The Vinyl Institute
A Division of The Society of The Plastics Industry, Inc.

~~REB-DFM~~ OCT 16 1987
JCL

October 6, 1987

To: Robert Hinderer, BFGoodrich Company
Paul E. Gurba, Occidental Chemical Corporation

RE: White Paper - Reproductive Outcome

Bob Oubre of Dow Chemical gave me the enclosed at the September 30th meeting of the Vinyl Institute Health, Safety and Environment Committee and asked me to forward it to you for your use in preparing the above-referenced white paper.

Sincerely,

Meredith N. Scheck
Assistant Director

MNS/pmb

cc: VI Health, Safety & Environment Committee
VI Technical/Medical Subcommittee

Wayne Interchange Plaza II • 155 Route 46 West • Wayne, NJ 07470 • (201) 890-9299

THE DOW CHEMICAL COMPANY
MIDLAND, MICHIGAN 48674

x Henderer
x Bennett
x Medical Subcom

September 22, 1987

TO: Bob Oubre
CEP Technology Center
OC 1120 Building
Texas Operations

FROM: Geary Olsen

Enclosed is an article by the late Dr. Stallones (Dean of the School of Public Health, University of Texas at Houston) which graphically depicts the subgroup problem of the Infante et al vinyl chloride/spontaneous abortion paper published in Lancet in 1976. See only pages 184 - 186 of the enclosed paper. Basically there was only one age range (25-29) between the exposed and nonexposed workers in the Infante et al paper that showed a higher risk for spontaneous abortions for the exposed workers' wives. However, the problem is that the aged 25-29 nonexposed subjects had a lower rate than what you should have expected, meaning that for this age range, the data are suspect. What Infante et al did in their original article is to report only the overall results which showed higher risk among the exposed. However, this perceived increased risk was completely due to the aberrant findings of the aged 25-29 exposed and nonexposed subjects. Quoting Dr. Stallones, " No answers to these questions are evident; a sensible conlusion is that something went wrong in the study (Infante's), resulting in aberrant findings, and that the study therefore should be discarded. If published at all, the data should appear in a textbook of epidemiology as a most pertinent example of the value of subgroup analysis in discovering a problem, internal to the research, which renders invalid the results obtained for the total group, and which might have been accepted had teh subgroup analysis had not been done." I am sure the Vinyl Institute would like to see this article and I trust you will forward it to the Institute.

ISSN 0091-7435

Volume 16, Number 2, March 1987

American Health Foundation

Preventive Medicine

An International Journal
Devoted to Practice and Theory

ACADEMIC PRESS, INC.
San Diego · Orlando · New York · Austin · Boston

Epidemiol 106, 462–469

lividually matched case–
le taking multiple values.

use–control studies. *Bio-*

veight change in women:

ntraceptive use and cer-
2 (1977).
and cervical carcinoma.

unding in epidemiologic
miol. 115, 10–18 (1982).
enesis. *Obstet. Gynecol.*

-term follow-up study of
t. *J. Biosoc. Sci.* 8, 373

. J., and Kohn, H. I. An
utritive sweeteners and

effects of birth control
Commonw. 79, 673–679

R. Neoplasia and dys-
t of the diaphragm. *Brit.*

A second look. *Cancer*

cancer: A case–control

PREVENTIVE MEDICINE 16, 183–194 (1987)

The Use and Abuse of Subgroup Analysis in Epidemiological Research[1]

REUEL A. STALLONES, M.D., M.P.H.†[2]

Faculty of Epidemiology, School of Public Health, The University of Texas Health Science Center, Houston, Texas 77225

Most epidemiology involves the analysis of subgroups. In observational studies, identification of associations within particular subgroups is the usual method of investigation, while in experimental studies the purpose is to compare the rate of occurrence of the outcome of interest in treated and control groups. Problems can arise through failure to specify the subgroups of interest a priori and through examining large numbers of subgroups after the fact, i.e., through multiple testing. Rules for interpretation of findings in subgroups are suggested, and the value of the systematic application of criteria for judgment of the causal significance of any associations that may be observed is noted.

INTRODUCTION

The primary method of epidemiological analysis is to dissect time and space and relevant characteristics of people to identify subsets with higher risk for disease. We track upward gradients of risk in these subsets, searching for causes with the avidity of a mosquito following an increasing concentration of CO_2 to find a blood meal. Epidemiologists must be careful, therefore, not to condemn subgroup analysis in general. As an exploratory method, it is the bread and butter of epidemiological research, and in an epidemiological experiment, the entire purpose is expressed in the comparison of two subgroups, the test subjects and the controls. Nearly everything that has been learned in epidemiology has been derived from the analysis of subgroups. This is an incisive, effective technique to which we owe our sustenance.

Subgroup analysis is especially appealing in situations where the associations uncovered are weak. A weak overall association may, of course, be weak because of dilution by irrelevant variables, and a search for a classification axis that identifies persons at higher risk can be very rewarding. Unfortunately, the process has a price: as the study population is divided and subdivided, the rapid deterioration in the numbers of people in the cells reduces our confidence in the rising relative risks.

Two classes of problems that are especially troublesome arise frequently. First is the subgroup that turns up in an observational study and appears exceptional. The issue is whether this represents a grand new answer to some important ques-

† Deceased.

[1] From the Workshop on Guidelines to the Epidemiology of Weak Associations, American Health Foundation, New York, December 4–5, 1985.

[2] Address reprint requests to Office of the Dean, School of Public Health, The University of Texas Health Science Center at Houston, Reuel A. Stallones Building, P.O. Box 20186, Houston, TX 77225.

0091-7435/87 $3.00

tion, or whether it is a bit of irrelevant noise in the system. The second special problem is the interpretation of the subgroup analyses within a controlled experiment where the subgroup categories were not defined by the original hypotheses and were not specified in the sample size calculations.

SUBGROUPS IN OBSERVATIONAL STUDIES

We must expect that most truly new research findings in observational research will be unexpected, and an unexpected result may well appear aberrant. To distinguish between those exceptional subgroups that have epidemiological significance and those that do not may be difficult. Statistical treatment is of little use in addressing the problem. A probability statement is of value in assessing the magnitude of an association in relation to the number of persons in the study; that is, it may help us restrain our enthusiasm over large differences in small groups. However, a P value of 0.01 in an observational study does not mean that a difference as great or greater than the one observed will occur no more often than 1 in 100 trials. By and large, statistical significance is a meaningless term in observational research. This is true for many reasons, but mainly because a P value does not know any epidemiology; it was born into a world of tossing pennies and urns full of black and white balls, and it never read a book about disease. It can't tell the difference between truth and bias, and it worships sample size more than truth.

Everyone who has spent much time with epidemiological data has noted a cell in a tabulation unaccountably containing more or fewer people than expected, or an arbitrary disruption of a smooth trend on a graph by one or more uncooperative points. These look like orphan subgroups, not related to their companions, and we must decide whether to adopt the orphan or to submerge its individuality by some technique of data smoothing. In the latter case, we usually whisper to ourselves that we are searching for the deeper truth concealed beneath the surface variation, as though truth were to be found like a submarine cruising a rowdy sea at periscope depth.

The Orphan Subgroup Submerged

In 1976, Infante *et al.* published an article (5) in which they concluded that persons exposed to vinyl chloride monomer (VCM) in the course of their employment had an excess risk of fetal loss: 8.8 and 15.8% of pregnancies associated with the unexposed and the exposed men, respectively. This was interesting; few people think well of VCM exposure, and fewer still would stand up to defend it. However, Paddle wrote a letter to the editor (11) seeking clarification, and the response by the authors (6) presented the data shown in Fig. 1, which had not appeared in the original publication. The pregnancies and the fetal deaths were presented according to age of father, and the risk differed substantively between the exposed and the unexposed employees in only one of four age groups. In persons ages 25 to 29, the proportion of pregnancies terminated by fetal death was very high for the exposed persons and very low for unexposed persons. Therefore the difference that was reported in the initial publication was due entirely to the difference in one of four age groups, which is puzzling. The risk of

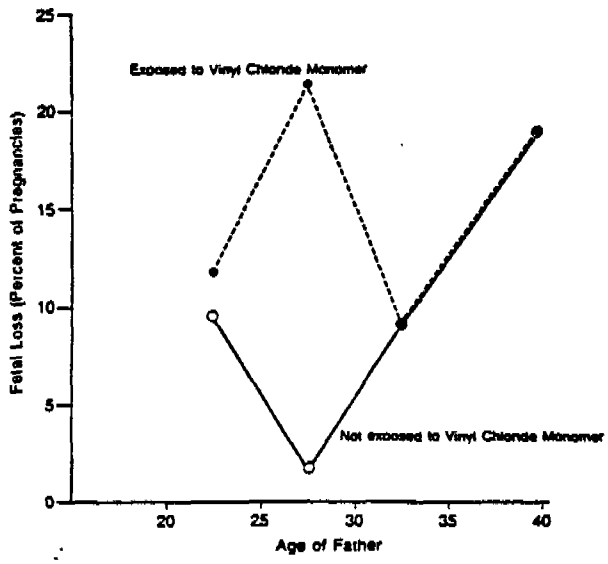

FIG. 1. Fetal deaths according to paternal age for men exposed and not exposed to vinyl chloride monomer [adapted from Ref. (6)].

fetal death should increase with increasing age of parent, but in this study, the 25- to 29-year age subsets of both the exposed and the nonexposed persons departed from that expectation in opposite directions, which is even more puzzling.

The analysis of these data is a bit of a problem. The authors found the total exposed risk, smoothed across age, to be (significantly) higher than the total unexposed risk. The problem may be viewed in different ways, but no matter how it is viewed, that analysis is wrong, for it implies that the hazard affects the entire age span, and it intentionally dilutes the strong association in persons ages 25–29 with the zero associations in the other age groups. The isolated high risk in the 25- to 29-year-old exposed persons might conceivably be explained by some kind of age–exposure interaction. That allows the hypothesis to be conserved, although one might wish for some external corroboration of such an interaction. If that were the premise, then the appropriate comparisons are between the several age-specific subgroups. In that case, however, the unnaturally low rate among the 25- to 29-year-old unexposed men becomes an insoluble problem. To propose an age-specific protection for unexposed persons is not at all credible; the rate for that nonexposed group is inexplicably too low, it is not a suitable referent, and the age-specific comparisons are not acceptable.

The findings raise the following questions: (a) Why was the harm limited to the 25- to 29-year age group? (b) Why was the rate higher in the exposed 25- to 29-year age group than in either of the two older exposed age groups? (c) Why was the rate lower in the nonexposed 25- to 29-year age group than in the younger, nonexposed age group?

No answers to these questions are evident; a sensible conclusion is that something went wrong in the study, resulting in aberrant findings, and that the study

therefore should be discarded. If published at all, the data should appear in a textbook of epidemiology as a most pertinent example of the value of subgroup analysis in discovering a problem, internal to the research, which renders invalid the results obtained for the total group, and which might have been accepted had the subgroup analysis had not been done.

The Orphan Subgroup Adopted

In 1973, Ramcharan *et al.* presented the results of a 7-year study of periodic health evaluation (13). Examinations conducted at intervals frequent enough to detect early, latent, and incipient illness seem so reasonable as hardly to require rigorous assessment. Nevertheless, in 1964 Morris Collen had initiated a carefully designed, random allocation trial based on the multiphasic screening examination offered to subscribers by the Kaiser Foundation Health Plan. Consultants to the study included Hardin Jones, Jerzy Neyman, Jacob Yerushalmy, and, briefly, me.

A randomly selected group of subscribers in the San Francisco/Oakland area was enrolled in the study and urged once each year to take advantage of the multiphasic examination. Another sample served as the control group, and it was not subjected to special pleading. Information concerning disease, disability, and medical care received was obtained from both groups by mailed questionnaire; the surveys were done in 1964, 1967, 1969, and 1971 for the experimental group and in 1966, 1967, 1969, and 1971 for the control group.

For analysis, the study groups were divided by gender and birthdate (1910–1919 and 1920–1929). No benefit was found in three of the four subgroups; they were dropped and the remaining group, the subset of older men, was adopted. Part of the analysis was a presentation of the results by time, the data points being the biennial recapitulation of the experience to date. Of the various measures of health and illness status, only the reports of disability and working status seemed importantly better for the experimental group. The authors noted that statistically significant differences were observed at the 1969 and 1971 evaluations (Fig. 2). They wrote, "The repeated observation of this significant difference in disability in two consecutive surveys decreases the likelihood that the difference is due to chance even though the two survey observations are not independent of each other." This interpretation completely ignores the information available in the slopes of the two lines. If the treatment were effective, it should reduce the rate of accumulation of disability, and that would be reflected in a change in slope, as occurred between 1967 and 1969. However, the slope for the treated group reverted to its original pattern after 1969, so that the apparent benefit was sharply localized in one subset defined by time, and no additional benefit accrued thereafter. If multiphasic screening is beneficial, its limited expression in time is inexplicable.

The logical problems that arise in an attempt to save the hypothesis that periodic examinations are beneficial are formidable and were not addressed by the authors. Ignoring three of the age–gender groups is not licit; the study actually had four age–gender cells, further subdivided into three time intervals, for a total of 12 cells, and only one showed a beneficial effect related to the multiphasic screening examination. The following questions are posed: (a) Why was the ben-

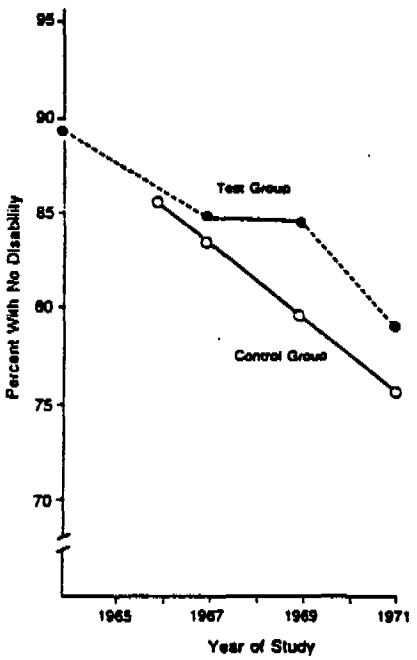

FIG. 2. Percentage of men reporting no disability by year of study for persons encouraged to participate in annual multiphasic screening examinations and for a control group [adapted from Ref. (13)].

efit confined to older men and not observed in three other age–gender groups? (b) Why was the benefit confined to the time interval between 1967 and 1969? (c) Why was the benefit limited to only very few health status indices?

Lacking satisfactory answers to any of these questions, the prudent course would be to discard the findings as an aberrant subgroup. Adopting a subgroup should be weighed carefully; not all ugly ducklings grow up to be beautiful swans: Some of them grow up to be ugly ducks.

The Orphan Subgroup May Have a Twin

In the Coronary Drug Project (3), the 5-year mortality for persons treated with clofibrate was not significantly lower than the mortality for the placebo control. However, within the clofibrate treatment group, the subgroup that adhered well to the regimen (taking 80% or more of the drug prescribed) had a considerably lower mortality. This naturally reopened the issue of the efficacy of the drug, since it could not be expected to work unless people took it. However, the placebo group also included people who were more and less compulsive in disposing of their pills, and analyzing the data in a neat fourfold table (Table 1) showed that the difference was due to selection and not to drug effect, for the persons who complied well with their regimens had lower mortality rates than did the nonadherers, whether they were in the drug group or the placebo group (3). Subgroup analysis may be either evil (misleading us as to the benefits of clofibrate treatment) or good (correcting the misapprehension).

TABLE 1
5-YEAR MORTALITY IN STUDY SUBJECTS GIVEN CLOFIBRATE OR PLACEBO: CORONARY DRUG PROJECT

Cumulative adherence	Treatment group: Clofibrate — Number	Clofibrate — Percentage mortality	Placebo — Number	Placebo — Percentage mortality
<80%	357	22.5	882	25.8
>80%	708	15.7	1,813	16.4

A Visit to an Orphanage

In 1974, the Boston Collaborative Drug Surveillance Program staff reported an association between breast cancer and rauwolfia alkaloids (2). The Surveillance Program had long been recognized as a spawning place for orphan subgroups. In this instance, information was obtained on 25,000 patients admitted to the hospital, and they were interviewed as to habits and medications. This led to an enormous matrix of drugs vs diagnoses with some thousands of cells, and to an expectation of a number of adventitious associations. To resolve this question with respect to rauwolfia and breast cancer has taken a decade; and the issue may not yet be laid to rest. Since the first report, 22 studies have been done, 5 using cohorts and 17 employing case–comparison designs. Of the case–comparison studies, 4 were positive, 2 were inconclusive, and 11 showed no association; of the cohort studies, 1 was inconclusive and 4 did not support the hypothesis (7, 8). Science does not live by body counts, but scientific findings should be replicable in different places and in different circumstances.

SUBGROUPS IN EXPERIMENTAL STUDIES

Experiments may be controlled or uncontrolled, and may be based on treatments applied to individuals or to communities. For the most part, problems with subgroup analysis arise in random-allocation, placebo-controlled experiments with individual study subjects.

In such a study, alpha and beta are selected formally, and sample size computations are based on them and on other pertinent considerations, such as expected occurrence rate of events, expected treatment benefit, response rates, attrition rates, and time required to realize benefit. Constraints of time and money usually limit the sample size to the minimum number necessary to detect a difference between the treated and the untreated persons in the incidence of the specific disease or, sometimes, of the cause-specific deaths. The numbers of study subjects in subgroups are often too small to support a demonstration of statistical significance, but even if the differences are large and the sample sizes are adequate, a significance test cannot properly be used as a test of an hypothesis unless the subgroup hypothesis was specifically and explicitly included in the initial study formulation. If this was not done, then the analysis of subgroups within the experimental population is exploratory, not confirmatory, and the same rules

apply to the interpretation of results as apply to other observational epidemiological research.

In the following, we do not assess the value of dietary intervention for the prevention of coronary heart disease (CHD), nor the cosmic value of carotid angioplasty or coronary artery bypass. The examples deal with those matters, but the issues addressed here have to do with procedure, rules of evidence, and how subgroup analysis is interpreted, and not with whether the interpretations reached in the various studies were correct.

Interactive Orphans

The Multiple Risk Factor Intervention Trial (MRFIT) was launched in 1972 amid high hopes that adoption of a few hygienic measures would markedly reduce the risk of ischemic heart disease. The confidence was so high that the planners embarked on a confounded study, mixing together dietary change, blood pressure reduction, smoking cessation, and whatever incidental effects these activities might carry in train. This was an experiment, planned to test an explicit, although somewhat sloppy, hypothesis. Roughly, the hypothesis was that the things that were done to the study subjects would reduce the risk of ischemic heart disease. MRFIT was designed to detect reductions in deaths from CHD, deaths from cardiovascular disease (CVD), deaths from any cause, and incident cases of myocardial infarction. Subgroup differences were incorporated into the hypotheses at some stage of the proceedings, especially with respect to persons who had demonstrable evidence of prior atherosclerotic disease.

Insofar as mortality was concerned, MRFIT had a null result (9); the differences in death rates between the test (special intervention; SI) and the control (usual care; UC) groups were modest and were not statistically significant. The null hypothesis could not be rejected. This was such a disappointment that the researchers rejected the result, contending that it was "inconsistent with most published scientific data." This conserves the hypothesis but does great damage to science.

The authors sought to explain the problem in several ways.

First, they noted that the reductions in cholesterol, blood pressure, and cigarette smoking among the controls were greater than had been expected, and that their illness experience was lower. This sharply diminished the difference that might otherwise have occurred between the test and the control groups. This occurred against a background of rapidly declining ischemic heart disease mortality in the general population of the United States. The directors of the MRFIT seem to wish to enjoy the advantage of this global benefit for the test group, but deny it to the control group. But, effects such as these are exactly the reason for maintaining a control group. In any case, the argument is irrelevant; what is important is what difference in risk was achieved by the intervention program and how that difference related to disease occurrence. Risk was assessed for each individual by application of a multiple logistic regression equation, and the original plan projected that a difference as great as 26% in the CHD mortality rate could be detected, corresponding to an achieved 26% difference in the estimates

of risk. As it turned out, despite the unexpected changes in the control subjects, the overall risk was 22% lower in the test group; this was not accompanied by a 22% lower disease experience, but rather one of 7%. The null result was not due to a failure to achieve an adequate differential, but rather due to a failure to enjoy a reduction in CHD deaths commensurate with the reduction in risk score. An appropriate conclusion is that the risk score is not synonymous with risk.

Next the investigators called attention to some aberrant subgroups, arguing that an excess of deaths occurred among members of the experimental group who were hypertensive and who also had abnormal ECGs at entry. They claimed that this matter could be considered legitimately because the abnormal ECG was evidence of prior clinically manifest atherosclerotic disease, and that this had been mentioned earlier as a special problem (although the subgroup hypothesis apparently did not specify an interaction between ECG abnormality and hypertension). Since intensive treatment with antihypertensive drugs was an integral part of the study design, the test group would be more likely to be treated than the control group, and the treatment then may have had unexpected, untoward results. Clearly, any research group has a responsibility to be alert to such hazards and to warn about them. However, the data presented in the MRFIT report offer alternative explanations. Table 2 illustrates the way in which subgroup analysis wearies the mind. Here we have fourfold tables of the combinations of hypertensive/not hypertensive and ECG abnormal/ECG not abnormal, for the test group and the control group, giving mortality rates for CHD and for all other causes. Each of the fourfold tables is arranged so that the group with the lowest mortality rate should appear in the upper left cell and that with the highest mortality rate in

TABLE 2

CHD AND NON-CHD DEATH RATES FOR SELECTED SUBGROUPS OF THE MRFIT STUDY POPULATION[a]

CHD death rates/1000			CHD death rates/1000		
ECG abnormal	Hypertension		ECG abnormal	Hypertension	
	No	Yes		No	Yes
	Set A: Test group			Set B: Control group	
No	13.2	15.8	No	16.1	20.7
Yes	18.6	29.2[b]	Yes	25.7	17.7[c]
	Set C: Test group			Set D: Control group	
No	26.9[d]	20.1	No	17.7	22.7
Yes	15.2	30.8	Yes	22.3	22.0[e]

[a] Multiple Risk Factor Intervention Trial Research Group (Ref. 9).

[b] The authors argue that 29.2 is too high.

[c] The data indicate that 17.7 is assuredly too low. If this is true, the difference between the test and the control groups is not explicable in terms that save the presumption that the interventions were beneficial.

[d] According to expectation, this should be the smallest value in the table.

[e] This should be the largest value in this table, according to expectation. The presence of two aberrations in two tables casts the gravest doubt on the wisdom of selecting particular results from particular subgroups.

the lower right. This expectation is realized for only one of the four tables. The authors aver that the rate in the lower right cell of Set A is unnaturally high, and that this accounts in part for the failure of the study to show a beneficial effect from the interventions. In fact, a much more compelling argument can be made that the SI/UC difference is not due to an excess of SI deaths but rather to the marked deficit of UC deaths shown in the lower right cell of Set B. The complexity of these matters is further attested by the observations of an apparent excess of non-CHD deaths in SI participants who were not hypertensive and who did not have ECG abnormalities (Set C), and by a deficit of non-CHD deaths in UC persons who had both (Set D).

Don Meredith has said that what is important is not whether you win or lose, but who gets the blame. One of the issues in interpreting subgroup analyses is how to apportion the blame for the subgroups in which the experimental subjects had an especially unfortunate result. The following questions need to be answered: (a) If the high CHD mortality in the subset of the test group with abnormal ECG and hypertension is due to a treatment effect and not an aberration, then why should the equivalent subset in the control group have a deficit of CHD deaths? (b) Why should the test group have an excess of non-CHD deaths in the subset without hypertension and without abnormal ECGs? (c) Why should the subset of control subjects with hypertension and ECG abnormalities have a deficit of non-CHD deaths?

The point is simple: The results come as a package: taking some pieces and throwing away others may be satisfying, but it is not good science. The issue that arises out of all of this is whether investigators have the privilege of sifting though a set of subgroups and selecting those they wish to support a particular position, ignoring others that do not. The answer is affirmative, with the understanding that these findings are not the result of a controlled experiment, and that probability statements attached to them are meaningless.

The Hidden Orphan

In the 1950s, carotid angioplasty was introduced as a prophylactic treatment for brain infarction. The rationale was reasonable and the results seemed satisfactory but, nevertheless, a random-allocation, controlled trial was proposed and was conducted, beginning in 1962. The data were analyzed after 42 months of observation, and, according to the rigid rules of hypothesis testing, no benefit could be attributed to the procedure (11). Indeed, after the initial surgical losses in the operated group, the life table survivorship curves were parallel for the treated and the control groups. This result was not acceptable, and subgroups were sorted through, as researchers looked for a ray of light. Subdivision of the study population into persons with and without significant neurological disability showed equal survival for the operated and nonoperated people in the healthier group. Further subdivision by angiographic findings revealed that in the healthier group, the people who had unilateral carotid artery stenosis had better survival if operated than if managed medically. The critical subgroup had come to light. The analysis also illustrated a long-standing principle that healthy people withstand surgery better than do sick people. The summary that introduced the paper did

not mention the overall result of the experiment, restricting attention to the results in the subgroups. All of the statements regarding the subgroup differences were bolstered by statements of statistical significance. I was a member of the committee of the National Heart Institute that had oversight responsibility for this study, and I do not recall any explicit hypotheses concerning the results in subgroups; the major concern was whether the sample sizes that were feasible would be large enough to provide a reasonable answer to the global question as to whether carotid angioplasty was beneficial.

To conduct another trial, selecting for random allocation only those persons with unilateral carotid artery stenosis, was not feasible. No one would pay for it, and the ethical problem of withholding treatment was overwhelming. The doctrine was adopted that carotid angioplasty was a good and useful procedure, when applied to those persons for whom it was good and useful. This may have been an absolutely correct judgment, but it cannot be said to have been affirmed by the ordeal of a controlled trial.

The Sibling Orphans

With this precedent the results of controlled trials of coronary artery bypass surgery should have been predictable. A report of the Veterans Administration (VA) study (10) excluded the persons with left main coronary artery disease and found that "at 36 months, 87 percent of the medical group and 88 percent of the surgical group were alive." The excluded group was reported to have shown benefit from surgery (15), although the difference was statistically significant at 24 months and not at 36 months. The European Coronary Surgery Study Group conducted a controlled trial beginning in 1973; as in the VA study, it found no statistically significant difference between the two groups treated surgically and medically, but the subgroup that did show a benefit from surgery turned out to comprise those patients with three-vessel disease (14). These two orphans look a lot alike, but they are not twins. Each study picks one as real and the other as not, and no criteria are easily evident to help us decide which one to believe. Evaluating the alternatives in a formal way in a controlled experiment is not likely.

RULES OF EVIDENCE

The difference between prediction and "postdiction" is often ignored. An experiment must start with an explicit hypothesis and an observational study may be just as firmly based. An hypothesis predicts a result, and the research either confirms the prediction or it does not, according to a predetermined set of rules as to what constitutes confirmation. This is a rigid game and should be played only by compulsive rule followers. An exploration of the differences among subsets of a study population without previously formulated hypotheses uncovers postdictions, i.e., associations that were not predicted in advance. In these instances, tests of significance are useful summary statistics, combining information about the number of persons under observation with the magnitude of the observed differences, but they are not hypothesis tests, because no hypothesis was proposed to be tested. In this context, an hypothesis is an explicit statement

of an expectation; a general statement that the investigator thinks something interesting might turn up if this information were obtained does not qualify. After the identifying data are accounted for, every item included on a questionnaire or in an examination may be said to owe its presence to an hypothesis linking those items to the disease under study. However, these expectations are not often framed formally. This does not mean that hypotheses cannot be tested in observational studies—they most certainly can be. It does mean, however, that hypothesis-testing statistics do not have the same meaning in observational research as they have in assessing experiments. Only in an experiment can one hope to fulfill the requirements for formal hypothesis testing, and many experiments fall far short.

One of the serious problems with the MRFIT was that it was pronounced a failure from the moment the first results appeared. This was, of course, totally incorrect, for the MRFIT was a major success. To bring such a massive study, protracted and logistically complicated, to a successful culmination is testimony to the skill and dedication of hundreds of people who participated in its conduct. Unfortunately, a confused and erroneous set of values has led some people to consider that an experiment with a positive result, that is, that tends to confirm the hypothesis under test, is a success, and a study with a negative or an inconclusive result is a failure. These equalities could be true only if the researchers believed so strongly in the hypothesis that they should not, for ethical reasons, have ever tried to test it. A strong argument has been made (12) that the most successful experiment is the one that produces a definitive negative result, that the strongest and most productive science derives from hypothesis elimination. Part of the problem in our field is the use of the term "negative study," for the word "negative" has unkind connotations. Perhaps if we referred to these results as "null studies," attitudes might be slightly desensitized.

In a research endeavor, after all the shouting is over, we are left with the data and perhaps a little quiet time to think about what they mean. This is the most important procedure in epidemiology (or any other science): to evaluate the body of evidence that is assembled around an important or interesting question. For the most part, we are attempting to determine whether a set of observations helps us understand the causes of a disease. The evaluation should incorporate as much of the relevant information as possible, including laboratory results and statistical analyses, but the evaluation is based on epidemiological knowledge, compounded of biology, behavioral science, and environmental science. This coalescence directed specifically toward questions of the reasons for the occurrence of disease in human communities constitutes the unique character of epidemiology.

In this assessment, the results derived from subgroups should be treated the same way in which any other information is processed. The systematic application of criteria for judgment of the causal significance of association may be useful, although an epidemiologist will apply these whether systematically or not. One frequently cited set from the Surgeon General's Report on Smoking and Health (14) includes magnitude, consistency, specificity, time sequence, and coherence. For the most part, the associations discovered in subgroup analysis are large enough to attract our attention; indeed the analysis was probably engineered

to maximize that value. For the others, based on the examples presented earlier, the tests of consistency and coherence would appear to be the most important. The aberrant subgroup that stands alone and is difficult to explain in relation to whatever else we know about the situation is not likely to represent a grand new contribution to our knowledge of disease causation. The result of the evaluation is not a number and not a statistic; it is a judgmental statement, and its merit is a function of the knowledge, skill, and insight of the researcher.

REFERENCES

1. Bauer, R. B., Meyer, J. S., Fields, W. S., Remington, R., MacDonald M. C., and Callen, P. Joint study of extracranial arterial occlusion. III. Progress report of controlled study of long-term survival in patients with and without operations. *JAMA* 208, 509–518 (1969).
2. Boston Collaborative Drug Surveillance Program. Reserpine and breast cancer. *Lancet* 2, 669–671 (1974).
3. The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *New Engl. J. Med.* 303, 1038–1041 (1980).
4. European Coronary Surgery Study Group. Coronary–artery bypass surgery in stable angina pectoris: Survival at two years. *Lancet* 1, 889–893 (1979).
5. Infante, P. F., Wagoner, J. K., McMichael, A. J., Waxweiler, R. J., and Falk, H. Genetic risks of vinyl chloride. *Lancet* 1, 734–735 (1976).
6. Infante, P. F., Wagoner, J. K., McMichael, A. J., Waxweiler, R. J., and Falk, H. Genetic risks of vinyl chloride. *Lancet* 1, 1289–1290 (1976) (letter).
7. Joyce, C. R. B. Rauwolfia derivatives and breast cancer: How do we know when we have the answers? Unpublished manuscript.
8. Labarthe, D. R. Methodologic variation in case–control studies of reserpine and breast cancer. *J. Chronic Dis.* 32, 95–104 (1979).
9. Multiple Risk Factor Intervention Trial Research Group. Multiple Risk Factor Intervention Trial: Risk factor changes and mortality results. *JAMA* 248, 1465–1477 (1982).
10. Murphy, M. L., Hultgren, H. N., Detre, K., Thomsen, J., and Takaro, T. Treatment of chronic stable angina: A preliminary report of survival data of the randomized Veterans Administration cooperative study. *New Engl. J. Med.* 297, 621–627 (1977).
11. Paddle, G. M. Genetic risks of vinyl chloride. *Lancet* 1, 1079 (1976) (letter).
12. Platt, J. R. Strong inference. *Science* 146, 347–353 (1964).
13. Ramcharan, S., Cutler, J. L., Feldman, R., Siegelaub, A. B., Campbell, B., Friedman, G. D., Dales, L. G., and Collen, M. F. Multiphasic checkup evaluation study. 2. Disability and chronic disease after 7 years of multiphasic health checkups. *Prev. Med.* 2, 207–229 (1973).
14. The Surgeon General's Advisory Committee on Smoking and Health. "Smoking and Health," PHS Publ. No. 1103. U.S. Govt. Printing Office, Washington, DC.
15. Takaro, T., Hultgren, H. N., Lipton, M. J., and Detre, K. The VA Cooperative Randomized Study of Surgery for Coronary Arterial Occlusive Disease. II. Subgroup with significant left main lesions. *Circulation* 54 (Suppl. III), 107–117 (1976).