Articles

Scientific Standards in Epidemiologic Studies of the Menace of Daily Life

ALVAN R. FEINSTEIN

Many substances used in daily life, such as coffee, alcohol, and pharmaceutical treatment for hypertension, have been accused of "menace" in causing cancer or other major diseases. Although some of the accusations have subsequently been refuted or withdrawn, they have usually been based on statistical associations in epidemiologic studies that could not be done with the customary experimental methods of science. With these epidemiologic methods, however, the fundamental scientific standards used to specify hypotheses and groups, get high-quality data, analyze attributable actions, and avoid detection bias may also be omitted. Despite peer-review approval, the current methods need substantial improvement to produce trustworthy scientific evidence.

THE EPISODES HAVE NOW DEVELOPED A FAMILIAR PATTERN. A report appears in a prominent medical journal; the conclusions receive wide publicity by newspapers, television, and other media; and another common entity of daily life becomes "indicted" as a "menace" to health—possibly causing strokes, heart attacks, birth defects, cancer. Sometimes the accused menace is a nonmedical substance—coffee, water, sugar, saccharin, alcohol—that people eat or drink on their own. Sometimes it is a pharmaceutical agent prescribed for such common phenomena as high blood pressure.

Regardless of whether the alleged menace is a medical or nonmedical entity, however, the reported evidence is almost always a statistical analysis of epidemiologic data, and the scientific tactics that produced the evidence are almost always difficult to understand and evaluate. Most people learn about science by studying the experimental methods of physics, chemistry, botany, or biology; but experiments are seldom possible in epidemiologic research. Because of barriers in ethics or feasibility, the investigators cannot do experiments in which healthy people are randomly assigned to receive or not receive long-term exposure to potentially noxious substances.

In the epidemiologic substitutes for experiments, the research methods seldom have the precautions, calibrations, and relative simplicity that are taken for granted in other branches of science. Groups of free-living people cannot be assembled and studied as easily as captive animals or inanimate material; data about nutrition, medical exposures, and life style are difficult to check for scientific quality; and the results often receive statistical analysis with methods that are unfamiliar and sometimes inscrutable.

Despite these problems, nonexperimental epidemiologic research has led to some outstanding health accomplishments, many of them in the field of infectious disease. During the period from about 1850 to World War II, epidemiologic studies—leading to sanitary methods for disposing sewage, purifying water, and preparing foods—had a more profound impact on public health, infectious disease, and individual longevity than any other contemporary advances in medical science.

Since World War II, infectious disease epidemiologists, using high-quality scientific methods to identify microbial agents, have had such dramatic successes as preventing poliomyelitis, eradicating smallpox, and demonstrating that rubella in early pregnancy can cause birth defects. In noninfectious disease, epidemiologic research has shown that a dietary deficiency leads to pellagra, the association between cigarette smoking and lung cancer, the protective dental effect of fluoridated water, and the role of thalidomide in phocomelia.

These splendid achievements, however, have also been accompanied by major uncertainties and controversies in other epidemiologic studies, particularly for cause-effect relationships in noninfectious disease. In a recent survey (*1*) of the problems, 56 different cause-effect relationships had conflicting evidence in which the results of at least one epidemiologic study were contradicted by the results of another. About 40 more conflicting relationships would have been added if the review had included studies of disputed associations between individual sexual hormones and individual birth defects.

None of these conflicting studies was done as an experiment, although cause-effect relationships have often been investigated during the past 40 years with the human experiments that are called randomized controlled clinical trials. These experiments, which have become widely employed, well accepted, and generally regarded as the "gold standard" of cause-effect evaluation, have mainly been applied, however, to appraise the short-term benefits of pharmaceutical treatment. Randomized trials have not been generally feasible or ethical for evaluating long-term risks of therapy, or for testing whether such public health agents as smoking, alcohol, nutrition, and occupational hazards have noxious effects in causing disease.

In substituting for randomized trials, many epidemiologic methods have had an all-or-none scientific approach. If the causal agents were assigned with randomization, the research would be done with all of the scientific principles that accompany an experiment. If randomization could not be used, however, the other scientific principles have also not been used. My purpose in this article is to indicate that the current problems arise because those other scientific principles, although needed and applicable, have not received

The author is professor of medicine and epidemiology and director of the Clinical Epidemiology Unit, Yale University School of Medicine, New Haven, CT 06510, and senior biostatistician, Cooperative Studies Program Coordinating Center, Veterans Administration Medical Center, West Haven, CT 06516.

adequate attention in epidemiologic studies of the "menace" of daily life.

I begin with an outline of the basic scientific principles employed experimentally in a randomized trial. The outline is followed by a brief description of epidemiologic substitutes for randomized trials. The main discussion thereafter is devoted to the application (or omission) of the cited scientific standards in three prominent epidemiologic accusations about "menacing" exposures in daily life. The accusations were that cause-effect relationships existed for reserpine with breast cancer, coffee with pancreatic cancer, and alcohol with breast cancer.

Scientific Standards in Human Experimental Research

When a therapeutic agent is suspected of having a causal action, the outcome can be a benefit, such as relieving pain or retarding death, or an adverse effect, such as an abnormal reaction in blood or tissue. For an etiologic agent, such as a dietary pattern or an occupational exposure, the outcome is the development of a particular disease in a healthy person.

In the ordinary activities of daily life or medical practice, the compared agents are selected by personal choices of individual people or their physicians. The choices often produce susceptibility bias (*2*) when the outcomes of the selected agents are later compared in the groups of recipients. A particularly striking clinical example of susceptibility bias occurs if routinely selected surgical and nonsurgical treatments are compared in patients with cancer. The comparison is biased because surgery is usually reserved for the relatively healthy "operable" patients, who have much better prognostic susceptibility to a favorable outcome (even if the proposed surgery is not done) than the relatively unhealthy "inoperable" patients who are relegated to receive nonsurgical therapy. In public health activities, a prominent example of susceptibility bias is the paradox called the "healthy worker effect." The preemployment criteria for certain hazardous occupations make the employees so relatively healthy that they have lower mortality rates than the rest of the population, despite the potential dangers of the occupational exposure.

The most obvious experimental scientific contribution of randomization is its role in helping avoid susceptibility bias. As the scientific counterpart of tossing a coin, random assignment of treatment will make the compared groups have similar prognostic susceptibility, except for chance inequities occurring during "the luck of the draw." To avoid the susceptibility bias produced by selective personal decisions, the Food and Drug Administration during the past 20 years has regularly required randomized assignment of therapy for any causal claims of beneficial effects.

In studies of deleterious effects, however, the suspected agents can seldom be assigned in a randomized experiment. When nonexperimental methods are substituted in the research, the baseline state of the compared groups is often statistically "adjusted" for possible prognostic imbalances (*3*). The statistical tactics often employ the unfamiliar analytic procedures—demographic "matching," "confounding scores," multi-categorical adjustments and stratifications, multivariable regressions—that make the results so difficult to understand and interpret.

If all of the appropriate prognostic or "risk" factors have been included, the statistical efforts will often succeed. For example, the adjustments often seem to work well in etiologic studies of cancer, where the adjusted data usually include the relatively few factors that are known or believed to affect biologic susceptibility to cancer. In etiologic studies of coronary heart disease, however, the adjustments are usually less successful because they seldom include suitable attention to such cogent susceptibility features as family longevity and personality traits. In studies of clinical therapy, the adjustments are often inadequate. Such prognostically important factors as the clinical and co-morbid severity of illness are regularly omitted because the data are deemed too "soft" and subjective (*2, 3*).

For the particular epidemiologic studies under review here, the investigators used conventional statistical adjustments for susceptibility bias. Those procedures will not receive further discussion because their impact is much less important than problems in the five other basic scientific standards. The five other standards precede or follow the use of randomization when agents and outcomes are examined experimentally. The standards are: (i) a stipulated research hypothesis, (ii) a well-specified cohort, (iii) high-quality data, (iv) analysis of attributable actions, and (v) avoidance of detection bias.

All of these standards are attained so readily and used so routinely in laboratory experiments that a nonmedical scientist may not even think about them as basic principles of scientific methods and may not recognize their frequent absence in epidemiologic research.

Methods Used in Epidemiologic Studies

The epidemiologic substitutes that replace randomized trials have a wide variety of methods and names. They include cross-sectional community surveys, prospective cohorts, retrospective cohorts, convenience cohorts, retrospective case-control studies, ecologic association studies, and many other statistical arrangements of groups and data. Because the methods and relative merits of these different epidemiologic structures have been discussed elsewhere (*2, 4, 5*), the rest of these comments are devoted to two commonly used techniques—the retrospective case-control study and the convenience cohort study—that were the sources of indictment for the three prominent "menaces" under discussion.

In a retrospective case-control study, the customary forward direction of scientific observation is reversed. The investigator begins at the end of the causal pathway, after everything has already occurred. Two groups are assembled, a "case" group, containing people known to have the target disease, and a "control" group, containing people in whom that disease has not been demonstrated. The control group, chosen arbitrarily, may be healthy or may have other diseases. Members of both groups are then checked (by personal interviews, telephone calls, mailed questionnaires, or reviews of appropriate previous records) to determine (or "ascertain") each person's previous exposure to the suspected causal agent.

In any cohort study, the investigator observes the exposed and nonexposed groups in a forward direction, following them from imposition (or nonimposition) of the suspected causal agent toward subsequent occurrence (or nonoccurrence) of the disease regarded as the outcome effect. In a convenience cohort, however, the observed groups were originally collected without deliberate, specific plans to investigate the hypothesis that becomes tested in the research. The cohort is available because it was assembled in a general manner from people who sent in responses to mailed questionnaires, or who were examined in research performed for other purposes. These people are seldom actually followed and reexamined at regular intervals thereafter. Instead, their outcome status is usually determined from their responses to subsequently mailed questionnaires.

With either the case-control or cohort structure, a causal suspicion is supported if an impressive statistical association appears in the 2 by 2 tabulation for subgroups of people reported as being exposed or nonexposed, diseased or nondiseased.

Of the three menacing relationships under discussion here, the first was reported about 14 years ago. Three retrospective case-

control studies (*6–8*), published in the same issue of the same journal, all found a positive association between breast cancer and reserpine, a medication that was then widely prescribed for treating hypertension. The risk of breast cancer for reserpine users was estimated as 2.0 to 3.9 times higher than in nonusers. In those three studies, the cases consisted, respectively, of 150, 438, and 708 women reported to have breast cancer. The corresponding control group in the first study (*6*) contained 1200 patients hospitalized with conditions that excluded "cancer or any form of cardiovascular disease" as a "first discharge diagnosis." In the second study (*7*), the controls were 438 patients "admitted for elective surgery" of conditions that excluded cancer, gall bladder disease, thyrotoxicosis, renal disease, or cardiovascular disease. In the third study (*8*), the control group originally had 1430 members with any other form of cancer that had been reported to a tumor registry; but when the results were not statistically satisfactory, certain cancers were removed from the original control group. The main conclusions were then drawn from the original cases and the reduced group of 963 controls.

The second menacing relationship was reported (*9*) in a case-control study in 1981, when the risk of pancreatic cancer was estimated to be about 2.5 times as high in people who drank coffee as in those who did not. The investigators estimated further that "slightly more than 50 percent" of pancreatic cancer was "potentially attributable to coffee consumption". The case group contained 369 patients with pancreatic cancer, and the control group had 644 patients who were under the care of the same physicians who had treated the cases.

The third "menace" was found as an association between alcohol and breast cancer in two convenience-cohort studies, reported in the same issue of the same journal in 1987. One study (*10*) contained a convenience cohort of about 122,000 "female, married, registered nurses, aged 30 to 55, who were living in 1 of 11 large U.S. states" and who had responded to mailed questionnaires in 1976. In subsequent questionnaires, information about "diet" was obtained in 1980; and about breast cancer in 1982 and 1984. In the 89,538 nurses who returned follow-up questionnaires "every two years", 601 breast cancers were reported. The other convenience cohort (*11*) was derived from the residue of a group of women who were originally willing, 10 years earlier between 1971 and 1975, to be interviewed for a National Health and Nutrition Evaluation Survey (NHANES). These surveys are periodically conducted, by the National Center for Health Statistics, to obtain cross-sectional data for descriptive reports of various health-related attributes in randomly selected members of the community (*12*). In this instance, 121 breast cancers were noted when 7188 (84%) of 8596 originally interviewed women were traced a decade later, interviewed again between 1981 and 1984, and found eligible for inclusion in the alcohol–breast cancer analyses. The two cohorts had substantially different occurrence rates of breast cancer: about 6.7 per thousand (601/89,538) in the nurses cohort and about 18.2 per thousand (131/7188) in the NHANES cohort. The relative risks of breast cancer in reported alcohol drinkers and nondrinkers, however, were similar: about 1.5 in both cohorts.

All of these highly publicized accusations of "menace" came from research that had been approved by the "peer review" of authoritative experts. The peer review process, however, provides assurance only that an act of research complies with accepted methods in a field of investigation. The process provides no assurance about the methods themselves, particularly if the reviewing experts also establish and maintain the very methods that they are asked to approve.

The rest of this essay is concerned with the research methods that revealed the "menacing" relationships and with the application or omission of the five basic scientific standards in those methods.

Application of Scientific Standards

Each scientific standard will be briefly described and then discussed for its application in the cited epidemiologic studies.

1) *A stipulated research hypothesis.* To plan an experimental trial, the investigator identifies the cause-effect comparison that will be tested as the research hypothesis. It may be the belief that treatment A relieves pain more effectively than a placebo, or that people exposed to agent B develop a particular disease more often than unexposed people.

Although an obvious activity in other branches of science, hypotheses are not always specified before an epidemiologic study begins. Instead, the hypotheses may be "generated" retrospectively, after the research data have been analyzed. This retrospective process seldom occurs for the concise information that is usually collected in laboratory experiments. In many epidemiologic studies, however, vast amounts of diverse information can be assembled. It can include demographic data (age, race, sex, socioeconomic status), data about individual agents (diet, smoking, alcohol, environmental exposures, pharmaceutical substances, other treatments), and data about individual outcomes (birth defects, stroke, heart disease, cancer, death).

With modern electronic computation, all this information is readily explored in an activity sometimes called "data dredging." A large number of statistical associations are explored in an automated manner for diverse individual groups, agents, and outcomes. The groups can consist of all the people under study, or demographic divisions of multiple subgroups having one, two, or more than two separating characteristics (such as men and women, old men and young women, or old poor black men and young rich white women). Within each group or subgroup, each of the multiple individual agents is statistically associated with each of the multiple individual outcomes. Whenever a "statistically significant" result emerges during the myriads of computations, the event may be proposed as a cause-effect relationship.

An advance hypothesis has the scientific virtue of avoiding these computer-generated conjectures. It also avoids the problem of choosing appropriate statistical adjustments for the many "significant" relationships that will emerge by chance alone when the multiple associations are tested (*2*).

In the first of the three case-control studies of reserpine and breast cancer (*6*), no advance hypotheses were stated. The association with reserpine appeared when tens of thousands of statistical relationships were checked in a computerized exploration of hospital data for multiple antecedent exposures and multiple subsequent diseases (*13*). The second and third studies (*7, 8*) were then instigated to confirm the hypothesis.

In the case-control study of pancreatic cancer (*9*), the originally suspected etiologic agents were cigarette smoking and alcohol. When these suspects did not yield a positive result, available data for many other agents were explored statistically. The total number of examined agents was not reported, but the association with coffee emerged from the exploratory process.

In the convenience-cohort studies, the positive association between alcohol and breast cancer was found in explorations conducted after the basic research data were assembled. The total number of tested associations has not been stated, but results of the nurses' cohort data have already been published for explorations of at least 11 other cause-effect topics (*14–24*).

Thus, all three of the menacing relationships were noted not from previously stipulated research hypotheses, but from statistically significant results in computerized multiple explorations.

2) *A well-specified cohort.* In randomized trials, the cohort under study is well specified by examinations done before the exposure (or

nonexposure) begins, and in subsequent follow-up to see whether the selected outcome event has occurred. The process has two prime virtues. Scientifically, each admitted person is checked for suitable eligibility for the study, and statistically, each person is accounted for thereafter.

In the nurses and NHANES studies, the baseline conditions of the admitted persons were identified mainly from their individual responses, not from any direct medical examinations. Consequently, the investigators had no assurance that all members of the cohorts were initially free of breast cancer. The hazards of this process were later noted when outcome data were being sought in the NHANES cohort: 12 women were discovered (*11*) to have already had breast cancers that were not mentioned in the initial interview 10 years earlier. In the nurses cohort, the published report had no comment about the problems of verifying baseline status.

Because a case-control study begins at the end of the causal pathway, a demarcated cohort is not assembled and checked before exposure (or nonexposure) to the compared maneuvers. Vigorous retrospective efforts are needed, but are seldom used, to determine that each case and control person was appropriately eligible for the study, and to avoid referral bias, exclusion bias, and other distorted depletions or augmentations of what would have been a suitable cohort (*2*).

The basic scientific principle of a well-specified cohort was not maintained in any of the studies under discussion.

3) *High-quality data.* While admitting and following the individual people studied in an experiment, the investigators can get relatively high-quality data because each person is directly examined with methods that can be carefully calibrated for their reproducibility and validity. This process prevents the errors and uncertainties that arise when the basic information comes from second-hand sources, such as health survey household interviews or death certificates, or from mailed questionnaires submitted by respondents whose reliability is not directly checked. A direct examination process, before the agents are imposed and while they are in progress, also helps prevent two major problems in identifying the agents and outcome events. For identifying agents, the ongoing examination process will avoid the difficulties and biases of retrospective memory (*2*) when members of a case-control study are asked, long after the outcome events have occurred, to recall exposures that may have taken place many years previously. For identifying development of the target disease, the ongoing monitoring and repeated examinations can help avoid both the "false positive" errors of diagnosing a disease when it has not occurred, and the "false negative" errors of failing to detect the disease when it is present (*2*).

In the convenience-cohort studies, data about alcohol—a substance whose recorded intake is notoriously inaccurate—depended in the nurses study (*10*) on answers to questions about the average frequency of usage of beer, wine, or liquor "over the past year." The investigators tried to "validate" the single-question reports of alcohol intake by asking a small group of participants to record additional data in special "diaries." No external sources—such as spouses or friends—were asked to confirm the reports in the original questionnaires or in the "diary" data. In the NHANES cohort (*11*), the interviewed women were asked about the daily, weekly, or other frequency of having at least one drink of beer, wine, or liquor during the previous year. Women who reported at least one drink during that year were asked to specify the amount usually consumed in 24 hours. No studies of reproducibility or validity were reported for the interview responses. Regardless of whether or how the data were checked, however, the single responses for intake in both cohorts were analyzed as representing long-term alcohol patterns, with no provision for changes in drinking habits before or after the initially recorded quantities.

In case-control studies, the information about antecedent exposure is usually obtained after the interviewed person's diseased or nondiseased status has been identified. Despite the subjective retrospective impact of knowledge of that status, few or no attempts may be made to "blind" the interviewers appropriately, to use equal efforts in prodding the distant memory of the interviewed case and control groups, or to establish special additional control groups aimed at the problem of "recall bias" (*2*). None of these (or other) precautions for avoiding bias in ascertaining the antecedent exposure were reported for the case-control studies of coffee and pancreatic cancer or for two (*7, 8*) of the three studies of reserpine and breast cancer. [In the other study (*6*), information about antecedent reserpine usage had been obtained routinely when all patients were admitted to the hospital, before any case or control groups were identified.]

In the two convenience cohorts, none of the women under study received any direct medical examinations initiated by the investigators. The data about breast cancer were obtained from responses (sometimes by next of kin) to a questionnaire, or from efforts to find death certificates for cohort members who died or who had not subsequently responded. Whenever breast cancer was reported, the investigators checked for false positive diagnoses by reviewing the available medical records and other appropriate evidence. No attempts were made, however, to check for false negative diagnoses by examining the patients or their medical records, or by determining whether their breasts had indeed been appropriately examined in search of cancer.

In case-control studies, the clinical diagnosis of the outcome disease in the cases is usually accepted as stated, but the investigators may sometimes check for false positive errors by reviewing the available diagnostic evidence. In the control group, which is chosen because the target disease was not diagnosed, evidence of the disease's absence is almost never verified. Even if the investigators wanted to check for false negative diagnoses, however, a proper review is often impossible because members of the control group may not have received the appropriate diagnostic tests (*2*).

Thus, in both the convenience cohort and the case-control studies, the investigators often sought "false positive" errors by trying to confirm reported diagnoses of the target disease, but seldom checked for the vital counterpart error of "false negative" diagnoses.

4) *Analysis of attributable actions.* An ideal experimental design should allow an observed agent to be held responsible for the outcomes that follow it, but few human agents are received in isolation, and many are maintained in an erratic manner with frequent changes in schedule. Beyond the main agents under study, people can regularly be "contaminated" by exposure to other pharmaceutical agents, as well as to the other smoking, dietary, and occupational "risks" of daily life. Even in a randomized trial, the people assigned to a particular agent may refuse to take it, exchange it for the comparison agent, or supplement it in diverse unauthorized ways.

In a randomized trial, the investigators can plan to get suitable data for analyzing or "adjusting" the contamination problem. In nonexperimental studies, however, the main agents themselves may be difficult to identify reliably, let alone the external sources of contamination. If the investigated people were not examined or followed directly, the agents may be identified merely from personal responses at a single point in time, with no information about intervening changes thereafter.

These uncertainties create two substantial problems in epidemiologic analyses of attributable actions. The first problem is to choose the amount of exposure required to classify someone as "exposed." To credit or blame agent X for outcome Y, how much of agent X

should have been received and for how long? In a randomized trial, the dose and duration of exposure are defined beforehand, as part of the experimental plans. In many epidemiologic studies, however, exposure is defined only after the data have been collected and analyzed. McDonald *et al.* (*25*) have shown how arbitrary changes in these definitions, before or during the analysis of data, can make the relative risk of a particular agent range from 0.9 to 5.1 for the selected outcome event.

A separate problem occurs when the investigators use a "dose-response" analysis to support the idea of causality. In a "confirmatory dose-response curve," the occurrence rates of the outcome event progress in a rising monotonic pattern as the amount of exposure progressively increases in dose, duration, or both. Aside from all the difficulties with contaminating agents and changing exposure over time, the interpretation of dose-response data requires a judgmental decision about whether the pattern indeed shows a progressive increase.

None of the three case-control studies (*6–8*) of the reserpine–breast cancer relation specified the amount of reserpine required for "exposure," and none reported documentary data for a dose-duration-response relation. In the coffee–pancreatic cancer study (*9*), exposure to coffee was not defined, but the main results were presented as a dose-response curve. People who drank no coffee were arbitrarily assigned a risk of 1; and the relative risks (actually, odds ratios) of pancreatic cancer were calculated at three levels of coffee drinking: one to two, three to four, and five or more cups per day. A distinctively monotonic dose-response curve was not found in either men or women. For men, the respective relative risks at the three levels of coffee drinking were 2.6, 2.3, and 2.6. For women, the corresponding values were 1.6, 3.3, and 3.1.

For the alcohol–breast cancer relation, "the adjusted relative-risk estimate" in the nurses cohort was set at 1.0 in the nondrinking group. The corresponding successive values were 1.0, 0.8, 1.3, 1.6, and 1.6 as alcohol intake rose progressively from 1.5 to more than 25 grams per day. In the NHANES cohort, the relative risk was set at 1.0 in the "none" group, and had values of 1.4, 1.5, and 1.6 as levels of alcohol rose upward in three categories to more than 5 grams per day. The relative flatness of these patterns indicates that neither cohort had the monotonically increasing or dramatically escalating rises of a true dose-response curve. Nevertheless, the investigators stated that the NHANES pattern was "compatible with a moderate dose-response relation," and that the nurses pattern had a "dose relation (that) lends further credence to a causal interpretation."

5) *Avoidance of detection bias.* The double-blinding process that keeps both investigators and recipients unaware of the assigned maneuvers has several important roles in a randomized trial (*2*). The avoidance of detection bias is essential if the outcome event is relief of pain or other symptoms whose subjective perception and reporting can be substantially altered when a placebo rather than "active" agent is knowingly received by a patient or prescribed by a physician.

In nonexperimental studies where the outcome event is the development of a disease, rather than a change in symptoms, a different challenge occurs in diagnostic detection. Many diseases, such as cancer, coronary disease, and other major ailments, are regularly first found at postmortem necropsy examination (*26–28*), having been undiagnosed while the patient was alive. The previously undiagnosed diseases were rarely fatal, and usually occurred as coexisting "silent" phenomena that escaped detection during life because they had not produced the overt manifestations that might evoke the appropriate diagnostic procedures in clinical or technologic examinations. In search of these silent diseases, many "screening" examinations are now done in public health or clinical practice. For example, silent breast cancers will regularly be found when women receive a screening mammography that was evoked by a public campaign, by their own solicitation, or during regular medical surveillance for treatment of hypertension or some other clinical condition.

The existence of these silent cases of disease constitutes a formidable difficulty in epidemiologic research because any therapeutic or etiologic agent that is associated with increased "medicalization" and increased use of diagnostic technology will also be associated with an increased detection of the silent cases. Since these cases will be overlooked in people who do not receive the same diagnostic attention, the apparent increase in occurrence of the diseases may then be erroneously attributed to the agents, rather than to the detection process (*2*).

To avoid the problem of detection bias, the research methods should offer assurance that the disease was sought with equally intense methods of surveillance and examination in the exposed and nonexposed groups. Although both of the outcome diseases under discussion here can be silent and undetected during life, pancreatic cancer has no simple screening tests; its diagnosis requires surgery or complex imaging technology. Silent breast cancer, however, is particularly easy to find if sought. The most simple routine screening procedure is to palpate the breast, a process often done by physicians and now often by many women themselves. An additional highly effective screening procedure is mammography, now widely publicized as desirable, which many women have begun to seek routinely.

In both of the outcome diseases, detection bias would arise if the exposed persons sought or received more screening and other diagnostic procedures than the nonexposed persons. Since coffee drinking provokes no pertinent symptoms, it would not be expected to produce an increased diagnostic search for pancreatic cancer. Reserpine treatment of hypertension, however, is prescribed by a physician, and the treatment would be accompanied by an increased medical surveillance that would raise the opportunity for finding silent breast cancers.

The apparent association of alcohol and breast cancer could easily be explained if women who drink in moderate "social" quantities are also more likely than abstainers to maintain a medical "life style" that brings routine palpation of the breast and mammography. Many studies of breast cancer have shown that it is more commonly found in women of higher socioeconomic status, where social drinking and routine screening examinations of the breast are also more common. Furthermore, women who drink heavily may develop alcohol-related illnesses that also bring increased medical attention and the opportunity to detect hitherto undiagnosed breast cancers. If these features of the increased detection process are ignored, the associated increase in breast cancer will be fallaciously attributed to the alcohol.

Despite these possibilities, detection bias was not considered in the basic plans for any of the cited studies. In the case-control studies of breast cancer (*6–8*), no effort was made to analyze the medical detection process in the compared case and control groups, or to choose an additional control group from patients with negative mammograms. For breast cancer as an outcome event in the NHANES and nurses cohorts, the interviews and questionnaires contained no attention to the frequency or intensity of the routine examination process for breast cancer.

In the nurses cohort, the investigators perceived that detection bias might occur, but no additional questionnaires were sent to get the data needed for checking this possibility. Instead, the investigators tried to exclude it by using other information that was conveniently available. They contended (*10*) that detection bias was unlikely because "the four-year follow-up rate was similar" for

persons at each level of alcohol intake and because similar percentages of positive lymph nodes were identified in the cases of breast cancer reported among drinkers and nondrinkers. Neither of these contentions is pertinent, however, for the problem of detection bias. The follow-up "rates" for returning questionnaires do not demonstrate the intensity or frequency of the antecedent medical examination process; and the occurrence of silent lymph node metastases does not indicate whether drinkers and nondrinkers were similarly examined before the breasts and lymph nodes were removed.

Additional Comments and Discussion

In two of the three cited topics, the status of the proposed cause-effect relation has been resolved. Despite the original support of three simultaneously published case-control studies, the reserpine–breast cancer association has now been discredited by the contradictions found in many subsequent case-control and cohort studies (*29, 30*). In a retrospective attempt to explain the error, one of the original investigators (*30*) said that the first reserpine–breast cancer association (*6*) was probably a "statistically significant" artifact of the multiple calculations done during the data dredging. This explanation, however, does not account for the erroneous "confirmation" obtained when the hypothesis was tested in the two subsequent studies. In one study (*8*), the error was probably produced when the original control group was altered. In the second study (*7*), the previous exposure to reserpine may have received biased ascertainment, and its proportionate usage in the control group was biased downward by the selective exclusion of patients with conditions for which reserpine might have been prescribed (*2, 31*).

The coffee–pancreatic cancer relation was refuted by several other studies and particularly when the same group of investigators did a second case-control study of the same topic at the same hospitals used in the original study. The striking contradictory results, reported in a letter to the editor (*32*), showed no relative risks, at any level of coffee drinking, that were significantly elevated above 1. Without reconciling the disparate results in the two studies or acknowledging any errors in either, the investigators concluded that if a risk existed, "it is not as strong as our earlier data suggested."

The two convenience-cohort investigations of the alcohol–breast cancer relation offer the most recent prominent suggestion about the menace of daily life. The accompanying editorial (*33*) made no comment about the hazard of detection bias, the lack of a true dose-response curve, the disparate occurrence rates of breast cancer in the two cohorts, or the absence of a plausible mechanism by which alcohol might cause breast cancer. In the editorial comments, several previous conflicting epidemiologic studies were dismissed as methodologically inadequate. The contradictory results of a large case-control study (*34*), conducted by the Centers for Disease Control, were regarded as "aberrant" and "difficult to explain." In subsequent research, a positive alcohol–breast cancer relation was not found when the original CDC study was extended (*35*) or when evidence was reviewed from two other cohort studies (*36, 37*).

In other branches of science, substantial distress would be evoked by conflicting results in different studies of the three relationships discussed here, and in the 56 other disputed associations that have been cited elsewhere (*1*). Authorities would clamor for special conferences or workshops intended to identify the methodologic defects and to institute suitable repairs. No such clamor and no such workshops have occurred, despite these conflicts and despite a prominent leader's public denunciation (*38*), 9 years ago, of the frequently poor basic scientific quality of epidemiologic data.

This apparent complacency about fundamental methodologic flaws is not a recent development. At least two outstanding problems in epidemiologic methods have been neglected for about 40 years. In 1943 a prominent American biostatistician (*39*) described a profound methodologic defect, now often called "Berkson's Bias," that threatens the validity of any case-control study done with hospitalized patients. Berkson's theoretical suggestion, however, was not accompanied by specific evidence, because his work at the Mayo Clinic did not give him access to the community data that might confirm or refute his proposal about bias in patterns with which different diseases are referred for hospitalization. The investigators who did have access to such information, however, did nothing to check Berkson's contention. More than 30 years later, it was finally tested and confirmed in several studies (*40–42*), but the confirmation has had no apparent effects on the methodologic status quo.

Another long-standing epidemiologic problem is the reliance on death certificates for information that is used not only in convenience-cohort studies, but also for statistical tabulations of the occurrence rates of individual diseases (*2, 43*). This information has two fundamental flaws. First, as shown in many studies of the accuracy of death certificates (*44*), the individual diseases listed on those certificates are often identified incorrectly or inadequately. Second, the general occurrence rates derived for individual diseases are much too low, because the rates depend on counting only one of the many diagnoses that can be cited on the death certificate; and the data do not include the many silent diseases that are first detected (if at all) at necropsy (*2, 28*). More than 35 years ago, the prominent British epidemiologist J. N. Morris (*45*) proposed a method of using necropsy data to estimate the true occurrence rates of undetected disease. His proposal received no further investigative attention until research using the "epidemiologic necropsy" began to appear about a year ago (*28*).

Although the scientific complacency may not be admirable, the currently underdeveloped state of epidemiologic science in noninfectious disease is an entirely reasonable phenomenon. Each field of science develops at its own appropriate pace, and the inanimate sciences of physics and chemistry could surely be expected to advance more rapidly than biology, in which the majestic achievements of molecular science have occurred only in the past 30 years. Because individual people are much more difficult to study than molecules or animals, and because groups of people are even more difficult to study than individuals, it is entirely reasonable for scientific methods to be less well developed in epidemiology than in other fields.

What is less reasonable, however, is the assumption that current epidemiologic methods for studying noninfectious disease have the same high standards (*46*) as the methods used in other branches of science, or even in infectious disease epidemiology. In other branches of science, the progress to modern standards occurred when defective old paradigms (*47*) were replaced by new concepts and methods. The flat earth became round; the sun replaced earth as the center of the universe; Vesalius' dissections and Harvey's demonstration of the circulation supplanted Galen's erroneous dogmas about anatomic structures; oxygen and modern chemistry replaced phlogiston and alchemy; a randomized trial of high-concentration oxygen therapy demolished the entrenched academic belief that a treatment so beneficial to lungs could not harm the eyes of premature babies. In each instance, the paradigm replacements did not occur without avid resistance from the "peer-review process" of the era: the authoritative experts who were knowledgeable, dedicated, and honest—but wrong.

Lewis Thomas has suggested (*48*) that epidemiologic studies of noninfectious disease have produced their own adverse side effect: an "epidemic of apprehension." The epidemic grows with each new

alarm about a new menace in daily life. Uncertain about how to distinguish the many false alarms from the few that may be true, the public and nonepidemiologic scientists are confronted by evidence that is peer group–approved but scientifically inadequate.

Like the achievements of modern molecular biology, which required antecedent progress in technology and other sciences, the opportunity to discern the scientific inadequacies of epidemiologic methods required the antecedent development of randomized trials. They have become widely used only in the past 25 years. During the next 25 years, the methodologic lessons taught by randomized trials can lead to new paradigms, concepts, and approaches that will achieve fundamental scientific standards when randomized trials are not possible. The investigators will have to focus more on the scientific quality of the evidence, and less on the statistical methods of analysis and adjustment.

Until the new paradigms, methods, and data are developed, however, nonepidemiologic scientists and members of the lay public will have to use common sense and their own scientific concepts to evaluate the reported evidence. If war is too important to be left to military leaders, and medicine to physicians, the interpretation of epidemiologic results cannot be relegated exclusively to epidemiologists. The people who struggle to understand those results can be helped by recalling the old adage that statistics are like a bikini bathing suit: what is revealed is interesting; what is concealed is crucial.

REFERENCES AND NOTES

1. L. C. Mayes, R. I. Horwitz, A. R. Feinstein, *Int. J. Epidemiol.* **17**, 680 (1988).
2. A. R. Feinstein, *Clinical Epidemiology. The Architecture of Clinical Research* (Saunders, Philadelphia, 1985).
3. ———, *Ann. Intern. Med.* **99**, 705 and 843 (1983).
4. A. M. Lilienfeld, *Foundations of Epidemiology* (Oxford Univ. Press, New York, 1976).
5. D. L. Sackett, *Arch. Intern. Med.* **146**, 464 (1986).
6. Boston Collaborative Drug Surveillance Program, *Lancet* **ii**, 669 (1974).
7. O. P. Heinonen, S. Shapiro, L. Tuominen, M. I. Turunen, *ibid.*, p. 675.
8. B. Armstrong, N. Stevens, R. Doll, *ibid.*, p. 672.
9. B. MacMahon, S. Yen, D. Trichopoulos, K. Warren, G. Nardi, *N. Engl. J. Med.* **304**, 630 (1981).
10. W. C. Willett, M. J. Stampfer, G. A. Colditz, B. A. Rosner, C. H. Hennekens, F. E. Speizer, *ibid.* **316**, 1174 (1987).
11. A. Schatzkin *et al.*, *ibid.*, p. 1169.
12. J. F. Jekel, *J. Chron. Dis.* **37**, 679 (1984).
13. Boston Collaborative Drug Surveillance Program, *Lancet* **i**, 1399 (1973).
14. M. J. Stampfer *et al.*, *N. Engl. J. Med.* **313**, 1044 (1985).
15. G. A. Colditz *et al.*, *ibid.* **316**, 1105 (1987).
16. J. E. Buring *et al.*, *Am. J. Epidemiol.* **125**, 939 (1987).
17. A. Green *et al.*, *J. Natl. Cancer Inst.* **79**, 253 (1987).
18. G. A. Colditz *et al.*, *Int. J. Epidemiol.* **16**, 392 (1987).
19. W. C. Willett *et al.*, *N. Engl. J. Med.* **317**, 1303 (1987).
20. G. A. Colditz *et al.*, *Am. J. Epidemiol.* **126**, 861 (1987).
21. M. J. Stampfer *et al.*, *N. Engl. J. Med.* **319**, 267 (1988).
22. M. J. Stampfer *et al.*, *Am. J. Epidemiol.* **128**, 549 (1988).
23. S. J. London *et al.*, *ibid.*, p. 914.
24. M. J. Stampfer *et al.*, *ibid.*, p. 923.
25. T. W. McDonald *et al.*, *Am. J. Obstet. Gynecol.* **127**, 572 (1977).
26. F. W. Bauer and S. L. Robbins, *J. Am. Med. Assoc.* **221**, 1471 (1972).
27. L. Goldman *et al.*, *N. Engl. J. Med.* **308**, 1000 (1983).
28. M. J. McFarlane, A. R. Feinstein, C. K. Wells, C. K. Chan, *J. Am. Med. Assoc.* **258**, 331 (1987).
29. D. R. LaBarthe, *J. Chron. Dis.* **32**, 95 (1979).
30. S. Shapiro and D. Slone, *ibid.*, p. 105.
31. R. I. Horwitz and A. R. Feinstein, *Arch. Intern. Med.* **145**, 1873 (1985).
32. C. Hsieh *et al.*, *N. Engl. J. Med.* **315**, 587 (1986).
33. S. Graham, *ibid.* **316**, 1211 (1987).
34. L. A. Webster, P. M. Layde, P. A. Wingo, H. W. Ory, *Lancet* **ii**, 724 (1983).
35. S. Y. Chu *et al.*, *Am. J. Epidemiol.* **128**, 912 (1988).
36. A. Schatzkin *et al.*, *ibid.*, p. 913.
37. P. Reynolds, T. Camacho, G. A. Kaplan, *ibid.*, p. 930.
38. L. Gordis, *ibid.* **109**, 21 (1979).
39. J. Berkson, *Biom. Bull.* **2**, 47 (1946).
40. R. S. Roberts *et al.*, *J. Chron. Dis.* **31**, 119 (1978).
41. L. M. Gerber *et al.*, *J. Am. Med. Assoc.* **247**, 43 (1982).
42. H. O. Conn, N. Snyder, C. E. Atterbury, *Yale J. Biol. Med.* **2**, 141 (1979).
43. A. R. Feinstein and J. M. Esdaile, *Am. J. Med.* **82**, 113 (1987).
44. A. Gittelsohn and P. Royston, Annotated bibliography of cause-of-death validation studies: 1958–1980 (National Center for Health Statistics, Hyattsville, MD, 1982); Vital and Health Statistics, series 2 (no. 89), DHHS publ. no. (PHS)82-1363 (Public Health Service, Washington, DC, 1982).
45. J. N. Morris, *Lancet* **i**, 1, 69 (1951).
46. A. R. Feinstein and R. I. Horwitz, *N. Engl. J. Med.* **307**, 1611 (1982).
47. T. S. Kuhn, *The Structure of Scientific Revolutions* (Univ. of Chicago Press, Chicago, ed. 2, 1970).
48. L. Thomas, *Discover* **4** (no. 11), 78 (1983).

49. Supported in part by a grant from the Andrew W. Mellon Foundation.