Methodologic Research Needs in Environmental Epidemiology: Data Analysis

Ross L. Prentice* and Duncan Thomas

*Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104; and Department of Preventive Medicine, School of Medicine, University of Southern California, 1420 San Pablo, PMB B-201, Los Angeles, CA 90033

A brief review is given of data analysis methods for the identification and quantification of associations between environmental exposures and health events of interest. Data analysis methods are outlined for each of the study designs mentioned, with an emphasis on topics in need of further research. Particularly noted are the need for improved methods for accommodating exposure assessment measurement errors in analytic epidemiologic studies and for improved methods for the conduct and analysis of aggregate data (ecologic) studies. — Environ Health Perspect 101 (Suppl 4):39-48 (1993).

Key Words: Aggregate data studies, analytic data studies, carcinogenic models, exposure measurement error, meta-analysis, relative risk regression, validation sub-studies

Introduction

Nearly all study of the health consequences of environmental and lifestyle exposures in human populations is purely observational. This means that the validity of the comparison of disease rates between more exposed and less or nonexposed persons is dependent on the assumption that disease rates in the two groups are comparable in the absence of such exposure. This comparability assumption can be weakened somewhat by the measurement and accommodation of other factors that are associated with disease risk and that have a different distribution in the compared exposure groups. If such confounding factors are accurately measured and adequately acknowledged in the data analysis, it is then sufficient that in the absence of the exposure of interest, the groups being compared have common disease rates conditional on the values of the confounding factors. Lack of validity (i.e., bias) in testing or estimation can be expected if there are unidentified confounding factors, if the recorded confounding factors are measured with error, or if the treatment of individual confounding factors is inadequate (e.g., linear allowance for confounders having effects that are substantially nonlinear). Bias also can be introduced if the exposure variables of interest or the health effects under study are not measured accurately. In practice these sources of bias can be reduced, but it is unlikely that they will be completely eliminated.

The sources of bias mentioned here are the principal reasons why epidemiologic cohort studies, among others, may yield inaccurate and conflicting results. Concern about residual, uncontrolled confounding can never be completely eliminated in any non-experimental study. Hence, such studies are most reliable for the detection of moderate to large health effects (e.g., increased in disease incidence by a factor of two or more among highly exposed persons) that are unlikely to be qualitatively affected by modest confounding. There is also a strong role for the replication of results in diverse populations that are presumed to have different potentials for severe confounding. It is worth noting that experimental studies also have important practical limitations in the context of environmental epidemiology. Data analysis methods for cohort studies with accurate and complete assessment of exposure variables, confounding factors, and potential health consequences are well developed, as summarized in "Exposure-Response Estimate in Cohort Studies," below.

Case-control studies in which exposure and confounding factors are assessed retrospectively are subject to all of the biases noted above, as well as to recall bias, which occurs when diseased individuals (cases) and disease-free individuals (controls) differentially recall their exposures, their confounders, or their health outcomes. Aggregate data studies, referred to later in this paper as ecologic studies, attempt to relate the exposure and confounding factor experience of groups to their corresponding disease rates. Such studies may be subject to additional biases if the statistical model for the group disease rate does not equal the average of valid disease rate models for the individuals being aggregated.

Apparent disagreement between environmental epidemiologic studies can also arise, not from bias, but from lack of power combined with attention to point estimates rather than confidence limits. The ability to detect an association between the levels of an exposure or exposure history and the risk of a disease depends primarily on the observed number of disease events in the study sample, on the range of exposures in the sample, and on the strength of association between exposure and disease. The distribution of exposures in the study cohort or in the cohort from which cases and controls are selected for a case-control study also has important influences on study power. While random measurement error in (univariate) exposure assessment will not invalidate, under weak conditions, a test for the hypothesis that no association between exposure and disease exists, test power may be reduced considerably by such measurement errors. Also, estimates of dose-response parameters may be substantially distorted (usually biased downward), including the possibility of a loss of monotonicity of dose-response trends (1). Thus, the proper analysis and interpretation of environmental epidemiologic studies rely heavily on the investigator's assessment of the magnitude of both potential biases and study power in the absence of such biases. For practical reasons, the power of specific studies will often be rather low, and knowledge of disease mechanisms and measurement
properties will be too limited to place useful bounds on potential biases. Hence, there are important uses for formal tests of the equality of exposure–disease associations from two or more studies in differing populations and for techniques used in combining the results of several studies. This topic will be discussed in the section titled "Comparing and Combining the Results of Several Studies."

The following section describes statistical- and biological-based models that can serve as the basis for exposure–disease analyses.

Models for Disease Occurrence

The simplest cohort studies occur when exposure takes place in one instant, as in Japanese atomic bomb survivors, or is constant over the individual's lifetime, as in some animal inhalation experiments. However, most exposures, and most confounders, are complex functions of time and demand a more complicated mathematical description. Our discussion of descriptive disease occurrence models begins with the over-simplified case.

Let \( \lambda_0(t) \) denote the instantaneous rate of occurrence of a study disease or other health-related event for subjects of age \( t \) who have not received the exposure of interest. This means that if \( N \) such persons, all at age \( t \), were observed for a short time \( dt \), the expected number of disease occurrences would be \( N \lambda_0(t) dt \). If a person of age \( t \) received an exposure \( Z \), the instantaneous occurrence rate would be altered from \( \lambda_0(t) \) to \( \lambda(t | Z) \), the (instantaneous) relative risk is \( \lambda(t | Z) / \lambda_0(t) \). These rates are nonnegative and, provided neither is zero, one can take the logarithm of this relative risk. It is often convenient and useful to assume the logarithm of the relative risk to be a linear function of exposure and confounding factor measurements. This is equivalent to modeling the relative risk as an exponential function, \( \exp(\chi \beta) \), where the vector \( \chi = (\chi_1, \ldots, \chi_p) \), which replaces the more general \( Z \), consists of carefully chosen (and usually incomplete) measures of exposure or confounding factors, with \( \chi = (0, \ldots, 0) \) corresponding to no exposure and standard values for confounders. The coefficients \( (\beta_1, \ldots, \beta_p) \), regression coefficients that comprise the vector \( \beta \) (or, more precisely, its transpose \( \beta^T \)), then tell us about the impact of each \( \chi_i \) on relative risk when the other \( \chi_s \) are fixed.

The result is a simple proportional hazards (or Cox) model

\[
\lambda(t | Z) = \lambda_0(t) \exp(\chi \beta) \tag{1}
\]

which is used widely in the analysis of failure time data (2). In order to deal with complications inherent in most environmental epidemiologic studies, one must generalize this discussion and complicate the appearance of some formulae, but be careful not to change the essentials of the approach. Such generalization follows in the next subsection.

Descriptive Relative-Risk Models

As above, let \( \lambda_0(t) \) denote the instantaneous rate of a study disease (or other health-related problem) at age \( t \) in the absence of the exposure of interest. A person of age \( t \) may have received exposures \( z(u) \) at certain ages \( u < t \). One can refer to \( Z(t) = \{ z(u), \, u < t \} \) as the person's exposure history up to age \( t \). Furthermore, one can allow the vector \( z(u) \) to include the values of confounding factors at age \( u \), so that \( Z(t) \) includes both exposure and confounding factor histories up to age \( t \). The disease rate at age \( t \) is \( \lambda(t | Z(t)) \), a function of this exposure and confounder history. The relative risk associated with history \( Z(t) \) is then the ratio \( \lambda(t | Z(t)) / \lambda_0(t) \). Because this ratio is nonnegative, it can be, and often is, modeled using an exponential function \( \exp(\chi(t) \beta) \), where \( \chi(t) = (\chi_1(t), \ldots, \chi_p(t)) \). This function consists of data-analytically defined functions of \( Z(t) \) and \( t \), with \( \chi(t) = (0, \ldots, 0) \) again corresponding to no exposure and standard confounder histories, while \( \beta^T = (\beta_1, \ldots, \beta_p) \) is a corresponding vector of relative-risk parameters to be estimated.

This relative risk (RR) regression model

\[
\lambda(t | Z(t)) = \lambda_0(t) \exp(\chi(t) \beta) \tag{2}
\]

also called the Cox-regression model or (inaccurately called) the proportional hazards model (2-4) or an approximation to these models, forms the basis for most descriptive analyses of environmental epidemiologic studies. As a simple example to illustrate the notation, consider the relationship between exposure to ionizing radiation and the rate of a certain cancer in the atomic bomb-exposed populations in Hiroshima and Nagasaki. One could define

\[
z(u_0)^T = \{ z_1(u_0), z_2(u_0) \}, \tag{3}
\]

as the gamma and neutron exposures for a person at age \( u_0 \) in 1945 when the exposure occurred and as \( z(u_0) = 0 \) otherwise. A specification \( \chi(t) \equiv z(u_0) \) then assumes a log-relative risk function that is linear in gamma and neutron exposure levels. The regression model can be relaxed to allow, for example, the relative risk to depend on age at exposure and time since exposure and to allow for non-linear dependencies of the log-relative risk on gamma and neutron exposure.

As noted above, the histories of potential confounding factors can also be included in \( Z(t) \), in which case \( \chi(t) \) will include functions of both the exposures of interest and other factors, while product terms between the two will allow the relative risk associated with a given exposure history to depend on the value of other variables. This allowance is termed effect modification in epidemiologic parlance. Confounding factors may also be controlled by means of stratification rather than, or in addition to, regression modeling using the descriptive model

\[
\lambda(t | Z(t)) = \lambda_0(t) \exp(\chi(t) \beta) \tag{4}
\]

where the baseline rate \( \lambda_0(t) \) is allowed to vary across a number of strata defined as functions of \( t \) and confounding factor values.

Relative-risk forms other than exponential also may be considered in the above models. In particular, the linear form \( 1 + \chi(t) \beta \) often is felt to be theoretically and empirically more appropriate for certain carcinogenic exposures and has been used widely in radiation literature, sometimes with the addition of quadratic terms. Absolute rather than relative-risk models, such as

\[
\lambda(t | Z(t)) = \lambda_0(t) + \chi(t) \beta \tag{5}
\]

also have been used in modeling radiation effects, although there is a consensus that it generally does not fit well without the addition of terms for the modifying effect of age at exposure and latency. It may also be useful for modeling certain rare diseases such as mesothelioma, for which the baseline rate in the absence of asbestos exposure is virtually zero. In all of these alternatives to the standard exponential relative-risk model, estimates of the relative-risk parameters and baseline rates are often found to have poor statistical properties. However, quite general programs that use likelihood-based methods to obtain appropriate confidence limits (5) are now available to fit a broad class of relative- and absolute-risk models with combinations of linear and exponential terms.

Suppose that the regression vector \( \chi(t) \) in the above unstratified model consists only of functions of the exposure variable under study, and let \( p^T \{ \chi(t) \} \) denote the probability density for value \( \chi(t) \) in the source population of the modeled regression vector. In addition to estimating the relative-risk function, one may be interested in the fraction of the disease incidence at age \( t \) that may be attributed to
A similar expression can be written for the attributable proportion under the stratified relative-risk model.

In some applications of these relative-risk models, it is convenient to define the basic time variable \( t \) to be chronological time or time from entry into a certain cohort rather than age, which is accommodated through stratification or regression modeling. For example, in a cohort study with covariate information collected at specified points in chronological time, such a definition can help ensure comparability of the covariate (i.e., exposure and confounding) information on all study subjects at a given value of \( t \).

There are distinct advantages in using hazard rates or instantaneous disease rates, \( \lambda(t) \), in our formulae rather than disease rates over some specified age or time period, in part because the interpretation of these latter rates will depend on the duration of the age period or time period in question, which will vary inevitably from study to study. Nevertheless, in some studies one observes only whether disease occurs in a certain time period rather than the actual times or ages of disease occurrence. Let \( D = 1 \) denote disease occurrence during a prescribed disease ascertainment period for a study and \( D = 0 \) denote lack of occurrence. Ignoring issues such as competing risks and losses to follow-up, one may choose to model the disease probabilities \( pr\{D = 1 \mid Z(t_0)\} \) by an exponential-form odds-ratio model in which

\[
\frac{pr\{D = 1 \mid Z(t_0)\}}{pr\{D = 1 \mid Z(t_0) = Z_0\}} \times \frac{pr\{D = 0 \mid Z(t_0)\}}{pr\{D = 0 \mid Z(t_0) = Z_0\}} = \exp\{\chi(t_0)\beta\},
\]

[7]

where \( Z(t_0) \) denotes a subject’s exposure and confounding factor history at age \( t_0 \) at the beginning of the ascertainment period, and \( Z_0 \) denotes the standard, or base, covariate history. This odds-ratio model can be rewritten as a logistic regression model

\[
pr\{D = 1 \mid Z(t_0)\} = \frac{1 + \exp\{a(Z_0) + \chi(t_0)\beta\}}{1 + \exp\{a(Z_0)\}}.
\]

[8]

where the function \( a(Z_0) \) may, for example, be defined to take value \( a \) whenever the study subject falls in stratum \( t \), which is defined as a function of potential confounding factor values at \( t_0 \).

The above relative-risk and odds-ratio models are purely descriptive models. Their application is intended as an aid for summarizing and displaying aspects of large, complex data sets. In some situations, such as a regulatory decision concerning the safe level of a certain exposure, it will be essential to bring to bear any available biologic or mechanistic knowledge on the inference problem. Such knowledge could be used, for example, to specify a form for the relative risk at age \( t \) as certain elements of \( X(t) \) approach zero, where these elements capture the dosage, duration, or other aspects of the exposure in question. Similarly, knowledge or assumption about the pertinent biological mechanisms could be used to derive models for \( \lambda(t) \) of forms other than those mentioned above. The next subsection overviews two classes of carcinogenesis models that have been proposed on mechanistic or biological grounds.

### Mechanistic and Biologically Based Models

Efforts to describe a disease process in terms of deterministic or stochastic models have focused mostly on models for the spread of infectious diseases in a population and models for carcinogenesis. Some of the work on carcinogenesis models, as outlined below, may be pertinent to other diseases.

Much of the early work on mathematical models for cancer was reviewed in a classic paper by Armitage and Doll (6). Whitemore and Keller (7) also provide a comprehensive review. A major contribution of the Armitage and Doll paper is the use of the multistage model of carcinogenesis. This model is based on the assumptions that cancer results from a single cell line undergoing a series of discrete, heritable changes (e.g., point mutations, chromosomal breaks or translocations, or other types of copying errors) in a particular sequence, and the rates of such transitions do not depend explicitly on age, although they may be affected by exposure to carcinogens or by factors that modify the rate of cell division.

As a consequence of these and some additional assumptions, it can be shown that the age-specific incidence rate is predicted to vary approximately as the \((k-1)^{th}\) power of age, where \( k \) is the number of transitions required (usually estimated to be about 5 to 7 for adult tumors). If a carcinogenic exposure occurs at a constant rate over time, the incidence will vary approximately as a polynomial function of dose rate of order equal to the number of dose-dependent transitions. If exposure is instantaneous or varies over time, the incidence rate will be modified by age at exposure and/or time since exposure, depending upon which stage(s) is dose-dependent.

Until recently, most of the empirical tests of these predictions have been done by fitting the model to aggregate data on population age-specific rates, or to broadly-grouped data on cohorts, stratified by dose, age at exposure, or time since exposure. A problem with this approach is the difficulty of separating the effects of dose rate, age at first exposure, duration of exposure, time since last exposure, and attained age, all of which influence the predictions of the model. Simple comparisons of one factor without controlling the other factors can be misleading. This is less of a problem when animal bioassay data are used, as these are usually limited to constant, lifetime, dose regimens. However, such data are not informative about whether the carcinogen acts at an early or late stage. Nevertheless, the approach has been used for risk-assessment purposes by many regulatory agencies. The default approach advocated by the U.S. Environmental Protection Agency (EPA) and others involves fitting the multistage model to available epidemiologic or toxicologic data and using an upper confidence limit on the estimated slope coefficient (scaled for species differences in weight and life span) to compute the lifetime excess risk in humans. The scientific and statistical validity of this approach is controversial (8,9).

With the development of general relative-risk models ("Descriptive Relative-Risk Models" above), it has become possible to test the multistage and other models by fitting them directly to data on individuals. This offers great advantages for dealing with time-dependent exposures, which are the most informative about the stage at which a carcinogen acts. This approach has been applied to data on occupational exposures to asbestos (10), arsenic (11), and benzene (12); on the atomic bomb survivors (13); and on smoking (14), with varying results. The three occupational applications all were consistent with a single stage of action (relatively late for asbestos and arsenic, early for benzene), while the radiation and smoking data both showed signs of two stages being affected.

The multistage model has several important limitations, including its inability to account...
for leukemia and childhood cancers, the genetics of cancer, and the distinction between mechanisms of initiation and promotion. It also has been criticized for its need for as many as 5 to 7 stages to account for the steep age dependence, when only two or three have been established in experimental systems. Moolgavkar and Knudson (15) have proposed an alternative model that addresses these issues. This model assumes that two mutational events are required and the cell lines that have experienced the first event may be at a competitive advantage (proliferation) or disadvantage (repair) relative to normal cells. Carcinogens might act by affecting either mutation rates or proliferation rates. Major gene effects are accounted for by assuming that individuals who inherit the gene begin life with all cells in the intermediate stage. This model has been successful in fitting epidemiologic data on smoking (15,16), breast cancer (17), and radon (18). In the latter example, data from an experimental study of rats exposed to radon were fitted to the model and radon was found to have an effect on both the mutation and proliferation rates. However, the interpretation of this result is complicated by the authors’ use, for both of these dependencies, of a power function dose–response relationship with a very low exponent rather than a simple linear dose–response. Thomas (19) has proposed a variant of this model that adds an additional stage to the process to try to explain the difference in the modifying effect of dose rate and the duration of exposure for different types of radiation; so far, no attempt has been made to test this model.

With the rapid growth in our understanding of the fundamental biology of cancer, further development of methods to validate these mechanistic ideas and, where appropriate, to incorporate them into the analysis of epidemiologic data would be worthwhile. Most of the models that have been considered seriously are sufficiently general that some parameter values can be found to provide an adequate fit to epidemiologic data sets. Thus, these models are not easily falsified as a class, and it is unlikely that one could choose among them on purely statistical grounds. Instead, their utility lies in the types of comparisons that can be made within the context of a particular model—whether a carcinogen acts at an early or a late stage in the multistage model or as an initiator or a promoter in the two-stage model, for example. Their real value, therefore, lies in their ability to organize a complex set of hypotheses into a unified framework and to suggest empirical tests, in populations of humans or animals, of mechanistic ideas suggested by observations at the cellular level. Research efforts to identify and measure the assumed biological entities on the pathway to cancer cell formation seem particularly well motivated.

**Exposure–Response Estimation in Cohort Studies**

**Relative-Risk and Odds-Ratio Estimation**

Consider the unstratified relative-risk regression model of “Relative-Risk Models.” A cohort study involves the selection of a sample of individuals from the population under study, succeeded by a follow-up to observe disease occurrence. The relative-risk parameter \( \beta \) can be estimated by maximizing a partial likelihood function \( L(\beta) \) that is a product over all disease occurrence times (ages) that appear in the sample of the ratio of the relative risk for the subject developing disease to the sum of the relative risks for all subjects at risk at that time \( t \). The corresponding likelihood function under the above stratified relative-risk model is simply the product over strata of the stratum-specific likelihood functions. Note that this estimation procedure is quite general in that exposure variables, confounding variables, and stratum assignments each can vary with follow-up time. The principal assumption underlying this estimation procedure requires the set of subjects at risk for disease at any follow-up time to be representative of the base population, conditional on the covariate history and stratum assignment. This assumption will be satisfied, for example, if study subjects are sampled randomly and independently from the study population, and if rates of censoring (e.g., losses to follow-up) at a given follow-up time depend most on the covariate histories and stratum assignments at that time. Also, under weak conditions, \( L(\beta) \) can be manipulated as if it were an ordinary likelihood function for asymptotic inference on \( \beta \).

The score statistic \( U(\beta_0) \), defined as the value at \( \beta = \beta_0 \) of the derivative with respect to \( \beta \) of \( \log L(\beta_0) \), can be used to test \( \beta = 0 \). If \( \beta_0 = 0 \) and \( X(t) \) consists only of indicator variables to distinguish exposure groups, then \( U(\beta_0) = U(0) \) is known as the log rank statistic. Other choices of \( X(t) \) yield other familiar, censored data test statistics, including generalizations of the Wilcoxon statistic.

Suppose now that there is no possibility of early censorship in the cohort study throughout the follow-up. The odds-ratio parameter \( \beta \) in the logistic regression model of “Relative-Risk Models,” along with the location parameters \( \alpha_i = \alpha_i(Z_{i0}) \), can be estimated by a likelihood function \( L(\beta) \) that is simply the product over all study subjects of the logistic regression probabilities \( pr(D = 1|Z(t)) \) for subjects developing disease, and one minus such probabilities for other study subjects. Computer programs are widely available for inference on \( \beta \) from this likelihood function. If there are few disease events in stratum \( s \), it is preferable to eliminate \( \alpha_i \) by conditioning on the number of such events prior to applying standard likelihood procedures for the estimation of \( \beta \).

The likelihood functions just described may seem esoteric to readers not having a statistical background. The main point to note, however, is that estimation of relative-risk and odds-ratio parameters in the very flexible models of “Relative-Risk Models” is now routine, and suitable software is available. Of course, the odds-ratio parameter will approach the relative-risk parameter if the disease acquisition period \( ds \) becomes short. This occurs because the odds of disease,

\[
pr\{D = 1|Z(t)\} / \left[ 1 - pr\{D = 1|Z(t)\} \right]
\]

then typically approaches \( \lambda(Z(t))dt \), from which the exponential-form odds ratio approaches a corresponding exponential-form relative risk with identical regression parameter.

Estimation of the relative-risk regression parameter \( \beta \) may be computationally demanding if there are many distinct disease incidence times and if the regression vector and stratum assignment depend on time. However, if each \( \lambda(Z)(t) \) is defined to be constant over a partition of the time axis and \( X(t) \) is restricted to be constant within the elements of this partition, then \( \beta \) can be estimated in a computationally simple fashion using Poisson regression methods. See Preston et al. (24) for application of such methods to radiation dose–response estimation from the Hiroshima and Nagasaki cohorts.

Particular care is required if these estimation procedures are applied to cohorts having few cases or if most cases occur within a small portion of the overall range of exposures. Asymptotic formulae for interval estimation on \( \beta \) may then be inaccurate and more specialized procedures (e.g., resampling methods) may be required. In fact, there has been little study of cohort data configurations under which such asymptotic formulae will provide adequate approximations.

Kalbfleisch and Prentice (3) and Cox and Oakes (4) provide detailed accounts of the theory and application of relative-risk regression models.
Disease Rate Estimation and Graphical Models

Denote by

\[ \Lambda_0(t) = \int_{t_0}^{t} \lambda_0(u) \, du, \quad [10] \]

the cumulative baseline disease rate in stratum \( s \) in the stratified model of “Relative-Risk Models” over the range of ages \( t \geq t_0 \), represented in the cohort. A simple nonparametric estimator of \( \Lambda_0(t) \) can then be defined as the sum over all disease occurrence times in stratum \( s \) of the ratio of the number of stratum \( s \) failures to the sum of the relative risks for all subjects at risk in stratum \( s \) at that time, with all relative risks evaluated at that \( \beta \) which maximizes \( L(\beta) \).

As with ordinary regression methods, model-checking procedures are important to the application of relative-risk and odds-ratio models. Such procedures naturally focus on the assumed relative-risk process, \( \exp \{ \lambda(t) \beta \} \), because other aspects of the model essentially are nonparametric. For example, the postulated relative-risk function can be generalized by adding well-selected additional elements to \( \lambda(t) \) and testing the hypothesis that corresponding coefficients equal zero. Computationally feasible methods also have been developed for approximating the influence of each study subject or each age group on \( \beta \)-estimation, in order to highlight questionable data points and to highlight vulnerabilities of the inference to model assumptions (25). Graphical procedures particularly are useful. In addition to the usual types of plots of influence (i.e., sensitivity) values and residuals, plots of separate estimates of \( \Lambda_0(t) \) for subsets of the cohort can provide useful visual checks on proportionality and other relative-risk assumptions (3).

The fact that the baseline rates \( \lambda_0(t; \beta) \) are unrestricted is an important source of robustness in respect to \( \beta \)-estimation. Specifically, relative-risk estimation is unlikely to be affected much if the intensity of ascertainment of disease events in the cohort varies somewhat across time or among strata. Similarly, location shifts in the modeled regression vector \( \lambda(t) \) across different values of \( t \) would not affect \( \beta \)-estimation in the exponential-form relative-risk model.

Moreover, more general measurement error in the ascertainment of \( \lambda(t) \) may have a profound effect on relative-risk estimation.

Measurement Error in Exposure Variables and Confounding Factors

Epidemiologists have long recognized that errors in the measurement of the study variables, including misclassification in the case of categorical variables, can lead to biased tests and estimates of the associations under study. Measurement error in the exposure histories or confounding factor histories may be of particular importance in environmental epidemiologic applications. Unfortunately, the methodology for avoiding bias due to measurement error is still at a rudimentary stage of development.

Consider the unstratified relative-risk regression model of “Relative-Risk Models” and suppose that rather than the covariate history \( Z(t) \) one observes an estimate \( W(t) \).

The disease rate function at age (or chronologic time) \( t \), given the observed covariate history \( W(t) \) can then be written (26)

\[ \lambda \{ t; W(t) \} = \lambda_0(t) E \left[ \exp \{ \lambda(t) \beta \} \mid W(t) \right], \quad [11] \]

where the expectation also is conditional on lack of disease occurrence or censorship prior to \( t \). In fact, this induced relative-risk model also requires

\[ \lambda \{ t; Z(t), W(t) \} = \lambda \{ t; Z(t) \} \quad [12] \]

so that the \( W(t) \) is unrelated to disease risk, given the true covariate history \( Z(t) \).

Unfortunately, the expectation in \( \lambda \{ t; W(t) \} \) generally depends on the baseline rates \( \lambda_0(u) \), \( u \leq t \), which complicates the estimation. However, in cohort studies in which the cumulative probability of disease occurrence is small, this dependence usually can be ignored and estimation of \( \beta \) can be based on a likelihood function in the form described above upon specifying a measurement error distribution for \( Z(t) \) given \( W(t) \), from which \( \lambda \{ t; W(t) \} \) can be calculated.

Specification of the distribution of \( Z(t) \) given \( W(t) \) would seem to be a hazardous undertaking unless there is a subsample in which both \( Z(t) \) and \( W(t) \) are available. In the presence of such a validation sample, simultaneous inference on relative-risk parameters and measurement error distribution parameters is possible (27), though further development is necessary before such estimation can be viewed as routine. More difficult issues arise if a true validation sample is not available. A reliability sample, in which separate estimates \( W_1(t) \) and \( W_2(t) \) of \( Z(t) \) are obtained on a cohort subsample at two (or more) points in time, permits insight into some aspects of measurement error distribution, but additional strong assumptions are required for the estimation of \( \beta \).

Even if the exposures under study are precisely estimated and pertinent confounding factors are identified, severe confounding may occur if confounding factor histories are measured with error (28), as is obvious if one considers an extreme situation in which measurement error produces a totally useless confounding factor estimate. This bias is likely to be more acute if the exposure and confounding factor values appearing in \( X(t) \) are highly correlated.

A hypothetical cohort study of prenatal exposure to passive smoking in relation to the risk of lower respiratory disease during the first 3 years of life provided illustration in Morgenstern and Thomas, this volume. Any elevation in the odds of lower respiratory disease among more heavily exposed neonates may be severely attenuated by inaccuracies in exposure assessment in such a study. An analysis that controls for passive smoke exposure during the first 3 years of life, an exposure that would often be highly correlated with prenatal exposure, may be dominated by measurement error and be totally unreliable. A more practical illustration of the impact of measurement error is seen in the analysis of the mortality rates of various cancers in relation to gamma and neutron exposures in the Hiroshima and Nagasaki cohorts. Individual exposure estimates were constructed based on each study subject’s location and shielding information as early as 1960. These estimates have continued to be refined in succeeding years through the use of improved models for the yields of the two bombs and more sophisticated models for the formation, transmission, and attenuation of gamma and neutron radiation. Many of the analyses of these cohorts simply combine gamma and neutron exposures into a single total dose estimate. The corresponding cancer mortality analyses have been affected somewhat by the changes in total dose estimates from one dosimetry system to the next (e.g., in the magnitude of elevated relative risks and the apparent shape of the dose-response curves), whereas analyses that attempt to estimate simultaneously the effect of gamma and neutron exposures on relative risk have been completely changed by dose estimate modifications. This illustrates the difficulty of reliably estimating exposure–disease associations when there are two or more exposure variables that are each measured with error (random or systematic) or, analogously, when there are exposure and confounding variables each measured with error. Very similar issues arise in epidemiologic studies of nonenvironmental factors; for example, they arise in attempts to separate the effects of fat and calories on cancer risk in nutritional epidemiology, or to separate the effects of...
types of fat by degree of saturation on cancer risk in nutritional epidemiology (29).

Some recent work has concentrated on developing methods to adjust associations for the effects of measurement errors when their distributions are known. A very general framework for attacking this problem has been outlined by Clayton (30), who specifies the problem in terms of component models: the disease model describes the dependence of disease risk on true exposures and other factors; the measurement error model describes the relationship between true and measured exposures and any modifying factors; and the exposure model describes the population distribution of true exposures. These three models are combined in a maximum likelihood framework, and approaches to estimating the parameters of the disease model are described. Unfortunately, the approach is mathematically intractable in its general form, but useful progress has been made in some special cases. For categorical variables, Greenland and Kleinbaum (31) described a method based on applying the inverse of a matrix of known misclassification rates to the subject counts by measured exposure and disease classifications. Hui and Walter (32) have considered the case in which replicate measurements of exposure are available, and they use a form of log-linear model for the resulting four-way contingency table (counting true exposure as an unobserved dimension). For continuous variables, Prentice (26), Pierce et al. (33), Whittemore (34), Sposto et al. (35), and others have discussed approaches that replace the measured doses with empirical Bayes estimates of the true dose and use these in standard analyses. For a general review of these approaches, see Armstrong (36) and Thomas et al. (37). Another recent development involves combining nonparametric density estimation techniques with a computational device known as Gibbs sampling to overcome the tractability problems in the Clayton approach and avoid the need for parametric assumptions about the distribution of true doses. This method has been applied to data on studies of leukemia and thyroid disease in Utah residents downwind of the Nevada Test Site (38,39). These approaches are in an early stage of development, but they offer the prospect of removing the bias due to misclassification, correcting the shapes of dose–response curves, adjusting for covariates, and examining interaction effects, all while allowing for the additional uncertainties due to uncertainties in exposure estimates. Further developments along these lines are highly desirable.

Most of the literature on correcting for measurement errors has assumed that the misclassification rates were known and were constant across subjects. In practice, only estimates of these error distributions are available, either from earlier validation studies, from replicate measurements, from gold standard measurements on a subset of the subjects, or from theoretical uncertainty analysis. Methods need to be developed to account for uncertainties in the estimates of these misclassification rates (40). As a design issue, the optimal allocation of resources between high-quality measurements on a subset and larger numbers of approximate measurements should be considered (41,42). A unique aspect of the Utah fallout studies is the availability of individual-specific uncertainty estimates based on elaborate sensitivity analyses of the exposure pathways. This has allowed subjects with more precise exposure estimates to be given heavier weight in the analysis. Whether such efforts are warranted in terms of improved precision needs to be considered.

In summary, covariate measurement errors can bias severely the results of environmental epidemiologic studies. Improved analytic methods for accommodating random, nondifferential covariate measurement errors are required. Such methodologic developments might naturally focus on the potential for obtaining a true validation sample, on validation study design, and on the incorporation of validation study data in the overall estimation procedure (27).

**Exposure–Response Estimation Under Case–Control and Other Sampling Procedures**

Relative-risk and odds-ratio estimation often can be carried out more economically by sampling only subjects developing the study disease (the cases) or a random sample thereof, along with a suitably matched sample of subjects without disease (the controls). Typically covariate histories \( Z(t) \), where \( t \) is the age (time) of case or control ascertainment, then have to be obtained retrospectively.

Consider the stratified relative-risk model of "Relative-Risk Models" and suppose that each case has one or more randomly selected controls that are matched on age at ascertainment (\( t \)) and stratum (\( s \)). Given the covariate histories \( \{Z_1(t), \ldots, Z_m(t)\} \) for a case and its \( (m-1) \) age- and stratum-matched controls, the probability that exposure history \( Z_1(t) \) corresponds to the case is simply the relative risk at \( t \) for the case divided by the sum of such relative risks for the \( m \)-matched subjects (including the case). Hence, the relative-risk parameter \( \beta \) can be estimated by maximizing the likelihood function \( L(\beta) \), which is formed by multiplying these ratios for all matched case–control sets (43). To avoid strict matching on \((t,s)\), relaxations of this sampling scheme are possible.

Similarly, the odds-ratio parameter \( \beta \) in the logistic regression model of "Relative-Risk Models" can be estimated under case–control sampling by maximizing the resulting logistic regression likelihood function by acting as though a prospective study had been conducted, though the estimates of \( \alpha \) no longer reflect disease incidence probabilities (23). In fact, the baseline rates \( \lambda_0(t) \) and \( \alpha(t) \) in the relative-risk and odds-ratio regression models of "Relative-Risk Models" cannot be identified from case–control data in the absence of additional information on case and control sampling fractions.

In general, relative-risk and odds-ratio parameter estimates from case–control studies will be subject to the same biases as cohort studies. They also may be subject to recall bias if exposures or other covariate histories are differentially recalled by cases and controls or if they involve measurements that are affected by disease occurrence or its sequelae. There are often various practical steps that can be taken to minimize bias in ascertaining the covariate histories \( Z(t) \) (e.g., interviewers blinded to case or control status), but usually it is not possible to identify residual recall bias because the requirement to obtain prediagnosis and postdiagnosis covariate histories on a sufficient sample of cases would often eliminate much of the efficiency of the case–control design.

As with the cohort study design, nondifferential measurement errors lead to the expectation

\[
E\left[\exp\{X(t)\} \mid \mathcal{W}(t)\right],
\]

where \( X(t) \) is the true \( \mathcal{W}(t) \) is the measured regression vector at age \( t \), as the identifiable relative-risk function under age- and stratum-matched case–control sampling. To the extent that a representative validation sample can be ascertained retrospectively, there will be a potential to conduct valid relative-risk estimation from this type of study without making further assumptions.

A case–cohort (case–base) sampling procedure can also be considered as a means of reducing the cost or simplifying the logistics of a cohort study. With this design, covariate histories \( Z(t) \) are assembled only
for cases and a (stratified) random sample of the study cohort. This sampling procedure has advantages if several ends points (diseases) are to be studied in relation to an exposure. Also, the subcohort may be used to monitor exposures and other variables during the study’s follow-up. However, estimation may be less efficient than estimation based on a case–control study with a comparable number of study subjects if cases and subcohort members are not well matched (44,45), and recall bias typically will be an issue. Prentice (46) has developed a procedure for estimating the relative-risk and odds-ratio parameters from case–cohort samples, and, in contrast to case–control sampling, baseline rates also can be estimated without external information. Comparisons and refinements of these sampling procedures are worthwhile research activities. Note also that the use of so-called two-stage designs (47,48) can lead to further valuable efficiency gains in some case-control study applications.

**Exposure–Response Estimation in Aggregate Data (Ecologic) Studies**

As discussed previously, sometimes it will be economical and convenient to examine an exposure–disease association by relating the disease rates among several groups of individuals to aspects of the exposure experience of each group. Such studies can be referred to as aggregate data studies since they involve the disease rates and exposures for the aggregate, rather than for individuals. These studies also are commonly referred to as ecological studies since groups having differing exposure histories are sometimes defined on an ecologic or geographic basis.

Denote by \( \lambda_k(t) \) the age- and sex-specific disease rate in the \( k \)th group during (chronological) time period \( t \). A multiple group study involves the analysis of estimates of \( \lambda_k(t) \), \( k = 1, \ldots, K \) during a fixed time period; a time trend study involves estimates of \( \lambda_k(t) \), \( t = 1, \ldots, T \) in a single population, while a mixed study involves estimates of \( \lambda_k(t) \) at several values of both \( k \) and \( t \). An exponential-form relative-risk model for \( \lambda_k(t) \) can be written, in the notation of “Relative-Risk Models,” as

\[
\lambda_k(t) = \lambda_{k0}(t) \exp \left\{ X_{k1}(t) \beta \right\},
\]

from which the average disease rate \( \lambda_k(t) \) for the \( n_k(t) \) individuals in group \( k \) during time period \( t \) is

\[
\lambda_k(t) = \lambda_{k0}(t) \left[ \sum_{i=1}^{n_k(t)} \exp \left\{ X_{ki}(t) \beta \right\} / n_k(t) \right]
\]

\[
= \lambda_{k0}(t) \exp \left\{ \bar{X}_k(t) \beta \right\}
\]

where

\[
\bar{X}_k(t) = n_k(t) X_{k1}(t)
\]

and

\[
d_{k1}(t) = X_{k1}(t) - \bar{X}_k(t)
\]

Let \( y_{ik}(t) \) denote the observed age- and sex-specific disease incidence rate in group \( k \) during time period \( t \) as may be available from a disease register or other administrative source. From the above expression for \( \lambda_k(t) \), one expects a regression of \( \log y_{ik}(t) \) on \( X_{ki}(t) \) for various values of \( k \) or \( t \) (or both) to yield biased estimates of the relative-risk parameter \( \beta \), because of the influences of the residuals \( d_{k1}(t) \), even if the logarithms of the baseline rates \( \lambda_{k0}(t) \) can be regarded as independent random variables with a common mean. This specification bias will be small if the \( d_{k1}(t) \) values are small, that is, if the exposure and other regression variables have little variation within groups. Such bias presumably can be reduced by extending the regression equation to include averages of squares and of higher powers of the \( d_{k1}(t) \) terms, though there does not appear to have been specific study of this approach. A closely related approach would replace the exponential-form relative-risk model by a linear-form model, so that

\[
\lambda_k(t) = \lambda_{k0}(t) \left\{ 1 + \lambda_{k1}(t) \beta \right\},
\]

and

\[
\lambda_k(t) = \lambda_{k0}(t) \left\{ 1 + \bar{X}_k(t) \beta \right\}
\]

from which the regression of \( y_{ik}(t) \) on \( X_{ki}(t) \), under certain random-effects assumptions on the baseline rates \( \lambda_{k0}(t) \), will yield valid estimates of the linear relative-risk parameters (49). Note, however, that an exponential-form relative-risk model often might be more parsimonious than a linear-form model in environmental epidemiologic applications so that the regression vector in a linear relative-risk model may need to be quite lengthy and involve, for example, the average of product terms between exposure and potential confounding factors in order to adequately describe the data. In a multigrouip study, it may be sensible to assume the \( \lambda_{k0}(t) \) terms are independent random variables with a common mean for \( k = 1, \ldots, K \) thought it often may be useful to allow for the possibility of correlation among groups in a similar geographic area. In time-trend and mixed studies, however, it will typically be essential to model, or otherwise accommodate, the correlation structure among \( \lambda_{k0}(t) \), \( t = 1, \ldots, T \) at any fixed \( k \). Inadequate modeling of the \( \lambda_{k0}(t) \) may lead to aggregation bias.

These types of data analysis methods have received very little attention in the scientific literature and constitute an important gap in the collection of methods pertinent to environmental epidemiologic applications. Aggregate data studies involving the simple linear regression of disease rates or the logarithm of disease rates on average exposures and average values of potential confounding factors can often be conducted quickly and cheaply and can play a useful role in hypothesis generation. It is obvious, however, that more comprehensive data sources and more sophisticated data analyses typically will be required if aggregate data studies are to contribute reliably to the identification and estimation of exposure–disease associations. Better data could come from randomly sampling each of the compared groups in order to obtain estimates, \( X_{ki}(t) \) of acceptable precision for use in a linear relative-risk model or to obtain estimates of the average of \( \exp \{ X_{ki}(t) \beta \} \), \( i = 1, \ldots, n_k(t) \) for use in an exponential relative-risk model. Random measurement error in the ascertainment of individual exposure and confounding factors could impact substantially survey design. Better data analyses may arise from the application of so-called marginal methods (50,51) to mean and covariance models for the set of \( y_{ik}(t) \) or \( \log y_{ik}(t) \) values being analyzed.

Most effort to date concerning aggregate data studies has been directed to identifying the biases that may arise from aggregation, confounding, and other sources (52,53). It seems timely to direct a major effort to the development of procedures to prevent (or greatly reduce) such biases and, hence, to evaluate whether aggregate data studies can play a more fundamental and useful role in environmental epidemiologic studies and in epidemiologic research more generally.

**Comparing and Combining the Results of Several Studies**

Studies of a certain exposure–disease association may, for a variety of practical reasons, be lacking in power, and they may be subject to biases that can differ according to the population under study, the type of study design, and the rigor of the investigation. It follows that tests of agreement among the results of various studies and the formal combining of results from pertinent studies can play an important role in an overall exposure–disease association assessment.

Under ideal conditions, each of the types of studies described above can yield a valid estimate \( \hat{\beta} \) of the logarithm of the relative risk associated with a specified exposure history, as well as an estimate \( \hat{\sigma}^2 \) of its variance. The

---

*Environmental Health Perspectives Supplements Volume 101, Supplement 4, December 1993*
logarithm is used here, because its estimate is likely to adhere more closely to a normal distribution (with mean $\beta$) than the estimate $\theta$ of the relative risk itself. Suppose $m$-independent studies yield (scalar) log-relative risk estimates $\beta_1, \ldots, \beta_m$ with corresponding variance estimates $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_m^2$.

$$\beta = \frac{\sum \hat{\sigma}_i^{-2} \beta_i}{\sum \hat{\sigma}_i^{-2}}$$  \hspace{1cm} [20]

estimates a weighted mean of $\beta_i$'s, which reduces to a common $\beta$ if all $\beta_i$'s are identical. To obtain the most stable estimate of this common mean, one can follow developments arising from Cochran's (54) introduction of partial weighting, thereby avoiding weights $\hat{\sigma}_i^{-2} \neq \sigma_i^{-2}$, which may be too small.

If all the $\beta_i$'s are the same and the $\beta_i$ are independent and normally distributed, then

$$X^2 = \sum_{i=1}^{m} \hat{\sigma}_i (\beta_i - \beta)^2$$  \hspace{1cm} [21]

will have a chi-square distribution with $m-1$ degrees of freedom, thereby giving a simple test of "all $\beta_i = \beta$" (assuming each $\beta_i$ is distributed normally). If the $\beta_i$'s are not identical, then a $t$-procedure can be used to set confidence limits for the weighted mean

$$\beta = \frac{\sum \hat{\sigma}_i^{-2} \beta_i}{\sum \hat{\sigma}_i^{-2}}$$  \hspace{1cm} [22]

Confidence limits on $\beta$ are approximately

$$\beta \pm t_v (\sum \hat{\sigma}_i^{-2})^{-1/2}$$  \hspace{1cm} [23]

where $t_v$ is a critical value of $t$ on $v$ (somewhat less than $m$) degrees of freedom. These limits are often conservative, particularly when the $\beta_i$ follow longer-tailed distributions.

There are various reasons why the chi-square test described may provide evidence of heterogeneity of the relative-risk estimates from the $m$ studies. For example, studies of the same type (e.g., $m$-cohort studies) may have differentially controlled for confounding or may have defined and measured exposure differently. Studies of different types (e.g., $m$-cohort, case-control and aggregate studies) have different sources of potential bias, for example, recall bias for case-control studies and aggregation bias in ecologic studies. Hence, it may be useful first to contrast and combine studies of the same type and then to examine whether the summary estimates of $\beta$ from each study type are heterogeneous. In respect to studies of the same type, the overview, or metaanalysis, may be strengthened by analyzing the raw data from each study in a uniform format, which would maximize their comparability in terms of confounding control and exposure modeling. A fundamental principle of such analyses is that the parameter estimate $\hat{\beta}$ is based only on the combination of within-study information, as is the case for the heterogeneity test and the log-relative risk estimate described above.

Measurement error in exposure and in covariate assessment may be a particularly important source of heterogeneity among relative-risk estimates. For example, random measurement error may attenuate severely or otherwise distort relative-risk estimates in a cohort or case-control study if, for example, exposure assessment is based on data provided by individual interviews (e.g., location and shielding information in the Hiroshima and Nagasaki cohorts), but such attenuation may not be an issue in an aggregate data study if the desired averages (see "Exposure–Response Estimation in Aggregate Data Studies") can be estimated precisely. In this circumstance, some effort to deattenuate the analytic study relative-risk estimates, or to attenuate equally the aggregate data relative-risk estimates, is essential prior to the comparison of these estimates. See Prentice and Sheppard (55) for a recent attempt to study the consistency of international disease rate, time-trend, case-control and cohort studies in the dietary carcinogenesis area. Note also that $\hat{\beta}$ will be biased as an estimator of $\beta$ if the available log-relative risk estimates $\hat{\beta}_1, \ldots, \hat{\beta}_m$ are a biased sample of estimates from existing studies, which may arise if there is so-called publication bias in which relative-risk estimates that are significantly different from unity are more likely to be reported in the scientific literature. See Yusuf et al. (56) for a discussion of some issues in the conduct of such metaanalyses.

Other Data Analysis Topics

The above presentation emphasized time to disease endpoints and corresponding relative-risk and odds-ratio models. In some areas of environmental epidemiologic research (e.g., respiratory epidemiology or neuroepidemiology), important endpoints are continuous. Much of the corresponding data analysis methodology is well established and does not need to be discussed here. However, methods for handling measurement error with continuous data (57) also require much additional development. Recent advances in the methods for analysis of longitudinal data (50) for discrete or continuous data are also quite relevant to the analysis of certain types of environmental epidemiologic data.

Preceding sections also have not addressed the simultaneous analysis of two or more endpoints. For example, in respiratory epidemiology, there may be several measures of lung function, and a data analysis goal may be to summarize exposure effects over several correlated measures of change in lung function. The estimating equation approaches mentioned above (50,51) provide an approach to such problems with discrete or continuous outcomes, but work could be done to compare these methods to univariate methods based on some summary endpoint. Methods for the analysis of correlated failure time data currently are not well established, though much statistical research is underway presently. See, for example, Clayton and Cuzick (58), Che et al. (59), and Prentice and Cai (60) for recent contributions. Correlated failure-time methods also are required for the investigation of genetic factors or gene–environment interactions under certain types of study designs. For example, in a pedigreed cohort study, it typically will be essential to allow for dependence between the disease occurrence times of family members when studying environmental exposure effects in relation to genetic indicators of susceptibility.

Morgenstern and Thomas, in this volume, mention certain designs other than those discussed thus far in this article, as well as the use of biomarker endpoints. Corresponding data analysis issues and methods will be mentioned only briefly here.

It was noted that experimental designs are practical occasionally in environmental epidemiologic research. The relative-risk and odds-ratio regression methods described above apply equally well for the comparison of disease incidence (or mortality) rates between randomization groups in individually randomized designs. However, a group-randomized design (e.g., with community as the unit of randomization) is more likely to be feasible, in which case it is essential to acknowledge the possibility of correlation among the responses (e.g., disease incidence times) of subjects in the same randomization group, which require the use of the type of correlated failure-time methods mentioned above.

In the discussion of ecologic designs it was noted that descriptive studies of the clustering of disease (e.g., in space or time) can play a useful role in the generation of environmental health hypotheses. These types of studies also have specialized data-analytic issues and methods. Statistical analysis has little to offer in the event of an isolated cluster discovered by ad hoc methods. Clusters within which the disease counts substantially
exceed expected counts perhaps are best addressed by direct fieldwork to identify a putative cause. On the other hand, hypotheses of a general tendency to cluster can be addressed statistically by using methods that compare the number of cases in certain neighborhoods of each case to the expected number of cases, while also taking account of population density. Local neighborhood tests also are available with case–control sampling. See Rothman (61) and other papers in this volume for discussions of disease-clustering methods.

The design chapter (in this issue) also emphasizes cross-sectional studies for the estimation of prevalence rates. The logistic regression methods outlined in "Relative-Risk and Odds-Ratio Estimation" may be used to relate prevalence probabilities to retrospectively obtained exposure and confounding factor histories. Of course, such prevalence probabilities reflect aspects of both disease incidence and disease duration, and therefore, may be difficult to interpret. Keiding (62) provides a comprehensive discussion of the relationships between prevalence probabilities, incidence rates, and disease durations and of the possibility of deriving estimates of age-specific incidence from cross-sectional studies.

As discussed previously, biomarkers may serve usefully as exposure indicators or as early indicators of disease (see Hatch and Thomas, this volume). An example of a biomarker as an intermediate endpoint is seen in chromosomal abnormalities in the radiation-exposed cohorts of Hiroshima and Nagasaki. The rates of such abnormalities among long-lived lymphocytes (usually 100 cells examined for each subject) have played a useful role in assessing the health effects of radiation exposure in these populations. The correlation among the chromosomal events in cells from the same study subject has a strong influence on dose–response analyses in this application (35, 63). Recent advances in the ability to study the cellular and molecular mechanisms involved when responding to exposure and disease pathogenesis will lead inevitably to greater use of biomarkers and biological measurement in environmental epidemiologic studies. Hence, data analysis methods that incorporate such measurements in a biologically meaningful fashion are required. Suitable methods for dose–response analysis with biomarker endpoints will vary according to the type of endpoint(s) involved. Recent estimating equation approaches (50, 51) often may be useful for such analyses. Circumstances under which a biomarker endpoint can substitute for disease occurrence and yield valid dose-response tests and estimates is also of considerable interest. See Prentice (64) for the introduction and discussion of such criteria.

Finally, it seems worth noting that the interpretation of relative-risk estimates from a study may depend on prior knowledge and on study goals. For example, if such estimation takes place in the context of a study specifically designed to confirm a particular association, the corresponding tests and confidence intervals are more appropriately taken at face value than if the relative risk is estimated in a purely exploratory context wherein various other exposures also are examined in relation to disease risk. In this latter situation, formal methods may be used to acknowledge the multiple hypotheses being examined, but precise statistical methods for doing so in a general way are not available. (So-called Bonferroni methods are widely available and may be precise enough.) Also, one is often neither in a purely exploratory nor in a purely confirmatory mode in data analysis.

Summary

Recommendations

Perhaps the single most important data analysis research need in environmental epidemiology concerns the development of improved methods to accommodate measurement errors in exposure assessment. Efforts aimed at the design and use of validation studies would be particularly useful, as would studies to document the scope and magnitude of measurement error influences.

A second important need concerns improved methods for the conduct and analysis of aggregate data (ecologic) studies. The development of strategies for controlling potential confounding, particularly by using individual surveys in multigroup studies, along with corresponding innovative data analysis methods, will be important. Empirical studies that illustrate various analytic and aggregate data analyses of real data sets also would be valuable.

Other pertinent topics for data analysis research include the development of improved methods for meta-analyses when studies of different types with differing potential for measurement error biases are available, the development of flexible data analysis methods, and the study of properties of analyses based on biomarker indicators of exposure or biomarker endpoints. Studies that evaluate and compare strategies for the control of confounding also merit continuing attention in environmental epidemiology as in other observational research areas. Further work on biologically based mathematical models for cancer and for other disease also would be well motivated.

REFERENCES

1. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 132:746–748 (1990).
2. Cox DR. Regression models and life tables (with discussion). J R Statist Soc B 34:187–220 (1972).
3. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: Wiley, 1980.
4. Cox DR, Oakes D. Analysis of survival data. London: Chapman and Hall, 1984.
5. Preston DL, Lubin JH, Pierce DA. Epicure: risk regression and data analysis software. Seattle, Washington: HiroSort International Corp, 1991.
6. Armitage P, Doll R. Stochastic models for carcinogenesis. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol 4 (Neyman J, ed). Berkeley: University of California Press, 1961; 19–38.
7. Whittmore AS, Keller JB. Quantitative theories of carcinogenesis. SIAM Rev 20:1–30 (1978).
8. Freedman D, Navidi W. Multistage models for carcinogenesis. Environ Health Perspect 81:161–188 (1989).
9. Freedman DA, Zeisel H. From mouse to man: the quantitative assessment of cancer risks (with discussion). Statistical Science 3:3–56 (1988).
10. Thomas DC. Statistical methods for analyzing effects of temporal patterns of exposure on cancer risks. Scand J Work Environ Health 13:353–366 (1983).
11. Brown CC, Chu KC. A new method for the analysis of cohort studies: implications of the multistage theory of carcinogenesis applied to occupational arsenic exposure. Environ Health Perspect 50:293–308 (1983).
12. Crump KS, Allen BC, Howe RB, Crockett PW. Time factors in quantitative risk assessment. J Chron Dis 40(2):101–111 (1987).
13. Thomas DC. A model for dose rate and duration of exposure effects in relation to carcinogenesis. Environ Health Perspect 87:163–171 (1990).
14. Brown CC, Chu KC. Use of multistage models to infer stage affected by carcinogenic exposure: example of lung cancer and cigarette smoking. J Chron Dis 40(2):171–179 (1987).
15. Moolgavkar SH, Knudson A. Mutation and cancer: a model for human carcinogenesis. J Natl Cancer Inst 66:1037–1052 (1981).
16. Moolgavkar SH, Dewanj A, Luebeck J. Cigarette smoking and lung cancer: a reanalysis of the British doctor's data. J Natl Cancer Inst 81:415–420 (1989).
17. Moolgavkar SH, Day NE, Stevens RG. Two-stage model for carcinogenesis: epidemiology of breast cancer in females. J Natl Cancer Inst 65:550–569 (1980).
18. Moolgavkar SH, Cross FT, Luebeck G, Dagle GE. A two-mutation model for radon induced lung tumors in rats. Radiat Res 121:28–37 (1990).
19. Thomas DC. A model for dose rate and duration of exposure effects in radiation carcinogenesis. Environ Health Perspect 87:163–171 (1990).
20. Cox DR. Partial likelihood. Biometrika 62:269–276 (1975).
21. Andersen PK, Gill RD. Cox's regression model with counting processes: a large sample study. Ann Stat 10:1100–1120 (1982).
22. Prentice RL, Self SG. Asymptotic distribution theory for Cox-type regression models with general relative-risk form. Ann Statist 11:804–813 (1983).
23. Breslow NE, Day NE. Statistical methods in cancer research. In: The analysis of case-control studies, vol 1. Lyon: International Agency for Research on Cancer, 1980.
24. Preston DL, Kato H, Kopecky KJ, Fujita S. Studies of the mortality of the A-bomb survivors: cancer mortality 1950–82. Radiat Res 111:151–178 (1987).
25. Barlow WE, Prentice RL. Residuals for relative risk regression. Biometrika 75:65–74 (1988).
26. Prentice RL. Covariate measurement errors and parameter estimation in Cox's failure time regression model. Biometrika 69:331–342 (1982).
27. Pepe MS, Fleming T. A nonparametric method for dealing with mismeasured covariate data. J Am Stat Assoc 86:108–113 (1991).
28. Kupper LL. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. Am J Epidemiol 120:643–648 (1984).
29. Prentice RL, Sheppard L. Dietary fat and cancer: rejoinder and discussion of research strategies. Cancer Causes and Control 23:53–58 (1991).
30. Clayton DG. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In: Statistical models for longitudinal studies of health (Dwyer JH, Lipper P, Feinstein M, Hoffmeister H, eds). Oxford: Oxford University Press, 1992; 301–331.
31. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. Int J Epidemiol 12:93–97 (1983).
32. Hui SSL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 36:167–171 (1980).
33. Pierce DA, Stratton D, Vaeth M. Allowing for random errors in radiation dose estimates for the atomic bomb survivors. Radiat Res 126:36–42 (1990).
34. Whittemore AS. Errors-in-variables regression using Stein estimates. Am Stat 43:226–228 (1989).
35. Spoto R, Stratton DO, Awa AA. An estimate of the magnitude of random errors in the DS86 dosimetry from data on chromosome aberrations and severe epilation. Radiat Res 128:157–169 (1991).
36. Armstrong BG. The effect of measurement error on relative risk regressions. Am J Epidemiol 132:1176–1184 (1991).
37. Thomas D, Stratton D, Dwyer J. Exposure measurement error: influence on exposure–disease relationships and methods of correction. Annu Rev Public Health 14:69–93 (1993).
38. Stevens V, Thomas DC, Lyon J, Tills JE, Kerber R, Simon SL, Lloyd RD, Elghany NA, Preston–Martin S. Leukemia in Utah and radioactive fallout from the Nevada Test Site: a case–control study. Jama 264:585–591 (1990).
39. Thomas DC, Gauderman J, Kerber R. A nonparametric Monte Carlo approach to adjustment for covariate measurement errors in regression (tech rep #16, USC Department of Preventive Medicine). Biometrics, (in press).
40. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. Stat Med 7:745–757 (1988).
41. Greenland S. Statistical uncertainty due to misclassification: implications for validation studies. J Clin Epidemiol 41:1167–1174 (1988).
42. Spiegelman D, Gray R. Cost–efficient study designs for binary response data with Gaussian covariate measurement error. Biometrics 47:851–869 (1991).
43. Prentice RL, Breslow NE. Retrospective studies and failure time models. Biometrika 65:153-158 (1978).
44. Langholz B, Thomas DC. Nested case–control and case–cohort methods of sampling from a cohort: a critical comparison. Am J Epidemiol 31:169–176 (1990).
45. Langholz B, Thomas DC. Efficiency of cohort sampling designs: some surprising results. Biometrics 47:1553–1571 (1991).
46. Prentice RL. A case–cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–11 (1986).
47. White JE. A two-stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 115:119–128 (1982).
48. Breslow NE, Cain KC. Logistic regression for two-stage, case–control data. Biometrika 74:11–20 (1988).
49. Prentice RL, Sheppard L. Validity of international, time-trend and migrant studies of dietary factors and disease risk. Prev Med 18:167–179 (1989).
50. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22 (1986).
51. Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47:825–839 (1991).
52. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 18:269–274 (1989).
53. Pantadosi S, Byar DP, Green SB. The ecological fallacy. Am J Epidemiol 127:893–904 (1988).
54. Cochran WG. Some methods for strengthening the common $\chi^2$ tests. Biometrics 10:417–451 (1954).
55. Prentice RL, Sheppard L. Dietary fat and cancer: consistency of the epidemiologic data and disease prevention that may follow from a practical reduction in fat consumption. Cancer Causes and Control 1:87–97 (1990).
56. Yuusuf S, Simon R, Ellenberg S. Proceedings of a workshop on methodologic issues in overviews of randomized clinical trials. Statist Med 6:217–409 (1987).
57. Fuller W. Measurement error models. New York: John Wiley and Sons, 1987.
58. Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model (with discussion). J R Statist Soc A 148:82–117 (1985).
59. Wei LJ, Lin DY, Wei S. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J Am Stat Assoc 84:1065–1073 (1989).
60. Prentice RL, Cai J. Covariance and survivor function estimation using censored multivariate failure time data. Biometrika 79:495–512 (1992).
61. Rothman KJ. A sobering start for the cluster buster’s conference. Amer J Epid 132:56–571 (1990).
62. Keiding N. Age–specific incidence and prevalence: a statistical perspective (with discussion). J R Statist Soc A 154:371–412 (1991).
63. Drake M, Prentice RL. The analysis of chromosomally aberrant cells based on a beta–binomial distribution. Radiat Res 98:456–470 (1984).
64. Prentice RL. Surrogate end points in clinical trials: discussion, definition and operational criteria. Stat Med 8:431–440 (1989).