New Techniques for the Analysis of Cohort Studies

Duncan Thomas

INTRODUCTION

Cohort studies involve the key element of follow-up of individuals over time to study an outcome in relation to some earlier exposure factor or a fixed host characteristic (such as genotype). While the outcome under study could be the change in some continuous variable, we shall restrict this review to studies of disease incidence. Subjects can be randomly assigned to different exposures, as in a clinical or prevention trial, or the exposure histories of free-living individuals can be passively observed, as in most epidemiologic cohort studies. Although the issues of confounding and comparability are very different in randomized and observational studies, the basic analysis methods are similar (except, perhaps, for a greater emphasis on adjustment for covariates in observational studies) and the distinction will be ignored. Similarly, follow-up can be conducted “prospectively” or “retrospectively” in real time, but this too has no significance for methods of analysis.

We begin with a brief description of several cohort studies with different types of data structures and different analysis problems that will be used to illustrate the statistical issues. Following a review of the basic approaches to the analysis of the different types of cohort data, we focus on empirical and mechanistic approaches to model specification. Some special problems, such as measurement error, dependent outcomes, and the unique problems of reproductive data, are addressed. We conclude with a more in-depth treatment of approaches to the analysis based on cohort sampling methods—the nested case-control and case-cohort designs, and variants thereof.

EXAMPLES OF COHORT STUDIES

Atomic bomb survivors

One of the largest cohort studies ever conducted is of the 120,128 survivors of the atomic bombing of Hiroshima and Nagasaki, Japan (1). The cohort comprises all those residents of the two cities at the time of bombing who survived to 1950. Passive follow-up has been conducted using the resources of the Japanese family registration (koseki) system and is planned to continue to the extinction of the cohort. Exposure was estimated for each survivor based on information obtained at entry about location at the time of bombing, combined with elaborate physical models for dose as a function of distance, position, and shielding. This study illustrates an important class of studies involving a single instantaneous exposure, as well as the analysis efficiencies that can result from the use of grouped data.

Diet cohorts

In contrast with the study of a single instantaneous exposure are several cohort studies involving relatively short-term follow-up of subjects in relation to their reported dietary habits at entry (2). Although several of these cohorts have now been followed for many years, with updated dietary information and substantial losses along the way, we use them here to illustrate analytical approaches where the exposure variables under study represent “usual” lifetime exposure (assumed constant over the period of follow-up) and where the prospective observation means subjects are all at risk for essentially the same period of time.

Uranium miners

Occupational cohort studies comprise a more complicated situation, typically involving extended and time-varying exposures and variable periods of time at risk. A good example is the US Public Health Service study of mortality in 3,347 uranium miners on the Colorado plateau (3–7). This study was initiated in the 1950s because of a concern about the risks of lung cancer from the high levels of radon and its daughter...
products. Exposure information was obtained from the mining companies' payroll records, combined with measurements, extrapolations, and "guessimates" of radon levels in the mines over time.

Smoking information was also obtained at entry and updated several times. The primary endpoint is lung cancer mortality, with 329 cases having occurred by 1987.

A reproductive cohort

Reproductive endpoints raise several new issues relating to the nature of the endpoints. In addition to continuous outcomes, such as birthweight, there are two main types of binary endpoints—those manifest only at birth, such as congenital malformations, and those that can occur throughout the pregnancy, such as spontaneous abortions. The former is normally treated as a dichotomous endpoint, with all subjects having been at risk for the same period of time (except for variation in gestational age, which is usually treated as a confounder), whereas spontaneous abortions need to be treated as censored event-time data. However, the two endpoints are closely interrelated since fetuses with severe malformations are likely to be aborted. Furthermore, the time of entry into the cohort (recognition of a pregnancy, not conception) will generally vary and may be difficult to pinpoint. These problems are well illustrated by a cohort of 7,450 pregnancies in San Francisco Bay Area women enrolled in a health maintenance organization who were exposed to varying degrees of aerial spraying of the pesticide malathion during the early 1980s (8).

Family cohort

Genetic studies usually involve family data, often including extended pedigrees. The analysis of genetic segregation and linkage models is beyond the scope of this review (see, for example, Ott (9) and Khoury et al. (10)), but some unique issues arise in analyzing the effects of family history or measured genes within families. These issues are well illustrated by analyses of breast cancer in the families of the cases and controls from the Cancer and Steroid Hormone study (11-13). In addition to many other "environmental" factors, data on the history of cancer in first-degree relatives of cases and controls were obtained. The standard case-control analysis (11) simply combines all this information into a classification of family history as positive or negative (and subcategories of positive); no special problems of dependency arise in this analysis, since only the cases and controls themselves are included in the analysis and they are independent. Claus et al. (12), however, excluded the original cases and controls (since their outcomes were determined by design) and treated their family members as a cohort of subjects exposed since birth to the risk factor of having an affected or unaffected proband; although more informative, these analyses must then deal with any residual dependency in outcomes within families not accounted by this risk factor. In a subsequent paper (13), such dependencies were addressed using segregation analysis to infer whether they could be explained in terms of a single major gene and/or polygenic effects.

BASIC ANALYTICAL APPROACHES

All these designs involve the collection of a set of data for each individual i = 1, . . . , I comprising an event time or censoring time ti, a censoring indicator di = 1 if the subject is affected, zero otherwise, and a vector of covariates Zi (exposures, confounders, and modifiers), which can be time-dependent.

Risk models are used to describe the incidence rate λ(t, Z) for times-to-event (disease diagnosis or death) as a function of time t and covariates Z = (Z1, . . . , Zp), which may themselves be time-dependent. A special class of risk models that has been widely used in epidemiology are known as "relative risk models," which are based on the proportional hazards assumption,

\[ \lambda(t, Z) = \lambda_0(t) r[Z(t); \beta] \quad (1) \]

where \( \beta \) represents a vector of parameters to be estimated and \( \lambda_0(t) \) is an unknown set of age-specific "baseline" rates for subjects with \( Z = 0 \). In the standard proportional hazards model, the relative risk term takes the log-linear form \( r[Z, \beta] = \exp[Z' \beta] \). This has the convenient property that it is positive for all possible covariate and parameter values, since the hazard rate itself must be non-negative. However, in particular applications, some alternative form of relative risk model may be more appropriate. Although time since entry to the study is commonly used as the time axis t in the analysis of clinical trial data, age is a more appropriate axis for most cohort studies (14). Other temporal factors, such as calendar date, or time since exposure began may also be relevant and can generally be handled either by treating them as covariates or by stratification. One might even consider some biologic time scale related to the underlying disease process; for example, Krailo et al. (15) fitted data on breast cancer using a model for the "breast tissue aging rate" proposed by Pike et al. (16) based on reproductive history. Note, however, that if any scale other than time since entry to the cohort is used, one must then deal with the problem of staggered entry times (e.g.,
age at first employment in the occupational example or gestational age at diagnosis of pregnancy in the reproductive example).

The process of specifying an analysis entails two distinct steps. First, one must choose a form of analysis appropriate to the particular data structure available. Second, one must specify a particular model for the relations amongst the variables. These two steps overlap in the sense that most any data structure can be fitted to most any model, although conventionally the two are often treated as linked in the sense that a particular data structure is often analyzed with a particular model. To avoid this narrow perspective, we have organized the following discussion by first discussing forms of analysis appropriate to the most common data structures and then focusing the rest of the review on models for disease incidence or mortality data.

**Likelihoods and data structures**

The appropriate likelihood depends on the sampling design and data structure. The key elements in determining the appropriate analysis are:

- whether the subjects are to be treated as individuals or grouped on the basis of their exposure histories in some way; for example, the basic analyses of the atomic bomb survivor cohort have all been based on grouped data (by dose, age, gender, city, and time) because the large size of the cohort essentially precludes extensive exploratory analyses on an individual basis;
- if grouped, whether a single exposure variable is of particular interest (the others being treated as confounders), or the joint effects of multiple variables are to be modeled; for example, malathion was the primary exposure of interest in the reproductive cohort, whereas unscrambling the effects of multiple dietary components is a key aim of the dietary cohort studies;
- whether subjects are considered to be at risk for an essentially constant period of time (as in a short-term cohort study or trial) with relatively little censoring, or the periods of observation vary considerably between individuals; for example, both the reproductive and dietary cohorts were followed for a relatively short period with few censoring events, whereas the two radiation cohorts entail lifetime follow-up;
- if observation time is extended, whether the covariates are constant or time-dependent (the atomic bomb survivor and uranium miner cohorts, respectively), and whether assumptions are to be made about the baseline risk as a function of time or age; and
- whether all subjects are to be included in the analysis or only a sample of them (we will illustrate cohort sampling options below using the uranium miner cohort).

These various questions will then influence the approach to the analysis. The most commonly used alternatives are briefly reviewed in the rest of this section. A more detailed treatment of the standard methods of analysis can be found in the standard text of Breslow and Day (17) and recent epidemiologic textbooks.

**Standardized mortality ratio analysis.** The simplest analysis of cohort data is a comparison of the numbers of observed events \( N \) with their corresponding expected numbers \( E \), via the standardized mortality ratio, \( SMR = N/E \). Expected numbers are usually estimated by multiplying a set of "standard" rates \( \lambda^* \) to the person-time at risk \( T_s \) in strata \( s \) defined by age, gender, calendar time, and perhaps other factors, and summing over strata to produce \( E = \sum_s T_s \lambda^* \). External rates (e.g., national) are commonly used to determine whether the cohort rates are different, but if the primary interest concerns internal comparisons between subcohorts with different exposures, the rates for the entire cohort can be used as the standard. The method described above, known as "indirect standardization," is but one of several ways of standardizing for the stratifying factors, but it is the most commonly used method and the one that is most closely related to the multivariate methods to be discussed below. Breslow and Day (17) provide a comprehensive treatment of the alternative methods of standardization, as well as methods for significance testing, confidence interval estimation, comparison between subcohorts, and modeling.

**Poisson regression.** If there are several risk factors under study, it may be more revealing to model their joint effects than simply to describe the effect of each, adjusting for the others. For large datasets, it may be more convenient to analyze the data in grouped form using Poisson regression (17). This technique provides the natural multivariate generalization of the standardized mortality rate method. For this purpose, the total person-time of follow-up is grouped into \( k = 1, \ldots, K \) categories on the basis of time and covariates, and the number of events \( N_k \) and person-time \( T_k \) in each category is recorded, together with the corresponding values of the person-time-weighted averages of age \( t_k \) and covariates \( z_k \). The proportional hazards model now leads to a Poisson likelihood for the grouped data of the form

*Epidemiol Rev* Vol. 20, No. 1, 1998
Analysis of Cohort Studies

\[ L(\lambda, \beta) = \prod_{k=1}^{K} \Pr(N_k|E_k) = \prod_{k=1}^{K} E_k^{N_k} \exp(-E_k)/N_k! \]  

(2)

where \( E_k = \lambda_i T_k r(z_i; \beta) \) and \( \lambda_i = \lambda_0(t_i) \) denote a set of baseline hazard parameters that must be estimated together with \( \beta \).

**Logistic regression.** For a clinical trial or cohort study with the same period of observation for all subjects, but where only the disease status, not the event-time itself, is observed, a logistic model for the probability of an event of the form \( \Pr(D = 0|Z) = \frac{1 + \alpha r(Z, \beta)}{1 + \alpha r(Z, \beta)^{-1}} \) might be used, where \( \alpha \) is the odds of the event for a subject with \( Z = 0 \). Again, the standard form is obtained using \( r(Z, \beta) = \exp(Z^T \beta) \). The likelihood for this design would then be

\[ L(\alpha, \beta) = \prod_{i} \Pr(D = d_i|Z = z_i; \alpha, \beta) = \prod_{i} \frac{[\alpha r(z_i; \beta)]^{d_i}}{1 + \alpha r(z_i; \beta)} \]  

(3)

**Survival analysis.** In a clinical trial or cohort study in which the event times are observed, the proportional hazards model (equation 1) leads to a full likelihood of the form

\[ L[\lambda_0(\cdot), \beta] = \prod_{i} \lambda_0(t_i)^{d_i} r[z_i(t_i); \beta]^{d_i} \times \exp \left\{ - \int_{t_n}^{t_i} \lambda_0(t) r[z_i(t); \beta] dt \right\} \]  

(4)

where \( s_i \) denotes the entry time of subject \( i \). Use of the full likelihood requires specification of the form of the baseline hazard, for example, constant (exponential survival), step function, Weibull, or Gompertz (18). Cox (19) proposed instead a “partial likelihood” of the form

\[ L(\beta) = \prod_{n=1}^{N} \frac{r[z_n(t_n); \beta]}{\sum_{j \in R_n} r[z_j(t_n); \beta]} \]  

(5)

where \( n = 1, \ldots, N \) indexes the observed event times, \( i_n \) denotes the individual who fails at time \( t_n \), and \( R_n \) denotes the set of subjects at risk at time \( t_n \). (When using a time scale like age, this may raise the issue of staggered entry times, discussed earlier, where \( R_n \) includes only subjects who have entered the cohort by time \( t_n \). Although some Cox regression programs do not explicitly allow for staggered entry times, it is often possible to deal with this by creating a time-dependent indicator for times when the subject is not in view and fixing its regression coefficient to a large negative value, thereby reducing its contributions at such times to zero.) This likelihood does not require any specification of the form of the baseline hazard; the estimation of \( \beta \) is said to be “semi-parametric,” as the relative risk factor is still specified parametrically (e.g., the loglinear model in the standard form).

**Nested case-control and case-cohort sampling.** This partial likelihood can also be used to fit relative risk models for nested case-control studies within a cohort, where \( n \) now indexes the cases and \( R_n \) indicates the set comprising the \( n \)th case and his/her matched controls. This approach is discussed more fully below.

**Models**

**Why model relative risks?** Before proceeding further, it is worth pausing to inquire why one might wish to adopt the proportional hazards model at all. Certainly, there are examples where some other form of model provides a better description of the underlying biologic process. Although any risk model can be reparameterized in proportional hazards form, it may be that a more parsimonious model can be found using some alternative formulation, such as an excess risk model

\[ r(t, Z) = \lambda_0(t) + Z^T \alpha. \]

In this case, whether the proportional hazards or excess risk model provides a more parsimonious description of the data depends on which is more nearly constant over time (or requires the fewest time-dependent interaction effects).

The advantages of relative risk models are both mathematical and empirical. Mathematically, the proportional hazards model allows “semi-parametric” estimation of covariate effects via partial likelihood without requiring parametric assumptions about the form of the baseline hazard. Furthermore, the asymptotic distribution theory for estimating confidence regions and significance testing generally seems to apply at smaller sample sizes than for most alternative models. Empirically, it appears that many survival-time processes do indeed show rough proportionality of the hazard to time and covariate effects, at least with appropriate specification of the covariates. Evidence of this phenomenon for cancer incidence is reviewed in Breslow and Day (20, chapter 2): age-specific incidence rates from a variety of populations have more nearly constant ratios than differences. Such considerations have led the view amongst most cancer epidemiologists that the relative risk model is the “right” one biologically for that endpoint, but this is not necessarily the case for other endpoints (14).

**Some alternatives.** Two alternative models that have received some attention are the excess risk model
and the accelerated failure time model. The excess risk model \( \lambda(t; Z) = \lambda_0(t) + Z'\alpha \) was often used in early work on the radiation field, in part because of the simplicity of the resulting risk assessment calculations. However, in addition to growing evidence that it did not fit radiation data as well as the relative risk model, it does not allow the types of semiparametric inference on exposure effects that is possible under the relative risk model, where no parametric assumptions about baseline risks are needed. Recent work (21, 22), however, provides quite a general framework for semiparametric inference in additive models that merits further consideration. In particular, the model allows the magnitude of the regression coefficients \( \alpha \) to vary over time in an arbitrary fashion, so that one can get a visual feel for whether a constant excess risk model would be appropriate.

The accelerated failure time model is generally written in the form \( \ln t = Z'\gamma + \epsilon \) (for uncensored observations) where the residuals \( \epsilon \) are assumed to have some common, but unspecified, distribution \( f(\epsilon) \). This expression provides a natural interpretation of the regression coefficients \( \gamma \) in terms of the effects of covariates on the mean survival times. The same model can also be expressed in terms of the incidence rate as \( \lambda(t; Z) = \lambda_0(t) e^{-Z'\gamma} \), where the baseline rate \( \lambda_0(t) \) is related to the distribution of residuals \( f(\epsilon) \). The model is easily fitted to uncensored event times under parametric assumptions about the distribution of residuals and can be extended in a straightforward manner to censored event times (18, chapter 3). Its extension to semiparametric (rank) regression for censored data is more complex, but methods are now available (18, chapter 6; 23; 24). Alternatively, semiparametric regression models have been recently developed for median, rather than mean, survival times, median \( m(t) = Z'\gamma + \epsilon \) where the error terms again have some common but unknown distribution (25).

**GENERAL MODELING ISSUES**

For any of these likelihoods, it suffices to substitute some appropriate function for \( r(Z; \beta) \) and then use the standard methods of maximum likelihood to estimate its parameters and test hypotheses. In the remainder of this section, we discuss various approaches to specifying this function. The major distinction we make is between empiric and mechanistic approaches. Empiric models are not based on any particular biologic theory for the underlying disease process, but simply attempt to provide a parsimonious description of it, particularly to identify and quantify the effects of covariates that affect the relative risk. Perhaps the best-known empiric model is the loglinear model for relative risks, but other forms may be appropriate for testing particular hypotheses or for more parsimonious modeling in particular datasets, as discussed in the following section. With a small number of covariates, it may also be possible to model the relative risk nonparametrically. Mechanistic models, on the other hand, aim to describe the observed data in terms of some unobservable underlying disease process, such as the multistage theory of carcinogenesis. A more mathematical treatment of risk modeling can be found in Thomas (26).

**Empiric models**

The log-linear model, \( r(Z; \beta) = \exp(Z'\beta) \), is probably the most widely used empiric model and is the standard form included in all statistical packages for logistic, Cox, and Poisson regression. As noted earlier, it is nonnegative and it produces a nonzero likelihood for all possible parameter values, which doubtless contributes to the observation that in most applications, parameter estimates are reasonably normally distributed, even with relatively sparse data. However, the model involves two key assumptions that merit testing in any particular application:

- For a continuous covariate \( Z \), the relative risk depends exponentially on the value of \( Z \); and
- For a pair of covariates, \( Z_1 \) and \( Z_2 \), the relative risk depends multiplicatively on the marginal risks from each covariate separately (i.e.,
  \[ r(Z; \beta) = r(Z_1; \beta_1) r(Z_2; \beta_2) \].

Neither of these assumptions is relevant for a single categorical covariate. In other cases, the two assumptions can be tested by nesting the model in some more general model that includes the fitted model as a special case, for example, by adding covariate transformations or interaction terms to a model of the same form.

If these tests reveal significant lack of fit of the original model, one might still be satisfied with the expanded model as a reasonable description of the data, but one should then also consider the possibility that the data might be more parsimoniously described by some completely different form of model. For example, a negative quadratic term might suggest that a linear model be tried, and a negative interaction term might suggest an additive model. Thus, one might be led to a model of the form \( r(Z; \beta) = 1 + Z'\beta \). In other circumstances, one might consider a linear-multiplicative or loglinear-additive model.

In a rich dataset, the number of possible alternative models can quickly get out of hand, so some structured approach to model building is needed. The key is to adopt a general class of models that would include all the alternatives one might be interested in as special cases, allowing specific submodels to be tested within

*Epidemiol Rev* Vol. 20, No. 1, 1998
nested alternatives. A general model that has achieved some popularity recently consists of a mixture of linear and loglinear terms of the form

\[ r(Z, W; \beta, \gamma) = \exp(W'_0 \gamma_0) \times \left[ 1 + \sum_{m=1}^{M} Z'_m \beta_m \exp(W'_m \gamma_m) \right] \]

where \( \beta_m \) and \( \gamma_m \) denote vectors of regression coefficients corresponding to the subsets of covariates \( Z_m \) and \( W_m \) included in the \( m \)th linear and loglinear terms, respectively. A special case that has been widely used in radiobiology (7, 27, 28) is of the form

\[ r(Z, W; \beta, \gamma) = 1 + \beta Z + \beta^2 Z^2 \exp(-\beta_3 Z + W' \gamma) \]

where \( Z \) represents radiation dose (believed from microdosimetry considerations to have a linear-quadratic effect on mutation rates at low doses multiplied by a negative exponential survival term to account for cell killing at high doses) and \( W \) comprises modifiers of the slope of the dose-response relation, such as attained age, sex, latency, or age at exposure. For example, including the log of latency and its square in \( W \) allows for a lognormal dependence of excess relative risk on latency.

Comparisons of alternative models that are nested within such a general class can be accomplished using standard likelihood ratio tests. Models that are of a fundamentally different form can always be nested within some more general class, such as the exponential mixture of linear-additive and loglinear-multiplicative models proposed by Thomas (29)

\[ r(Z; \beta, \theta) = (1 + Z' \beta)^{1-\theta} \exp(\theta Z' \beta) \]

which produces the linear model when \( \theta = 0 \) and the loglinear model with \( \theta = 1 \). Several alternative mixtures have been proposed, of which the Guerrero-Johnson (30) mixture

\[ r(Z; \beta, \theta) = \begin{cases} \exp(Z' \beta), & (\theta = 0) \\ (1 + \theta Z' \beta)^{1/\theta}, & (\theta \neq 0) \end{cases} \]

appears to have the most satisfactory statistical properties (26, 31, 32). However, a word of warning is needed concerning inference on the parameters of most nonstandard models. Their likelihood is generally far from normal (32), leading to highly skewed confidence regions and Wald tests that are seriously weakened (33, 34). Thus, inference should be based on the likelihood ratio test and likelihood-based confidence limits. For example, Lubin and Gaffey (35) describe an application of the exponential mixture of linear-additive and linear-multiplicative models (29) to testing the joint effect of radon and smoking on lung cancer risk in uranium miners; the point estimate of \( \theta \) was 0.4, apparently closer to additivity than multiplicativity, but the likelihood ratio tests rejected the additive model \( (\chi^2 = 9.8) \) but not the multiplicative model \( (\chi^2 = 1.1) \). A linear mixture showed an even more skewed likelihood, with \( \hat{\theta} = 0.1 \) (apparently nearly additive) but with very similar likelihood ratio tests that rejected the additive but not the multiplicative model.

**Extended exposure histories.** Chronic disease epidemiology often involves measurement of an entire history of exposure \( \{X(u), u < t\} \) which we wish to incorporate into a relative risk model through one or more time-dependent covariates \( Z(t) \). How this is done depends upon one's assumptions about the underlying disease mechanism.

Most approaches to exposure-response modeling in epidemiology are based on an implicit assumption of dose additivity, i.e., that the excess relative risk at time \( t \) is a sum of independent contributions from each increment of exposure at earlier times \( u \), possibly modified in some fashion by temporal factors. In a relative risk model, this hypothesis might be expressed generally as

\[ r[t, X(.); \beta, \gamma] = R \left( \int_0^t f[X(u); \alpha] g(t,u; \gamma) du; \beta \right) \]

where \( R(Z; \beta) \) is some known relative risk function such as the linear or loglinear models discussed above, \( f \) is a known function describing the modifying effect of dose-rate, and \( g \) is a known function describing the modifying effect of temporal factors. For example, the choice \( f[X] = X \) and \( g(t,u) = 1 \) leads to the standard relative risk model based on cumulative exposure, presumably the most widely used exposure index in epidemiology. For many diseases with long latency, such as cancer, it is common to use lagged cumulative exposure, corresponding to a weighting function of the form \( g(t,u; \gamma) = 1 \) if \( t - u > \gamma \), zero otherwise. Other simple exposure indices might include latency-weighted exposure \( \int_0^t X(u) (t - u) du \) or age-weighted exposure \( \int_0^t X(u) u du \). Similarly, the function \( f \) can be used to test dose-rate effects (the phenomenon that a long, low-intensity exposure has a different risk from a short, high-intensity exposure for the same cumulative dose). For example, letting \( f[X(u); \alpha] = X(u)^\alpha \) and \( R(Z; \beta, \alpha) = 1 + \beta Z^\alpha \) generates a family of exposure-response functions, ranging from a cumulative linear relation (for \( \alpha = 1 \)) to those which show conventional dose-rate (\( \alpha > 1 \)) and inverse dose-rate (\( \alpha < 1 \)) effects. Unfortunately, the additivity assumption has
seldom been tested, although in principle this could be done by nesting the dose-additive model in some more general alternative. (See Thomas (26) for further details and a discussion of fitting methods.) For example, in the uranium miner data, we have tested this hypothesis by adding a covariate of the form \( \int_0^1 \int_0^1 \sqrt{X(u)X(v)} f(u) g(v-u) h(t-v) \, dv \, du \) to the equation, but found no significant improvement in the fit for \( \alpha \) of several simple choices of the weight functions \( f, g, \) or \( h \), suggesting that the dose additivity assumption is reasonable for these data.

**Nonparametric models.** The appeal of Cox’s partial likelihood is that no assumptions are needed about the form of the dependence of risk on time, but it remains parametric in modeling covariate effects. Even more appealing would be a nonparametric model for both time and covariate effects. For categorical data, no parametric assumptions are needed, of course, although the effects of multiple covariates are commonly estimated using the loglinear (i.e., multiplicative) model, with additional interaction terms as needed. Similarly, continuous covariates are frequently categorized to provide a visual impression of the exposure-response relation, but the choice of cutpoints is arbitrary. However, nonparametric smoothing techniques are now available to allow covariate effects to be estimated without such arbitrary grouping.

One approach relies only on an assumption of monotonicity. Thomas (36) adapted the technique of isotonic regression to relative risk modeling, and showed that the maximum likelihood estimate of the exposure-response relation under this constraint was a step function with jumps at the observed covariate values of a subset of the cases. The technique has been extended to two dimensions (37), but in higher dimensions the resulting function is difficult to visualize and can be quite unstable.

Cubic splines and other means of smoothing provide attractive alternatives which produce smooth, but not necessarily monotonic relations. The generalized additive model (38) has been widely used for this purpose. For example, Schwartz (39) described the effect of air pollution on daily mortality rates using a generalized additive model, after controlling for weather variables and other factors using similar models. A complex dependence on dew point temperature was found, with multiple maxima and minima, whereas the smoothed plot of the particulate air pollution was seen to be almost perfectly linear over the entire range of concentrations.

**Mechanistic models**

In contrast with the empiric models discussed above, there are circumstances where the underlying disease process is well enough understood to allow it to be characterized mathematically. Probably the greatest activity along these lines has been in the field of cancer epidemiology. Two models in particular have dominated this development, the multistage model of Armitage and Doll (40) and the two-event model of Moolgavkar and Knudson (41). For thorough reviews of this literature, see Whittemore and Keller (42), Moolgavkar (43), and Thomas (44); here, we merely sketch the basic ideas.

The Armitage-Doll multistage model postulates that cancer arises from a single cell that undergoes a sequence of \( k \) heritable changes, such as point mutations, chromosomal rearrangements, or deletions, in a particular sequence. The model further postulates that the rate of one or more of these changes may depend on exposure to carcinogens. Then the model predicts that the hazard rate for the incidence of cancer (or more precisely, the appearance of the first truly malignant cell) following continuous exposure at rate \( X \) is of the form

\[
\lambda(t,Z) = \alpha t^{k-1} \prod_{i=1}^{k} (1 + \beta_i X).
\]

Thus, the hazard has a power-function dependence on age and a polynomial dependence on exposure rate with order equal to the number of dose-dependent stages. It further implies that two carcinogens would produce an additive effect if they act at the same stage and a multiplicative effect if they act at different stages. If exposure is instantaneous with intensity \( X(u) \) at age \( u \), its effect is modified by the age at and time since exposure: if it acts at a single stage \( i \), then the excess relative risk at time \( t \) is proportional to \( Z_i(t) = X(u) u^{i-1} (t-u)^{k-i-1}/k^{i-1} \), and for an extended exposure at varying dose rates, the excess relative risk is obtained by integrating this expression over \( u \) (45, 46). Analogous expressions are available for time-dependent exposures to multiple agents acting at multiple stages (47). Note, however, that the expressions given above are only approximations to the far more complex exact solution of the stochastic differential equations (48); the approximate expressions given above are valid only when the mutation rates are all small.

The Moolgavkar-Knudson two-stage model postulates that cancer results from a clone of cells from which one descendent has undergone two mutational events at rates \( \mu_1[Z(t)] \) and \( \mu_2[Z(t)] \), either or both of which may depend on exposure to carcinogens. The clone of intermediate cells is subject to a birth-and-death process with net proliferation rate \( \rho[Z(t)] \) that may also depend on carcinogenic exposures. The number of normal stem cells at risk \( N(t) \) varies with age, depending on the rate of development of the target.
Finally, in genetically susceptible individuals (carriers), all cells carry the first mutation at birth. An approximate expression for the resulting incidence rate at age $t$ is then

$$
\lambda(t, Z) = \begin{cases} 
\mu_2[Z(t)] \int_0^t N(u) \mu_1[Z(u)] du & \text{noncarriers} \\
\mu_2[Z(t)] N(0) \exp\left[\int_0^t \rho[Z(v)] dv\right] & \text{carriers.}
\end{cases}
$$

(10)

Again, note that this expression is only an approximate solution to the stochastic process (49), the validity of which depends upon all the rates being small.

There have been a number of interesting applications of these models to various carcinogenic exposures. For example, the multistage model has been fitted to data on lung cancer in relation to asbestos and smoking (47), arsenic (50), coke oven emissions (51), and smoking (52, 53), as well as to data on leukemia and benzene (54) and nonleukemic cancers and radiation (55). The two- stage model has been fitted to data on lung cancer in relation to smoking (56), radon (57, 58), and cadmium (59), as well as to data on breast (60) and colon cancers (61). Few of these reports have provided any formal assessment of goodness of fit, focusing instead on comparisons between alternative models. This can be done, however, by grouping the subjects in various ways and comparing the numbers of observed and predicted cases; for example, Moolgavkar et al. (58) grouped uranium miners by the temporal sequence of their radon and smoking exposure histories and reported good agreement with the predictions of their two-stage model.

As in any other form of statistical modeling, the analyst should be cautious in interpretation. A good fit to a particular model does not, of course, establish the truth of the model. Instead the value of models, whether descriptive or mechanistic, lies in their ability to organize a range of hypotheses into a systematic framework in which simpler models can be tested against more complex alternatives. The usefulness of the multistage model of carcinogenesis, for example, lies not in our belief that it is an accurate description of the process but, rather, in its ability to distinguish whether a carcinogen appears to act early or late in the process or at more than one stage. Similarly, the importance of the Moolgavkar-Knudson model lies in its ability to test whether a carcinogen acts as an "initiator" (i.e., on the mutation rates) or a "promoter" (i.e., on proliferation rates). Such inferences can be valuable, even if the model itself is an incomplete description of the process, as must always be the case.

**Special problems**

**Measurement error.** The above treatment has assumed either that the covariates $Z$ are accurately measured or that the exposure-response relation that is sought refers to the measured value of the covariates, not to their true values. There is a large and growing literature on methods of adjustment of relative risk models for measurement error, which is beyond the scope of this review (62). However, some general observations are worth making:

- It is well known that the usual effect of measurement error is to bias the relative risk towards the null and weaken power. However, there are some important exceptions. First, in multivariate models with correlated exposures and possibly correlated errors, the bias is not necessarily towards the null; instead, there is a general tendency for the more precisely measured variables to absorb proportionally more of the effect of variables with which they are correlated, but the magnitude and direction of the effects depends upon the correlational structure. Second, one must carefully consider whether the errors are independent of the true values ("classic error"), the measured values ("Berkson error"), or neither, as the effect of measurement error will differ; in linear models, for example, Berkson error does not tend to produce any bias in relative risk estimates. Third, measurement error can distort the shape of an exposure-response relation in various ways, particularly for nonlinear models or error variances that are proportional to the true values.

- Many methods have been proposed for correcting for measurement errors; most involve some form of replacement of the measured values by estimates of the corresponding true values. For example, if validation data are available, one might use them to build a model for true given measured values, and then use this model to impute "true" values for subjects in the main study; adjustments to the standard errors of the relative risks are needed to allow for the uncertainty in this imputation (63, 64). If only a summary estimate of the variance of the error distribution is available, a Bayesian estimate of the expectation of true given measured values can be used instead. (This method has been applied to the analysis of the
atomic bomb survivor data, to show that the slope of the dose-response relation may have been underestimated by about 15 percent if the dose errors were lognormally distributed with a coefficient of variation of 35 percent (65). These methods are considerably simpler than the full likelihood methods, which entail integration of the likelihood over the unobserved true exposure variables, but can be seen as approximations to these more sophisticated methods which will be valid if the error variances are not too large.

- Monte Carlo methods can be very useful for more complex problems where likelihood methods are intractable and these approximate methods are dubious. Essentially, one would randomly impute values for the true exposures of each subject, given their measured values and all other relevant factors, and then analyze the resulting data to obtain a point estimate of the relative risk; this process is then repeated many times to build up an entire distribution of risk estimates, which incorporates the uncertainty in the various imputations. This approach is currently being applied to the data on the Colorado plateau uranium miners.

**Modeling baseline risks.** If one adopts a parametric assumption for the baseline risk function \( \lambda_0(t) \) (for example, a simple step function dependence on age \( t \) and perhaps a small number of additional stratification variables), then estimation of the parameters of this function together with the regression coefficients involves no unusual complexities. In the semiparametric approach of Cox, however, the estimated baseline hazard rate is discrete, involving infinite "spikes" at each of the observed event times, zero elsewhere. The cumulative baseline hazard remains finite, however, and provides a natural extension of the now familiar Kaplan-Meier (66) survival curve to models involving covariates. (See Langholz and Borgan (67) for a discussion of estimation of baseline hazards in excess risk models.)

**Reproductive outcomes.** As noted earlier, the analysis of reproductive outcomes entails two types of endpoints, those manifest only at birth and those that can occur throughout the pregnancy. If one ignores their interdependency, then the former can be analyzed in a straightforward manner by comparing risks (with fetus, not fetus-time, denominators) between exposure groups or using unconditional logistic regression for multivariate analysis, assuming all fetuses have been at risk for essentially the same duration. Gestational age is a common risk factor for many malformations, but this is more appropriately handled as a covariate than as a time scale in survival analysis, since the true time at which the malformation developed is unobserved. Exposures, however, are likely to be time-dependent, and it is important to examine such exposures during the critical periods of organogenesis for each malformation type. For example, in the malathion study, limb and orofacial malformations were found to be more strongly associated with first trimester exposures, whereas gastrointestinal anomalies were more associated with second trimester exposures; the latter observation is plausible for the seven pyloric stenosis cases in that group, but not for the four tracheoesophageal fistulas.

Spontaneous abortion data require survival analysis techniques, since the set of fetuses to be used for comparison will vary over time because of the variable times of recognition of pregnancy and because of the elimination of earlier spontaneous or induced abortions. Since the malathion study was conducted within a health maintenance organization, the entry time to the cohort could be easily defined in terms of the date of the pregnancy confirmation visit. This was important, since it is possible that any causal effect of malathion exposure might be strongest for very early spontaneous abortions, which would never have been observed by this study; if gestational age at pregnancy diagnosis was also associated with malathion exposure (e.g., through socioeconomic correlates), then an analysis which included the fetus-time prior to pregnancy diagnosis would have produced biased estimates of the relative risk. Spontaneous abortions also need to be treated as event-time data with time-dependent covariates, since it would be inappropriate to compare the exposures of late abortions with those of fetuses who had aborted earlier and did not have the same opportunity for exposure. Although a crude comparison suggested that the spontaneous abortion group tended to be less exposed than the live births, this difference disappeared when the data were properly analyzed allowing for the shorter opportunity for exposure among the spontaneous exposure group.

Conceptually, the close relation between the processes leading to spontaneous abortions and congenital malformations cries out for a joint analysis of the two endpoints, but statistical methods remain undeveloped in this area. Such an analysis would be considerably strengthened if data could be obtained on the characteristics of aborted fetuses. Such data are not routinely available, but have been obtained in special studies.

**Dependent outcomes.** Dependent outcomes can arise in various ways. The endpoint may be a recurrent event (accidents or heart attacks, for example). Dependency can arise either because the occurrence of the first event alters the risk of subsequent events or because individuals differ in the underlying risks (e.g., "accident proneness"). Next, there may be several
correlated endpoints under study: for mortality data, only the first of the possible competing risks is observed, so such dependency cannot be studied; for incidence data, however, it may be desirable to consider related events (e.g., multiple congenital anomalies) jointly, although the usual practice is to restrict the analysis to the first event. Finally, there may be correlations between the outcomes of different individuals. This most commonly arises in the context of family studies, due to shared genetic or unmeasured environmental factors. Several approaches to this problem have been considered. Setting aside the more specialized genetic models, which are beyond the scope of this review, the three most commonly used alternatives are regressive models, latent variable models, and marginal models; all three are generally applicable to any form of dependency, but we limit our discussion to the case of family data.

Regressive models are based on an ordering of the subjects within a family in some natural order, such as parents \((f,m)\) before offspring \(1, \ldots, s\), older sibs before younger, and postulates a direct dependence of the outcomes of the later members on the outcomes of the earlier \((68, 69)\). For example, for binary outcomes, one might add to a logistic regression model describing the dependence of each subject’s outcome on his or her own risk factors additional covariates for the outcomes of their spouse, parents, and older sibs:

\[
\logit P(d_f = 1|Z_f) = \alpha + Z_f' \beta
\]

\[
\logit P(d_m = 1|Z_m, d_f) = \alpha + Z_m' \beta + \gamma_{sp} d_f^*
\]

\[
\logit P(d_1 = 1|Z_1, d_f, d_m) = \alpha + Z_1' \beta + d_f^* + d_m^*
\]

\[
\gamma_{2p} \frac{d_f^* + d_m^*}{2} + \gamma_{sib} d_1^*
\]

\[
\logit P(d_2 = 1|Z_2, d_f, d_m, d_1) = \alpha + Z_2' \beta + d_f^* + d_m^*
\]

\[
\gamma_{2p} \frac{d_f^* + d_m^*}{2} + \gamma_{sib} d_1^*
\]

\[
\vdots
\]

\[
\logit P(d_s = 1|Z_s, d_f, d_m, d_1, \ldots, d_{s-1}) = \alpha + Z_s' \beta + \gamma_{sp} d_f^* + d_m^*
\]

\[
+ \frac{d_1^* + \ldots + d_{s-1}^*}{s - 1} + \gamma_{sib}
\]

where \(d_i^* = d_i - \Pr(d_i = 1|Z_i, d_1, \ldots, d_{i-1})\) or zero if \(d_i\) is unknown.

The latent variables approach assumes that dependencies arise because members of a family share one or more unobservable risk factors. In the context of survival data, such a factor has come to be called "frailty." The simplest frailty model assumes that all members of the family share a common frailty which has a gamma distribution and acts as a constant relative risk. Methods of fitting frailty models have been described by Clayton \((70)\) and others; recent work is aimed at relaxing the assumption of a particular parametric distribution for the frailties \((71)\) and allowing for more complex models of sharing individual frailties within families \((72, 73)\).

The marginal models approach treats the outcomes of all the members of a family as a vector of observations with some simple covariance structure. By using generalized estimating equations methods \((74)\), estimates of the parameters of the relative risk model for the measured risk factors can be obtained which are robust to misspecification of the covariance structure \((75, 76)\). Using higher moments, it is also possible to obtain robust estimates of the parameters in the covariance structure as well, which can be of interest for testing hypotheses about residual familial risks not explained by the measured factors \((77)\).

**COHORT SAMPLING REVISITED**

For most chronic diseases, the number of events expected during the period of observation is small in relation to the size of the cohort. Thus, most of the study resources, both in terms of data collection and, perhaps, biologic sample collection as well as data analysis, would normally be devoted to subjects who will have relatively little influence on the final results. For this reason, Liddell et al. \((78)\) introduced the nested case-control design, and, subsequently, Prentice \((79)\) introduced the case-cohort design. Both designs involve comparison of the cases in the cohort with controls sampled in different ways from within the cohort, thus requiring risk factor information to be available only on the cases and the selected controls. Although the entire cohort must still be followed to identify the cases, the burden of data collection and analysis is considerably reduced.

The nested case-control design entails matched selection of controls from the "risk sets" for each case, comprising those who are at risk and disease free at the time the case occurred. The analysis of the nested case-control design uses standard conditional logistic regression methods that are identical to those used for any matched case-control study \((80)\) for the relevant statistical theory). The case-cohort design entails selection of a single unmatched control sample at random from the entire cohort at entry, and uses a form of Cox regression to compare each case.
with the subset of controls who are still at risk at the time that case occurred. The analysis of the case-cohort design is more complex, owing to the dependency between the contributions from each case-subcohort comparison. There are practical and statistical issues in choosing between the two designs (81, 82): for example, the case-cohort design is more convenient for studying multiple diseases, because the same control group can be used for each one, but for long-term cohort studies, the case-cohort design may leave few controls at risk for the later cases and subtleties arise in its application to studies with variable entry times. Generally speaking, however, the differences in statistical efficiency between the two designs are modest.

Much greater efficiency gains are possible, however, if one exploits information on exposure (or surrogates thereof) that are readily available for the entire cohort. The original two-stage designs were developed for population-based case-control studies (83, 84), but are equally applicable to case-control studies nested within cohorts, where the cohort essentially plays the role of the first-stage sample. The basic idea of the unmatched two-stage case-control design is to select different sampling fractions for the two-way classification of subjects defined by case-control status and the surrogate exposure variable, and then assess the exposure variable of primary interest only in this subgroup. The known sampling fractions are then used in the analysis to obtain unbiased estimates of the relative risk for the primary exposure variable. By appropriate selection of the sampling fractions, considerable efficiency gains (per subject included in the second stage) are possible relative to simple random sampling of cases and controls.

This basic idea would, in principle, be applicable to the case-cohort design, although the statistical theory has not yet been developed. However, a variant of this design, known as “counter matching,” has been developed for the nested case-control design (85). The basic idea is to select a matched control for each case drawn from the subset of the risk set that is discordant for the surrogate exposure, and to incorporate the corresponding sampling fractions into the usual conditional likelihood for matched case-control designs. For example, supposing the surrogate exposure variable were dichotomous, then each exposed case would be matched with an unexposed control from the case’s risk set, and vice versa. This approach ensures a high degree of variability in the primary exposure variable within matched sets, thereby producing great efficiency for the main effect of exposure and its interactions with other variables (but generally at the cost of some loss of efficiency for estimating the effect of confounders) (86, 87).

In the uranium miner cohort, a 1:1 nested case-control design counter-matched on radon produced an estimate of the standard error of the radon effect that was only 27 percent larger than from the analysis of the full cohort, compared with 82 percent larger for the standard 1:1 matched case-control study. A 1:3 counter-matched study was nearly fully efficient (18 percent larger standard error than from the cohort) compared with 45 percent larger for the standard 1:3 matched study. In contrast, the standard errors for the estimate of the smoking effect were very similar for both designs (88). In an application of this method to a cohort study of gold miners, Steenland and Deddens (89) found that a 1:3 counter-matched case-control study provided efficiency approximately equivalent to a standard nested case-control study with 10 controls per case.

These analyses provide an idea of the potential efficiency gains that are possible by efficient selection of controls in nested case-control studies, although in the miner study the basic data was already collected. In other studies, such as the on-going International Nuclear Worker Study (90), in which a summary dose estimate is available for all nuclear workers, but extensive efforts are underway to characterize the exposure measurement errors over time, the cost savings from having to do this only for a small sample of highly informative cohort members could be very substantial indeed.

CONCLUSIONS

Although the basic statistical methods for cohort studies have been well established for many years, new methods are continuing to be developed. The proportional hazards model has proven to be a very useful framework for unifying approaches to the analysis of time-to-event data for individuals as well as for grouped count and person-year data. However, particular applications may call for imagination in the development of relative risk models that are biologically appropriate and fit the available data. Although most routine analyses proceed with empirical model building techniques using standard relative risk regression models, models more strongly grounded in biologic theory would be appropriate where there are strong effects, the data are of high quality, and there is a solid biologic theory. Beyond the proportional hazards model, future work may well be usefully directed toward such alternatives as the excess risk and accelerated failure time models. Methods for dealing with exposure measurement error and dependent outcomes have become a very active area of statistical research recently, but applications are in their infancy. But
perhaps the most important methodological development in recent years has been in the area of cohort sampling methods, where considerable cost savings are possible by efficient study design. A close interaction between epidemiologists and statisticians will be needed to fully realize the potential of these new methodological developments.

ACKNOWLEDGMENTS

This work was supported by grants CA42949 and CA52862 from the National Institutes of Health.

REFERENCES


89. Steenland K, Deddens JA. Increased precision using counter-matching in nested case-control studies. Epidemiology 1997;8:238–42.