Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease [1]

NATHAN MANTEL *and* WILLIAM HAENSZEL, *Biometry Branch, National Cancer Institute,* [2] *Bethesda, Maryland*

Summary

The role and limitations of retrospective investigations of factors possibly associated with the occurrence of a disease are discussed and their relationship to forward-type studies emphasized. Examples of situations in which misleading associations could arise through the use of inappropriate control groups are presented. The possibility of misleading associations may be minimized by controlling or matching on factors which could produce such associations; the statistical analysis will then be modified. Statistical methodology is presented for analyzing retrospective study data, including chi-square measures of statistical significance of the observed association between the disease and the factor under study, and measures for interpreting the association in terms of an increased relative risk of disease. An extension of the chi-square test to the situation where data are subclassified by factors controlled in the analysis is given. A summary relative risk formula, R, is presented and discussed in connection with the problem of weighting the individual subcategory relative risks according to their importance or their precision. Alternative relative-risk formulas, R_1, R_2, R_3, and R_4, which require the calculation of subcategory-adjusted proportions of the study factor among diseased persons and controls for the computation of relative risks, are discussed. While these latter formulas may be useful in many instances, they may be biased or inconsistent and are not, in fact, averages of the relative risks observed in the separate subcategories. Only the relative-risk formula, R, of those presented, can be viewed as such an average. The relationship of the matched-sample method to the subclassification approach is indicated. The statistical methodology presented is illustrated with examples from a study of women with epidermoid and undifferentiated pulmonary carcinoma.—**J. Nat. Cancer Inst. 22: 719–748, 1959.**

Introduction

A retrospective study of disease occurrence may be defined as one in which the determination of association of a disease with some factor is based on an unusually high or low frequency of that factor among diseased persons. This contrasts with a forward study in which one looks instead

[1] Received for publication November 6, 1958.

[2] National Institutes of Health, Public Health Service, U.S. Department of Health, Education, and Welfare.

for an unusually high or low occurrence of the disease among individuals possessing the factor in question. Each approach has its advantages. Among the desirable attributes of the retrospective study is the ability to yield results from presently collectible data, whereas the forward study usually requires future observation of individuals over an extended period (this is not always true; if the status of individuals can be determined as of some past date, the data for a forward study may already be at hand). The retrospective approach is also adapted to the limited resources of an individual investigator and places a premium on the formulation of hypotheses for testing, rather than on facilities for data collection. For especially rare diseases a retrospective study may be the only feasible approach, since the forward study may prove too expensive to consider and the study size required to obtain a respectable number of cases completely unmanageable.

In the absence of important biases in the study setting, the retrospective method could be regarded, according to sound statistical theory, as the study method of choice. This follows from the much reduced sample sizes required by this approach and may be illustrated by the following extreme example. If a disease attack rate of 10 per 100,000 among 50 percent of the population free of some factor were increased tenfold among the other half of the population subject to the factor, a retrospective study of 100 cases and 100 controls would, with high probability, reveal this significantly increased risk. On the other hand, a forward study covering 2,000 persons, half with and half without the factor, would almost certainly fail to detect a significant difference. For comparable ability to find the type of increased risk just indicated, a forward study would need to cover about 500 times as many individuals as the corresponding retrospective study. The disparity in the required number of persons to be studied could, of course, be reduced by lengthening the follow-up period for forward studies to increase the experience in terms of person-years observed. The larger sample size required for the forward study reflects principally the infrequent occurrence of the disease entity under investigation. In the example illustrated, uncovering 100 cases of disease in a forward study would require either 100,000 individuals with the factor or 1,000,000 without. For diseases with a higher probability of occurrence the disparity in required size between retrospective and forward studies would be progressively reduced.

The retrospective study might be looked upon as a natural extension of the practice of physicians since the time of Hippocrates, to take case histories as an aid to diagnosis. Its guise has varied with respect to the means of measuring the prevalence of the suspect factor among diseased persons and the criteria for determining unusual departures from normal experience. When an association is so marked, as in Percival Pott's observations on the representation of chimney sweeps among cases of scrotal cancer, no further quantitative data are required to perceive its significance.

The retrospective approach has often been employed in studies of communicable diseases, one illustration being Snow's observations (*1*) on a common water supply for cholera cases in an area served by several sources (there would have been no element of unusualness had there been but one water supply). When a disease is epidemic in a circumscribed locality, the disease-free population in the same area offers a natural contrast. The method may be used successfully for endemic diseases as well. Holmes, in reaching his conclusions on the communicable nature of puerperal fever (*2*), noted particularly that a large number of women with puerperal fever had been attended by the same physicians. In this context it should be emphasized that communicable disease investigations have often combined retrospective and forward study methods. For example, Snow supplemented his retrospective observations on water supply by a contrast of cholera rates among subscribers of the Southwark and Vauxhall water company with the experience of persons served by the Lambeth water company within the same area.

When a disease occurs sporadically, or its occurrence is not confined to a well-defined group (such as women at childbirth), a choice of controls is not immediately evident. For cancer and other diseases characterized by high fatality rates, a study restricted to decedents might use persons dying from other causes as controls. Rigoni Stern adopted this technique in deducing the relationship of cancer of the breast and of the uterus to pregnancy history (*3*). Some contemporary studies have also used deaths from other causes as controls (*4, 5*).

The present-day controlled retrospective studies of cancer date from the Lane-Claypon paper on breast cancer published in 1926 (*6*). This report is significant in setting forth procedures for selecting matched hospital controls and relating them to a consideration of study objectives. Retrospective techniques have since been applied in several investigations of cancer, including the following partial list of current references for a few primary sites: bladder (*7–10*), breast (*11–13*), cervix (*13–16*), larynx (*17, 18*), leukemia (*19*), lung (*18, 20–27*), and stomach (*18, 28–30*).

Statisticians have been somewhat reluctant to discuss the analysis of data gathered by retrospective techniques, possibly because their training emphasizes the importance of defining a universe and specifying rules for counting events or drawing samples possessing certain properties. To them, proceeding from "effect to cause," with its consequent lack of specificity of a study population at risk, seems an unnatural approach. Certainly, the retrospective study raises some questions concerning the representative nature of the cases and controls in a given situation which cannot be completely satisfied by internal examination of any single set of data.

Only a few published papers have treated the statistical aspects of retrospective studies. Cornfield discussed the problem in terms of estimated measures of relative and absolute risks arising from contrasts of persons with and without specified characteristics (*31*). His paper was concerned with the simple situation of a homogeneous population of cases and controls, presumably alike in all characteristics except the one under

investigation, which could be represented by a single contingency table. In a later contribution he handled the problem of controlling for other variables by adjusting the distribution of controls to the observed distribution of cases (*16*). Dorn briefly mentions retrospective studies with emphasis on such topics as sources of data, choice of controls, and validity of inferences (*32*).

This paper presents a method for computing relative risks for retrospective study contrasts, which controls for the effects of other variables by use of the basic statistical principle of subclassification of data. The related problem of significance testing is also considered. Since details of statistical treatment are conditioned by study objectives, data collection methods, choice of a control series, and the use of matched or unmatched controls, these topics are also discussed briefly.

Objectives

Retrospective studies are relatively inexpensive and can play a valuable role as scouting forays to uncover leads on hitherto unknown effects, which can then be explored further by other techniques. The effects may be novel and not suggested by existing data, as in the pioneer work on the association of smoking and lung cancer or the association of blood type and gastric cancer, or they may represent refinements of current knowledge. The latter category might include collection of lifetime residence and/or work histories to elaborate differences in incidence and mortality which appear when some diseases are classified by last place of residence or last occupation of the newly diagnosed case or decedent.

With diseases of low incidence the controlled retrospective study may be the only feasible approach. Here emphasis should be placed on assembling results from several studies. Before accepting a finding and offering an interpretation, scientific caution calls for ascertaining whether it can be reproduced by others and in other administrative settings having their own peculiar biases.

- *A primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done.* Even when observations for a forward study have been collected, a supplementary retrospective approach to the same body of material may prove useful in collecting more data on points not covered in the original study design or in amplifying suggestive associations appearing in the initial forward-study results.

The findings of a retrospective study are necessarily in the form of statements about associations between diseases and factors, rather than about cause and effect relationships. This is due to the inability of the retrospective study to distinguish among the possible forms of association—cause and effect, association due to common causes, etc. Similar difficulties of interpretation arise in forward studies as well. A forward study, to avoid these difficulties, would need to be performed with the preciseness of a laboratory experiment. For example, such a study of associations with cigarette smoking would require that an investigator

randomly assign his subjects in advance to the various smoking categories, rather than simply note the categories to which they belong. The inherent practical difficulties of such an enterprise are evident.

In addition to the failings shared with the forward study, the retrospective study is further exposed to misleading associations arising from the circumstances under which test and control subjects are obtained. The retrospective study picks up factors associated with becoming a diseased or a disease-free *subject*, rather than simply factors associated with presence or absence of the disease. The difficulties in this regard may be most pronounced when the study group represents a cross section of patients alive at any time (prevalence), including some who have been ill for a long period. Inclusion of the latter may lead to identification of items associated with the course of the illness, unrelated to increased or decreased risk of developing the disease. The theoretical point has been raised that factors conducive to longer survival of patients may be found in "prevalence" samples and interpreted erroneously as being associated with excess liability to the disease (*33*). Loopholes of this type are minimized when investigations are restricted to samples of newly diagnosed patients (incidence).

A partial remedy for these uncertainties lies in employing a conservative approach to interpretation of the associations observed. Recognizing the ease with which associations may be influenced by extraneous factors, the investigator may require not only that the measure of relative risk be significantly different from unity but also that it be importantly different. He may, for instance, require that the data indicate an increased relative risk for a characteristic of at least 50 percent, on the assumption that an excess of this magnitude would not arise from extraneous factors alone. However, the use of such conservative procedures emphasizes a corresponding need to pinpoint the disease entity under study. A strong relationship between a factor and a disease entity might fail to be revealed, if the entity was included in a larger, less well-defined, disease category. After the event from data now at hand, we know that a study of the association of cigarette smoking with epidermoid and undifferentiated pulmonary carcinoma is more revealing than an inquiry covering all histologic types of lung cancer.

Multiple Comparison Problem

The present-day retrospective study is usually concerned with investigating a variety of associations with a disease, little effort being involved in acquiring, within limits, added information from respondents. The results may be analyzed in a number of ways: the various factors may be investigated separately, without regard to the other factors; they may be investigated in conjunction with each other, a particular conjunction being considered a factor in its own right; or, more commonly, a factor may be tested with control for the presence or absence of other factors. Thus, if the role of cigarette smoking and coffee drinking in a given disease are under study, the possible comparisons include the relative

risk of disease for individuals who both smoke and drink as opposed to all other persons, or as opposed to those who neither smoke, nor drink coffee. In addition, the relative risk associated with smoking might be obtained separately for drinkers and nondrinkers of coffee, with a weighted average of these two relative risks constituting still another item. Conversely, risks associated with coffee drinking, with adjustments for cigarette smoking, could be computed.

The potential comparisons arising from a comprehensive retrospective study can be large. Almost any reasonable level of statistical significance used to test a single contrast, when applied to a long series of contrasts, will, with a high degree of probability, result in some contrasts testing significant, even in the absence of any real associations. The usual prescription for coping with this multiple comparison problem—requiring individual comparisons to test significant at an extreme probability level to reduce the number of associations incorrectly asserted to be true—would result only in making real associations difficult to detect.

However, the multiple comparison problem exists only when inferences are to be drawn from a single set of data. If the purpose of the retrospective study is to uncover leads for fuller investigation, it becomes clear there is no real multiple significance testing problem—a single retrospective study does not yield conclusions, only leads. Also, the problem does not exist when several retrospective and other type studies are at hand, since the inferences will be based on a collation of evidence, the degree of agreement and reproducibility among studies, and their consistency with other types of available evidence, and not on the findings of a single study.

Nevertheless, it would be wise to employ testing procedures which do not lead to a superabundance of potential clues from any one study. This may be achieved by employing nominal significance levels in testing factors of primary interest incorporated into the design of an investigation and applying more stringent significance tests to comparisons of secondary interest or to comparisons suggested by the data. For the usual problem of multiple significance testing, this would be equivalent to allocating a large part of the desired risk of erroneous acceptance of an association as real to a small group of comparisons where fruitful results were anticipated, and parceling out the remainder of the available risk to the large bulk of comparisons of a more secondary nature. This minimizes the risk of diluting, through inclusion of many secondary comparisons, the chances for detecting an important primary effect.

Representative Nature of Data

The fundamental assumption underlying the analysis of retrospective data is that the assembled cases and controls are representative of the universe defined for investigation. This obligates the investigator not only to examine the data which are the end product but also to go behind the scenes and evaluate the forces which have channeled the material to his attention, including such items as local practices of referral to specialists and hospitals and the patient's condition and the effect of these items on the probability of diagnosis or hospital admission. We re-emphasize that this requires the exercise of judgment on the potential magnitude of biases and as to whether they could result in factors seeming to be related to a disease, in the absence of a real association of the factor with presence or absence of the disease. The danger of bias may be greatest in working with material from a single diagnostic source or institution.

Among the more important practical considerations affecting retrospective studies is that they are ordinarily designed to follow the line of least resistance in obtaining case and control histories. This means that cases and controls will often be hospital patients rather than persons in the general population outside hospitals. As a result, any factor which increases the probability that a diseased individual will be hospitalized for the disease may mistakenly be found to be associated with the disease. For example, Berkson (*34*) and White (*35*) have pointed out that positive association between two diseases, not present in the general population, may be produced when hospital admissions alone are studied, because persons with a combination of complaints are more likely to require hospital treatment. In theory, bias might also be produced in reverse manner, if the suspect factor diminished the probability of hospitalization for other diagnoses used as controls. The difficulties are not unique for hospital patients. Similar loopholes in interpretation may be advanced for any special groups used as sources of cases and controls.

However, a mere catalogue of biases arising from the possibly unrepresentative nature of a sample of cases and controls should not *ipso facto* invalidate any study findings. This is a substantive issue to be resolved on its merits for a specific investigation. Collateral evidence may provide information on the potential magnitude of bias and the size of spurious associations which could result. In some situations the difference between cases and controls may be so great that postulation of an unreasonably large bias would be required. Whether he consciously recognizes it or not, the investigator must always balance the risks confronting him and decide whether it is more important to detect an effect, when present, or to reject findings, when they may not reflect the true situation. If opportunities for further testing exist, one should not be too hasty in rejecting an association as an artifact arising from the method of data collection, and in foreclosing exploration of a potentially fruitful lead.

Because of the important role retrospective studies play in studies of human genetics, mention may be made of a bias frequently encountered in studies dealing with the familial distribution of diseases. A frequently used procedure takes a group of diagnosed cases for a disease in question and a group of controls and compares the prevalence of this disease among relatives of the probands and controls. The bias arises from the unrepresentative nature of the probands with respect to familial distribution and is known in other fields as "the problem of the index case" or "the effect of method of ascertainment." It has long been recognized that the

characteristics for a random sample of families will differ from those for families to whom the investigator's attention has been directed because the family rosters include individuals selected for study on the basis of a specified attribute. For example, data on family size (number of children) obtained from siblings, rather than parents, are biased, since two or three potential index cases are present in the population for two- and three-child families as opposed to one for one-child families and none for childless couples. The analogy for disease occurrence is apparent. Families with two or three cases of the disease under study may have double or triple the probability of being represented by individuals in source material and having a representative selected as a proband than families with only one case. An appropriate analysis for this situation in studies of family size and birth order has been discussed by Greenwood and Yule (*36*), which takes account of the probability of family representation in proband data. Haenszel (*37*) has applied their correction to gastric-cancer data reported by Videback and Mosbech (*38*) and found the correction to reduce the originally reported fourfold excess of gastric cancer among relatives of probands, as compared to relatives of controls, to one of about 60 percent.

One remedy for the weakness of the retrospective approach to problems involving association of diseases and familial distribution would be to place greater reliance on forward observations of defined cohorts for data on these topics.

Controls

While easier accessibility to and lesser expense of hospital controls are important considerations, they should not deter one from collecting control data for a sample representing a more general population, if the latter are demonstrably superior. Some of the uncertainties about the superiority of hospital or general population controls arise from the need to maintain comparability in responses. The dependence of retrospective studies on comparability of responses from cases and controls cannot be overemphasized. When more accurate answers can be obtained from controls in a medical-care environment, the gain in comparability of responses for these controls could outweigh the other advantages to be derived from the more representative nature of general population controls. The difficulties may be illustrated by the experience with smoking histories. Hospital controls invariably yield a higher proportion of smokers for each sex than controls of comparable age drawn from the general population (*27*). Does this mean more complete smoking histories are collected in hospitals or does it imply that smokers have higher hospital admission rates? If the first alternative is correct, hospital controls are the appropriate choice for measuring the association of smoking history with a given disease. The second alternative calls for general population controls and in this situation the use of hospital controls yields underestimates of the degree of association.

Dual hospital and general population controls would have some merit. If control data from the two sources were in agreement, this would rule

out some alternative interpretations of the findings. In the event of disagreement, its extent could be measured and alternate calculations made on the degree of association between an event and a suspect antecedent characteristic. Where the two sets of controls lead to substantially different results, a cautious and conservative interpretation is indicated.

Some topics, such as those bearing on sex practices and use of alcohol, may be amenable to study only within a clinical setting, and the collection of general population data on these items may prove impractical. The limitations of general population controls in this regard may have been overstressed, and empirical trials to test what information can be collected in household surveys should be encouraged instead of dismissing the possibility with no investigation whatsoever. Whelpton and Freedman, for example, have reported some success in collecting histories of contraceptive practices in interviews of a random sample of housewives (*39*).

When hospital controls are chosen, some precautions may be built into the study. Within limitations on the nature of controls imposed by a study hypothesis, controls drawn from a wide variety of diseases or admission diagnoses should be preferred. This permits examination of the distribution of the study characteristics among subgroups to check on internal consistency or variation among controls. This affords protection against two sources of error: *a*) attributing an association to the disease under investigation, when the effect is really linked to the diagnosis from which controls were drawn, and *b*) failure to detect an effect because both the study and control diseases are associated with the suspect factor. The latter is far from impossible. Both tuberculosis and bronchitis have exhibited association with smoking history and the use of one disease or the other as a control could easily lead to missing the association with smoking history. Similarly, patients with coronary artery disease would not constitute suitable controls for a study of the relationship of smoking and bladder cancer and *vice versa*, since the investigator would probably conclude that smoking was not related to either disease, when in truth it appears related to both. When there is definite evidence that two diseases are associated, for example, pernicious anemia and stomach cancer, the use of one as a control for the other is contraindicated, unless the study is specially designed to elucidate some aspects of the relationship.

It is always advantageous to include several items in a questionnaire for which general population data are available. This could be considered a partial substitute for dual hospital and general population controls. Disparity among cases, hospital controls, and general population controls on several general characteristics unrelated to the study hypothesis may be regarded as warning signals of the unrepresentative nature of the hospital cases and controls.

Where possible, interviews should be conducted without knowledge of the identity of cases and controls to guard against interviewer bias, although administrative reasons will often prevent attainment of "blind" interviews. In cooperative studies employing several interviewers, the

magnitude of interviewer bias may be diminished, since it is unlikely that all interviewers will share the same bias in concert. In special circumstances, such as those prevailing at Roswell Park Memorial Institute, admissions may be interviewed before diagnosis, and hence before the identity of cases and controls is established. This feature requires a comprehensive, general purpose interview routinely administered to all admissions, which may restrict its use to publicly supported institutions diagnosing and treating neoplastic diseases or other specialized disease entities. Several epidemiological contributions for specific cancer sites have been based on the unique control data available from Roswell Park Memorial Institute (*9, 11, 12, 30, 40–43*), which are particularly valuable for collation with studies depending on more conventional sources of controls to evaluate interviewer bias and related issues.

Some patients interviewed as diagnosed cases will subsequently have their diagnoses changed. This may be turned to advantage. If scrutiny of the data for the erroneously diagnosed group reveals they had histories resembling those for the control rather than the case series, as Doll and Hill found in their study of smoking and lung cancer (*21*), this would constitute evidence against interviewer bias.

In investigations of a cancer site the association of a factor may often be restricted to a specific histologic type or a well-defined portion of an organ. The finding that epidermoid and undifferentiated pulmonary carcinoma is more strongly related to smoking history than adenocarcinoma of the lung is now well established. The range of explanations for the observed deficit of epidermoid carcinoma of the cervix in Jewish women as compared to other white women is greatly circumscribed by the presence of about equal numbers of adenocarcinoma of the corpus in both groups. When these finer diagnostic details or their significance are unknown to the interviewer, another check on interviewer bias is provided. Furthermore, the confirmation in repeated studies of an association limited to a specific histologic type or a detailed site will lend credence to an etiological interpretation of the association. Repeated confirmation is an essential element. Otherwise, a very specific association may be a reflection of the multiple comparison problem; if enough contrasts are created by fractionation of a single set of data, some apparently significant result is likely to appear. For this reason it would be desirable to reproduce such provocative results as Wynder's finding that use of alcohol was more strongly associated with cancer of the extrinsic larynx than of the intrinsic larynx (*18*), and Billington's report that prepyloric and cardiac neoplasms of the stomach were associated with blood group A and those located in the fundus with blood group O (*44*).

Discussion of matched controls in relation to the analysis and the computation of relative risks is deferred to a later section. One consideration on matched controls arising in the planning and development of a study should be mentioned here. Obviously, if the risk of disease changes with age an apparent association of the disease with other age-related factors may result. Other apparent associations with race, sex,

nativity, etc., may arise in a similar manner. In devising rules for selecting controls, those factors known or strongly suspected to be related to disease occurrence should be taken into account if unbiased and more precise tests of the significance of the factors under investigation are desired. A sensible rule is to match those factors, such as age and sex, the effect of which may be conceded in advance and for which strong evidence is available from other sources, such as mortality data and morbidity surveys. When a factor is matched, however, it is eliminated as an independent study variable; it can be used only as a control on other factors. This suggests caution in the amount of matching attempted. If the effect of a factor is in doubt, the preferable strategy will be not to match but to control it in the statistical analysis. While the logical absurdity of attempting to measure an effect for a factor controlled by matching must be obvious, it is surprising how often investigators must be restrained from attempting this.

When a minimum of matching is involved, the importance of establishing, precisely and in advance, the method by which controls are selected for study increases. The rule should be rigid and unambiguous to avoid creating effects by subconscious selection and manipulation of controls. The problem is similar to that encountered in therapeutic trials where a protocol spelling out all the contingencies and actions to be taken in advance is, along with random assignment of cases and controls, the major bulwark against bias.

To reduce interview time and expense there are advantages in procedures for selecting controls which permit a case and the corresponding controls to be interviewed in a single session, particularly if travel to several institutions is involved. In practice, this favors selecting controls from a hospital patient census rather than from hospital admission lists. The difficulty with hospital admissions is that there is no guarantee that the controls will be available in the hospital at the time the diagnosed case is interviewed. This point seems more important than the fact that patients with diagnoses requiring long-term stays are overrepresented in a current hospital census (*45*). If the latter is an important issue, it may be handled in analysis through subclassification of controls by diagnosis.

Normally there will be little difficulty in reconciling these considerations into a harmonious set of rules. The items to be matched often lend themselves to a procedure for specifying controls. In a recent study on female lung cancer we found that the definition of two controls as the next older and the next younger women in the same hospital service, present on the day the case was interviewed, met the requirements just outlined (*27*). The controls were uniquely defined, the records establishing their identity were readily available on the service floor, interviews could be completed in one day, and a provision for balancing ages of cases and controls was incorporated. Simultaneous interviews of cases and controls may be more than an administrative convenience. If the prevalence of the associated factor is rapidly shifting over time,