Environmental Health Perspectives
Vol. 87, pp. 43–49, 1990

Statistical Problems in Epidemiologic Studies of the Natural History of Disease

by Ron Brookmeyer*

The development of effective disease prevention and treatment programs depends on an understanding of the natural history of disease. A conceptual framework is presented for disease natural history and consists of an asymptomatic period of disease followed by a period of symptomatic disease. The focus is on epidemiologic studies for identifying risk factors of the onset of asymptomatic disease, for identifying cofactors of progression to symptomatic disease, and for estimating the duration of the asymptomatic period. The strengths and limitations of various epidemiologic study designs and sources of epidemiologic data are considered for characterizing disease natural history. Issues in the interpretation and analysis of natural history parameters of disease estimated from cross-sectional, prevalent cohort, cohort, and matched case-control studies are considered. The issues and analytic methods are illustrated with studies of the acquired immunodeficiency syndrome (AIDS) and cervical cancer. Based on these analytic methods, an estimate of the incubation period distribution of AIDS is given.

Introduction

An understanding of the natural history of disease is important for developing effective disease prevention and treatment programs. The objective of this paper is to consider the strengths and limitations of various epidemiologic study designs and sources of epidemiologic data for characterizing the natural history of disease.

A simple conceptual framework for the natural history of a disease is a two-stage model (Fig. 1). An individual is free of disease (healthy) until the onset of stage 1 disease. Stage 1 refers to preclinical or asymptomatic disease. It is assumed there is a diagnostic screening test that can detect the presence of stage 1 disease. The individual with stage 1 disease may eventually progress to stage 2, which is the clinical or symptomatic period. It is assumed that individuals enter stage 1 before the onset of stage 2 disease. The focus of this paper is on the natural history of disease up to the onset of symptomatic disease (stage 2).

Two types of covariates affect the natural history of disease. The first type, $\underset{\sim}{X}_1$, are those that affect the risk of stage 1 disease. The hazard (or incidence) of onset of stage 1 at time s is call $\lambda(s; \underset{\sim}{X}_1)$, where the time scale s may refer either to calendar time or chronological age. The second type of covariates, $\underset{\sim}{X}_2$, are those that effect risk of progression to stage 2 from stage 1 disease. The duration of time spent in stage 1 has been termed the preclinical duration, the incubation period and the sojourn time (*1*). The distribution function of stage 1 durations, $F(t; \underset{\sim}{X}_2)$, is the probability an individual with stage 1 disease progresses to stage 2 within t years of onset of stage 1. The corresponding hazard and density functions are called $h(t; \underset{\sim}{X}_2)$ and $f(t; \underset{\sim}{X}_2)$, respectively. The distribution function may be improper as not all individuals may eventually progress to stage 2.

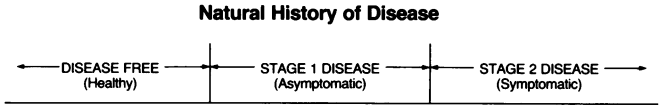

FIGURE 1. Two-stage model for disease natural history. Covariates, $\underset{\sim}{X}_1$, effect risk of stage 1 (asymptomatic) disease. Covariates, $\underset{\sim}{X}_2$, effect risk of progression to stage 2 (symptomatic) disease.

Two examples illustrate this conceptual framework: cervical cancer and the acquired immunodeficiency syndrome (AIDS). The natural history of cervical cancer consists a very long asymptomatic period (stage 1), which may consist of histological abnormalities ranging from cervical intraepithelial neoplasia (CIN I, II, and III) to preclinical invasive disease (*2*). The asymptomatic period may be followed by the onset of symptoms (stage 2) by which point the lesion has become invasive cancer. Cervical cytology (the PAP test) can detect the presence of stage 1 disease. Risk factors ($\underset{\sim}{X}_1$) which may be related to risk of stage 1 disease include human papillomavirus (HPV) infection and certain contraceptive practices. A cofactor, $\underset{\sim}{X}_2$ which has been suggested to possibly accelerate progression to stage 2 from stage 1, is infection with HPV Type 18 (*3*).

The natural history of AIDS begins with infection with the etiologic agent, the human immunodeficiency virus (HIV) (*4,5*). In our framework, the onset of stage 1 refers to HIV infection (actually the development of

*Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205.

HIV antibodies or seroconversion) while stage 2 refers to clinically defined AIDS. The screening test for detecting stage 1 is the AIDS antibody test (Elisa or Western Blot). Risk factors ($\underset{\sim}{X}_1$) for infection (stage 1) include high-risk behaviors (such as intravenous drug use and large numbers of sexual partners), hemophilia, and previous blood transfusions. It has been suggested that a cofactor ($\underset{\sim}{X}_2$) for progression to clinical AIDS from the asymptomatic HIV infected state is age at infection (*6*).

There are important public health reasons for studying the natural history of disease. Identification of risk factors for stage 1 disease is crucial for developing effective prevention programs. Identification of cofactors for progression to stage 2 is important for the development of treatment and intervention programs: individuals at higher risk of progression to stage 2 may be monitored more closely or treated more aggressively. An estimate of the distribution function $F(t; \underset{\sim}{X}_2)$ is important for two reasons. First, it is useful in developing recommendations on screening frequency and the time interval between screens; second, it is useful for predicting future cases of clinical disease. For example, assume that the time of onset of stage 1 and the duration spent in stage 1 are independent given the covariates $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$. Then the cumulative probability an individual with covariates ($\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$) develops stage 2 disease at or before calendar time (or chronological age) t is

$$D(t; \underset{\sim}{X}_1, \underset{\sim}{X}_2) = \int_0^t g(s; \underset{\sim}{X}_1)F(t - s; \underset{\sim}{X}_2)ds \qquad (1)$$

where $g(s; \underset{\sim}{X}_1) = \lambda(s; \underset{\sim}{X}_1)\exp[-\int_0^s \lambda(s; \underset{\sim}{X}_1)ds]$ is the probability density of onset of stage 1 at time s. Eq. (1) has been used to project the course of the AIDS epidemic (*7–9*). Information is available both on the numbers of AIDS cases diagnosed over calendar time and $F(t)$ (the incubation period distribution). Thus estimates of the numbers of individuals previously infected can be obtained through the technique of back-calculation (*8*). These numbers infected are then projected forward to obtain short-term projections of AIDS incidence.

Epidemiologic Study Designs for Characterizing Natural History

It is useful to consider the ideal epidemiologic study for characterizing the natural history of disease. The ideal study would consist of a disease-free cohort defined at (chronological or calendar) time $s = 0$. The cohort would undergo continuous surveillance and screening in order to determine the exact times that individuals develop stage 1 and stage 2 disease. The covariates $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ would be ascertained on all individuals. The screening test to detect stage 1 disease would have negligible error (specificity = sensitivity = 1.0). Further, individuals detected in stage 1 would be monitored for onset of stage 2. There would be no treatment intervention for these individuals that could alter the natural history.

However, such a study is usually impossible to perform for many reasons. First, if an effective treatment exists, individuals detected with stage 1 must be treated, which interrupts the natural history of disease. Second, we cannot perform continuous screening tests for stage 1 disease but, at best, perform only periodic screens. Third, the errors associated with the screening test may not be negligible. Fourth, a very large cohort would be required for a rare disease in order to identify a sufficient number of individuals with incident stage 1 disease. Fifth, the follow-up period would need to be long for diseases with long incubation periods (stage 1 durations).

In the next sections, we consider the strengths and limitations of alternative epidemiologic designs for characterizing the natural history of disease. We outline analytic approaches to estimate parameters that describe disease natural history. We assume in the next sections that the errors associated with the screening test are negligible and can be ignored, although in the last section, some consideration is given to situations in which screening test errors are not negligible.

The Cross-Sectional Study

One of the simplest study designs is the cross-sectional study. Consider a large cohort of individuals defined at time $s = 0$. A random sample of these individuals is chosen at a point in time ($s = Y$). We test each individual in the sample for presence of stage 1 disease and obtain information on a covariate, Z. We assume, for simplicity, Z is dichotomous. A common practice is to cross classify individuals according to presence or absence of stage 1 disease and the two levels of the covariate ($Z = 0$ and $Z = 1$). This was the design of a recent study to investigate the relationship between HPV infection and early stages of cervical cancer (*10*). What are the limitations of the cross-sectional study for characterizing natural history?

The most serious limitation is that the time sequence of events cannot be established. We cannot determine if an individual was exposed ($Z = 1$) before or after onset of stage 1 disease. This is an important limitation with cross-sectional studies of HPV infection and cervical cancer, because individuals with stage 1 (CIN) disease may be more (or perhaps less) prone to acquire HPV infection. This problem is not unique to the cross-sectional study and occurs in the case-control study as well.

The issue of the time sequence does not arise if Z is a fixed covariate, that is, the value of the covariate is determined at time $s = 0$ for each individual. However, even if Z is a fixed covariate, the interpretation of commonly used parameters of association such as the odds ratio must be modified.

Table 1 displays the classification probabilities associated with the 2×2 table which results from the cross-sectional study, that is, the joint probability distribution of stage 1 disease (presence or absence) and the covariate value ($Z = 0$ or $Z = 1$). These probabilities

depend on the following: the incidence of stage 1 disease, $\lambda(s; Z)$; the distribution of stage 1 durations, $F(t; Z)$; the probability distribution of the covariate Z [i.e., $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$)]; and the time $s = Y$ that the cross-sectional study is conducted. The odds ratio is the cross product of these cell probabilities, and is independent of p. Under the assumption that stage 1 disease is rare with constant incidence rate, that is $\lambda_1(s; Z) = \exp(\alpha_0 + \alpha_1 Z)$, then the hazard $\lambda_1(s; Z)$ is approximately the density of $g(s; Z)$. The odds ratio from the cross-sectional study, ω_{cs}, is approximately

$$\omega_{cs} \approx \exp(\alpha_1) \cdot \left[\frac{\int_0^Y (1 - F(t; Z = 1))dt}{\int_0^Y (1 - F(t; Z = 0))dt}\right]$$

For example, if the distribution of stage 1 durations is exponential, namely $F(t) = 1 - \exp(-\exp(\beta_0 + \beta_1 Z)t)$, then

$$\omega_{cs} \approx \exp(\alpha_1 - \beta_1) \cdot \left[\frac{1 - e^{-Y \cdot \exp(\beta_0 + \beta_1)}}{1 - e^{-Y \exp(\beta_0)}}\right] \quad (2)$$

If the study is conducted at a time $s = Y$ sufficiently large so that then the term in brackets in Eq. (2) is nearly 1, then

$$\omega_{cs} \approx \exp(\alpha_1 - \beta_1) \quad (3)$$

Under the above assumptions for which Eq. (3) is valid, consider the following example: suppose an individual with $Z = 1$ is at twice the risk of stage 1 disease [i.e., ($e^{\alpha_1} = 2$)]; further, suppose that among those with stage 1 disease an individual with $Z = 1$ is at twice the risk of progression to stage 2 disease ($e^{\beta_1} = 2.0$). Using Eq. (3), the odds ratio obtained from a cross-sectional study (aside from sampling variation) would be 1.0. The naive odds ratio obtained from the cross-sectional study would suggest the covariate Z is not associated with disease, when in fact, it is positively associated with both incidence of stage 1 disease and progression to stage 2 disease.

In summary, the odds ratio obtained from the cross-sectional study is determined by the joint effects of the covariate on both the risk of stage 1 disease and the risk of progression to stage 2 disease. It is not possible to separate out these effects solely from cross-sectional data.

It may be possible to separate the effects with some additional information. For example, we could supplement the cross-sectional study with an incident case-control study. The case-control study would consist of incident cases of stage 2 disease and a sample of controls. Covariate information, Z, would be ascertained on all cases and controls. Under the same assumptions which led to Eq. (3), the odds ratio from the incident case-control study would be (aside from sampling variation)

$$\omega_I \approx e^{\alpha_1}.$$

Then the ratio of the odds ratios obtained from the incident case-control and cross-sectional studies is $\omega_I/\omega_{cs} \approx e^{\beta_1}$. Thus, in this case, the supplementary data allows separate estimation of the two effects, e^{α_1} and e^{β_1}.

The Prevalent Cohort Study

The prevalent cohort study consists of a cohort of individuals, each of whom has stage 1 disease at entry into the cohort. The prior time of onset of stage 1 disease is a random, unknown quantity. The cohort is followed for onset of stage 2 disease. The objective is to identify cofactors, $\underset{\sim}{X}_2$, of disease progression and to estimate $F(t; \underset{\sim}{X}_2)$. This design was recently used in several natural history studies of the acquired immunodeficiency syndrome, in which patients with prevalent HIV infection were enrolled in a cohort and followed for onset of clinical AIDS.

The prevalent cohort study has advantages but also serious limitations. One advantage is that it is not necessary to follow a large cohort of disease-free individuals in order to identify incident (newly onset) stage 1 disease. Rather, individuals identified with prevalent stage 1 disease from a cross-sectional survey could be enrolled into the cohort. The savings in terms of sample size and follow-up time could be substantial if the disease is rare, (i.e., $\lambda(s; \underset{\sim}{X}_1)$ is small). However, the fact that individuals with prevalent stage 1 disease are enrolled in the cohort rather than incident stage 1 disease is also a serious limitation. There are biases inherent in estimating the relative risk of a cofactor and the distribution function $F(t; \underset{\sim}{X}_2)$ from prevalent cohorts (*11*).

For example, let $F^*(t, \underset{\sim}{X}_2)$ be the cumulative probability of onset of stage 2 disease within t years of follow-up for an individual prevalent with stage 1 disease at the beginning of follow-up. In general $F^*(t, \underset{\sim}{X}_2)$ will not equal $F(t; \underset{\sim}{X}_2)$. The direction of the bias ($F^*(t; \underset{\sim}{X}_2) - F(t; \underset{\sim}{X}_2)$), depends on whether the hazard, $h(t; \underset{\sim}{X}_2)$, of onset of stage 2 disease t years after entering stage 1 is increasing or decreasing. If the hazard is increasing then $F^*(t; \underset{\sim}{X}_2) > F(t; \underset{\sim}{X}_2)$. The intuition for this result is that with an increasing hazard of onset of stage 2, an individual with prevalent stage 1 disease has a worse prog-

Table 1. Classification probabilities in a cross-sectional study conducted at time $s = Y$.[a]

	Stage 1 disease	Disease-free
$Z = 1$	$p \cdot \int_0^Y g(s; Z = 1)\,[1 - F(Y - s; Z = 1]\,ds$	$p \cdot \exp(-\int_0^Y \lambda(s; Z = 1)ds$
$Z = 0$	$(1 - p) \int_0^Y g(s; Z = 0)\,[1 - F(Y - s; Z = 0]\,ds$	$(1 - p) \cdot \exp(-\int_0^Y \lambda(s; Z = 0)\,ds$

[a] The covariate Z is dichotomous taking values 0 and 1 with probabilities $(1 - p)$ and p, respectively.

nosis than an individual with newly onset stage 1 disease, because the prevalent individual has had stage 1 disease for a longer period of time. The bias can be substantial. For example, an estimate of the cumulative probability of developing AIDS within 3 years of follow-up based on a prevalent cohort, was 0.34 (*12*). More recent estimates based on cohorts of newly infected individuals suggest this cumulative probability is less than 0.05 (*6,13*).

Assuming a proportional hazards model $h(t) = h_o(t)\exp(\beta Z)$ where t is the time since onset of stage 1 disease, then estimates of the relative risk $\exp(\beta)$ derived from a prevalent cohort will be biased if the proportional hazards analysis is performed using follow-up time. There are two reasons for the bias. The first reason is that the distribution of the prior times of onset of stage 1 disease for two subgroups ($Z = 0$ and $Z = 1$) may be different ($\lambda(s; Z = 0) \neq \lambda(s; Z = 1)$. For example, one prevalent cohort study reported a higher cumulative proportion of AIDS in New York than in Washington (*12*). The most plausible explanation is the New York cohort was infected earlier in calendar time than the Washington cohort, and not that geography is a cofactor of disease progression. The second reason for bias occurs even if $\lambda(s; Z = 0) = \lambda(s; Z = 1)$. This bias is due to the differential effects of length-biased sampling in the two subgroups ($Z = 0$ and $Z = 1$). The direction of the bias also depends on whether the hazard $h(t)$ increases or decreases over time. For example, if Z is a real cofactor ($\beta > 0$) and the hazard is increasing, then the individuals with $Z = 0$ will tend to have been in stage 1 longer at the beginning of follow-up then individuals with $Z = 1$. This biases the relative risk β toward unity. Fortunately, as shown in Brookmeyer and Gail (*11*) this bias is not of sufficient magnitude to reverse the direction of an effect [that is, make a real cofactor ($\beta > 0$) appear protective].

Cohort Studies of Serially Screened Populations without Treatment Intervention

In this section, we consider cohort studies of a serially screened population without treatment intervention. Suppose a cohort of disease-free individuals is defined at calendar time (or chronological age) $s = 0$. Individuals in the cohort are periodically screened for presence or absence of stage 1 disease. All individuals are followed for onset of stage 2 disease. It is assumed there is no treatment intervention for individuals detected with stage 1 disease and, further, the errors associated with the screening test are negligible. Then, it is possible to determine the onset time of stage 1 disease up to an interval. This type of study has considerably more information than either the cross-sectional or prevalent cohort study. Unlike the cross-sectional and prevalent cohort study, this design allows estimation of the separate effects of a covariate on risk of stage 1 disease and risk of progression to stage 2 from stage 1.

This was the design of a recent epidemiologic study of the natural history of AIDS among hemophiliacs (*14*), the National Cancer Institute Multicenter Hemophilia Cohort Study. Hemophiliacs were at risk of HIV infection from the mid-1970s in the United States because of contamination of replacement clotting factors. The study consisted of hemophiliacs who regularly visited treatment centers. Serum samples which were obtained at these visits were stored and subsequently tested for presence of HIV infection. The following information is recorded on each individual (the subscript i denotes information obtained from the ith individual): *a*) an indicator variable ϵ_i that indicates whether the individual had a positive screening test during follow-up (in which case we set $\epsilon_i = 1$, or otherwise $\epsilon_i = 0$); *b*) if the individual had a positive test, then the interval in which onset of stage 1 disease occurred is recorded as (L_i, R_i) where L_i is the calendar (or chronological age) time of the most recent negative test and R_i is the calendar (or chronological age) time of the earliest positive screening test; *c*) an indicator, δ_i, that indicates if the individual had onset of stage 2 disease by last follow-up (in which case $\delta_i = 1$ or otherwise $\delta_i = 0$); *d*) the time t_i of last follow-up or onset of stage 2 disease whichever comes first; and *e*) covariates $\underset{\sim}{X}_{1i}$ and $\underset{\sim}{X}_{2i}$. The analysis must account for the fact that the time of infection is known only up to an interval.

We assume independence between onset time of stage 1 and the duration spent in stage 1 conditionally on the covariates $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$. Assuming parametric models for the probability density functions of stage 1 disease, $g(s; \underset{\sim}{X}_1)$ and stage 2 disease, $f(t, \underset{\sim}{X}_2)$, the full likelihood function can be derived. Each individual contributes one of four possible factors to the likelihood, corresponding to the four values of (ϵ_i, δ_i). These factors can be expressed in terms of convolutions. For example, the likelihood contribution for an individual with $\epsilon_i = 1$ and $\delta_i = 1$ is

$$\int_{L_i}^{R_i} g(s; \underset{\sim}{X}_1) f(t_i - s; \underset{\sim}{X}_2) ds \tag{4}$$

Brookmeyer and Goedert describe this approach (*14*). Modified Newton-Raphson algorithms can be used to find the maximum likelihood estimates of the parameters of the stage 1 and stage 2 disease incidence functions. The analysis produces not only estimates of relative risk of covariates but also estimates of the incubation period distribution, $F(t; \underset{\sim}{X}_2)$. The analysis (*14*) of the National Cancer Institute Multicenter Hemophilia Cohort Study found that age was a cofactor ($\underset{\sim}{X}_2$) of disease progression. Table 2 gives the estimates of $F(t)$ for hemophiliacs over the age of 20. The estimate of the 3-year cumulative probability of AIDS was only 0.033, which is considerably less than prior estimates obtained from prevalent cohorts.

Cohort Studies of Serially Screened Populations with Treatment Intervention

A major analytic complication of most cohort studies of disease natural history is that if an effective treat-

Table 2. Estimated cumulative probability of AIDS within t years of seroconversion, $\hat{F}(t)$, and 95% confidence intervals for individuals over age 20 based on hemophilia cohort.

t years	$\hat{F}(t)$	95% Confidence interval[a]
1	0.002	(0.002, 0.006)
2	0.012	(0.011, 0.024)
3	0.033	(0.025, 0.065)
4	0.066	(0.045, 0.095)
5	0.113	(0.088, 0.168)
6	0.174	(0.128, 0.238)
7	0.245	(0.180, 0.320)

[a]Confidence intervals computed for each fixed t separately by inversion of a likelihood ratio test.

ment exits, then individuals who are detected in an asymptomatic state (stage 1) must be treated. Consider a cohort of disease-free individuals defined at calendar time (or chronological age) $s = 0$. Individuals are periodically screened for stage 1 disease. If an individual is detected with stage 1 then he is treated. At the time treatment begins, the individual no longer contributes information about natural history because the natural course of the disease is altered.

Assuming parametric models for the stage 1 and stage 2 disease incidence functions, the full likelihood function can be developed under the assumption of independence between onset time and duration of stage 1 disease. In the notation of the preceding section, individuals contribute one of three possible factors to the likelihood corresponding to ($\epsilon_i = 1$), ($\epsilon_i = 0$, $\delta_i = 0$), and ($\epsilon_i = 0$, $\delta_i = 1$). For example, the likelihood contribution for an individual detected with stage 1 disease ($\epsilon_i = 1$) at time R_i is Eq. (4) with the stage 2 probability density function $f(t; \underset{\sim}{X}_2)$ replaced by the survival function $1 - F(t; \underset{\sim}{X}_2)$. The stage 1 durations for individuals detected by the screen ($\epsilon_i = 1$) are right censored because of treatment intervention. Accordingly, important information for estimating the parameters of $F(t_j \underset{\sim}{X}_2)$ is derived from individuals with onset of both stage 1 and stage 2 disease between two successive screening tests ($\epsilon_i = 0$, $\delta_i = 1$).

Under a rare disease assumption with constant incidence of stage 1 disease, $\lambda(s; \underset{\sim}{X}_1) = \exp(\alpha_0 + \underset{\sim}{\alpha}_1\underset{\sim}{X}_1)$, an approximate Poisson likelihood can be constructed as in Day and Walter (*15*). Suppose the screening tests occur at fixed times $S_1, S_2, \ldots, S_k$ for all individuals. Then the number of individuals, d_i, who are diagnosed with stage 2 disease between the ith and $(i+1)^{st}$ screen has an approximate Poisson distribution with mean

$$e^{\alpha_0 + \underset{\sim}{\alpha}_1\underset{\sim}{X}_1} \cdot F(s_{i+1} - s_i)$$

The number of individuals detected with stage 1 disease at the ith screen is also approximately Poisson distributed with mean

$$e^{\alpha_0 + \underset{\sim}{\alpha}_1\underset{\sim}{X}_1} \int_0^{s_i - s_{i-1}} [1 - F(u)]du$$

A modification of this approach has been used in a breast screening program (*1,15*) to account for errors in the screening test.

Case-Control Studies of Serially Screened Populations with Treatment Intervention

There are a number of disadvantages of the cohort study described in the preceding section. If the disease is rare ($\lambda(s; \underset{\sim}{X}_1)$ small) a very large cohort would be required and the screening program would have to be centrally organized. An alternative design is the case-control study. The advantage, of course, is follow-up on a large cohort is not required. However, an important limitation of the case-control design is that the absolute incidence of stage 1 is not estimable because the numbers of cases that are sampled are pre-fixed. Nevertheless, the case-control approach can be useful, and we briefly consider analytic approaches for gleaning information about natural history from matched case-control studies.

In case-control studies of a serially screened population, there are two types of cases. The first type is cases with incident stage 2 disease, and the second type are cases that are screen detected with stage 1 disease. Cases that are diagnosed between screens with incident stage 2 disease are called interval cases to emphasize that the cases are diagnosed in the interval between screens. Cases detected by the screening test with stage 1 disease are called screen detected cases. Each interval case is matched to R controls on the basis of specified matching criteria (R may vary across matched sets). These controls are required to be free of stage 2 disease at the time of diagnosis of the case. Each screen detected case is also matched to R controls. These controls are required to have screened negative at the time the screen detected case was found with stage 1 disease.

We assume an underlying cohort defined as $s = 0$. Individuals are screened at random times, beginning at some sufficiently large time, s^*, so that $F(s^*) \approx 1$.

We assume stage 1 disease is rare with constant incidence rate. We allow for the fact that the incidence of stage 1 disease may depend upon covariates not included in the matching criteria. We assume the incidence of stage 1 disease for the jth individual in the ith matched set with covariate vector $\underset{\sim}{X}_{ij}$ is $\lambda_{ij} = \lambda_i\exp(\underset{\sim}{\alpha}\underset{\sim}{X}_{ij})$. In this model, we are allowing for the possibility that the baseline incidence of stage 1 disease, λ_i, may vary across matched sets. The covariate vector, $\underset{\sim}{X}_{ij}$, includes covariates that are not included in the matching criteria that effect incidence of stage 1 disease. It is further assumed for simplicity that there is a common distribution function of stage 1 durations, $F(t)$, which does not depend on any covariate. The time interval between selection as a case or control and the last prior screening test is ascertained and is called t_{ij}, where i indexes the matched set and $j = 0, \ldots, R$ indexes the individuals in the set. By convention we let $j = 0$ refer to the case. If the individual did not have a prior screen we set $t_{ij} = +\infty$.

We assume a parametric model for $F(t)$ and that F is

a proper distribution function; that is we assume a progressive disease model. We can construct a conditional likelihood (*16,17*) that involves the parameters α and the parameters of $F(t)$. Matched sets (based on interval cases) contribute the following factor to the conditional likelihood

$$\frac{e^{\alpha \mathbf{X}_{i0}} F(t_{i0})}{\sum_{j=0}^{R} e^{\alpha \mathbf{X}_{ij}} F(t_{ij})} \tag{5}$$

We note the term λ_i appears in the numerator and denominator of Eq. (5) and cancels out. Matched sets based on screen detected cases contribute analogous contributions as Eq. (5) except F is replaced by F_B, the backward recurrence time distribution

$$F_B(t) = \frac{1}{\mu} \cdot \int_0^t (1 - F(u))du$$

where $\mu = \int_0^\infty (1 - F(u))du$ is the expected stage 1 duration. The backward recurrence time distribution is the distribution function of durations spent in stage 1 for an individual who is known to be prevalent with stage 1 disease at a fixed point in time. The conditional maximum likelihood estimates are found by maximizing the product of the likelihood contributions from the interval and screen-detected matched sets. The conditional likelihood is independent of the parameters λ_i, the baseline incidence of stage 1 disease in the ith matched set. This serves to emphasize again that it is not possible to estimate absolute incidence from a case-control study.

A modification of this analytic approach was used in the analysis of a matched case-control study of PAP smear screening for cervical cancer in Northeast Scotland (*17*). The modification accounted for screening test errors, and also accounted for the fact that not all preclinical lesions (stage 1 disease) progress to clinically (symptomatic) invasive cervical cancer (stage 2).

Errors in the Screening Test

We have considered the strengths, limitations, and analytic approaches associated with various epidemiologic designs for studies of disease natural history. An underlying assumption of the preceding sections was that diagnostic errors of the screening test were negligible and could be ignored. Often the test errors are not negligible; thus, the analytic approaches must be modified accordingly. There are two types of errors. The false positive error occurs when an individual without stage 1 disease falsely tests positive. The false negative error occurs when an individual with stage 1 disease falsely tests negative.

In order to develop analytic procedures in the presence of test errors, additional probabilistic assumptions are required. For example, one can assume the probability of a false positive error is 0, the probability of a false negative error is ϵ, and further errors on successive screens are independent. These were the assumptions employed in an analysis of natural history from a breast cancer screening program (*15*) and a cervical cancer screening program (*16*).

An alternative to the independence assumption is to assume a proportion of individuals with stage 1 disease always falsely test negative. Another alternative is to assume the probability of a false negative error changes over the course of stage 1 disease. For example, the preclinical stage (stage 1) of cervical cancer can be divided into a noninvasive preclinical disease phase and an invasive preclinical disease phase. One can assume that the false negative probabilities are different for the two phases and that test errors conditional of the disease state (i.e., preclinical noninvasive or preclinical invasive) are independent.

Under model assumptions for test errors such as those described above, the likelihood function for the various epidemiologic designs could be derived. The likelihood function would include an additional parameter, ϵ, which is the probability of false negative error. This parameter can be estimated jointly along with the other natural history parameters (*1,17*). However, an important caveat is that parameter estimates may be highly correlated. For example, under the independence assumption it was found that the estimate of ϵ, the probability of a false negative error, and μ, the mean stage 1 duration, were highly correlated (*15*). Clearly, it would be preferable to use a reliable external estimate of ϵ, if available, rather than to jointly estimate ϵ along with the other natural history parameters.

However, even if an estimate of ϵ was available, assumptions about the joint distribution of successive test results, conditional on the true disease state, would be required. An important issue in studies of disease natural history concerns the development of plausible assumptions about screening test errors and the sensitivity of estimates of natural history parameters to alternative model assumptions about these errors.

R. Brookmeyer was partially supported by Public Health Service grants CA-48723 from the National Cancer Institute and AI-16959 from the National Institute of Allergy and Infections Diseases.

REFERENCES

1. Walter, S. D., and Day, N. E. Estimation of the duration of a preclinical disease state using screening data. Am. J. Epidemiol. 118: 865–885 (1983).
2. Campion, M. J., McCance, D. J., Cuzick, J., and Singer, A. Progressive potential of mild cervical atypia: prospective cytological colposcopic and virologic study. Lancet ii: 237–240 (1986).
3. Kurman, R. J., Schiffman, M. H., Lancaster, W. O., Reid, R., Jenson, A., Temple, R., and Lorincz, A. Analysis of individual human papillomavirus types in cervical neoplasia: a possible role for type 18 in rapid progression. Am. J. Obstet. Gynecol. 159: 293–296 (1988).
4. Barre-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., Daugvet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. Isolation of T-lymphotropic retrovirus from a patient at risk for

acquired immune deficiency syndrome (AIDS). Science 220: 868–871 (1983).

5. Gallo, R. C., Salahuddin, S. Z., Popovic, M., Shearer, G., Kaplan, M., Haynes, B., Palker, T., Redfield, R., Oleske, J., Safai, B., White, G., Foster, P., and Markham, P. Frequent detection and isolation of cytopathic retrovirus (HTLV-III) from patients with AIDS and at risk for AIDS. Science 224: 500–503 (1984).
6. Eyster, M. E., Gail, M. H., Ballard, J. O. et al. Natural history of human immunodeficiency virus infections in hemophiliacs: effects of T-cell subsets, platelet counts and age. Ann. Intern. Med. 107: 1–6 (1987).
7. Brookmeyer, R., and Gail, M. Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States. Lancet 2: 1320–1322 (1986).
8. Brookmeyer, R., Gail, M. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. J. Am. Stat. Assoc. 83: 301–308 (1988).
9. Gail, M. H., and Brookmeyer, R. Methods for projecting the course of the AIDS epidemic. J. Natl. Cancer Inst. 80: 900–911 (1988).
10. Lorincz, A. T., Temple, G. F., Kurman, R. J., Jenson, A. B., and Lancaster, W. D. Oncogenic association of specific human papillomavirus types with cervical neoplasia. J. Natl. Cancer Inst. 79: 671–676 (1987).
11. Brookmeyer, R., and Gail, M. H. Biases in prevalent cohorts. Biometrics 43: 739–749 (1987).
12. Goedert, J. J., Biggar, R. J., Weiss, S. H., Eyster, M., Melbye, M., Wilson, S., Ginzburgh, H., Grossman, R., Di Gioia, R., Sanchez, W., Giron, J., Ebbesen, P., Gallo, R., and Blattner, W. Three-year incidence of AIDS among HTLV-III infected risk group members. Science 231: 992–995.
13. Hessol, N. A., Rutherford, G. W., O'Malley, P. M., Doll, L., Darrow, W., and Jaffe, H. The natural history of human immunodeficiency virus infection in a cohort of homosexual and bisexual men: a seven year prospective study. Presented at the Third International Conference on AIDS, Washington, DC, June 1987.
14. Brookmeyer, R., and Goedert, J. J. Censoring in an epidemic with an application to hemophilia associated AIDS. Biometrics 45: 325–335 (1989).
15. Day, N. E., and Walter, S. D. Simplified models of screening for chronic disease from mass screening programs. Biometrics 40: 1–14 (1984).
16. Brookmeyer, R., Day, N. E., and Moss, S. Case-control studies for estimation of the natural history of preclinical disease from screening data. Stat. Med. 5: 127–138 (1986).
17. Brookmeyer, R., and Day, N. E. Two-stage models for the analysis of cancer screening data. Biometrics 43: 657–670 (1987).