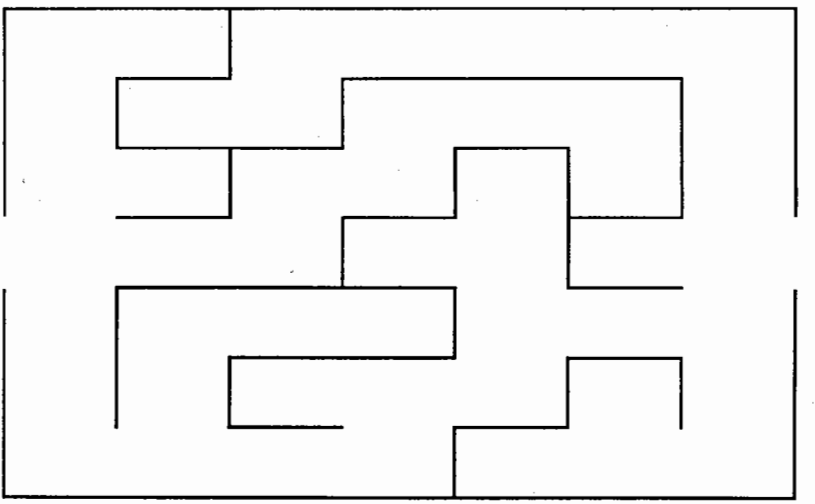
10

Reporting Results



Throughout the conception, design, and analysis of epidemiologic studies, there are points at which anticipating the report that must follow can lead to a more lucid final product. The reader will have to feel in the end that he has been able to disentangle truth from serendipity and bias. 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. The purpose of this chapter is to lay out guidelines for communicating results to a reader.

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 that suggest an exposure-disease relation. Since the details of exposure and disease definition 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 that 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, for 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, proportions, and functions of these 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 permit any given person's inclusion in the source population at any given in time to be determined.

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 to factors under study, such as ethnic group or income. An investigator's inability to specify his source population in operational terms presages uncertainties throughout design, analysis, and interpretation.

In case-control studies only a sample of the source population (the control series) is actually studied, and its characteristics are taken to represent those of the source population or population time at risk. The sampling mechanism must be described in sufficient detail for the reader to judge whether the selection procedure is likely to have produced a control series that reflects the distribution of characteristics under study in the source population. 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.

However controls are selected, one guiding principle should be evident: within the sampling frame, controls provide 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 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 that could not be clearly presented in a few tables of raw data. 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 2x2 or 2x*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 analysis 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 widely available; multivariate techniques that were once reserved for the last stage of analysis are 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 replace the scanning of tables to get a "feel" for the variability and interrelations in the data.

Whether the analyst's insight 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. Confidence intervals provide estimates of a gamut of relations consistent with a given set of observations. They may allow reconciliation of 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. 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. For the most part, p values should not be presented in isolation or with a point estimate alone, much less in the degraded form of a statement such as "significant" or "NS." Epidemiologists study and estimate the magnitude of biologic relations, and the dichotomizing effect of an uprooted report of significance is generally out of place.

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 clinical trials, where the chance mechanism allocates exposure. In an observational study, we hypothesize that measured exposures are distributed 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 exposures are distributed in a random fashion, conditionally upon other 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 subjects 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 a random 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 in the analysis or in the discussion of results.

Multiple hypotheses. Unanticipated results are common when 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 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.

The multiple comparisons problem concerns the expectation that tests of a large number of independent hypotheses will lead predictably to statistically significant findings for some of the tests, even under a null hypothesis. When many tests are done at the same time, and when any one of them being statistically significant would constitute an important finding, a frequent recommendation is to decrease the size of the p value required to declare any single finding significant. The extent of the decrease is a function of the number of comparisons being made. Fewer findings are declared significant by the more stringent criterion, but 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. 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 that 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 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. It follows that proper interpretation of a study will change with time; it will continue to evolve long after publication, as new understanding is brought to bear on old data.

Subgroup analysis is a variant of the multiple comparisons problem in which a single hypothesis is multiplied by separate investigation in many subpopulations. Except when strata are few and heavily populated, tests for heterogeneity have low power against many interesting alternatives. Insisting on significant tests of heterogeneity to justify subgroup analysis protects the analyst from being distracted by the inconsequential; the cost is an almost total inability to recognize true variation. Relevant observations external to the study are usually crucial to the decision to take an observed subgroup effect seriously.

Often, control series that are not chosen by random sampling from a well-defined population are tailored to specific studies. Hospital controls 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. The simplest case holds when classical survey sampling methods generate the controls for population-based cases.

Implications

The impetus for epidemiologic studies may come from many disciplines, and the implications of an observation may be broad. Nonetheless, it is rare that epidemiologic results themselves justify lengthy mechanistic discussion. The commentary should place the

results in the context of relevant work, drawing parallels where possible, and highlighting points of conflict. Findings inconsistent with previous hypotheses are more likely than confirmatory results to lead to new insight, and divergences should be explored with as much care as the data permit.