Different Views

Reporting the Results of Epidemiologic Studies

ALEXANDER M. WALKER, MD, DRPH

Introduction

Epidemiologic research seems regularly to yield data which resist simple encapsulation. Take epidemiology broadly to mean the study of the determinants of health in human populations: the most common hurdles in such work arise from the nonrandom allocation of exposures in the general population and of measurement error in the study subjects. In clinical trials, randomization anticipates and to some extent defines the analyses that are intended to follow, and it permits quantitative assessment of the distortions which may have been introduced by unknown or poorly measured causes of a condition under study. By contrast, when exposure may be a natural concomitant of the forces that produce and prevent disease, and when the quality of observation may be affected by those same powers, then both the practice and the reporting of research have to take on an emphasis that delves beneath the formalism of preplanned comparison.

Throughout the conception, design, and analysis of epidemiologic studies, there are points at which proper anticipation of the report that must follow can lead to a more lucid presentation. The reader will have to feel in the end that he has been able to disentangle the skeins of serendipity, bias, and intent that run throughout epidemiologic research. He will have to be given sufficient information to judge the quality of the data collection, and the pertinence of those data to the conditions and events that are the real objects of study. He must be guided through a tiny fraction of all the observable relations in the data and yet be convinced that he has grasped the essential information therein. The purpose of the observations that follow is to point out ways in which clarity in the report may be achievable through attention to the methods of research and to the underlying substantive issues in the execution and reporting of a study. These ideas represent a collation of what I believe to be successes in presentation that I have observed in other people's work and occasionally in my own.

Context

Begin a report by stating the relations that are to be addressed and the motivations for considering those relations to be of interest. The observations that stimulated the present work may be the product of laboratory efforts, but more often they will stem from case series, correlational studies, or previous formal epidemiologic analyses which suggest the presence of an exposure-disease relation of scientific or public health importance. Since the details of exposure and disease definition may vary across populations, it may be desirable to do little more than replicate an earlier design. In any case, give the testable implications of other studies as a simply stated, positive hypothesis. Previous findings which are not testable, that is to say refutable, in the work being reported have little relevance.

An important aspect of the study's context, quite different from its scientific antecedents, is its logistical setting. State whether the current work is based on *ad hoc* data collection, is part of a series of studies carried out in the same population, or is an offshoot of a larger study, a multipurpose study, or a surveillance system. Catch phrases to be used later in the text, such as "specially trained interviewers," will take on the coloring of their surroundings, so be sure to offer evidence, even circumstantial, on the extent to which the data emerging from your work can be expected to reflect the reality of the universe under observation. If well-known relations have been reproduced in the data, report as much concisely.

Study Subjects and the Source Population

Rates, fractions, and functions of these two elemental measures underlie all meaningful reports of epidemiologic findings. Each entails the definition of a source population within which the observations are made. The source population may be enumerated, or its definition may be only implicit in the choice of study subjects. In either case, it should be described in terms of those features that affect disease frequency and feasibility of data collection. Characteristics that are known prior to initiating the study may be described in the methods section, together with a specification that would in principle permit that any given person at any given point in time could be determined to be either in or out of the population under study.

Occasionally an attempt to describe the source population implied by a case selection procedure will highlight an underlying weakness of a study. The source population of cases admitted to a single hospital is a frequently cited example. Several hospitals may serve a single catchment area, and the propensity of patients to choose one or another facility may be related in an unquantified way to factors of direct or surrogate interest in an etiologic study, such as ethnic group or income. An investigator's inability to specify his source population in adequate descriptive or operational terms presages uncertainties through the remainder of design, analysis, and interpretation.

Address reprint requests to Alexander M. Walker, MD, DrPH, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. This paper, solicited by the Journal editor, was received April 29, 1985, revised, and accepted for publication December 17, 1985.

Editor's Note: Concerns have been voiced about the way results have been expressed in recent issues of the Journal. For that reason we solicited this paper from Dr. Alexander Walker, and invited Dr. Joseph Fleiss to present a different view which follows. Another exchange of views on pages 587–588 in the Letters to the Editor section of this issue of the Journal (see Fleiss, Foxman, and Editors' Note) may also help clarify the issues for readers. We welcome further comments.

Although most cohort designs and many case-control designs require that all members of the source population be identified and enumerated, in most case-control studies only a sample group representing the source population or population time at risk (the control series) is actually studied. In this situation, the sampling mechanism must be described in sufficient detail for the reader to understand its relation to the source population and to judge whether the selection procedure is likely to have produced an appropriate control series, one that accurately reflects the distribution of characteristics under study in the source population. When control selection is achieved through direct sampling of an operationally defined group, response frequencies and characteristics of nonrespondents are of interest. Sometimes control series are chosen in a multistep process that may be only partly under the investigator's control. A diseased control may be "selected" from the source population by his disease, then by the investigator from among all similarly diseased persons, then by himself in agreeing to participate in the study; each step should be examined for possible dependency on exposure status.

Whether controls are selected without prior stratification, after matching on broad criteria (e.g. age and sex), or on the narrowest ones (as in a neighborhood-matched study), a common principle holds and should be evident to the reader: within the sampling frame, controls give unbiased information about the population giving rise to the cases. A helpful way to describe the control series is by presenting the control selection process as homologous to the case selection process: each defines a source population. The two source populations must be identical with respect to determinants of exposure; the simplest practical device to ensure the identity of characteristics is to make the source populations for the two selection processes the same. If the investigator does not have a clear idea as to just which population gives rise to the cases, neither he nor the reader can be expected to judge the adequacy of the controls.

Data Collection

In describing data sources, provide detail as to who collects and provides the information, how the data are recorded, and the route by which the initial information reaches the form finally analyzed. Note any quality control procedures and methods for detecting obviously wrong or inconsistent responses. When the methods used are routine, be brief; when they are novel, be ample (and circumspect). When the detail of the available data puts prior limitations on the questions that can be asked, say so.

Results

Communicate the Substance of the Data

There has never been an important epidemiologic observation which could not be clearly presented in a few tables of raw data with simple summary statistics. Tables of cases and populations at risk or of case and control counts cross-classified by exposure status serve a double purpose of conveying both the substantive message of a set of observations and the uncertainty that may result from small numbers. Often a single 2 × 2 or 2 × k table captures a result, sometimes stratification by an important confounder is needed, seldom is anything more complex required.

Simplicity in data presentation does not mean that the analysis that leads to the selection of a few key tables should be obtuse. Factors which potentially confound or influence a result must be examined through stratification and appropriate calculation of summary measures, or through statistical modeling. Packages for statistical analysis of epidemiologic data are now so widely available that multivariate techniques that were once reserved for the last stage of analysis are now being used to sift through large numbers of potentially interacting and confounding terms. When this practice is followed responsibly, the analyst's monitoring of changes in parameter estimates, the covariance matrix, and goodness-of-fit measures replaces the scanning of tables to get a "feel" for the variability and interrelations in the data.

Whether the analyst's insight into the relations between the variables under study derives from the perusal of scores of tables or dozens of regression equations, he has an understanding of the data which cannot be fully communicated under the normal constraints of journal publication; he must accordingly choose the central themes to be presented. While a reader should understand the strategy employed to sort through the data, there is no reason for him precisely to relive the analyst's exploration. An increasingly common and useful practice is to present the simplest tables that capture an effect together with effect estimates based on the most comprehensive feasible analysis.

Certainty of the Estimates

P values can be useful when no direct estimate of effect is available or readily interpretable, as is sometimes the case with higher order terms in statistical models of rich data sets. P values should not, however, be presented in isolation or with a point estimate alone, much less in the degraded form of a statement such as "significant" or "not significant". Epidemiologists study and estimate the magnitude of biologic relations, and the dichotomizing effect on an uprooted report of significance is generally out of place. Confidence intervals provide estimates of the gamut of true relations consistent with a given set of observations. As such they may allow reconciliation of apparently divergent results, and they generally (since confidence intervals are almost always wider than one would wish) introduce an appropriate note of caution into the interpretation of "clear" findings.

Neither p values nor confidence intervals provide a full accounting of the uncertainty inherent in the analysis of epidemiologic results. The distinction between observational and experimental data in this respect is that the analyst substitutes a working hypothesis about the nature of unmeasured variables for the physical act of randomization. Both the confidence interval and the p value have simple operational definitions in the clinical trial, where the chance mechanism allocates determinants of outcome in a manner whose behavior is understood. In an observational study we hypothesize that unmeasured determinants are distributed between comparison groups as if by chance and we apply techniques proper to the analysis of truly probabilistic phenomena to assess the possible contribution of "chance" to a study's findings. The proposition that the unmeasured determinants are distributed in an arbitrary fashion, conditionally upon the measured factors, is not testable. Its plausibility should be reviewed in the discussion section of a report under the general heading of uncontrolled confounding.

Missing Data

Even after subjects have successfully participated in a study, certain items of information may remain missing. Respondents occasionally give uninformative answers to the most carefully posed questions; routine records are commonly incomplete. The frequency with which data are missing for

any reason is an important piece of information about the quality of a study and ought to be presented explicitly. A common assumption that permits the simple removal from analyses of cases with missing data (or occasionally the estimation of what the missing data would have been had they been available) is that the loss of data is an arbitrary event, unrelated to the true values of observed quantities. The proposition is sometimes patently false, as when a value is missing because it is out of the range of recordable characteristics, but will more often be subtly wrong as when a crucial variable is censored as a function of a predictor of risk. For example, histological verification of a difficult-to-diagnose tumor may be poor in the very old or unusually accurate in the affluent. In these cases an analysis of the relation between tumor type and any correlate of age or social class will be in error. A minimal safeguard is to present unknowns in every table, and to include "unknown" in multivariate procedures as a distinct category of risk or disease. While serious distortions cannot always be prevented, their presence may be signaled in associations between "unknown" status and disease or risk factors. If unknown responses are common, some consideration of their impact must appear, either formally in the analysis or informally in the discussion of results.

Multiple Hypotheses

Unanticipated results are common in studies in which large numbers of factors have been investigated. Within limits imposed by the subjects' ability to provide meaningful responses, the goal of extracting as much information as possible from costly interviews is worthwhile, but a number of problems present themselves, particularly when the number of cases is not large. The principal difficulties imposed by "too rich" data are multiple comparisons, subgroup analysis, and invalidation of control representativeness. Each of these demands special care in presentation.

As it is generally posed, the multiple comparisons problem concerns the expectation that tests of a large number of independent hypotheses will lead to statistically significant findings in the proportion of instances specified by the Type I error of the test, and a frequent recommendation is to deflate the size of the rejection region in compensation. The proposed remedy highlights an unfortunate aspect of dependence on p values, in that it leads to an inability to detect any effect as the required significance level drops toward zero with increasing numbers of hypotheses. More serious is that the suggestion often cannot be implemented in any consistent way: the number of independent hypotheses that could be tested in a set of richly interrelated observations may not be determinable from the data at hand, and those hypotheses that might reasonably be tested differ as a function of information external to the study. Should I discount an interesting finding because the investigator tested some hypotheses which I consider absurd? A preferable alternative is to present unanticipated findings and their unadjusted confidence intervals with an appropriate comment identifying the corresponding hypotheses as ones not entertained at the beginning of the study, and to test if possible further implications of the new hypotheses in the data at hand. The interpretation of unanticipated results depends heavily on external criteria of biologic plausibility and of consistency with other findings.

Subgroup analysis is a variant of the multiple comparisons problem in which a single hypothesis is multiplied by separate investigation in many subpopulations. Insistence on significant tests of heterogeneity of effect over subgroups as a prerequisite for subgroup analysis protects the analyst from distraction by spurious minor variation, but at a cost of almost total inability to recognize variability that has its roots in the populations under study. Except when strata are few and heavily populated, tests for heterogeneity have low power against many interesting alternatives. As a result, relevant observations external to the study may be crucial in deciding whether to take an observed subgroup effect seriously, and they should be presented along with the data.

Often control series which are not chosen by random sampling from a well-defined population are tailored to specific studies. Hospital controls for example might be selected from persons with diagnoses thought not to be associated with alcohol or tobacco consumption in a study that addresses the effects of those exposures. Such a series may provide valid estimates of alcohol and tobacco use, yet highly biased estimates of the prevalence of other habits related to diagnoses used to specify controls. One way to reduce risk of error in this situation is to choose control diagnoses by inclusion (rather than by exclusion) and to present exposure frequencies within control categories. In general, however, it is wise to limit exploratory case-control analyses to studies in which the process that generates controls has a small number of well defined and quantifiable steps in its remove from the general population, as may be the case when classical survey sampling methods are used to generate controls for population based case series.

Implications

Although the impetus for epidemiologic studies may come from many disciplines, and although the ramifications of an observation may similarly extend into many areas, it is rare that epidemiologic results themselves are sufficiently detailed to justify any lengthy discussion of proposed mechanisms of action. The discussion should place the results clearly in the context of other relevant epidemiologic work, drawing parallels where possible, and highlighting points of apparent conflict with the results of earlier studies. Findings inconsistent with previous hypotheses are more likely than confirmatory results to lead to new scientific insight, and divergences should be explored with as much care as the data merit.

ACKNOWLEDGMENTS

Many of the ideas presented here are the product of extensive and helpful comments from reviewers and from colleagues at the Harvard School of Public Health, the International Agency for Research on Cancer, and Epidemiology Resources, Inc. While preparing this work, I have been supported by the International Agency for Research on Cancer as a staff member and subsequently by a grant from the Mellon Foundation.