Multiple Factors: Discussion of Statistical and Epidemiologic Aspects

WILLIAM HAENZHEL
National Cancer Institute, USPHS, Bethesda, Maryland

SUMMARY

One can be optimistic about the contributions which the statistical-epidemiologic approach will make to future studies of multiple factors in cancer etiology. The epidemiologic method can lend itself to the study of both host characteristics and environmental factors. However, cancer epidemiology has not completely left the descriptive phase behind, and emphasis still needs to be placed on identification of factors associated with gradients in risk. As the fund of information on variations in cancer risk among population groups grows, more and more questions will be posed that will lead to investigations of multiple factors. Epidemiologic observations should foster certain lines of experimental work, the findings from which can later be incorporated into more refined epidemiologic studies, thus continuing and sustaining the interaction between epidemiology and experimental carcinogenesis.

Above all, investigators should not be deterred by possible imperfections in the data they can collect. It is better to do what can be done today, but with an eye to the future and recognizing that current, imperfect data are to some degree self-perfecting as provisional findings are introduced into the design and pursuit of new studies.

Multiple factors is an inclusive, all-embracing term. Dr. Foulds, in his address, has ranged widely and considered under this heading (a) the combined action of factors (genetic, hormonal, virus, chemical, and physical) in the induction of cancers; (b) the sequence of discrete events in the transformation from a normal to a cancer cell; (c) the spread of cancer cells in the host following initiation of a cancer cell. My task is to supplement Dr. Fould's review of developments in experimental carcinogenesis bearing on multiple factors with a few comments on related statistical and epidemiologic investigations; this may call for some modifications in the conceptual approach. For example, I would like to refer to his description of the induction of bladder cancer by aromatic amines and to suggest that the label "multiple" or "single" factor depends in part on definition and convention. Either the chain of events or the individual reactions might be regarded as the primary unit, and the choice should be the one best suited to understanding of the data within a particular frame of reference.

OBSERVATIONAL ASSOCIATIONS

The epidemiologic approach has its limitations, which are inherent in the nature of the data available. Information on cancer in man comes mainly from observations and not from planned experiments. Most results are thus reported as observational associations, usually stated in the form of a rate (incidence, mortality, etc.) or probability that an event will occur in a defined population. The time sequence for the defined event and associated factors remains indeterminate, a deficiency stressed by many authors. This point can be overemphasized, since collateral information will often rule out some interpretations of observed associations. For example, discussions of the association of cigarette smoking and lung cancer have not entertained seriously such hypotheses as "lung cancer causes cigarette smoking perhaps because a precancerous condition sets up a process which leads to a craving for tobacco" (4, 9).

The interpretation of observational associations is too large a subject to review here, but a few points may be worth noting:

a) The virtue of epidemiologic methods is that they provide information on man. Their prime role should be to provide leads which can be exploited in investigations combining the epidemiologic and experimental approaches.

b) Observational associations are crude measures of biologic effects and can be so analyzed, if need be, without assumptions as to the means by which the effects are produced. There is an obvious analogy to thermodynamics, which can be thought of as a form of "energy" bookkeeping. Its pursuit requires knowledge of the conversion values among various forms of energy, but the accounting per se does not depend on any assumptions about how energy is transformed. Neither thermodynamics nor epidemiology denies the utility of conceptual
models; when combined with such information, they become even more powerful study tools.

c) The use of observational associations requires only that factors be identified for which the events and population at risk can be classified in a consistent, reproducible manner. Crudity of classification constitutes no barrier so long as the classifications lead to results on which further studies to sharpen and pinpoint differences in risk can be based. From an operational point of view the joint manipulation of 2 or more classifications can be regarded as an exercise in multiple factors even if the classifications should later turn out to represent a common component, such as the possibility once advanced by Stocks that benz[a]pyrene might be the active agent in the combined effects of cigarette smoke and polluted urban atmosphere (24).

d) Statements concerning cause-effect relationships are inferences; the drawing of inferences from either experimental or observational data presents problems which differ quantitatively rather than qualitatively. Improvement in inferences by synthesis of information from all relevant sources is the appropriate goal, and we would agree with the assessment that "the epidemiologic method, when coupled with clinical or laboratory observations, can provide the basis from which judgments of causality may be derived" (6).

ANALYSIS OF OBSERVATIONAL ASSOCIATIONS

It has been noted that epidemiologic studies rest on the computation of rates whose determination requires a numerator (count of specified events) and a denominator (population at risk). The situation remains unchanged by the introduction of multiple-factor considerations. The events and population at risk are amenable to cross classification, and rates can be calculated for any combination of factors. Cross classification of variables is an elementary, but powerful, statistical tool that has not been outdated by recent innovations in statistical techniques. Flexibility in analysis of joint effects without premature commitment to a specific hypothesis is best maintained by insisting on preparation of data in a form which will permit the inspection of a schedule of rates specific for any prescribed combination of 2 or more factors. Such a schedule contains essentially all the available information. From examination of the gradients in the rates along all axes, the analog of what might be called a dose-response surface can be characterized.

The crux of the problem remaining is the transition from pure description to an analysis of the interaction of multiple factors reflected in the rates. One can go astray here on the borderline between observation and inference. The point to be stressed is that interaction is relative to the analytic model adopted. Depending on the model postulated—additive effects, multiplier effects, etc.—different estimates of interaction can be obtained. The situation parallels that encountered in reporting on combinations of drug therapy. Numerous pitfalls in the labeling of synergistic action, potentiation, etc., have been discussed (10) which have counterparts here. The investigator can best stay out of trouble by keeping the relation between his data and the analytic model fixed firmly in mind.

Positive prescriptions for an appropriate analysis are difficult to specify in the abstract, but a few warning notes can be sounded. Beware of statisticians who propose to use standard computer programs to make covariance and multiple regression analyses of your data. Such analyses may stipulate that interaction effects are absent and thus preclude inspection of the basic data for information on interaction, supposedly a primary goal. The problem has been settled by assumption rather than by observation. A recent paper on cervical cancer provides an example of inappropriate use of the additivity assumption, in the presence of strong evidence for intercorrelated study variables (22). Investigators would be well advised to eschew factor analysis, which calls for departing from the axes on which the observations have been measured and employing certain mathematical properties to arrive at new axes for positioning the observations. The stumbling block is the characterization of the units of measurement for the new axes. What do they represent? In our present state of knowledge a preferable strategy would be to remain in the world of observational data and to retain the ability to specify the properties of the classification schemes. The observed results can be used to sharpen the classification of study variables, and in my opinion reliance on observation rather than mathematics will prove the more fruitful approach.

Presentation of data in unhelpful forms should also be avoided. For example, discriminant function analyses which rely on weighted scores of several study variables to categorize and discriminate between groups of cases and controls will not normally lead to useful conclusions about interaction effects.

The emphasis placed on a simple-minded, cross classification approach to observational data does not imply that more sophisticated analyses incorporating assumptions on interaction effect are always inappropriate. Given a limited set of data or a calculated decision to test a specific hypothesis against the data, an analysis which shortcuts detailed inspection of factor-specific rates can be defended. What should be reiterated is that such procedures do not replace or have virtues superior to direct inspection of the schedule of rates.

STATUS OF "MULTIPLE-FACTOR" STUDIES

Until recently, the epidemiologic-statistical literature has had little to contribute on multiple factors. Most studies have reported on the association of a disease with one factor under scrutiny. Other factors, when introduced, have usually been treated as variables to be controlled in the main comparison. It is not difficult to understand why deliberate, planned investigations of the joint action of 2 or more factors have not been implemented. In the early phases emphasis is placed on the detection and demarcation of large effects, and investigators have been content to look for them one at a time. Only when an association has been established does one raise the issue of its effect in concert with other factors on disease response. To investigate ab initio the combined effects of 2 or more factors without knowledge of their individual (uncontrolled) contributions would appear highly speculative. Furthermore, the chances of detect-
ing joint effects of 2 or more factors to which the investigator would not ultimately be led via investigation of either factor singly seem remote. Joint effects would usually escape discovery only if the high- or low-risk population with attributes AB was small compared to populations AB and AB, for whom “normal” risks prevailed.

Sample size requirements have also deterred studies of multiple factors; when the association with one factor is very strong (cigarette smoking and lung cancer), the number of observations required to delineate the pattern of joint effects increases very rapidly. Interrelationship of variables available for study can also complicate matters. For example, it is well known that cervical carcinoma is associated with age at first intercourse, social class, marital status, and fertility (2, 7, 18, 19, 23, 27), but the interrelationships are so strong that analyses of this complex of factors have not really advanced our knowledge and understanding of this disease.

Nevertheless, the evolution of epidemiologic studies of cancer is leading to greater emphasis on multiple factors through the gradual accumulation of a fund of information delineating the variety of effects for individual sites. Work continues on the nature of the interaction between cigarette smoking and air pollution as it affects lung cancer risks (13, 14, 21, 24). Investigations of the interaction of cigarette smoking with established occupational carcinogens could be undertaken in employee populations with high lung cancer risks, such as chromate workers (1). Cancers of the esophagus and larynx, for which associations with both cigarette smoking and alcohol consumption have been noted (25, 26), should prove amenable to a multiple-factor approach to elaborate the uncertain inferences as to the nature of interaction between these variables. The combination of cytology screening with conventional retrospective and forward study technics opens the way to more subtle investigations on the interaction of sex and reproductive histories, circumcision of marital partner, and intercurrent infections of the genital tract on cervical cancer risks. Breast cancer will provide further opportunities to study the conjunction of host characteristics and environmental factors; the strong association of reproductive and breastfeeding history with breast cancer risks might suggest further work on such host characteristics as hormonal status (5, 17), despite some reservations expressed about hypotheses related to hormonal factors (11, 20). The last problem is closely related to the identification of susceptible and nonsusceptible individuals.

STUDY MATERIALS

Observations on cancer in man come from 6 major sources: (a) mortality data, (b) cancer registers, (c) cancer morbidity surveys, (d) retrospective studies of diagnosed cases and matched controls, (e) forward studies of defined cohorts, and (f) screening of populations for precursive signs (15). The first 3 are unlikely vehicles for the study of multiple factors, since data collection must conform to standard classifications of demographic variables. The controlling conditions are availability of data rather than relevance to specific objectives. We must rely on the latter 3 methods to elicit information on the joint effects of 2 or more factors. Here, the scope and content of the data collected are determined by their relevance to study objectives, technical factors—reproducibility and reliability of items collected—and costs.

Two rationales can be advanced for forward observations on cohorts:

a) The accumulation of observed events over time is recorded in the absence of information on interim changes in status of individuals. This leads to an analysis as of a cross section of time in which the longer period is used solely to increase the number of observations.

b) Observed events and changes in status of individuals are both recorded. This permits the estimation of transitional probabilities between specified states, and these opportunities will occur most often in studies which introduce screening of populations for precursive signs. The estimation of transitional probabilities is crucial to some of the multiple-factor problems considered by Dr. Foulds.

STUDY SETTINGS AVAILABLE FOR EXPLOITATION

In principle, the familiar retrospective studies of cases and controls could be adapted to investigations of multiple factors. Given a large enough study, risks for specific combinations of factors can be estimated by brute force via cross classification, such as was attempted in relating smoking and residence histories to lung cancer (13). Alternatively, information from 2 or more variables could be incorporated into a single comprehensive classification scheme and applied to both case and control series. Conventional machinery for estimating relative risks could then be used without change (3). However, there are severe limitations to what can be expected in practice. Unless the studies were done on a large scale, the confidence limits of the relative risks would be so large as to defeat the drawing of inferences on patterns of interaction.

Another defect of case-control studies is their unsuitability for estimation of transitional probabilities, since accurate determination long after the fact of the sequence in which events occurred is difficult. We should be content with the more modest goal of conjoint classification by Factors A and B and relative risk estimations for the 2 factors combined. One should press for separate estimates of $P(A)$ and $P_A(B)$ or $P(B)$ and $P_B(A)$ only if the sequence of events can be established by independent evidence.

Forward studies, particularly when large populations are available, are more suitable for multiple-factor investigations. The American Cancer Society 25-state study with a cohort of over 1 million persons provides unusual opportunities in this direction (16). A wide range of data on host characteristics and environmental exposures has been assembled, and the stage has already been set for inquiries into multiple factors. In this situation multiple factors can be investigated by the cross-classification approach utilizing information collected when the cohort was assembled or in subsequent queries. Supplementation through ad hoc queries for items thought to be particularly relevant to a specific study of multiple factors could be easily accomplished.
Following the lead of the American Cancer Society study and the earlier prospective studies of the cigarette smoking-lung cancer era, the future should witness the establishment of new cohorts and extension of data collection activities for existing cohorts. Study cohorts come into being by 2 routes. One avenue depends on the source of information. For example, the American Cancer Society study in its present form could not exist without the volunteer field workers, who constitute a unique study resource. In similar fashion, prospective studies of Seventh Day Adventists, Veterans Administration insurance policy holders, populations covered by comprehensive medical care plans, and employee groups were made possible by capitalizing on administrative schemes with built-in potential for continued surveillance.

Other cohorts are generated by attributes possessed by their members: patients with pernicious anemia, achlorhydria, other diseases; women with atypical or suspicious cytology; nonsmokers; children immunized by Salk vaccine contaminated with SV40; persons born in England and Norway who have migrated to the United States. Their special characteristics, when coupled with additional information collected by questionnaire or interview, could be employed in a variety of multiple-factor studies.

HOST AND ENVIRONMENT

The preceding comments have been couched in general terms and apply with equal force to studies of 2 or more host characteristics, 2 or more environmental factors, or a combination thereof. Questions concerning the relative contributions of host and environment to determination of cancer risks have been raised in the context of ongoing work, and investigations directed to this end should receive high priority in the immediate future. This statement holds true even for cancer sites displaying very strong associations with environmental factors. In this connection the literature on lung cancer contains suggestions that genetic and constitutional factors might be important determinants of risk (4). It is a curious fact that although several studies have indicated some differences between smokers and nonsmokers in personality traits and psychologic attributes, no one has taken the further step of carrying out a retrospective or forward study to see whether such variables do produce gradients in lung cancer risk (12). Such investigations are worth undertaking to see if any predictive factors for classifying individuals rich in the United States because the diverse ethnic groups display widely different backgrounds of site-specific cancer risks prevailing in the country of origin. Present evidence indicates that the patterns of displacement in cancer risks among the native-born vary greatly by site, but that the several migrant groups often display similar features in the direction and nature of shifts for a specific site. The results show sufficient promise to warrant comparative studies of migrants calling for collection of data in the United States and the country of origin. Unfortunately, information on this point can be sought from groups which in effect represent experiments of nature—persons who have migrated from their place of birth to a different environment. For many migrant groups, including foreign-born residents in the United States, information on site-specific cancer risks in the country of origin and in the host country is available as a base line against which to contrast the migrant experience. These relationships, when supplemented by data on the displacement in risks among the native-born, may offer leads which will direct further research to endogenous or environmental factors.

A substantial body of data on cancer risks by site for migrant groups has been assembled for the United States (11), and work along similar lines has been reported from New Zealand (8). The opportunities are particularly rich in the United States because the diverse ethnic groups display widely different backgrounds of site-specific cancer risks prevailing in the country of origin. Present evidence indicates that the patterns of displacement in cancer risks among migrants to the United States vary greatly by site, but that the several migrant groups often display similar features in the direction and nature of shifts for a specific site. The results show sufficient promise to warrant comparative studies of migrants calling for collection of data in the United States and the country of origin. Studies of cancer risks among persons born in Japan, England, and Norway are now underway in the United States. Collection of additional data covers a wide range of demographic variables, including age at time of migration. Information on cancer risk by age at migration should play a key role in determining whether later phases of these investigations should concentrate on genetic factors associated with differences in risk be attacked in a way that will not bankrupt us? Do we introduce simple observations on physical characteristics (body build, hair distribution), questionnaires (Cornell Medical Inventory), or routine laboratory tests (blood groups) into case-control studies and use suggestive results as a launching pad for proposing more expensive and sophisticated studies of hormonal status, antigen-antibody reactions, etc.? If the effects are linked with an attribute which does not change over time, findings of this nature might be elaborated in case-control studies. However, if the status of individuals changes with time, or even worse, the appropriate classification for an individual requires measurement of departures from an earlier bench-mark reading, a complex forward study would be required, but it would have a low probability of revealing important effects. This line of approach is handicapped by the fact that collation of the usual sources of cancer incidence and mortality data provides few clues on the relative importance of endogenous and environmental factors.
in this particular study setting, which permits much greater variety and flexibility in analysis.

REFERENCES


Multiple Factors: Discussion of Statistical and Epidemiologic Aspects

William Haenszel


Updated version  Access the most recent version of this article at:  
http://cancerres.aacrjournals.org/content/25/8/1356

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.