitted survival plots, and estimated median lifetimes resolves these problems. Some computer programs provide procedures for recovering fitted profiles from parameter estimates. By appropriately substituting back into the hazard model, a researcher can generate fitted hazard profiles at substantively interesting values of the predictors for the range of time values spanning the data collection period. As we show later, fitted hazard profiles are clear, comprehensive, and intuitively meaningful. They demonstrate the effect of predictors on risk and pinpoint when these effects rise, fall, or remain constant with the passage of time.

Researchers should consider their original questions and analytic findings when selecting predictor values for constructing fitted plots. Questions to ask include, "Which predictors did I emphasize in my research questions?" and "Which predictors were significantly associated with hazard?" Use predictors that are substantively and statistically important when generating

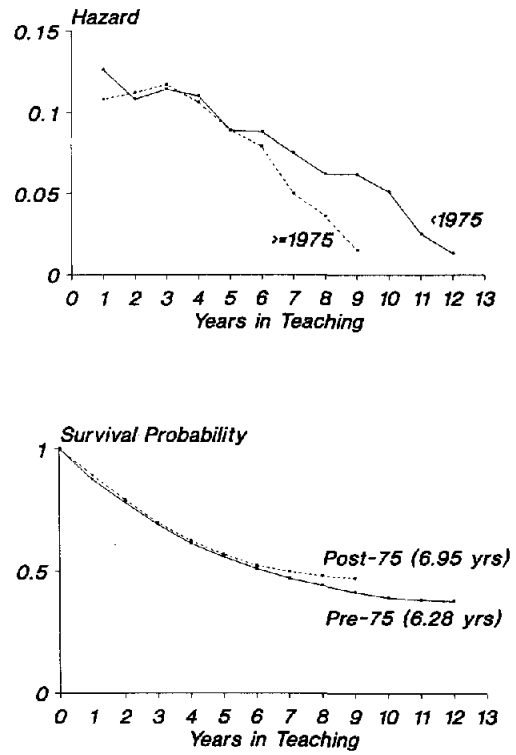

Figure 6. Fitted hazard (top) and survivor (bottom) functions for special education teachers by entry cohort (Singer, in press).

the fitted profiles; lesser variables can be included as controls by equating their value to their sample averages.

We illustrate these ideas in the top panel of Figure 6, which presents a pair of fitted hazard profiles generated in our own research on employee turnover. These profiles describe the career durations of 3,941 special education teachers hired in Michigan between 1972 and 1978. The teachers are divided into two cohorts: 1972–1974 and 1975–1978. We divided the sample at 1975 because in that year Congress passed PL 94-142, the Education for All Handicapped Children Act, the national special education law that dramatically altered the demand for and responsibilities of special educators (Singer & Butler, 1987). Research interest centers on identifying how long special educators stay in teaching, and whether the risk of leaving teaching differs between teachers hired before the passage of PL 94-142 and teachers hired after.

Using discrete-time hazard models, we found a main effect of COHORT (0 = 1972–1974, 1 = 1975–1978) and an interaction between COHORT and time. The interaction violates the proportional hazards assumption. The shapes of the fitted hazard profiles differ considerably; they are neither magnifications nor diminutions of each other; they virtually coincide in Years 1 through 5 and diverge thereafter. The interpretation of these plots is nevertheless straightforward: The two cohorts of teachers have virtually identical early risks and distinctly different later risks. Early after hire, the two cohorts of teachers behave comparably; after a few years in the classroom, however, those hired before the passage of PL 94-142 were more likely to leave than teachers hired after. By presenting fitted hazard functions, we need not struggle to capture these effects using abstract scaling factors and percentage increases that ignore these important interactions with time.

Fitted survivor functions and estimated median lifetimes can also be reconstructed from the fitted hazard profiles, although we believe that fitted hazard profiles are more informative for identifying when events occur (bottom panel of Figure 6). Comparison of the two survivor curves supports our earlier conclusion: The fitted survivor functions are almost identical for Years 1 through 5 and diverge thereafter. Notice that it is more difficult to discern differences between the fitted survivor profiles than between the fitted hazard profiles because the survivor function cumulates the risks. Even the usually helpful summary statistic of estimated median lifetime can mislead. Because of the near coincidence in hazard during teachers' early years on the job, the estimated median lifetimes for the two cohorts are almost identical (6.3 and 6.9), failing to capture the critical late career differences in risk.

Discussion

Survival analysis allows researchers to answer research questions about whether and if so, when critical events occur. The method is powerful, flexible, and applicable to many research questions arising in psychology. Although researchers are exploring the utility of these methods, we believe that many other questions about response time, waiting time, career duration, recidivism, relapse, and time to life events remain unasked and unanswered because researchers have yet to learn how to use this more sensitive analytic tool. Nor do these easy-to-implement methods require an initial investment in dedicated computer software.

To encourage the application of survival analysis in psychological research, we conclude with *20 guidelines for good survival analysis:*

1. *Gather longitudinal data on representative samples from well-defined target populations.* Do not exclude censored cases. Define your target population using delimiters unrelated to time.
2. *Define event states precisely.* Clearly describe the behaviors, responses, or scores that define each state.
3. *Identify an appropriate beginning to time.* Imprecise start times lead to imprecise event times. The clock may not always start at birth; other times may be just as good (e.g., date of diagnosis, randomization, treatment).
4. *Gather data for a long enough time so that more than half the sample will experience the target event.* The longer the study, the less censoring and the more powerful the study.
5. *If you can only collect data on a few occasions, take measurements more frequently during periods of high hazard.* Continuous time is best, but discrete time can be almost as good.
6. *When possible, collect data prospectively.* Retrospective data are subject to recall errors.
7. *Minimize and model attrition.* The longer the data collection period, the greater the attrition. Individuals lost to follow-up have censored event times, but they can differ systematically from continuing participants. Use sensitivity analysis to evaluate the impact of attrition.
8. *With repeated events, record the spell number and examine its effect.* The natural course of the first spell can differ from that of second and subsequent spells.
9. *Follow participants for longer periods to increase statistical power.* Doubling the length of follow-up can give the same statistical power with one-third fewer people.
10. *Use the survivor function to describe the cumulative probability that an event will occur on each of several successive occasions.* The survivor function incorporates both censored and uncensored cases. It can be summarized easily by the estimated median lifetime.
11. *Use the hazard function as a sensitive lens to detect when the event of interest is most likely to occur.* Hazard is high when the slope of the survivor function declines precipitously.
12. *Perform exploratory analyses using sample hazard and survivor profiles computed within groups.* Choose the groups based on characteristics of substantive interest (e.g., stratify by the predictors).
13. *Build statistical models of the hazard profile.* Hazard modeling should be used to explore the effects of several predictors simultaneously, to evaluate interactions among predictors, and to make inferences about the population.
14. *Do not ignore the shape of the baseline hazard profile.* Baseline hazard describes the overall level of risk and reveals the main effect of time.
15. *Always check the tenability of the proportional hazards assumption.* Violations of the assumption are commonplace and can dramatically affect parameter estimate interpretation.
16. *Include both time-invariant and time-varying predictors in hazard models.* Many interesting predictors vary over time; incorporate this variation into hazard models.

17. *Check interactions between time and the other predictors.* The effect of a predictor can vary over time; when it does, the proportional hazards assumption is violated.

18. *Consider a competing-risks formulation.* Lifetimes can be terminated by different competing events.

19. *Try discrete-time survival analysis.* It requires no dedicated software. It is simple to apply, easily incorporates time-varying predictors, and facilitates the estimation of the baseline hazard profile.

20. *Use fitted hazard and survivor functions to display effects of key predictors.* One picture is worth a thousand numbers.

Researchers rarely ask questions that they do not know how to answer. We believe that many psychologists interested in the timing of events have altered their research questions because they did not know how to analyze data from censored and noncensored observations simultaneously. We hope that our presentation of survival analysis will help researchers frame their questions appropriately, and provide them with strategies for answering those questions as simply and directly as possible.

References

Adler, I., & Kandel, D. B. (1983). Adolescence in France and Israel: Application of survival analysis to cross-sectional data. *Social Forces, 62,* 375–397.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 61–98). San Francisco: Jossey-Bass.

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data* (Sage University paper series on quantitative applications in the social sciences, Number 07-046). Beverly Hills, CA: Sage.

Allison, P. D., & Liker, J. K. (1982). Analyzing sequential categorical data on dyadic interaction: A comment on Gottman. *Psychological Bulletin, 91,* 393–403.

Amenson, C. S., & Lewinsohn, P. M. (1981). An investigation into the observed sex difference in prevalence of unipolar depression. *Journal of Abnormal Psychology, 90,* 1–13.

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *Journal of the American Statistical Association, 83,* 204–212.

Baltes, P. B., & Nesselroade, J. R. (1972). Cultural change and adolescent development: An application of longitudinal sequences. *Developmental Psychology, 7,* 244–256.

Benfari, R. C., & Eaker, E. (1984). Cigarette smoking outcomes at four years of follow-up, psychosocial factors, and reactions to group interventions. *Journal of Clinical Psychology, 40,* 1089–1097.

Berk, R. A., & Sherman, L. W. (1985). Data collection strategies in the Minneapolis domestic assault experiment. In L. Burstein, H. E. Freeman, & P. H. Rossi (Eds.), *Collecting evaluation data: Problems and solutions* (pp. 35–48). Beverly Hills, CA: Sage.

Berk, R. A., & Sherman, L. W. (1988). Police responses to family violence incidents: An analysis of an experimental design with incomplete randomization. *Journal of the American Statistical Association, 83,* 70–76.

Bernstein, D., & Lagakos, S. W. (1978). Sample size and power determination for stratified clinical trials. *Journal of Statistical Computing and Simulation, 8,* 65–73.

Blossfeld, H. P., Hamerle, A., & Mayer, K. U. (1989). *Event history analysis: Statistical theory and application in the social sciences.* Hillsdale, NJ: Erlbaum.

Bloxom, B. (1984). Estimating response time hazard functions: An exposition and extension. *Journal of Mathematical Psychology, 28,* 401–420.

Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika, 50,* 383–397.

Blumstein, A., & Cohen, J. (1987). Characterizing criminal careers. *Science, 237,* 985–991.

Bolger, N., Downey, G., Walker, E., & Steininger, P. (1989). The onset of suicide ideation in childhood and adolescence. *Journal of Youth and Adolescence, 18,* 175–189.

Bradburn, N. (1983). Response effects. In P. H. Rossi, J. D. Wright, & A. A. Anderson (Eds.), *Handbook of survey research* (pp. 289–328). San Diego, CA: Academic Press.

Brownell, K. D., Marlatt, G. A., Lichtenstein, E., & Wilson, G. T. (1986). Understanding and preventing relapse. *American Psychologist, 41,* 765–782.

Burke, K. C., Burke, J. D., Regier, D. A., & Rae, D. S. (1990). Age at onset of selected mental disorders in five community populations. *Archives of General Psychiatry, 47,* 511–518.

Campbell, R. T., Mutran, E., & Parker, R. N. (1987). Longitudinal design and longitudinal analysis: A comparison of three approaches. *Research on Aging, 8,* 480–504.

Coelho, R. J. (1984). Self-efficacy and cessation of smoking. *Psychological Reports, 54,* 309–310.

Cohen, J. (1990). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, S., & Lichtenstein, E. (1990). Partner behaviors that support quitting smoking. *Journal of Consulting and Clinical Psychology, 58,* 304–309.

Condiotte, M. M., & Lichtenstein, E. (1981). Self-efficacy and relapse in smoking cessation programs. *Journal of Consulting and Clinical Psychology, 49,* 648–658.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, 34,* 187–202.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data.* London: Chapman & Hall.

Crider, D. M., Willits, F. K., & Bealer, R. C. (1971). Tracking respondents in longitudinal surveys. *Public Opinion Quarterly, 35,* 613–620.

Crider, D. M., Willits, F. K., & Bealer, R. C. (1973). Panel studies: Some practical problems. *Sociological Methods and Research, 2,* 3–19.

Diamond, I. D., McDonald, J. W., & Shah, I. H. (1986). Proportional hazards models for current status data: Application to the study of differentials in age at weaning in Pakistan. *Demography, 23,* 607–620.

Diekmann, A., & Mitter, P. (1983). The sickle-hypothesis: A time dependent poisson model with applications to deviant behavior. *Journal of Mathematical Sociology, 9,* 85–101.

Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials: A review. *Statistics in Medicine, 3,* 199–214.

Dornbusch, S. M., Carlsmith, J. M., Gross, R. T., Martin, J. A., Jennings, D., Rosenberg, A., & Duke, P. (1981). Sexual development, age, and dating: A comparison of biological and social influences upon one set of behaviors. *Child Development, 52,* 179–185.

Drass, K. A. (1986). The effect of gender identity on conversation. *Social Psychology Quarterly, 49,* 294–301.

Dupont, W. D., & Plummer, W. D., Jr. (1990). Power and sample size calculations: A review and computer program. *Controlled Clinical Trials, 11,* 116–128.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association, 83,* 414–425.

Ehrenberg, A. S. C. (1982). *A primer in data reduction.* New York: Wiley.

Farrington, D. P., Gallagher, B., Morley, L., St. Ledger, R. J., & West, D. J. (1990). Minimizing attrition in longitudinal research: Methods of tracing and securing cooperation in a 24-year follow-up study. In D. Magnusson & L. R. Bergman (Eds.), *Data quality in longitudinal research* (pp. 122–147). New York: Cambridge University Press.

Farrington, D. P., Ohlin, L. E., & Wilson, J. Q. (1986). *Understanding and controlling crime: Toward a new research strategy.* New York: Springer-Verlag.

Featherman, D. L., & Lerner, R. M. (1985). Ontogenesis and sociogenesis: Problematics for theory and research about development and socialization across the lifespan. *American Sociological Review, 50,* 659–676.

Felmlee, D., & Eder, D. (1983). Contextual effects in the classroom: The impact of ability groups on student attention. *Sociology of Education, 56,* 77–87.

Felmlee, D., Eder, D., Tsui, W.-Y. (1985). Peer influence on classroom attention. *Social Psychology Quarterly, 48,* 215–226.

Fergusson, D. M., Horwood, L. J., & Dimond, M. E. (1985). A survival analysis of childhood family history. *Journal of Marriage and the Family, 47,* 287–295.

Fergusson, D. M., Horwood, L. J., & Shannon, F. T. (1984). A proportional hazards model of family breakdown. *Journal of Marriage and the Family, 46,* 539–549.

Fichman, M. (1988). Motivational consequences of absence and attendance: Proportional hazard estimation of a dynamic motivation model. *Journal of Applied Psychology, 73,* 119–134.

Fichman, M. (1989). Attendence makes the heart grow fonder: A hazard rate approach to modeling attendence. *Journal of Applied Psychology, 74,* 325–335.

Flinn, C. J., & Heckman, J. J. (1982). New methods for analyzing individual event histories. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 99–140). San Francisco, CA: Jossey-Bass.

Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the log rank test. *Statistics in Medicine, 1,* 121–129.

Furby, L., Weinrott, M. R., & Blackshaw, L. (1989). Sex offender recidivism: A review. *Psychological Bulletin, 105,* 3–30.

Gallagher-Thompson, D., Hanley-Peterson, P., Thompson, L. W. (1990). Maintenance of gains versus relapse following brief psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 58,* 371–374.

Gardner, W., & Griffin, W. A. (1989). Methods for the analysis of parallel streams of continuously recorded social behaviors. *Psychological Bulletin, 105,* 446–455.

Gibb, G. D., & Millard, R. J. (1981). Research on repeated abortion: State of the field: 1973–1979. *Psychological Reports, 48,* 415–424.

Greenhouse, J. B., Stangl, D., & Bromberg, J. (1989). An introduction to survival analysis methods for analysis of clinical trial data. *Journal of Consulting and Clinical Psychology, 57,* 536–544.

Grey, C., Osborn, E., & Reznikoff, M. (1986). Psychosocial factors in outcome in two opiate addictions. *Journal of Clinical Psychology, 42,* 185–189.

Griffin, W. A., & Gardner, W. (1989). Analysis of behavioral durations in observational studies of social interactions. *Psychological Bulletin, 106,* 497–502.

Gross, A. J., & Clark, V. A. (1975). *Survival distributions.* New York: Wiley.

Guilkey, D. K., & Rindfuss, R. R. (1987). Logistic regression multivariate life tables. *Sociological Methods and Research, 16,* 276–300.

Hachen, D. S., Jr. (1988). The competing risks model: A method for analyzing processes with multiple types of events. *Sociological Methods and Research, 17,* 21–54.

Hagan, J., & Zatz, M. S. (1985). The social organization of criminal justice processing: An event history analysis. *Social Science Research, 14,* 103–125.

Hall, S. M., Havassy, B. E., & Wasserman, D. A. (1990). Commitment to abstinence and acute stress in relapse to alcohol, opiates, and nicotine. *Journal of Consulting and Clinical Psychology, 58,* 175–181.

Hall, S. M., Rugg, D., Tunstall, C., & Jones, R. T. (1984). Preventing relapse to cigarette smoking by behavioral skill training. *Journal of Consulting and Clinical Psychology, 52,* 372–382.

Harrell, F. E., & Lee, K. L. (1990). Verifying assumptions of the Cox proportional hazards model. In *SAS users group international conference guide.* Cary, NC: SAS Institute.

Harrison, D. A., & Hulin, C. L. (1989). Investigations of absenteeism: Using event history models to study the absence-taking process. *Journal of Applied Psychology, 74,* 300–316.

Heckert, A., & Teachman, J. D. (1985). Religious factors in the timing of second births. *Journal of Marriage and the Family, 47,* 361–367.

Heckman, J., & Singer, B. (Eds.). (1985). *Longitudinal analysis of labor market data.* New York: Cambridge University Press.

Hembroff, L. A., & Myers, D. E. (1984). Status characteristics: Degrees of task relevance and decision processes. *Social Psychology Quarterly, 47,* 337–346.

Hogan, D. P. (1984). Cohort comparisons in the timing of life events. *Developmental Review, 4,* 289–310.

Hunt, W. A., Barnett, W., & Branch, L. G. (1971). Relapse rates in addiction programs. *Journal of Clinical Psychology, 27,* 455–456.

Hunt, W. A., & Bespalec, D. A. (1974a). An evaluation of current methods of modifying smoking behavior. *Journal of Clinical Psychology, 30,* 431–438.

Hunt, W. A., & Bespalec, D. A. (1974b). Relapse rates after treatment for heroin addiction. *Journal of Community Psychology, 2,* 85–87.

Hunt, W. A., & General, W. R. (1973). Relapse rates after treatment for alcoholism. *Journal of Community Psychology, 1,* 66–68.

Hunt, W. A., & Matarazzo, J. D. (1970). Habit mechanisms in smoking. In W. A. Hunt (Ed.), *Learning mechanisms in smoking* (pp. 65–90). Chicago: Aldine.

Hutchison, D. (1988a). Event history and survival analysis in the social sciences, I: Background and introduction. *Quality and Quantity, 22,* 203–219.

Hutchison, D. (1988b). Event history and survival analysis in the social sciences, II: Advanced applications and recent developments. *Quality and Quantity, 22,* 255–278.

Johnson, D. (1988). Panel analysis in family studies. *Journal of Marriage and the Family, 50,* 949–955.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data.* New York: Wiley.

Kandel, D. B., & Yamaguchi, K. (1987). Job mobility and drug use: An event history analysis. *American Journal of Sociology, 92,* 836–878.

Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of Economic Literature, 26,* 646–679.

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods.* Belmont, CA: Lifetime Learning Publications.

Klerman, G. L. (1978). Long-term maintenance of affective disorders. In M. A. Lipton, A. DiMascio, & K. Killam (Eds.), *Psychopharmacology: A generation of progress.* (pp. 1303–1311). New York: Raven Press.

Kraemer, H. C., & Theimann, S. (1988). *How many subjects?* Beverly Hills, CA: Sage.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled clinical trials, 2,* 93–113.

Lachin, J. M., & Foulkes, M. A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance and stratification. *Biometrics, 42,* 507–519.

Laird, N., & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association, 76,* 231–240.

Lavori, P. W., Keller, M. B., & Klerman, G. L. (1984). Relapse in affective disorders: A reanalysis of the literature using life-table methods. *Journal of Psychiatric Research, 18,* 13–25.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data.* New York: Wiley.

Lee, E. T. (1980). *Statistical methods for survival data analysis.* Belmont, CA: Lifetime Learning Publications.

Lessler, J., Tourangeau, R., & Salter, W. (1989). *Questionnaire design in the cognitive research laboratory: Results of an experimental prototype.* (Series 6, No. 1). Washington, DC: National Center for Health Statistics.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design.* Cambridge, MA: Harvard University Press.

Lilienfeld, A. M., & Lilienfeld, D. E. (1980). *Foundations of epidemiology* (2nd ed.). New York: Oxford University Press.

Litman, G. K., Eiser, J. R., & Taylor, C. (1979). Dependence, relapse and extinction: A theoretical critique and a behavioral examination. *Journal of Clinical Psychology, 35,* 192–199.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Makuch, R. W., & Simon, R. M. (1982). Sample size requirements for comparing time-to-failure among k treatment groups. *Journal of Chronic Diseases, 35,* 861–867.

Marini, M. M. (1987). Measuring the process of role change during the transition to adulthood. *Social Science Research, 16,* 1–38.

Marlatt, G. A., Curry, S., & Gordon, J. R. (1988). A longitudinal analysis of unaided smoking cessation. *Journal of Consulting and Clinical Psychology, 55,* 715–720.

Mausner, J. S., & Bahn, A. K. (1974). *Epidemiology.* Philadelphia: W. B. Saunders.

McFall, R. M. (1978). Smoking-cessation research. *Journal of Consulting and Clinical Psychology, 46,* 703–712.

McLanahan, S. (1988). Family structure and dependency: Early transitions to female household headship. *Demography, 25,* 1–16.

Means, B., Nigam, A., Zarrow, M., Loftus, E. F., & Donaldson, M. S. (1989). *Autobiographical memory for health-related events: Enhanced memory for recurring incidents.* (Series 6, No. 2). Washington, DC: National Center for Health Statistics.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 103,* 156–166.

Miller, R. G. (1981). *Survival analysis.* New York: Wiley.

Milner, J. L. (1987). An ecological perspective on duration of foster care. *Child Welfare, 66,* 113–123.

Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin, 86,* 493–522.

Moen, P., Dempster-McClain, D., & Williams, R. M., Jr. (1989). Social integration and longevity: An event history analysis of women's roles and resilience. *American Sociological Review, 54,* 635–647.

Morgan, S. P., Lye, D. N., & Condran, G. A. (1988). Sons, daughters, and the risk of marital disruption. *American Journal of Sociology, 94,* 110–129.

Morita, J. G., Lee, T. W., & Mowday, R. T. (1989). Introducing survival analysis to organizational researchers: A selected application to turnover research. *Journal of Applied Psychology, 74,* 280–292.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Moussa, M. A. A. (1988). Planning the size of survival time clinical trials with allowance for stratification. *Statistics in Medicine, 7,* 559–569.

Murnane, R. J., Singer, J. D., & Willett, J. B. (1988). The career paths of teachers: Implications for teacher supply and methodological lessons for research. *Educational Researcher, 17,* 22–30.

Murnane, R. J., Singer, J. D., & Willett, J. B. (1989). The influences of salaries and "opportunity costs" on teachers' career choices: Evidence from North Carolina. *Harvard Educational Review, 59,* 325–346.

Murnane, R. J., Singer, J. D., Willett, J. B., Kemple, J. J., & Olsen, R. J. (1991). *Who will teach?: Policies that matter.* Cambridge, MA: Harvard University Press.

Murphy, M. (1990). Minimizing attrition in longitudinal studies: Means or end? In D. Magnusson & L. R. Bergman (Eds.), *Data quality in longitudinal research* (pp. 122–147). New York: Cambridge University Press.

Nathan, P. E., & Lansky, O. (1978). Common methodological problems in research on the addictions. *Journal of Consulting and Clinical Psychology, 46,* 713–726.

Neter, J., & Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association, 59,* 18–55.

Oldridge, N. B., Donner, A. P., Buck, C. W., Jones, N. L., Andrew, G. M., Parker, J. O., Cunningham, D. A., Kavanaugh, T., Rechnitzer, P. A., & Sutton, J. R. (1983). Predictors of dropout from cardiac exercise rehabilitation: Ontario Exercise–Heart Collaborative Study. *American Journal of Cardiology, 51,* 70–74.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., & Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: Introduction and design. *British Journal of Cancer, 34,* 585–612.

Prien, R. F., Kupfer, D. J., Mansky, P. A., Small, J. G., Tuason, V. B., Voss, C. B., & Johnson, W. E. (1984). Drug therapy in the prevention of recurrences in unipolar and bipolar affective disorders. *Archives of General Psychiatry, 41,* 1096–1104.

Rice, J., Reich, T., Andreasen, N. C., Endicott, J., Van Eerdewegh, Fishman, Hirschfeld, & Klerman. (1987). The familial transition of bipolar illness. *Archives of General Psychiatry, 44,* 441–447.

Rubinstein, L. V., Gail, M. H., & Santner, T. J. (1981). Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases, 34,* 469–479.

Ryan, J., & Dent, O. (1984). An introduction to survival analysis: Factors influencing the duration of breastfeeding. *Australian and New Zealand Journal of Sociology, 20,* 183–196.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64,* 92–107.

Schmidt, P., & Witte, A. D. (1988). *Predicting recidivism using survival models.* New York: Springer-Verlag.

Schoenfeld, D. A., & Richter, J. R. (1982). Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics, 38,* 163–170.

Seltzer, M. M., Seltzer, G. B., & Sherwood, C. C. (1982). Comparison of community adjustment of older vs. younger mentally retarded adults. *American Journal of Mental Deficiency, 87,* 9–13.

Shapiro, D. R., Quitkin, F. M., & Fleiss, J. L. (1989). Response to maintenence therapy in bipolar illness: Effect of index episode. *Archives of General Psychiatry, 46,* 401–405.

Sherman, L. W., & Berk, R. A. (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review, 49,* 261–271.

Shiffman, S. (1982). Relapse following smoking cessation: A situational analysis. *Journal of Consulting and Clinical Psychology, 50,* 71–86.

Simons, A. D., Murphy, G. E., Levine, J. L., & Wetzel, R. D. (1986). Cognitive therapy and pharmacotherapy for depression. *Archives of General Psychiatry, 43,* 43–48.

Singer, J. D. (in press). Are special educators' career paths special? *Exceptional Children.*

Singer, J. D., & Butler, J. A. (1987). The Education for All Handicapped Children Act: Schools as Agents of Social Reform. *Harvard Educational Review, 57,* 125–152.

Singer, J. D., & Willett, J. B. (1988). Detecting involuntary layoffs in teacher survival data: The year of leaving dangerously. *Educational Evaluation and Policy Analysis, 10,* 212–224.

Singer, J. D., & Willett, J. B. (in press). Using discrete-time survival analysis in educational research. *Journal of Educational Statistics.*

Soothill, K. L., & Gibbens, T. C. N. (1978). Recidivism of sexual offenders: A re-appraisal. *British Journal of Criminology, 18,* 267–276.

Sorenson, A. (1977). Estimating rates from retrospective questions. In D. R. Heise (Ed.), *Sociological methodology* (pp. 209–222). San Francisco: Jossey-Bass.

Sorensen, A., & Tuma, N. B. (1981). Labor market structures and job mobility. *Research in Social Stratification and Mobility, 1,* 67–94.

Stancer, H. C., Persad, E., Wagener, D. K., & Jorma, T. (1987). Evidence for homogeneity of major depression and bipolar affective disorder. *Journal of Psychiatric Research, 21,* 37–53.

Stevens, V. J., & Hollis, J. F. (1989). Preventing smoking relapse using an individually tailored skills training technique. *Journal of Consulting and Clinical Psychology, 57,* 420–424.

Sudman, S., & Bradburn, N. (1974). *Response effects in surveys: A review and synthesis.* Chicago: Aldine.

Sudman, S., & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Sutton, S. R. (1979). Interpreting relapse curves. *Journal of Consulting and Clinical Psychology, 47,* 96–98.

Taber, M. A., & Proch, K. (1987). Placement stability for adolescents in foster care: Findings from a program experiment. *Child Welfare, 66,* 433–445.

Teachman, J. D. (1982). Methodological issues is the analysis of family formation and dissolution. *Journal of Marriage and the Family, 44,* 1037–1053.

Teachman, J. D., & Polonko, K. A. (1984). Out of sequence: The timing of marriage following a premarital birth. *Social Forces, 63,* 245–260.

Telles, J. L., & Spreat, S. (1985). Client progress toward referral for discharge to the community: An illustration of an evaluation technique. *Evaluation and Program Planning, 8,* 155–160.

Tognetti, J. (1990). *The role of employment versus the role of attitudes and beliefs in maternal infant feeding behavior in Bangkok, Thailand.* Unpublished doctoral thesis, Harvard University, Graduate School of Education.

Trussel, J., & Hammerslough, C. (1983). A hazards-model analysis of the covariates of infant and child mortality in Sri Lanka. *Demography, 20,* 1–26.

Tuma, N. B. (1986). Invoking RATE (3rd ed.). Menlo Park, CA: SRI International.

Tuma, N. B., & Hannan, M. T. (1979). Approaches to the censoring problem in analysis of event histories. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 1–60). San Francisco: Jossey-Bass.

Tuma, N. B., & Hannan, M. T. (1984). *Social dynamics: Models and methods.* San Diego, CA: Academic Press.

Tuma, N. B., Hannan, M. T., & Groeneveld, L. P. (1977). Dynamic analysis of event histories. *American Journal of Sociology, 84,* 820–854.

Turnbull, B. W. (1974). Non-parametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association, 69,* 169–173.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, 38,* 290–295.

Wainer, H. (1977). Speed vs. reaction time as a measure of cognitive performance. *Memory and Cognition, 5,* 278–280.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology.* San Francisco: Freeman.

Willett, J. B., & Singer, J. D. (1988). *Doing data analysis with proportional hazards models: Model building, interpretation and diagnosis.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 292 684)

Willett, J. B., & Singer, J. D. (1989). Two types of question about time: Methodological issues in the analysis of teacher career path data. *International Journal of Educational Research, 13,* 421–437.

Willett, J. B., & Singer, J. D. (1991). How long did it take . . . ?: Using survival analysis in psychological research. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 309–326). Washington, DC: American Psychological Association.

Willett, J. B., & Singer, J. D. (in press). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research.*

Yamaguchi, K., & Kandel, D. B. (1985a). Dynamic relationships between premarital cohabitation and illicit drug use: An event-history analysis of role selection and role socialization. *American Sociological Review, 50,* 530–546.

Yamaguchi, K., & Kandel, D. B. (1985b). On the resolution of role incompatibility. *American Journal of Sociology, 90,* 1284–1325.

Yamaguchi, K., & Kandel, D. B. (1987). Drug use and other determinants of premarital pregnancy and its outcome: A dynamic analysis of competing life events. *Journal of Marriage and the Family, 49,* 257–270.

Zatz, M. S. (1985). Los Cholos: Legal processing of Chicano gang members. *Social Problems, 33,* 13–30.

Zis, A. P., & Goodwin, F. K. (1979). Major affective disorder as a recurrent illness: A critical review. *Archives of General Psychiatry, 36,* 835–839.

Zito, J. M., Craig, T. T., Wanderling, J., & Siegel, C. (1987). Pharmacoepidemiology in 136 hospitalized schizophrenic patients. *American Journal of Psychiatry, 144,* 778–782.

Received September 10, 1990
Revision received September 28, 1990
Accepted March 19, 1991 ■