Epidemiologic Reviews
Vol. 20, No. 1

Printed in U.S.A.

# Evolution of the Cohort Study

Jonathan M. Samet and Alvaro Muñoz

## INTRODUCTION

The occurrence of events over time unifies epidemiologic research. Regardless of the study hypothesis and design, the disease-causing actions of exposures and modifying factors are formulated as antecedent to the occurrence of the outcome. All study designs inherently acknowledge time and represent alternative approaches for sampling populations as exposed and nonexposed persons develop disease over time.

The cohort design explicitly incorporates the passage of time. In cohort studies, participants are followed over an interval defined by the study's beginning and end, and observations are made on outcome measures of interest: death, incidence of disease, change in a biologic measure, or health status. The study's purpose may be focused—to test a specific (or general) hypothesis—to gather data for descriptive purposes or to facilitate the testing of multiple hypotheses concerning disease. During the follow-up experience of participants in a cohort study, the factors determining the health of the participants may be continually changing as the participants age and exposures to environmental agents start, stop, increase, or decrease (figure 1). The dynamic nature of many risk factors and their relations in time to disease occurrence can only be captured in the cohort design; this temporal interplay is inherently absent from cross-sectional data and often investigated only with difficulty using the case-control design (1).

Epidemiologists may be challenged in investigating the relation between multiple risk factors, some changing in time, and disease occurrence, as displayed schematically in figure 1. During the follow-up of a cohort, participants age, temporal trends may affect the participants, and exposures to risk factors of primary interest may change, as may potential confounding and modifying factors. Consider, for example, a prospective cohort study of cigarette smoking and lung cancer. Several determinants of lung cancer risk would change as the cohort was followed, including the participants' age (relevant because of the rise of lung cancer incidence with age), the cumulative amount smoked, the duration of smoking, and perhaps the characteristics of the cigarettes themselves. Some participants might choose to reduce the number of cigarettes smoked or to stop smoking, and some might start smoking. The study's data might be further complicated by a temporal trend in the validity of smoking information, if, for example, the social acceptability of smoking declined during the follow-up interval and participants began to underreport the extent of their smoking. Possible modifying or confounding factors (e.g., occupational exposures) might also change over time. An optimal study design would incorporate periodic assessment of smoking by the participants, as assessing smoking only at the study's start would not capture the temporally dynamic nature of the exposure.

Many examples of contemporary cohort studies show that the challenge of temporally varying exposures and disease risk can now be satisfactorily met using modern epidemiologic approaches for the design and analysis of cohort studies. In parallel with the increasing design sophistication of many cohort studies, new biostatistical methods now make possible longitudinal analyses that can incorporate temporally-varying exposures. Application of these analytic methods has been facilitated by the availability of hardware and software, which make possible analyses that could not have been contemplated one or two decades previously.

In spite of the central role of the cohort design in epidemiologic research, it has been the focus of few monographs. The book by Breslow and Day (2), published in 1987, represents a pioneering synthesis on design and analysis of cohort studies on cancer. A number of statistical texts address analysis of longitudinal data (3–5). The history of the cohort study was comprehensively addressed in a 1988 review by

Received for publication December 29, 1997, and accepted for publication June 12, 1998.

Abbreviations: AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus.

From the Department of Epidemiology, School of Hygiene and Public Health, The Johns Hopkins University, Baltimore, MD.

Reprint requests to Dr. Jonathan M. Samet, Professor and Chairman, Department of Epidemiology, School of Hygiene and Public Health, The Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205.

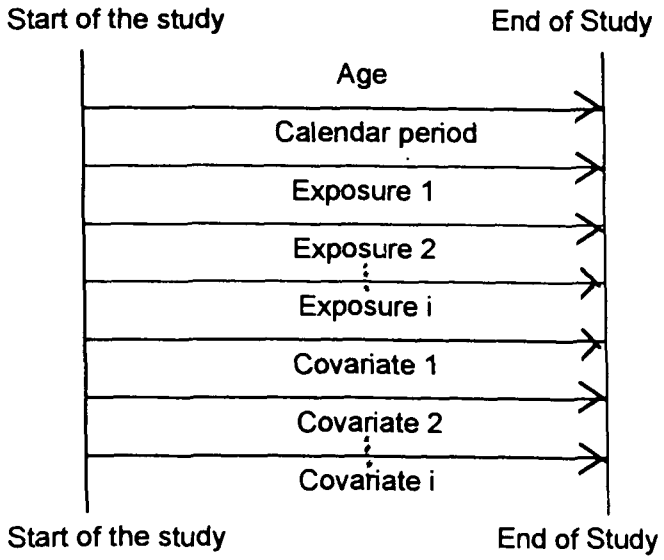


**FIGURE 1.** The multiple dimensions of time in a cohort study.

Liddell (6) and in papers published from a 1983 American Cancer Society workshop on cohort studies (7). In this presentation, which introduces this special volume of *Epidemiologic Reviews* on cohort studies, we first discuss the terminology, the definitions, and the evolution of the cohort design. We then review the analytical issues in cohort studies, outlining the overall objectives of analytical methods, followed by an historical perspective on their evolution. We close the paper by illustrating the power of the cohort design in epidemiology: the selected study has been key in providing comprehensive data on the epidemiology of the acquired immunodeficiency syndrome (AIDS).

## TERMINOLOGY AND DEFINITIONS

While "cohort study" has been variably defined, all definitions incorporate the concept that nonexposed and exposed, or variably exposed, individuals are observed over time for the outcome(s) of interest. The word "cohort" has its origin in the Latin *cohors*, referring to warriors and the notion of a group of persons proceeding together in time. In the past, various terms, including prospective, follow-up, and longitudinal, were sometimes used to reflect the temporal sequence of exposure and disease in study participants, compared with the retrospective sequence of case-control studies (8, 9). These terms have largely been abandoned in the epidemiologic literature, and the term "cohort study" is used most often.

Cohort studies are also designated by the timing of data collection, either prospectively or retrospectively, in the investigator's time. Studies collecting data on events that have already occurred have been labeled as historical, retrospective, and nonconcurrent. The 1995 edition of *A Dictionary of Epidemiology* (10) offers a definition for "historical cohort study" and the synonyms of "historical prospective study, nonconcurrent prospective study, and prospective study in retrospect." To describe these types of studies, the labels retrospective cohort study and nonconcurrent cohort study are widely used at present. Among epidemiologists, there is now close to uniform use of the term prospective cohort study to refer to studies in which the investigators observe the events as they occur in time. Any lingering debate with regard to terminology should be set aside to avoid unneeded confusion.

## EVOLUTION OF THE COHORT DESIGN

The origins of the cohort study can be traced to the need for information on the length of life and the course of disease—such information being central in medical practice and public health and of fundamental interest to the general population. The earliest life tables, developed by Graunt and Halley from cross-sectional mortality data, were intended to project deaths with aging, inherently acknowledging the passage of time. In the 19th century, Farr advanced the use of life tables as an indicator of population health (11). The emergence of the modern insurance industry several centuries ago, created the still-existent profession of the actuarialist whose job is to project risks (12, 13). The experience of policyholders provided data for this purpose, and over a century ago insurance companies pooled data to gain more precise and specific descriptions of mortality among the insured. A report published in 1904, the *Specialized Mortality Investigation*, pooled the experience of policyholders in 34 of the largest companies in the United States and Canada for the years 1870 through 1899. The report gives tables listing 141,977 deaths along with expected numbers by classes of impairment, medical history, and occupation (14, 15). Singer (14) documents other early studies carried out by the insurance industry, directed, for example, at asthma and mortality, hypertension and mortality, and obesity and mortality.

At the turn of the century, tuberculosis was the leading cause of mortality in the United States. Treatment was accomplished in sanatoriums with a variety of therapies; exposure to outdoor air and sunshine was central in the management of the disease. Follow-up studies, using the design now known as the prospective cohort study, were carried out to describe the natural history of the disease in tuberculosis patients and the consequences of the therapies then in use (16, 17). In one of the earliest of these studies of tuberculosis, Brown and Pope (18) traced over 1,000 persons

discharged from the Adirondack Cottage Sanitarium (later to become the famed Trudeau Sanitarium). Brown and Pope used life-table methods and compared patient survival to Farr's English Life Table no. 3. Sartwell (16) noted that this study was the first application of the life-table method in a clinical follow-up study. A similar study was reported in 1910 by Elderton and Perry (19) who described mortality among patients discharged from two sanatoriums in England, calculating observed and expected deaths based on one of Farr's life tables.

Wade Hampton Frost made several methodological advances related to cohort studies in his investigations of tuberculosis. He has been credited with development of cohort analysis of vital statistics data, that is, the separation of age, period, and cohort effects by stratification of vital statistics or other data on these time-varying factors. This development arose from his report on age- and birth-cohort specific patterns of tuberculosis mortality in Massachusetts for the years 1880 through 1930 (20). In 1930, Andvord (21) had published a similar but overlooked analysis.

In describing the risk of tuberculosis in 132 black families in Kingsport, Tennessee, Frost (22) pioneered the use of the retrospective cohort design. He interviewed family members to reconstruct the household composition from the time of establishment and then retrospectively followed the household members to the present. Frost estimated person-years of "life-experience" and calculated the age-specific attack rates for tuberculosis. He also initiated a prospective cohort study, the Williamson County Tuberculosis Study; the study was started in 1931 and follow-up continued through 1955, long after Frost's death (23).

The more contemporary era of cohort research can be traced to the late 1940s and early 1950s when several landmark prospective cohort studies were implemented. Some of these studies continue today: the Framingham Study, the study of the Japanese atomic bomb survivors, the study of British physicians, and the Public Health Service study of Colorado Plateau uranium miners. These studies were distinguished by their size and the richness of the data collected and by the sustained follow-up of participants over decades. All were implemented to address pressing public health concerns: the causes of heart disease, the consequences of smoking, and the risk of radiation.

Dawber (24) has chronicled the origins of the Framingham Study, which was implemented in the late 1940s to address the rising occurrence of cardiovascular disease. The long-term success of the study can be attributed to the selection of a small and cooperative community, sustained support from the National Institutes of Health which maintained the study as an intramural project, and to the prescience of the original investigators who established rigorous and standardized protocols for data collection. Data were collected relevant to testing the principal extant hypotheses concerning etiology, which were listed at the study's beginning. As a result, much of our initial understanding of risk factors for cardiovascular diseases was based on evidence from this study. Supplementary studies of other diseases capitalized on the opportunity afforded by having the Framingham population under follow-up, and offspring of the original cohort have now been enrolled in a new cohort study that should be informative on familial factors affecting cardiovascular disease risk. The longitudinal data on multiple risk factors necessitated methodological advances, as appropriate multivariate methods had not been available. For example, Truett and colleagues (25) described application of discriminant analysis in a 1967 paper. They predicted 12-year probability of developing coronary heart disease based on levels of seven risk factors.

Another landmark cohort study, the investigation of the atomic bomb survivors in Hiroshima and Nagasaki, addressed the consequences of ionizing radiation exposure. This population, a group having a unique and instantaneous exposure, contrasts with the Framingham Study, a general population study of several diseases that tests multiple hypotheses. The two bombs were dropped in 1945, and investigation of the medical consequences began almost immediately thereafter (26). By 1946, a decision had been made to conduct long-term studies, and the Atomic Bomb Casualty Commission was established in 1947. In 1975, this organization was replaced by the Radiation Effects Research Foundation, still operating, which is funded jointly by the United States and Japan. Acute and chronic effects of radiation exposure were already known but there was little quantitative information available on the risks of radiation. Radiation doses from the blasts were eventually reconstructed and selected survivors were entered into a cohort study that included periodic medical examinations.

This study has become one of the principal sources of evidence on the cancer risks of acute radiation exposure; its findings are the underpinnings of radiation standards throughout the world. Soon after the blast, the occurrence of acute leukemia rose in the survivors, but after peaking around 1952, the excess began to decline. By 1960, excesses of solid tumors were noted and current studies by the Radiation Effects Research Foundation emphasize risks of these cancers. In addition to the study's contributions to the evidence on radiation risks, the challenges of the longitudinal data have prompted substantial methodolog-

ical work directed at such issues as the time- and age-dependence of radiation risk, the joint effect of radiation with other factors, and the consequences of measurement error (27).

In the early 1950s, another key study of a radiation-exposed special population was implemented: the US Public Health Service study of Colorado Plateau uranium miners (28). The participants comprised underground uranium miners exposed to radon progeny released from the uranium ore. Over 3,000 men were enrolled in this study by 1960, and follow-up, as of 1998, is continuing. Unlike the single exposure of the atomic bomb survivors, the uranium miners received continuing exposure throughout employment at a rate that depended on mining conditions. Consequently, both the cumulative exposures and the rates of exposure varied over follow-up. This study has been significant for its definitive demonstration of the cancer risk of radon progeny and for the methodological advances it has fostered on time-varying aspects of exposure and exposure rate (29). The dataset also includes information on smoking, and methods have been developed and applied for characterizing the joint effects of the two causes of lung cancer—radon and smoking (30–32).

A number of now well-known prospective cohort studies were implemented during the 1950s in follow-up of the initial observations from case-control studies of a remarkably strong association between cigarette smoking and lung cancer. In starting the study of British physicians, Doll and Hill (33) commented on the need for more credible, prospectively collected data in follow-up of the mistrusted retrospective data of the case-control studies. The prospective cohort study of British physicians, initiated in 1951, continues as of 1998. A 1994 report (34) provides the findings after 40 years of follow-up. The success of the study reflects the investigators' foresight in selecting a cooperative population that could be readily followed for mortality, and in using a simple, mailed questionnaire to periodically assess smoking by the participants. Other key investigations on smoking included the American Cancer Society's Nine-State Study of approximately 188,000 persons (35) and the study of US veterans (36). By 1964, findings from eight prospective cohort studies on smoking and disease were available for review by the Surgeon General's advisory committee (37).

Case and colleagues (38) reported the prototype retrospective cohort study in 1954. Following up on the hypothesis that aniline-based dyes increased risk for bladder cancer, Case et al. developed a roster of exposed workers in the United Kingdom from 1920 forward and identified bladder cancer cases and deaths. This study showed the feasibility of retrospective cohort studies when the needed records are available; this design was soon to be widely used for worker groups for whom records documenting employment and exposures were available.

Many cohort studies, both prospective and retrospective in timing, were implemented widely beginning in the 1950s and 1960s. These early studies often proved to be landmarks for particular diseases and exposures: Doll's study (39) of workers in an asbestos textile factory, the study of Selikoff et al. (40) of US insulation workers, the American Cancer Society's Cancer Prevention Study I (41), and the study of Fletcher et al. (42) of lung function in workers in London are examples. In conducting the Cancer Prevention Study I, the American Cancer Society used volunteers to enroll one million participants, establishing feasibility and demonstrating the strength of evidence from large cohort studies. Further general population studies of multiple chronic diseases, following the model of Framingham, were implemented in Tecumseh, Michigan (43); Evans County, Georgia (44); Alameda County, California (45); and Washington County, Maryland (46). In the Washington County study, serum was stored so that serologic markers could be examined as predictors of disease risk; this study model represents an early application of the research approach now referred to as "molecular epidemiology."

The current era of large, focused cohort studies dates to the 1970s. With deepening understanding of risk factors for chronic disease, cohort studies were typically designed to collect extensive information on exposures and to follow participants rigorously for outcomes. In the Tucson, Arizona, study of respiratory diseases, for example, participants visited a central clinic annually for an evaluation that included questionnaires, lung function testing, skin testing, phlebotomy, and other evaluations (47). In the United States, the National Institutes of Health has taken the lead in establishing multicenter prospective cohort studies, particularly in the area of cardiovascular disease—for example, the Atherosclerosis Risk in Communities Study, (48), the Cardiovascular Health Study (49), and the Strong Heart Study (50). These multisite studies gain external validity by drawing participants from communities across the United States. Data collection is standardized and data are accumulated, evaluated, and managed at central coordinating centers.

Opportunities for data linkage have now facilitated the conduct of cohort studies. Using record linkage approaches, researchers can match lists of exposed individuals for outcome against death indexes and disease registries. Pioneering cohort studies based on

this approach were conducted in Canada, where a mortality register of deaths back to 1950 has been available for matching and establishing vital status and cause of death (51). The National Health and Nutrition Examination Survey, conducted by the National Center for Health Statistics, has been given a longitudinal component by linkage against death certificates and additional follow-up data collection (52).

Contemporary reminders of the strength of evidence from cohort studies are abundant. The Nurses' Health Study, started in 1976 to investigate risks of oral contraceptives, has become one of the principal sources of observational data on diet and disease (53). Drawing on the lessons learned in the study of British physicians, the Nurses' Health Study incorporates a cooperative participant group familiar with completing questionnaires, a mailed approach for data collection, and active and passive follow-up for outcome. The investigators have even been able to collect biologic specimens from participants.

Shortly after the recognition of the disease now referred to as AIDS, plans were made for a cohort study to characterize the natural history of the disease and the determinants of prognosis. The Multicenter AIDS Cohort Study, implemented in 1984, comprised a cohort of close to 5,000 homosexual men in four cities (54). The implementation of the study antedated the identification of the causal virus, human immunodeficiency virus 1 (HIV-1), but the study included the collection of complete clinical data and the storage of blood specimens every 6 months. Over time, this repository of information and specimens has been repeatedly used to address determinants of risk for AIDS and prognostic factors. For example, a 1997 publication addresses viral load and the course of disease, an application unanticipated at the study's beginnings (55). In parallel to the evolution of cohort studies for cardiovascular diseases mentioned above, in 1994 a major cohort study on women, the Women's Interagency HIV Study, was assembled to characterize the natural history of HIV so that comparison and gender-specific inferences could be reached.

Methodological advances have been motivated by the complexity of analyzing data from cohort studies. Many of these advances, reviewed below, have come from collaborations between epidemiologists and biostatisticians, as they addressed challenging longitudinal data generated by cohort studies. A key advance in design was the development of sampling methods for efficiently assessing the relation between exposure and outcome. These designs are particularly valuable for relatively infrequent outcomes. The nested case-control study compares exposures of cases of the disease of interest with controls drawn from the remainder of the cohort at the time the case developed the disease (56, 57). In the case-cohort design, covariate data are fully developed for cases and for a random sample of the full cohort, drawn at the start of the study (58). When properly analyzed, these designs yield unbiased estimates of the relative risk, even though the full suite of covariate data needs to be developed only for a sample of the participants.

Cohort studies, of course, have potential limitations. The use of the retrospective design is possible only if historical data of adequate quality are available. The prospective design is successful only if adequate follow-up of participants can be maintained. Repeated data collection, often warranted for scientific purposes, may be constrained by feasibility concerns, costs, and participant burden. Bias that is differential over time may complicate the interpretation of findings of a cohort study. Information bias may vary in its effect over the cause of data collection due to the sometimes subtle drifting of the quality of data collection. Apparent time-dependent effects may occur as a result. Selection bias may also be differential over time from losses to follow-up (59). To date, there has been little systematic consideration of the time-specific biases that may affect cohort studies.

## OBJECTIVES OF ANALYTICAL METHODS FOR COHORT STUDIES

The analysis of data collected in cohort studies is determined by 1) the specification of the substantive question to be answered; 2) the consideration of the nature of the outcome or measure of disease occurrence (e.g., time-to-event, change in marker measured repeatedly); and 3) the nature of the exposures and covariates of concern and their relations with the outcome. An analytical model needs to be selected that is appropriate for the substantive question and for the nature of the outcome data, including the specific parameters quantifying the association between exposure and disease. Once a model is selected, the analyst uses statistical procedures for the extraction of information contained in the data at hand. Methods based on the likelihood principle are widely used for this purpose. Likelihood-based procedures estimate the unknown parameters with the values that make the data at hand the most likely to have been observed (i.e., maximum likelihood principle); they also permit the determination of how deviant the likelihood under the estimated parameters is from the likelihood of the data under the assumption that no relation exists between exposure and disease (60). A large deviance between the maximum likelihood based on the data and the likelihood under the null assumption of no exposure/disease association casts doubt on this null hypothesis.

*p*-Values have been widely used in epidemiology to quantify how unlikely the data observed are under the null hypothesis. The more unlikely, then the stronger is the doubt as to the validity of the null hypothesis. Most of the regression methods reviewed here are based on the maximum likelihood principle or an extension, e.g., quasi-likelihood.

Data for a cohort study include, as a minimum, the follow-up experience of each participant and the status of each participant with regard to the occurrence of the event(s) of interest. Even the most basic cohort data include multiple dimensions of time: calendar time, age, and time on study. Exposures may vary over time as may confounding and modifying factors (figure 1). The contemporary study may thus include multiple time-dependent variables, and even the outcome measure may vary over time or occur multiple times, as in the example of some infections. The analytical challenges of such data have been addressed by innovative biostatistical methods developed mostly during the last two decades.

Common to other epidemiologic study designs, the primary objectives of the analysis of data from cohort studies are 1) to summarize and 2) to compare. One fundamental measure of disease occurrence in a cohort study is the incidence rate. We summarize main features of cohort data by graphic and tabular displays of measures of location and scale. Summary measures of location include the mean and median, and summary measures of scale include the standard deviation and interquartile range, which may be summarized within exposure groups and compared across groups. The procedures that are linked to proper summarization are the extensive methods developed under the rubric of estimation in the statistical literature (61). Another useful, although frequently overlooked, procedure for summarizing is the graphic depiction of data, a method now strengthened by current hardware and software (62).

Comparison of the frequency of disease occurrence in exposed and unexposed individuals is a primary objective of cohort studies. An underlying and fundamental assumption for valid epidemiologic inference is that the exposed and unexposed groups are comparable with respect to any other factors that may explain the heterogeneity of disease occurrence and that are related to the exposure and the disease. In clinical trials, randomization leads to comparability, but in cohort (observational) studies, comparability needs to be achieved by design or in analysis by stratification and/or regression. Advances in statistical methods have greatly augmented our capacity to meet this primary analytic goal while fully considering the temporal structure of data.

## EVOLUTION OF METHODS FOR THE ANALYSIS OF COHORT STUDIES

In this section, we review the advances in analytical methods for the basic outcome measures in cohort studies: time-to-event and repeated measures of markers of disease progression. The past 20 years have witnessed the development and wide application of multivariate methods for the elucidation of factors explaining the variability of hazard of disease (i.e., survival analysis), and of trajectories of markers of disease progression measured repeatedly over time (i.e., longitudinal data analysis). For studies using time-to-event as the primary outcome, the longitudinal data on markers are treated as covariates, with a typical substantive question being the distribution of event-free times based on marker values (e.g., AIDS-free times according to amount of HIV in blood and level of immune deficiency (55). Conversely, for studies using change of markers as the primary outcome, the time-to-event is treated as a covariate, with a typical substantive question being the effect that the occurrence of an event (disease) has on the trajectory of a marker (e.g., loss of homeostasis of total T-cell count with the imminent onset of AIDS (63)). The close interrelations between time-to-event and repeated measures of markers in cohort studies have opened an active area of current research to combine the two sets of information into a unified framework.

Prior to the 1970s, analyses of cohort data were based primarily in life-table methods and stratified approaches for handling confounding and evaluating effect modification. Binary variables were the principal outcome measures of concern. The methods first applied to cohort data for multivariate analyses, discriminant analysis, and logistic regression, while appropriate for binary data, did not explicitly incorporate time. These methods, now known to be more appropriate for data that are cross-sectional in time, are not considered further in this review.

In 1972, D. R. Cox published a seminal paper on regression methods for time-to-event data (64), providing the basis for what is now widely known as proportional hazard regression models in survival analysis. This method has the strength of needing no assumption as to the form of the hazard of disease in the unexposed reference group. The hazards of other groups under different exposures are modeled as multiples of the hazard (relative hazards) of the reference group. Measures of relative hazards between groups at all times at which events occur are combined into an overall estimate of the relative hazard (3).

The 1970s also witnessed the full development of Poisson regression methods for the analysis of events-in-person-years data (2). These methods are particu-

larly useful for the analysis of trends and changes in incidence of disease over calendar time; they are of great utility for data in which a specific time origin is not well defined or not of interest. These methods are suitable for data obtained in cohort studies, which provide the number of events and the person-years at risk for the event(s) of interest. Cox and Poisson regression methods are closely linked and, in most cases, they give similar results.

The decade of the 1980s witnessed the development of methods for the analysis of markers of disease progression observed repeatedly for participants in cohort studies (e.g., forced expiratory volume, a lung function measure, in cohort studies of respiratory diseases; blood pressure in cohort studies of cardiovascular disease; and CD4 cell count in cohort studies of infectious diseases). The methods, developed for the analysis of levels of markers over time and of trajectories of change, are now widely used and defined as the methods for the analysis of longitudinal data (5).

Methods for the analysis of longitudinal data can be broadly classified into three groups: marginal, transition, and random effects models which address distinct epidemiologic questions. The marginal approach combines the multiple cross-sections corresponding to data collected at cohort study visits to provide the most efficient summary of the relations between prevalence of disease (binary outcome) or other mean response and the prevalence of exposure. In this approach, the longitudinal element is typically incorporated by including age or time since baseline as a covariate in a regression model. Approaches for the incorporation of the correlation between repeated measurements within individuals include parametric (65) and nonparametric methods, the latter handling the correlation as a nuisance (66).

Transition models regress current outcome on past values of the outcome, and on current and previous exposures. Classic Markovian models for binary outcome data were introduced to epidemiologists in 1979 (67), and applied, for example, in 1980, to a study of air pollution and asthma (68). Extensions for the continuous outcome were used for the study of the effect of cigarette smoking on respiratory function (69).

Random effects models allow each individual to have unique regression parameters (e.g., intercept and slope) according to components of variance, and provide direct averages of rates of change across individuals. Methods for random effects models have been provided for Gaussian outcomes (70), binary outcomes (71), and for event-in-person-years outcomes (72).

Another advance in methods for cohort studies during the 1980s was the extension of regression trees methodology to survival data (73, 74). Regression trees are extremely flexible for handling interactions and are very effective for communicating epidemiologic inferences to wide audiences. The primary concern is that their flexibility may result in inferences that are too specific to the data at hand and, therefore, are of limited generalizability.

The first quinquennium of the 1990s witnessed the dissemination of a unified framework for linear models (75) under which linear, logistic, Poisson, and many survival regression models could be viewed as specific cases of generalized linear models. Extension of this framework also allowed for relaxation of assumptions and triggered the development of quasi-likelihood methods. These new methods can account for different variance structures and can handle nuisance correlations using robust methods for appropriately estimating standard errors. In parallel to these advances, graphic procedures have been substantially improved with the availability of smoothing algorithms and the development of additive models that free regression models from the usual linear assumptions (76). These generalized additive models are especially useful for summarizing data but are somewhat limited for comparison and determination of measures of differences in a probabilistic framework.

During the early 1990s, methods were also developed that allowed for late or staggered entries into observation in a cohort study. Using these methods, the analyst can select the most appropriate time scale from a biomedical perspective, not only time-on-study, and can describe the occurrence of disease in person-years and thereby address incidence—the fundamental measure of disease—directly. The unit of analysis becomes individual-periods-at-risk, as opposed to the individual.

These new methods for late entries have been used for juxtaposition of incident and prevalent cohorts (77), and analysis of time-varying exposures whereby follow-up time is partitioned into as many individual periods as changes in exposure are recorded (78). They control for the survival bias that can be introduced by classifying persons as never or ever exposed. They also facilitate proper inferences regarding intermediate events that actually increase the hazard of the event but that may appear protective under improper analysis because the intermediate events occur only after some period of time. In another application, the methods for late entries can be used in evaluating effectiveness of therapies over time by considering calendar time as an external time-dependent exposure. The hazards can be compared for individuals who reach the same duration of time at risk in different calendar periods (79).

Beginning in the 1980s, Cox regression has been

widely used to estimate the relative hazard. It was initially applied to cancer clinical trials which have the objective of assessing efficacy of therapies in the setting of a high underlying hazard and, consequently, there was little interest in describing the underlying hazard. In this context, proportional hazard methods were ideal because the underlying hazard is allowed to be arbitrary. However, in cohort studies the underlying hazard itself is generally of interest, particularly in studies describing the natural history of diseases; for example, the hazard of AIDS at different intervals since infection with HIV. Regression under parametric models (e.g., lognormal) provides direct measures of the underlying hazard and, more importantly, of relative percentiles (78, 80, 81). Relative percentiles or relative times compare exposure groups according to the ratio of the times over which a given percent of individuals in the groups under different exposures develop the disease. These methods are also consonant with the renewed interest in quantifying the disease-free years at the population level that an intervention may produce (79).

The use of nested designs and the development of related analytic techniques is another major advance. The primary objective of nesting substudies within a cohort is to use all the cases of interest but only a subsample of the noncases so that validity is not compromised and precision is adequate. The two most widely used approaches are the nested case-control and case-cohort substudies; the primary distinction between the two is the timing of the selection of a comparison group from the noncases. Analytical methods for the analysis of nested studies are readily available; namely, conditional logistic regression for nested case-control studies and Cox regression with staggered entries and robust methods for calculation of standard errors for nested case-cohort studies (82). An alternative method for nesting studies is based on trajectories of markers of disease progression (i.e., stable versus fast progressors) (83).

Table 1 provides a summary of the analytical methods for cohort studies reviewed here. Software is widely available for the implementation of different methods, and several statistical packages, including SAS (SAS Institute, Cary, North Carolina), Splus (Statistical Sciences, Inc., Seattle, Washington), STATA (Stata Corporation, College Station, Texas), and EGRET (Cytel Software Corporation, Cambridge, Massachusetts), provide procedures, functions, and commands to carry out analysis of data from cohort studies. For example, analysis of cohort studies with staggered entries can be equally accomplished by the PROC PHREG of SAS; the survfit, coxph and Surv functions of Splus; the stset with the t0 option and stcox functions of STATA; and the menu-driven options of the Kaplan-Meier and Cox regression modules of EGRET (84).

An important advance in the last 10 years has been the development of methods to incorporate measurement error into the analysis of cohort data (85, 86). The application of these methods requires the appropriate design of validation/reproducibility substudies within cohort studies. Unfortunately, in many cases these studies are not properly designed or their results are not properly incorporated in the analysis of the core questions in cohort studies. These methods hold promise as a partial solution to the persistent problem of measurement error.

In this section we have reviewed the advances in analytical methods for time-to-event and repeated measures of markers as two separate fields. An area of active methodological research since the mid 1990s has been the unification of time-to-event (survival analysis) and repeated measurement methods (longi-

**TABLE 1. Overview of analytical methods for cohort studies**

| Outcome | Summary measure | Comparison | | Measure of association |
|---|---|---|---|---|
| | | Exposed/unexposed (2-sample) | Multiple (regression) | |
| Events in person-years | Incidence rate | $(O-E)^2/var$ | Poisson | Relative incidence |
| Time to event | Kaplan-Meier/maximum likelihood estimates | Logrank or Mantel-Haenszel/likelihood ratio test | Proportional hazards/ parametric | Relative hazard/relative percentile or time |
| Time to event; exposures changing | Extended Kaplan-Meier | Extended logrank | Proportional hazards, staggered entries | Relative hazard |
| Case in nested case-control | Proportion exposed | Paired chi-square or McNemar | Conditional logistic | Odds ratio |
| Case in nested case-cohort | Proportion exposed | (Robust) logrank | Proportional hazards, staggered entries | Relative hazard |
| Intermediate outcome repeatedly measured | Change | | Regression for correlated data; marginal, conditional, random effects | Differences in change over time |

tudinal data analysis). Both types of data have incompleteness or missing data, and the nature of incompleteness is often informative (e.g., individuals with low values in markers cease to provide longitudinal data due to imminence of disease onset). The developments for the handling of missing data have generated promising approaches for a unified framework (87–92). These methods are certain to advance.

## AN EXAMPLE: NATURAL HISTORY OF HIV INFECTION AND THE MULTICENTER AIDS COHORT STUDY

In this section, we illustrate the application of analytical methods for cohort studies in the context of the natural history of HIV infection, drawing on the Multicenter Aids Cohort Study. It is included in this review as an example of the great utility of cohort studies contributed by epidemiologic research to the overall goals of science and public health. There are numerous comparable examples in cohort studies of cancer, cardiovascular disease, and occupational and environmental agents.

Figure 2 depicts the key events in the natural history of HIV infection that take place during a cohort study of an at-risk population. On enrollment, some individuals would enter with antibodies to HIV (i.e., seroprevalent), and of those who enter seronegative, some would become infected during follow-up (i.e., seroconverters). Both seroprevalent individuals and seroconverters are subject to progressive immune suppression and opportunistic infections as a consequence of infection with HIV. These key natural history events shown can be linked to specific epidemiologic aims for which different analytical methods are pertinent (figure 2, table 2). Although the table is specific to the natural history of HIV infection as elucidated by a particular cohort study, it illustrates the scope of analytic endpoints in a cohort study and gives examples of the methods used for the different types of data pertinent to specific scientific aims.

In 1984–1985, a cohort of 4,954 men was recruited into the Multicenter AIDS Cohort Study in Baltimore, Maryland, Chicago, Illinois, Los Angeles, California, and Pittsburgh, Pennsylvania. To increase minority enrollment, an additional 625 men were recruited from 1987–1991, of whom 433 (69.3 percent) were non-Caucasian, and an additional 43 seroconverters from Pittsburgh were also recruited at the same time. The entire Multicenter AIDS Cohort Study cohort, therefore, consists of 5,622 men, of whom 2,195 (39 percent) were seroprevalent for HIV at entry. All men were followed up every 6 months, and serologic tests for HIV antibody were routinely done at each visit. Up to July 1, 1997, 551 men had known dates of last negative and first positive visits for HIV (i.e., seroconverters). Through July 1997, 1,400 and 244 AIDS

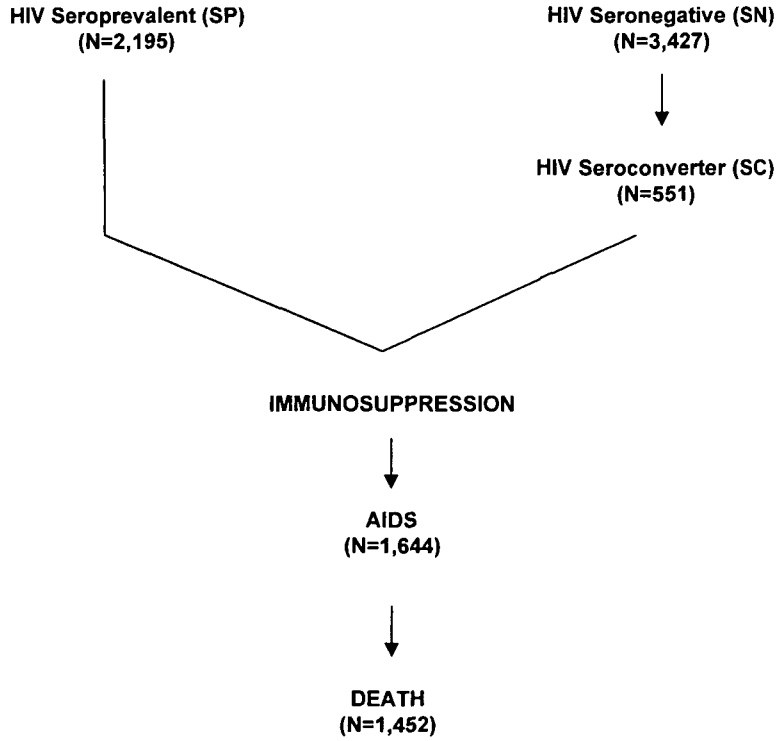


**FIGURE 2.** Descriptive statistics of participants in the Multicenter AIDS Cohort Study, 1984–1997.

**TABLE 2. Analytical methods for cohort studies; epidemiology of human immunodeficiency virus (HIV) infection in gay men, Multicenter AIDS* Cohort Study, 1984–1997**

| Aim | Method | Reference(s) |
|---|---|---|
| Prevalence of infection (SP*) | Logistic regression | Chmiel et al. (93) |
| Incidence of infection (SC*) | Poisson regression | Kingsley et al. (94) |
| Immunosuppression (CD4+) | Regression for continuous correlated data | Margolick et al. (95)<br>Mellors et al. (55) |
| Incubation (SC to AIDS) | Lognormal regression<br>Juxta-analysis: incident plus prevalent subcohorts | Muñoz and Xu (78)<br>Muñoz et al. (77, 97, 98)<br>Taylor et al. (99) |
| Markers and disease progression (Viral load, CD4 to AIDS and death) | Cox regression<br>Regression trees | Fahey et al. (100)<br>Enger et al. (101)<br>Mellors et al. (55) |
| Survival (AIDS to death) | Cox regression | Jacobson et al. (102) |
| Incidence of specific AIDS diagnosis by period | Poisson regression | Muñoz et al. (103) |
| Risk factors for fast progression to AIDS | Conditional logistic regression | Phair et al. (104) |
| High risk behavior | Regression for categorical correlated data | Gange et al. (105) |
| Effectiveness of AIDS therapies | Cox regression with staggered entries | Muñoz and Hoover (106)<br>Detels et al. (79) |
| Long-term survivors and pathogenesis of HIV | Nested studies based on markers | Muñoz et al. (107)<br>Gange et al. (108) |

* SP, seroprevalent; SC, seroconverters; AIDS, acquired immunodeficiency syndrome.

cases had been observed among the seroprevalent individuals and seroconverters, respectively. Among the 1,644 AIDS cases in HIV-positive individuals, 1,452 have died. Table 2 provides references based on the Multicenter AIDS Cohort Study data using specific analytical methods for corresponding scientific aims.

In an initial cross-sectional analysis, Chmiel et al. (93) used logistic regression to relate the odds of being seroprevalent for HIV to a constellation of putative risk factors, including sexual behavior, demographic characteristics, and history of other infectious diseases. It is noteworthy to point out that the serostatus at baseline in 1983–1984 could only be assessed *retrospectively*, since the test for HIV first became available in late 1985 and testing was performed on blood samples kept in a national repository. This example shows the great utility of storing samples collected *prospectively* in cohort studies.

To describe the incidence of infection, Kingsley et al. (94) modeled the number of seroconversions as a Poisson variable in strata defined by calendar, age, and ethnicity. At the time of publication in 1991, there was a suggestion of rising incidence over time, but subsequent follow-up did not confirm a trend and the rates of seroconversion remained low up to 1995, when follow-up of seronegative participants was confined to a small portion of the full cohort of seronegative participants.

The longitudinal data collected on CD4-cell count, a marker of immunodeficiency caused by infection with HIV, were used to characterize CD4 trajectories of individual participants. Margolick et al. (95), using the simplest case of random effects models (i.e., random intercept), documented that the total T-cell count remains fairly constant during the course of HIV infection up to approximately 1–5 years prior to the occurrence of AIDS, when it declines precipitously. These observations led to the postulate that homeostasis of the total T-cell count fails prior to the onset of AIDS. Mellors et al. (55) used random regression (i.e., intercept and slope follow a bivariate normal distribution) to show the very close relation between viral load at one time point and subsequent decline of CD4 cell count. These data were used to define the principles of HIV therapy now endorsed by the United States Public Health Service (96).

As for all infectious diseases, the incubation period of AIDS is of central interest. To describe the incubation period of AIDS and the corresponding hazard of AIDS at different durations of infection, Muñoz and Xu (78) showed that the lognormal model and regression were appropriate. This and several other reports

(97–99) on the AIDS incubation period in Multicenter AIDS Cohort Study participants combine data from incident and prevalent cases. These data can be combined by using an extension of Kaplan-Meier and Cox regression methods so as to consider at risk for AIDS only those individuals who have already entered into observation. Since this actually corresponds to a juxtaposition of the two subcohorts, the authors have suggested the name *juxta-analysis* for this type of approach (77).

Cox regression methods have been widely used to describe predictors of disease progression, generally measured by the length of AIDS-free survival time. In a seminal report, Fahey et al. (100) evaluated the prognostic value of three cellular and five serologic markers that are affected by infection with human immunodeficiency virus. As new methods were developed in 1995 to reproducibly quantify plasma viral load, Multicenter AIDS Cohort Study investigators used samples stored in the third semiannual visit (around September 1985) to assess viral load and natural history; Mellors et al. (55) showed that plasma viral load was the single best predictor of progression to AIDS and death; and using regression trees methodology (73, 74), the investigators showed that the prognosis of HIV-infected persons is more accurately defined by combined measurements of plasma HIV RNA and of CD4 lymphocytes. Enger et al. (101) estimated the expected survival time by calendar period before (1985–1988) and after (1989–1993) the widespread availability of AIDS treatments, and by stage of HIV disease quantified by the CD4 cell count at the beginning of each of the periods. In addition, Jacobson et al. (102) documented the changes in survival after AIDS in periods covering the years between 1984 and 1991.

Cohort studies have the substantial advantage of describing the incidence of different outcomes of interest. Muñoz et al. (103) used Poisson regression methods to describe the incidence of six groupings of the conditions that define AIDS. The investigators documented the effectiveness of *Pneumocystis carinii* pneumonia prophylaxis, showing a significant decline of the incidence of *P. carinii* pneumonia during follow-up of Multicenter AIDS Cohort Study participants. This decline was concomitant with upward trends of other opportunistic infections.

Phair et al. (104) nested a case-control study within the Multicenter AIDS Cohort Study to explore factors that may identify seroconverters who rapidly progress to AIDS. Consonant with the matched design of the study, they used conditional logistic regression to analyze the data, finding that high-risk behavior prior to seroconversion was related not only to the risk of infection but also to the risk of fast progression after infection. High-risk behavior is of interest as an outcome itself. Analytical challenges posed by behavioral data over time have led to methodological developments for regression methods for categorical correlated data (105).

Methods for cohort studies with staggered entries are also useful for the evaluation of effectiveness of AIDS therapies. Muñoz and Hoover (106) and Detels et al. (79) have analyzed the Multicenter AIDS Cohort Study data to determine if therapies as used by participants increase disease-free periods and/or survival. These analyses use calendar period as an external time-dependent covariate and as a proxy measure of relative intensity of exposure to antiretroviral therapies.

Based on the trajectories of markers of disease progression, cohort studies offer the possibility of comparing subgroups of individuals who exhibit different trajectories of markers in spite of starting at the same level (107). Furthermore, individuals who exhibit a stable profile during the first part of a cohort study could subsequently exhibit heterogeneous progression toward disease. These data offer the possibility of comparing late progressors with consistent nonprogressors, using a case-control study, in which cases and controls are matched longitudinally, thus presenting a hybrid of the case-control and cohort designs (108).

## EPILOGUE

We have reviewed the evolution of the cohort study from the substantive and methodological perspectives, and have illustrated how different methods have been useful for the key epidemiologic aims in cohort studies, using studies of HIV infection as our example. Because of their longitudinal nature, cohort studies offer an invaluable resource for the elucidation of disease pathogenesis. Cohort studies have provided fundamental knowledge for prevention strategies and have been a cornerstone of public health and policy. Methodological advances continue to strengthen this design and facilitate our understanding of how multiple factors acting over time can determine the etiology and natural history of disease.

## ACKNOWLEDGMENTS

This research was supported in part by grants UO1-AI-35043 and UO1-AI-42590 from the National Institutes of Health. The authors dedicate this review to B. Frank Polk, a mutual friend and colleague, whose contributions were seminal to the creation of the Multicenter AIDS Cohort Study.

## REFERENCES


1. Armenian HK (ed). Applications of the case-control method. Epidemiol Rev 1994;16:1–164.
2. Breslow NE, Day NE. Statistical methods in cancer research. Vol II—The design and analysis of cohort studies. Lyon, France: International Agency for Research on Cancer, 1987. (IARC scientific publications no. 82).
3. Cox DR, Oakes D. Analysis of survival data. New York, NY: Chapman and Hall, 1984.
4. Dwyer JH, Feinleib M, Lippert P, et al. Statistical models for longitudinal studies of health. New York, NY: Oxford University Press, 1992.
5. Diggle PJ, Liang KY, Zeger SL. Analysis of longitudinal data. New York, NY: Oxford University Press, 1994.
6. Liddell FDK. The development of cohort studies in epidemiology: a review. J Clin Epidemiol 1988;41:1217–37.
7. Changes in cigarette-related disease risks and their implication for prevention and control. Bethesda, MD: US Department of Health and Human Services, National Institutes of Health, National Cancer Institute, 1997. (NIH publication no. 97–4213).
8. Bailar JC III, Mosteller F. Medical uses of statistics. 2nd ed. Waltham, MA: NEJM Books, 1992.
9. MacMahon B, Trichopoulos D. Epidemiology: principles and methods. 2nd ed. Boston, MA: Little, Brown, 1996.
10. Last JM. A dictionary of epidemiology. 3rd ed. New York, NY: Oxford University Press, 1995.
11. Eyler JM. The conceptual origins of William Farr's epidemiology: numerical methods and social thought in the 1830s. In: Lilienfeld AM, ed. Times, places, and persons: aspects of the history of epidemiology. Baltimore, MD: Johns Hopkins University Press, 1980:1–21.
12. Lew EA. Actuarial contributions to life table analysis. Natl Cancer Inst Monogr 1985;67:29–36.
13. Bernstein PL. Against the gods: the remarkable story of risk. New York, NY: Wiley, 1996.
14. Singer RB. Examples of early mortality follow-up studies. Natl Cancer Inst Monogr 1985;67:15–21.
15. Specialized mortality investigation. New York, NY: Actuarial Society of America, 1903.
16. Sartwell PE. Editorial: The time factor in studies of the outcome of chronic disease. Am Rev Tuberc 1951;63: 608–12.
17. Comstock GW. Early studies of tuberculosis. Natl Cancer Inst Monogr 1985;67:23–7.
18. Brown L, Pope EG. The post-discharge mortality among the patients of the Adirondack Cottage Sanitarium. Am Med 1904;8:879–82.
19. Elderton WP, Perry SJ. Drapers' company (London, England). Research memoirs. Studies in national deterioration VI. A third study of the statistics of pulmonary tuberculosis: the mortality of the tuberculous and sanatorium treatment. London, England: Dalau, 1910.
20. Frost WH. The age selection of mortality from tuberculosis in successive decades. Am J Hyg 1939;30A:91–6.
21. Andvord KF. What can we learn by studying tuberculosis generations? (English translation by Gerard Wijsmuller). Norsk mag Lægevidensk 1930;91:642–60.
22. Frost WH. Risk of persons in familial contact with pulmonary tuberculosis. Am J Public Health 1933;23:426–32.
23. Zeidberg LD, Gass RS, Dillon A, et al. The Williamson County Tuberculosis Study: a twenty-four-year epidemiologic study. Am Rev Respir Dis 1963;87(no.3, part 2):1–88.
24. Dawber TR. The Framingham study: the epidemiology of atherosclerotic disease. Cambridge, MA: Harvard University Press, 1980.
25. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis 1967;20:511–24.
26. Schull WJ. Effects of atomic radiation: a half-century of studies from Hiroshima and Nagasaki. New York, NY: Wiley-Liss, 1995.
27. Samet JM. Epidemiologic studies of ionizing radiation and cancer: past successes and future challenges. Environ Health Perspect 1997;105(Suppl 4):883–9.
28. Lundin FE, Wagoner JK, Archer VE. Radon daughter exposure and respiratory cancer, quantitative and temporal aspects. National Instutute for Occupational Safety and Health-National Institute of Environmental Health Sciences joint monograph no. 1. Washington, DC: US Department of Health, Education, and Welfare, Public Health Service, 1971.
29. National Research Council Committee on Health Risks of Exposure to Radon. Report of the committee on health risks of exposure to radon (BEIR VI). Washington, DC: National Academy Press, 1997.
30. Whittemore AS, McMillan A. Analyzing cohort data: application to US uranium miners. In: Prentice RL, Whittemore AS, eds. Environmental epidemiology: risk assessment. Philadelphia, PA: SIAM, 1982:65–81.
31. Lubin JH, Gail MH. On power and sample size for studying features of the relative odds of disease. Am J Epidemiol 1990;131:552–66.
32. Thomas DC, McNeill KG, Dougherty C. Estimates of lifetime lung cancer risks resulting from Rn progeny exposure. Health Phys 1985;49:825–46.
33. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits: a preliminary report. Br Med J 1954;1: 1451–5.
34. Doll R, Peto R, Wheatley K, et al. Mortality in relation to smoking: 40 years' observations on male British doctors. BMJ 1994;309:901–11.
35. Hammond EC, Horn D. Smoking and death rates report on forty-four months of follow-up of 187,783 men. I. Total mortality. II. Death rates by cause. JAMA 1958;166: 1159–72, 1294–308.
36. Dorn HF. Tobacco consumption and mortality from cancer and other diseases. Public Health Rep 1959;74:981–93.
37. US Department of Health Education and Welfare, Public Health Service. Smoking and health; report of the advisory committee to the Surgeon General of the Public Health Service. Washington, DC: US GPO, 1964.
38. Case RAM, Hosker ME, McDonald DB, et al. Tumors of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry. Part I: The role of aniline, benzidine, alpha-naphthylamine, and beta-naphthylamine. Br J Ind Med 1954; 11:75–104.
39. Doll R. Mortality from lung cancer in asbestos workers. Br J Ind Med 1955;12:81–6.
40. Selikoff IJ, Hammond EC, Seidman H. Mortality experience of insulation workers in the United States and Canada, 1943–1976. Ann NY Acad Sci 1979;330:91–116.
41. Hammond EC. Smoking in relation to the death rates of one million men and women. Natl Cancer Inst Monogr 1966;19: 127–204.
42. Fletcher CM, Peto R, Tinker C, et al. The natural history of chronic bronchitis and emphysema: an eight year study of early chronic obstructive lung disease in working men in London. New York, NY: Oxford University Press, 1976.
43. Higgins MW, Keller JB. Predictors of mortality in the adult population of Tecumseh. Arch Environ Health 1970;21: 418–24.
44. Heyden S, Cassel JC, Bartel A, et al. Body weight and cigarette smoking as risk factors. Arch Intern Med 1971;128: 915–19.
45. Berkman LF, Syme SL. Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents. Am J Epidemiol 1979;109:186–204.
46. Menkes MS, Comstock GW, Vuilleumier JP, et al. Serum beta-carotene, vitamins A and E, selenium, and the risk of lung cancer. N Engl J Med 1986;315:1250–4.
47. Lebowitz MD, Burrows B. Tucson epidemiologic study of

obstructive lung diseases. II. Effects of in-migration factors on the prevalence of obstructive lung diseases. Am J Epidemiol 1975;102:153–63.
48. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol 1989;129:687–702.
49. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1991;1: 263–76.
50. Lee ET, Welty TK, Fabsitz R, et al. The Strong Heart Study: a study of cardiovascular disease in American Indians: design and methods. Am J Epidemiol 1990;132:1141–55.
51. Newcombe HB. Handbook of record linkage: methods for health and statistical studies, administration and business. New York, NY: Oxford University Press, 1988.
52. Cox CS, Rothwell ST, Madans JH, et al. Plan and operation of the NHANES I Epidemiologic Followup Study, 1987. National Center for Health Statistics. Vital Health Stat [1] 1992;(27):1–190.
53. Colditz GA, Willett WC, Hunter DJ, et al. Family history, age, and risk of breast cancer: prospective data from the Nurses' Health Study. JAMA 1993;270:338–43.
54. Polk BF, Fox R, Brookmeyer R, et al. Predictors of the acquired immunodeficiency syndrome developing in a cohort of seropositive homosexual men. N Engl J Med 1987; 316:61–6.
55. Mellors JW, Muñoz A, Giorgi JV, et al. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. Ann Intern Med 1997;126:946–54.
56. Mantel N. Synthetic retrospective studies and related topics. Biometrics 1973;29:479–86.
57. Liddell FDK, McDonald JC, Thomas DC. Methods of cohort analysis: appraisal by application to asbestos mining. J R Stat Soc 1977;140:469–91.
58. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73: 1–11.
59. Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol 1977;106:184–7.
60. Royall RM. Statistical evidence: a likelihood paradigm. New York, NY: Chapman and Hall, 1997.
61. Lehmann EL. Theory of point estimation. New York, NY: Springer-Verlag, 1997.
62. Tufte ER. The visual display of quantitative information. Cheshire, CT: Graphics Press, 1983.
63. Gange SJ, Muñoz A, Chmiel JS, et al. Identification of inflections in T-cell counts among HIV-1 infected individuals and relationship with progression of clinical AIDS. Proc Natl Acad Sci 1998 (in press).
64. Cox DR. Regression models and life tables. J R Stat Soc 1972;B34:187–202.
65. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 1986;42:805–20.
66. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121–30.
67. Korn EL, Whittemore AS. Methods for analyzing panel studies of acute health effects of air pollution. Biometrics 1979;35:795–802.
68. Whittemore AS, Korn EL. Asthma and air pollution in the Los Angeles area. Am J Public Health 1980;70:687–96.
69. Rosner B, Muñoz A. Conditional linear models for longitudinal data. In: Dwyer JH, Feinleib M, Lippert P, et al., eds. Statistical models for longitudinal studies of health. New York, NY: Oxford University Press, 1992:115–37.
70. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–74.
71. Kupper LL, Haseman JK. The use of a correlated binomial model for the analysis of certain toxicological experiments. Biometrics 1978;34:69–76.
72. Breslow N. Extra-Poisson variation in log-linear models. Appl Stat 1984;33:38–44.
73. Segal MR. Regression trees for censored data. Biometrics 1988;44:35–48.
74. Ciampi A, Lawless JF, McKinney SM, et al. Regression and recursive partition strategies in the analysis of medical survival data. J Clin Epidemiol 1988;41:737–48.
75. McCullagh P, Nelder JA. Generalized linear models. New York, NY: Chapman and Hall, 1989.
76. Hastie TJ, Tibshirani RJ. Generalized additive models. New York, NY: Chapman and Hall, 1990.
77. Muñoz A, Sabin CA, Phillips AN. The incubation period of AIDS. AIDS 1997;11(Suppl A):S69–76.
78. Muñoz A, Xu J. Models for the incubation of AIDS and variations according to age and period. Stat Med 1996;15: 2459–73.
79. Detels R, Muñoz A, McFarlane G. Effectiveness of highly active antiretroviral therapy on the natural history of HIV/AIDS in a large cohort of seroconverters. JAMA 1998 (in press).
80. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med 1992;11:1871–9.
81. Muñoz A, Sunyer J. Comparison of semiparametric and parametric survival models for the analysis of bronchial responsiveness. Am J Respir Crit Care Med 1996; 154(Suppl):S234–9.
82. Wacholder S, Gail MH, Pee D, et al. Alternative variance and efficiency calculations for the case-cohort design. Biometrika 1989;76:117–23.
83. Muñoz A, Gange SJ. Methodological issues for biomarkers and intermediate outcomes in cohort studies. Epidemiol Rev 1998;20:15–28.
84. Lamarca R, Alonso J, Gomez G, et al. Left-truncated data with age as time scale: an alternative for survival in the elderly. J Gerontol A Biol Sci Med Sci (in press).
85. Armstrong BK, White E, Saracci R. Principles of exposure measurement in epidemiology. New York, NY: Oxford University Press, 1992.
86. Carroll RJ, Ruppert D, Stefanski LA. Measurement error in nonlinear models. London, England: Chapman and Hall, 1995.
87. Tsiatis AA, De Gruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. J Am Stat Assoc 1995;90:27–37.
88. De Gruttola V, Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics 1994;50:1003–14.
89. Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. Appl Stat 1994;43:49–93.
90. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 1995;90:1112–21.
91. Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. Stat Med 1997;16:239–57.
92. Hogan JW, Laird NM. Model-based approaches to analysing incomplete longitudinal and failure time data. Stat Med 1997;16:259–72.
93. Chmiel JS, Detels R, Kaslow RA, et al. Factors associated with prevalent human immunodeficiency virus (HIV) infection in the Multicenter AIDS Cohort Study. Am J Epidemiol 1987;126:568–77.
94. Kingsley LA, Zhou SY, Bacellar H, et al. Temporal trends in human immunodeficiency virus type 1 seroconversion 1984–1989: a report from the Multicenter AIDS Cohort Study (MACS). Am J Epidemiol 1991;134:331–9.
95. Margolick JB, Muñoz A, Donnenberg AD, et al. Failure of T-cell homeostasis preceding AIDS in HIV-1 infection: the Multicenter AIDS Cohort Study. Nat Med 1995;1:674–80.
96. Centers for Disease Control and Prevention. Report of the NIH panel to define principles of therapy of HIV infection and guidelines for the use of antiretroviral agents in HIV-

infected adults and adolescents. MMWR 1998;47(RR-5): 1–82.
97. Muñoz A, Wang MC, Bass S, et al. Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type 1 (HIV-1) seroconversion in homosexual men: Multicenter AIDS Cohort Study Group. Am J Epidemiol 1989;130:530–9.
98. Muñoz A, Carey V, Taylor JM, et al. Estimation of time since exposure for a prevalent cohort. Stat Med 1992;11: 939–52.
99. Taylor JM, Muñoz A, Bass SM, et al. Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation: Multicenter AIDS Cohort Study. Stat Med 1990;9:505–14.
100. Fahey JL, Taylor JM, Detels R, et al. The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type 1. N Engl J Med 1990;322:166–72.
101. Enger C, Graham N, Peng Y, et al. Survival from early, intermediate, and late stages of HIV infection. JAMA 1996; 275:1329–34.
102. Jacobson LP, Kirby AJ, Polk S, et al. Changes in survival after acquired immunodeficiency syndrome (AIDS): 1984–1991. Am J Epidemiol 1993;138:952–64.
103. Muñoz A, Schrager LK, Bacellar H, et al. Trends in the incidence of outcomes defining acquired immunodeficiency syndrome (AIDS) in the Multicenter AIDS Cohort Study: 1985–1991. Am J Epidemiol 1993;137:423–38.
104. Phair J, Jacobson L, Detels R, et al. Acquired immune deficiency syndrome occurring within 5 years of infection with human immunodeficiency virus type-1: the Multicenter AIDS Cohort Study. J Acquir Immune Defic Syndr 1992;5: 490–6.
105. Gange SJ, Jacobson LP, Muñoz A. Parametric models to assess tracking of categorical high-risk sexual behaviour for HIV infection. Stat Med 1996;15:2323–36.
106. Muñoz A, Hoover DR. Use of cohort studies for evaluating AIDS therapies. New York, NY: Wiley, 1995.
107. Muñoz A, Kirby AJ, He YD, et al. Long-term survivors with HIV-1 infection: incubation period and longitudinal patterns of CD4+ lymphocytes. J Acquir Immune Defic Syndr Hum Retrovirol 1995;8:496–505.
108. Gange SJ, Muñoz A, Schrager LK, et al. Design of nested studies to identify factors related to late progression of HIV infection. J Acquir Immune Defic Syndr Hum Retrovirol 1997;15(Suppl 1):S5–9.