Chapter 7

Case-Control Studies

Case-control studies are closely related to prevalence or cross-sectional studies (discussed in Chap. 6). However, because they generally involve fewer and more readily accessible subjects, case-control studies are much more often carried out. Among analytic studies, they are usually the first approach to determining whether a particular personal characteristic or environmental factor is related to disease occurrence.

How Case-Control Studies Are Carried Out

Identification and Collection of Cases Once the study objectives and methods have been clearly defined, the first step in a case-control study is the identification of the cases or diseased persons to be studied. (Many rightfully object to the use of the term "case" to refer to a sick human being. Although this dehumanizing term should be avoided in the clinical setting, its use facilitates clear communication about research. In this context it does not imply any lack of sympathy or concern about the ill.)

As mentioned previously in connection with prevalence studies, it is important to set up criteria for the diagnosis and inclusion of cases in the investigation and to describe these criteria carefully when the study is finally reported. It is usually advisable to require objective evidence and documentation of the disease, even if, as a result, some cases will have to be omitted and the size of the case group reduced. Thus, for a study of renal calculi, it may be best to insist that all included cases have stones documented by x-ray evidence or removal by surgery, not diagnosed only by the presence of renal colic. By accepting less well-documented cases, the investigator runs the risk of diluting his case group with some noncases and lessening his chances of finding differences between the case group and the control group.

This recommendation, of course, applies to disease identification for all types of studies, not just case-control studies. However, as was stressed in the last section of Chap. 3, *misclassification* of a few nondiseased persons as cases and of a few diseased persons as controls, no matter how distressing to the clinician, will probably not prevent the discovery of major case-control differences.

The cases may be identified or "ascertained" by a community-wide search, but more often, they are limited to those found in one, or perhaps a few, hospitals, clinics, or medical centers. The case group will usually be limited to those seen or diagnosed during a particular time period. For example, one may decide to study all cases of well-documented renal stones seen at a particular hospital during the 2-year period, January 1, 1974 through December 31, 1975.

Usually, it will not be possible to include in the study *all* the patients who meet the diagnostic criteria and the time and place specifications. There will be a variety of reasons for this. Some patients will have moved away, died, or will refuse to cooperate; or, some hospital records may be lost so that certain essential information is not available to the investigator. He or she, in turn, should report how many cases met the initial criteria for inclusion and how many were finally included. The reasons why some cases had to be omitted and the number of cases omitted for each reason should be stated.

Selection of Control Subjects The decision as to who will constitute the control group or groups is perhaps the most difficult one to be made in planning a case-control study, and it requires a good deal of skill and judgment. In a prevalence study this problem does not arise since the cases may be compared with the entire nonaffected portion of the population. By settling for the simple low-cost case-control study instead of the large community-wide prevalence study, the investigator gives up the chance of comparing all the diseased and nondiseased persons in the community. However this is done in the hope that almost as much can be learned about the relationship of the disease to other variables by studying a group of cases and a group of controls. Sometimes a relatively small sample derived randomly from the entire population can be utilized as a control group. However obtaining the desired participation of this kind of representative control group is difficult and often not feasible.

General Principles One of the most important considerations in selecting controls involves the information to be collected concerning study variables or potential etiologic factors. There should be no major differences between case and control groups as to the quality or availability of this information. Availability of information implies both (1) how much information is obtained concerning each case and control, and (2) what proportions of the case and control groups will, or can, supply it. Equal access to important information previously recorded in a similar fashion for both cases and controls—for example, birth weight recorded in the same hospital—may strongly favor the use of a particular control group. If data have to be obtained by interview, then one worries that quality or availability of information may differ due to differences between cases and controls in emotional state, knowledge of the disease studied, educational or socioeconomic status, and location of the interview (e.g., at home or in a hospital).

Consideration of the *known* sources of bias in quality and quantity of information about cases and controls and of the fact that there are often biases which are *unknown* usually leads the investigator to attempt to find controls that are similar in a general way to the cases, except for the essential difference in whether the disease under investigation is present or absent. Yet, this striving for general similarity should not be carried to the point where there is little or no hope of finding case-control differences in the factors under study. For example, by selecting the controls so that they are of similar educational background to the cases, one will minimize case-control differences in the understanding of a written questionnaire. But this selection procedure will also preclude the study of the relation of educational level to the disease and may seriously impair case-control comparisons of factors related to education, such as socioeconomic status.

In selecting a control group two major questions must be answered

1 From what source(s) will controls be drawn?
2 What will be the method of selection of controls from each of these sources?

These decisions must take into account the need, mentioned above, for controls that are generally similar, but not too similar, to the cases, plus some very practical considerations—in particular, the control groups that are potentially available, and the human and financial resources that can be used for the study.

Selecting a Source of Controls Many sources of controls have been used, including:

1 Patients within the same medical-care facility
  a Without regard to their diagnosis
  b Excluding those with certain diseases
  c Including only those with certain diseases such as mild or "act-of-God" conditions (e.g., hernias, accidental injuries)
  d Examined and found to be healthy
2 Persons drawn from outside the facility
  a Sample of general community
  b Friends or acquaintances
  c Fellow employees
  d Neighbors
  e Family members such as spouses or siblings

When one is faced with the practical decision as to which source of controls to use, reasons for and against any potential source can usually be mustered, and the reasons why the source chosen might have given biased results will be heard from critics after the study is reported. For example, the investigator may decide to select controls for hospitalized renal calculus cases from herniorrhaphy cases in the same hospital, since that hospital serves a particular socioeconomic and ethnic segment of the community, and since, after the acute pain has subsided, the mental status of a kidney stone patient should not be very different from that of a hernia patient (as contrasted with a patient, say, with a stroke or terminal cancer). Yet if an important difference between kidney stone patients and their hernia controls is found, there will usually be the lingering question of whether the difference is related to kidney stones or to hernias. Therefore, it is frequently helpful to have a diagnostically heterogeneous control group, or more than one control group, if possible. Similarly, repetition of the study by other investigators in other settings will usually reveal whether or not some underlying truth about renal calculi has been discovered. MacMahon and Pugh (1970) have thoroughly discussed many of these important issues and other factors to be considered in selecting controls.

Selecting Control Subjects from the Source Selection of the control group from the chosen source usually involves sampling. If resources are limited, the control group will usually be equal in size to the case group or smaller than the case group, if necessary. If resources permit the inclusion of more study subjects and no more cases are available, the control group may be enlarged to decrease sampling variation by having, for example, twice or three times as many controls as cases, or even more.

As already noted, selecting a source places some general limitations on the nature of the control group. In addition, when individual controls are chosen from the source, the investigator will often *match* the controls to the cases with regard to some important characteristics such as age or sex. By matching on a particular characteristic, the investigator immediately eliminates a case-control difference in this characteristic as a possible contributor to a case-control difference in a study variable. For example, if the cases and controls are matched for age and it is subsequently found that they differ in blood pressure, age could not be the explanation for this blood pressure difference. In the unusual instance that nothing is known about the disease, not even, say, its age and sex distribution, then no matching would be desired since matching precludes any case-control comparison of the matched variable.

Controls are usually picked individually, in a "paired" fashion. That is, for each case, one or more controls is picked in some systematic fashion according to preset rules or criteria. In a study of renal calculi, it may be decided to include as controls other urological patients who have no urinary-tract stones or obvious mental impairment due to uremia or other cause and who are matched to the cases with regard to age, sex, race, and date of admission. The paired selection of a matched control for each case might involve selecting the first patient admitted to the urological service after the case, who meets the diagnostic and mental status criteria, who is of the same sex and race as the case, and whose age differs by no more than 5 years from that of the case. Some leeway is necessary in matching for quantitative variables such as age and admission date, or else no match will be found for most cases. Failure to find matched controls will also occur frequently if matching is attempted on more than a few characteristics.

If the disease being studied is known to be uncommon in the group serving as a source for controls, then little, if any, diagnostic effort or documentation is needed to rule out the disease in the selected controls. However, if the disease could occur commonly in controls, at least some attempt to rule it out, such as an interview question or a quick review of the medical chart, is desirable to minimize misclassification.

Data Collection Any source of data about the study variables may be used. As has been mentioned, accurate information collected on both cases and controls before the disease developed is ideal. Collecting information after the disease develops may be necessary, but every effort should be made to avoid qualitative and quantitative case-control differences in the data gathered. For example, if possible, the research assistant(s) recording laboratory data for all study subjects should do so without knowing whether particular individuals are cases or controls. Similarly it may often be desirable to structure data-collecting interviews to avoid discussing disease status altogether, or at least until the questions about etiologic variables have been asked.

Data Analysis Normally, the basic case-control comparison is expressed in terms of the proportion of cases versus the proportion of controls who show a particular characteristic. If the characteristic is quantitative rather than a qualitative "yes-or-no" attribute, then its distribution in cases and controls can be compared, as can the more general descriptions of the distribution, such as the mean, standard deviation, and the median.

Interpretation

If the cases show a higher proportion with an attribute than do the controls or if the distributions or mean levels of an attribute differ, then there is an observed association between the attribute and the disease. Interpreting whether this association implies a cause-and-effect relationship is another matter, involving a number of considerations to be discussed in Chap. 11.

It may seem more convenient or natural to think about the study results expressed, as is usually done in a prevalence or incidence study, as the rate of disease occurrence in persons with a particular attribute compared to the disease rate in those either without that attribute or with a different attribute. In case-control studies the results of comparisons are usually expressed in the converse manner, that is, as the relative frequency of the attribute in the diseased versus the nondiseased. Fortunately, the results of case-control studies can be converted mathematically to comparisons of disease rates, or at least to an expression of relative risk of disease, under certain conditions. These are, that cases and controls are reasonably representative of persons with and without disease in the underlying population and that the disease prevalence rate of the underlying population is known, or at least known to be small. The interested reader should refer to MacMahon and Pugh (1970) for a description of these methods.

As with prevalence studies, case-control studies usually involve

existing disease cases which, as discussed in Chap. 6, p. 80, may differ in a variety of ways from all cases that develop. One way to try to overcome this problem is to include only those cases that first develop or are first diagnosed during the period of data collection. By using only new cases and selecting controls to be representative of the population at risk for developing the disease, the case-control study then aims more directly at determining factors responsible for disease *development*, much like an incidence study. Paradoxically, although this should provide a broader and more representative *spectrum* of cases, it may limit the *number* of cases available for study, resulting in a sample size that is too small to provide reliable data.

It should also be emphasized that the source of cases for the study may be more apt to provide medical care to one type of case than another. For example, cases derived only from a hospital and not from outpatient clinics as well, may have the most severe disease. Thus, while we have emphasized the problems and vagaries of control groups, the characteristics of the case group must also be carefully considered in study design and interpretation.

Example 1: Oral Contraceptives and Thromboembolic Disease

Millions of women now take oral contraceptive tablets to prevent pregnancy. Several questions concerning the safety of these agents have arisen. One of the major areas of concern has been whether or not oral contraceptives predispose to thromboembolic conditions, particularly thrombophlebitis and its possibly fatal sequela, pulmonary embolism. Following the publication of some clinical case reports in the early 1960's it became apparent that epidemiologic studies were necessary to determine whether women who take oral contraceptives are indeed at greater risk of developing these diseases.

Thrombophlebitis and pulmonary embolism *not* secondary to trauma, surgery, or childbirth, develop rather rarely in women during the reproductive years. Thus a prevalence or incidence study of this question seemed impractical, at least as a first approach, since many thousands of women would have to have been studied in order to find an adequate sample of cases. Case-control investigations were therefore undertaken, both in Great Britain and the United States. The U.S. study by Sartwell and his associates is an excellent example of the case-control method.

The investigators decided to include as cases, women, ages 15–44, hospitalized with thromboembolic conditions and discharged alive within the previous 3 years. It was necessary to collect the cases from a large number of hospitals to obtain an adequate sample size. All told, there were 48 participating hospitals in five large eastern cities: Baltimore, New York City, Philadelphia, Pittsburgh, and Washington, D.C. Cases were excluded from the study if they also had a chronic condition possibly predisposing to thromboembolism, such as diabetes mellitus or hypertension, or a recent precipitating event such as surgery, pregnancy, trauma, localized infection, or prolonged inactivity. Reasonable medical evidence for thromboembolism was required, and all cases were reviewed independently by two physicians.

The derivation of the final study group of 175 cases was carefully described by the authors and clearly shows the marked attrition that often occurs between *potential* and *actual* numbers of study subjects. In all, 2,648 women in the desired age range with thromboembolic conditions within 3 years were identified and their hospital records were abstracted. The vast majority of these cases, 2,288, were immediately rejected because of having possibly predisposing conditions, and another 99 were rejected for other reasons, such as sterility (which obviates contraceptive use), death, or having moved from the area. Of the 261 women selected as suitable cases, 72 had to be dropped because the interview could not be obtained and another 14 were excluded because no interview could be obtained from their matched control subjects.

Two matched controls were selected for each case with the expectation that if one could not be interviewed the alternate control would still be available, thus yielding data on one control per case. Matching was done on several criteria:

Hospital	:	same as case
Sex	:	all women
Discharge date	:	same 6-month interval as that of case
Discharge status	:	all alive
Age	:	same 5-year span
Marital status	:	same
Residence	:	(not stated but presumably the same metropolitan area)
Race	:	same
Parity	:	same general class, i.e., no pregnancies, one or two pregnancies, three or more pregnancies
Hospital pay status	:	ward, semiprivate, or private room

Also, controls were excluded in the same manner as the cases, i.e., for chronic diseases possibly predisposing to thromboembolism or for sterility. Most control subjects turned out to have acute medical and surgical illnesses, conditions treated by elective surgery, or traumatic injuries.

Cases and controls were interviewed at home. A variety of questions were asked so as to provide data concerning pertinent variables such as religion, educational level, and smoking habits. To elicit information about contraceptive usage, cases and controls were asked to select from a list of thirteen methods those which they had used within the 2 years before they were hospitalized.

Data analysis showed that the overall frequency of employment of *any* birth-control method was similar in the 175 cases and controls—114 and 101 users of at least one method, respectively—and many women had used more than one method during the 2-year period. While the case-control differences in proportions using each of the other methods were small and not statistically significant, cases did report using oral contraceptives significantly more often than did controls—67 versus 30 women or 38 percent versus 17 percent.

Using a simple formula to compute relative risk, the investigators found that users of oral contraceptives were about four times as likely as nonusers to develop thromboembolic conditions. Furthermore it could be shown that about one-fourth of the total cases would be attributable to oral contraceptive usage if a cause-and-effect relationship were involved. It was, of course, carefully pointed out that the cases studied were a highly selected group, that is, free of predisposing conditions, unlike most thromboembolism cases.

Further analysis showed that the case-control differences in oral-contraceptive use were present in the major subgroups of the study subjects, when the total group was subdivided by such variables as age and marital status. The case-control differences were found for several different thromboembolic conditions including deep thrombophlebitis of the lower extremity, pulmonary embolism, and intracranial vascular conditions.

Example 2: Pedestrians Fatally Injured by Motor Vehicles

In their concern with learning about the diseases which present complex diagnostic or pathophysiologic problems, medical personnel are apt to forget that injuries and death due to gross physical trauma are one of the chief health problems in affluent industrialized societies as well as in "less developed" areas. In particular, accidents are the leading cause of death in children and young adults in the United States. Automobile accidents lead all other types as a cause of death.

The word "accident" implies that physical injuries produced by automobiles and other energy sources are haphazard and uncontrollable. Among those arguing against this fatalistic concept, Haddon has advocated the use of carefully designed and implemented epidemiologic studies as a means of identifying factors responsible for traumatic injuries, so that appropriate preventive measures can be instituted. His research group's interesting study of the characteristics of pedestrians fatally injuried by motor vehicles in New York City is an example of the imaginative use of the case-control method to attack a serious and poorly understood problem (Haddon et al., 1961).

At the time of the study in 1959, little was known about pedestrian-associated or "host" factors related to being struck and killed by a car. Substantial funds were being expended for public education programs and other means of "pedestrian control," without much evidence that these were effective preventive measures. The previous findings that many fatally injured pedestrians had been drinking heavily had not been evaluated in comparison to

the alcohol consumption of the population at risk or, more simply, to that of *noninjured pedestrians*. Likewise, the age distribution of killed pedestrians, with relatively high percentages of young children and elderly adults, had not been compared with the age distribution of all or of nonkilled pedestrians, to determine whether the ***mortality rate***, or risk of being killed, is actually greater in very young and very old pedestrians. Thus, age and blood-alcohol concentration were included among several characteristics that were measured in fatally injured pedestrians and their matched controls in the study to be described.

New York City was a very appropriate place for this investigation. Pedestrian deaths were relatively frequent, and they accounted for about 70 percent of all fatalities in motor vehicle accidents. The case series consisted of 50 adults (18 years of age and older) who were struck and killed by automobiles in Manhattan between May 3, 1959 and November 7, 1959. Autopsy confirmation of the cause of death was required. Of 57 cases initially considered, the 7 omissions consisted of 2 who were killed by bicycles, 1 who was purposely pushed into the path of a car, 1 with unknown site or time of the accident, 1 who died of a coronary occlusion while convalescing from the accident, and 2 who were omitted because of clerical errors.

Four matched controls were selected for each case by visiting each accident site at a later date, but on the same day of the week and as close as possible to the time of day when the accident occurred. All but eight site visits for control selection were completed within 6 weeks of the accident. Thus, controls were matched to the cases for accident site and time. In addition, controls were matched to the accident cases for sex and were limited, as were the cases, to adults.

The practical problems involved in this form of "shoe-leather" epidemiology can best be communicated by the investigators' own description of the control selection and interview procedures:

> The site visits were made by a team of two or three of the authors and one to four medical students working at each location with one or two uniformed members of the Police Department Accident Investigation Squad (A.I.S.).

In visiting each site one of three basic approaches was used. In the first type, that used in many busy neighborhoods, for example, opposite Grand Central Station on a weekday at 6:10 P.M., the entire team arrived and immediately stopped the *first* 4 adult pedestrians of the same sex as the deceased. At such busy sites the group arrived and accomplished its purposes in 5 minutes or less from start to finish.

When the accident site was in a neighborhood in which it was suspected that the group might be seen and avoided, a second approach was used. Under such circumstances, for example, at sites in the Bowery, the group arrived and 'swept the block' stopping successively the *first* 4 adult pedestrians of the required sex who were headed toward or away from the accident site. By pedestrian here and throughout this report is meant a person progressing by walking, not lounging stationary, sitting, or lying down.

In the third approach, used where pedestrian traffic was very light, for example at 108th Street and the East River (F.D.R.) Drive at 1:40 A.M., the group would lounge nearby or sit in a car at or near the site watching for approaching pedestrians, and as each of the *first* 4 of these came into view he, or, where appropriate, she, was quickly approached and stopped.

The site visited was the sidewalk point closest to the exact location of the accident as described on the police or medical examiner's report. For example, one report indicated that the deceased had been crossing the street 40 feet from a given corner. This was found to be directly in front of a 'rathskeller', and it was at that point that the first 4 pedestrians were stopped.

Great care was taken to avoid any attempt at matching for the characteristics of the deceased, except in so far as sex and adulthood were concerned. In addition, for methodologic uniformity, at all sites the same investigator pointed out to the accompanying police each individual to be stopped. Although the exact details varied with the circumstances, the person was immediately approached and told by the policeman, 'Please step over for a minute while the doctors ask you a few questions.' A nearby member of the team immediately stepped up and began talking uninterruptedly: 'I don't want to know your name; I merely want to ask you a few questions. Do you live in Manhattan?' The interview was usually easily begun in this manner, although 12 refusals occurred (for each of which the next pedestrian was substituted) . . . .

This investigation was carried out without publicity of any kind. With one exception it was invariably possible to stop the members of each pedestrian sample prior to the formation of the substantial group of watchers which sometimes formed thereafter. The exception, in a 'tough' neighborhood at 2:30 A.M., involved the only site at which 2 persons had been fatally injured in the same accident. On arrival, it was possible to obtain quickly the first 7 but not the eighth interview and specimen of breath, a small, hostile crowd quickly forming from an adjacent bar. As a result, only the first 4 of the 7 interviews and specimens obtained at this site were used, being counted twice in the analyses of the data.

The interview included questions as to: place and length of residence; place of birth; age; present occupation; and marital status. Sex, apparent race, appearance and apparent sobriety, date, location, time of interview, and weather were also recorded.

Immediately on finishing the interview the interviewer stated approximately as follows, 'I only have one more thing for you to do (and then you can go) and that is to blow up this bag for me.' Simultaneously he removed a Saran bag from an envelope and showed the pedestrian how to place one of its two ends in his mouth and blow until told to stop. This finished, the pedestrian was thanked and told that the interview was over.

A large percentage of those interviewed were foreign born, and many of these admitted to no knowledge of English. Rather than weaken the investigation by omitting these pedestrians when no member of the team knew a common language, passersby were stopped and asked to serve as interpreters. Apparently because those walking in the same neighborhoods or, in some cases, accompanying those stopped (many of the latter being interviewed themselves) tended to know the same languages, this procedure proved very satisfactory. With its use no one failed to be interviewed because of a language barrier and interviews were completed in Armenian, German, Greek, Spanish, and other languages and dialects.

As implied above, blood-alcohol concentrations were measured by analysis of breath specimens and the other data concerning the controls were recorded as described. Data concerning the cases were obtained chiefly from official records describing the accidents.

Postmortem blood-alcohol measurements were studied in those cases who survived less than 6 hours after the accident.

Data analysis for the case-control comparison revealed that, indeed, fatally injured pedestrians were older than the controls, their mean ages being 58.8 years and 41.6 years, respectively. Additional data collected later showed nonfatally injured pedestrians to be intermediate in age, with a mean of 48.4 years. Thus, advancing age appeared to increase the pedestrian's risk both of being struck by a car and of dying once struck.

Regarding the effects of alcohol, significantly higher blood-alcohol concentrations were found in cases than controls. Appreciable increases in risk were noted even at the relatively low levels of 10 to 40 mg/100 cc. Putting together the age and alcohol data it appeared that there were two relatively discrete high-risk groups—the elderly who had been drinking little if any alcohol and the middle-aged who had been drinking heavily.

It was also found that the case group was more often foreign-born and of lower socioeconomic status than the controls, and less often married. However these differences could be explained by age differences between the case and control groups. Weather conditions, rain in particular, did not appear to be associated to any substantial degree with traffic deaths.

In addition to the case-control comparisons, information about the fatally injured group itself was of interest and importance. Only a small percentage lived outside of Manhattan and were commuters or out-of-town visitors. While the accidents were scattered about the city, most occurred outside of major business and shopping areas. The accidents occurred most frequently in the evening and night hours, suggesting the importance of emergency medical care during this time of day.

Evaluation and Role of the Case-Control Method

Case-control studies are the most readily and cheaply carried out of all analytic epidemiologic studies. For rare diseases they may be the only practical approach. Yet the problems involved in selecting appropriate control groups and collecting comparable information on cases and controls are often of such magnitude that the results of case-control studies are open to a variety of legitimate questions and objections, generally more so than the results of prevalence and incidence studies.

Case-control studies have played a vital role in the development of many fruitful lines of study. For example the relationship of cigarette smoking to lung cancer was demonstrated in case-control studies before any incidence studies of this question were carried out. Because of their low cost, case-control studies should often be the first approach to the testing of a hypothesis. Similarly, they are useful for an exploratory study of a variety of variables (sometimes referred to as a "fishing expedition") to find clues and leads for further study.

REFERENCES

Haddon, W., Jr., P. Valien, J. R. McCarroll, and C. J. Umberger. 1961. A controlled investigation of the characteristics of adult pedestrians fatally injured by motor vehicles in Manhattan. *J. Chron. Dis.*, **14**:655–678.

MacMahon, B., and T. F. Pugh, *Epidemiology: Principles and Methods.* (Boston: Little, Brown, 1970), Chap. 12.

Sartwell, P. E., A. T. Masi, F. G. Arthes, G. R. Greene, and H. E. Smith. 1969. Thromboembolism and oral contraceptives: An epidemiologic case-control study. *Am. J. Epidemiology,* **90**:365–380.

Chapter 8

Incidence or Cohort Studies

Of the various types of observational epidemiologic studies, *incidence* or *cohort* studies are generally thought to provide the most definitive information about disease etiology. They do provide the most direct measurement of the *risk of disease development.* However, if carried out prospectively, they can be expensive and time-consuming, requiring a long-term commitment of funds and dedicated personnel. Furthermore, as will be discussed, they are not free of potential biases and other scientific problems.

How Incidence Studies Are Carried Out

Defining the Study Population Initially, a study population or cohort is identified. This population is to be followed up over a period of time for the development of the disease(s) under investigation. The cohort chosen may be a rather general population group,

such as the residents of a community, or a more specialized population that can readily be studied such as an occupational group or group of insured persons. Or, the cohort may be selected because of a known exposure to a suspected etiologic factor such as a source of ionizing radiation or a drug or pesticide. If exposure to the suspected factor characterizes all or virtually all cohort members, then a similar but unexposed cohort or some other standard of comparison is required to evaluate the experience of the exposed group.

The incidence study focuses on disease development. In order for a disease to develop, it must, of course, be absent initially. Thus the study population must be shown, in some way, to be free of the disease, that is, to be a population at risk for disease development. For a rare, rapidly fatal disease such as acute leukemia, a few cases initially present in the population will probably be self-evident. For a more common disease such as coronary heart disease in middle-aged men, an initial examination of the potential study population may be required to find and exclude existing cases of disease. As illustrated by the Evans County study (Chap. 6), this initial examination may be part of a prevalence study.

An initial examination may serve another important purpose. In it, some or all of the potential etiologic factors and other pertinent study variables may be measured. Nevertheless, some cohort studies with certain specific objectives do not require an initial examination since the data necessary to characterize the study subjects are available from other sources.

Follow-up Once the population is initially defined and the appropriate characteristics of its members have been assessed, the population must be followed up for the development of the disease. Follow-up procedures vary from study to study both in intensity and completeness, depending on the disease manifestations to be measured.

Simple, relatively complete follow-up is available for life-insurance-company investigations of factors affecting mortality. For their purposes, death is the only end-point of importance, and it must be reported to the company in order for the policy benefits to be paid.

On the other hand, follow-up to detect all new cases of coronary heart disease or stroke may require several different procedures, including periodic reexaminations, surveillance of deaths, hospitalizations, and physicians' office visits, and correspondence with subjects who have moved from the area. However, limitations on available resources may dictate that only a portion of all possible follow-up procedures be used, perhaps just hospitalizations and deaths, for example. Even though incomplete, such partial follow-up may be perfectly adequate for the purposes of the study.

The duration of follow-up required is determined primarily by the number of disease cases needed to provide reliable, statistically significant answers to the specific questions under study. This can usually be determined in advance, once the study population size and the disease incidence rate is known. For example, if the study population contains 1,000 persons and the incidence rate is 1 percent per year, about 10 new cases may be expected during each year of follow-up. If 100 cases are needed to provide answers with a certain degree of reliability, then the study may be expected to last about 10 years.

This example is somewhat oversimplified and does not take into account such factors as a possible reduction over the years in the number of new cases per year, due to losses of subjects to follow-up, or a possible increase in new cases per year as the population ages, if the incidence increases with age. Although it is often most practical to keep follow-up as short as possible, a study may be designed specifically with a long follow-up period in mind to assess factors which cause or predict disease in the distant future.

During the follow-up period it may be possible to repeat the initial measurements of population characteristics. In this way disease development may be studied in relation both to initial characteristics and to *changes* in these characteristics. For example, it is not only of interest to know whether serum cholesterol level is related to subsequent coronary heart disease, but also whether a *rising* level or a *falling* level adds additional predictive information.

There are other reasons for reassessing population characteristics in the follow-up period. During a long-term study there may be technological improvements in the measuring devices that were used initially. Also, new scientific information about the disease may indicate the importance of measuring additional variables that were not included at first.

Data Analysis As in a prevalence study, the population is subdivided or classified according to the variables that are to be related to the disease. The disease incidence rate is determined for each subgroup, and the rates are compared to see whether the presence or absence (or differences in level, if quantitative) of the variable is related to subsequent disease development. If the study population is a special cohort exposed to a suspected etiologic factor, then its disease incidence is compared to that in a similar nonexposed cohort or to that in the general population.

If all or virtually all study population members are followed up for the same period of time, then a simple overall incidence rate can be used. For example, if the period is uniformly 3 years, then the 3-year incidence rate may be computed for each subgroup. If there are substantial differences among study subjects in length of follow-up, these will have to be taken into account in the data analysis. Follow-up durations may differ markedly when subjects are lost to follow-up before the study is complete—if, for example, they move out of the area or die. Also, some investigations require that new subjects be added to the study population over a relatively long period of time. As a result, if disease incidence is determined up to a specific point in time, subjects will have been followed up for different durations from their time of entry into the study.

The standard method of handling variable follow-up periods involves the use of "person-years" of observation in the denominator of the incidence rate (or person-months or person-days, etc., if more appropriate or convenient). With this approach, each subject contributes only as many years of observation to the population at risk as he is actually observed; if he leaves after 1 year, he contributes 1 person-year; if after 10, 10 person-years.

The assumption involved in adding all subjects' person-years into one denominator is that the disease risk remains relatively constant over time. That is, the third year of observation, for example, is not appreciably different as to disease risk from the first; or, stated in another way, following up three persons for 1 year is equivalent to following up one person for 3 years. The validity of this assumption for any particular study should be considered in evaluating the person-years approach.

Another feature of the person-year method is that one person may contribute person years of observation to more than one subgroup. Suppose, for example, that in a 5-year study, disease incidence is determined for age-decade subgroups. A person entering the study population at age 48 will contribute two person-years of observation to the 40–49-year-old subgroup and three person-years of observation to the 50–59-year-old subgroup. This may also happen with other measurements if they change over time. A person may spend a few years in a particular quartile of serum cholesterol and then shift to a higher or lower quartile.

Interpretation and Evaluation of Incidence Studies

The emphasis in incidence studies is on the prediction of disease development. This type of investigation clearly demonstrates the time sequence between the presence or absence of a factor and the subsequent occurrence of the disease. However, even the prediction of disease does not necessarily imply a cause and effect relationship, as will be discussed in Chap. 11. Furthermore, as has been pointed out, factors associated with a disease can be shown to precede and thus predict the disease in prevalence and case-control studies as well.

A problem that has been emphasized with prevalence and case-control studies is the likelihood of overrepresentation of cases of long duration. This will not be a problem with incidence studies having complete and comprehensive follow-up; the full spectrum of the disease should be available for study.

Despite their good reputation, incidence studies can be subject to important biases. We have mentioned how, in a prevalence or case-control study, the presence or absence of disease may affect the factor under investigation or the measurement of that factor, using the example of cancer and its effects on one's emotional state. In a somewhat analogous fashion, the converse problem may be present in an incidence study. That is, the presence or absence of a study factor may affect the subsequent assessment of disease. This may be especially prone to occur if the decision as to the presence or absence of disease is made by persons who are aware of the subject's status with regard to the study factor.

In a stroke study, for example, it is clearly possible for knowledge of a subject's prior blood pressure to influence, consciously or unconsciously, the decision as to whether or not a stroke has occurred. If this happens, the study will have a built-in correlation between blood pressure and stroke incidence. Similarly, if in a study of cancer, disease detection depends partly upon the initiative or cooperation of the subjects in seeking an examination, those with a family history of cancer or those who smoke might be especially motivated to have a checkup. This can result in bias or in a built-in correlation of the disease with a family history of cancer or with smoking. Thus, every effort should be made to ensure that disease development is detected or decided upon independently of the possible etiologic factors under investigation.

Incidence studies are also subject to possible biases due to loss of study subjects. Such losses may occur initially, if a portion of the target study population does not participate, or later on as members of the study population are lost to follow-up. Marked losses of either type do not necessarily invalidate the study. However, the investigators should consider whether the reasons for loss of subjects might reasonably have affected the study outcome. Sometimes it is possible to gather outside information concerning lost subjects, particularly whether they left due to illness or death or for any reason that might be related to the variables and the disease under investigation.

Example 1: The Framingham Study

Considering the barrage of information about "coronary risk factors" to which the public has been subjected, it may come as a surprise to health-care personnel now in training that only a few decades ago, atherosclerosis and its clinical consequences were generally viewed by the medical profession as degenerative changes that were an inevitable consequence of aging. However, by the late 1940's, descriptive epidemiologic findings and clinical observations were beginning to convince public health authorities that environmental factors might be playing an important role in the disease and that, as a result, prevention was a real possibility. Because of the major importance of coronary heart disease as a cause of disability and death in this country, the U.S. Public Health Service decided to undertake a major long-term incidence study to better define the factors producing this disease.

When the Framingham Study began, around 1950, Framingham, Massachusetts was a town of about 28,000 inhabitants. There were several reasons for selecting this location for the study. At the time, it was a relatively self-contained community with both industrial and rural areas. In this and other ways it was not obviously atypical. There were sufficient numbers of residents in the desired age range to provide an adequate study group. There was evidence, both from a successful previous study of tuberculosis in the community, and from discussions with medical and lay residents, that the townspeople would be cooperative. The area of the town was sufficiently small that the residents could come to one central examining facility. Follow-up of hospitalizations would be relatively easy since most occurred at one central hospital in the town. Furthermore, Framingham was only 20 miles from major medical centers in Boston; thus, medical and scientific consultation would be readily available.

The study was planned to last for 20 years, in view of the slow development of atherosclerosis and its consequences. A long "incubation period" is believed to characterize many of the chronic noninfectious diseases and argues for a long-term study to identify predisposing factors early in life.

The lower and upper age limits of the study population were set at 30 and 60 years. It was felt that older persons should be excluded since many of them already had extensive coronary atherosclerosis and, as a result, to study them would reveal only immediate precipitating factors for clinical events. Persons under thirty were excluded primarily because their incidence of coronary heart disease would be very low and they were a more mobile, hard-to-follow group.

In selecting the study sample, the goal was a group of about 5,000, since this size sample in the 30–60-year age range would produce adequate numbers of cases over the 20-year follow-up period. Knowing that there would be some nonresponse, the investigators selected a larger systematic sample comprising two-thirds of the 10,000 residents of the appropriate ages. The list of town residents was arranged according to precinct, and within each precinct by family according to family-size groups (one member, two members, three or more members, ages 30–60). Two out of every three families were selected. Selection of *families* rather than individuals was a wise decision since (1) one member of a family in the study's age range would not be denied an examination service offered to another member of the same family, (2) many reluctant men received examinations because of being "persuaded" by their more cooperative wives to go to the clinic at the same time, and (3) studies of spouse pairs and familial aggregation of characteristics would be fostered.

The 6,507 members of the sample were invited to participate in the study by town residents who recruited subjects living in their own neighborhoods. These recruiters were part of a group of volunteers who were given a cardiovascular examination at the clinic before the study officially began. Having experienced the examination that was to be given in the study, the volunteer recruiters would be able to describe it to the invited subjects on the basis of personal experience.

Despite this personal approach only 4,469, or about two-thirds of the sample, participated. A group of 740 volunteers were added, yielding a total of 5,209 subjects. The initial examination revealed that 82 subjects already had clinically evident coronary heart disease. These were excluded from the population at risk, leaving a total of 5,127.

This study population has been offered a relatively complete examination every 2 years since the study began. The examination has included a medical history, physical examination, and pertinent laboratory tests such as electrocardiogram, chest x-ray, and serum lipid levels. It has been directed primarily at detecting the development of coronary heart disease and other atherosclerotic conditions such as stroke and peripheral vascular disease. Variables to be related to disease development have also been measured every 2 years. As new types of measurements have acquired importance in this area of research, they have been added to the examination. Thus the investigators have not been limited to the first examination as their only source of information about possible etiologic variables.

Every effort has been made to maintain rapport with the community and with the medical profession in the town. Subjects are kept waiting as little as possible during the examination. A complete report of the examination findings has been sent to each subject's personal physician. No medical care or advice is given by the study's examining physicians except that persons with newly discovered serious abnormalities are advised to contact their own physicians.

Although the biennial examinations at the clinic have been the chief source of follow-up information, disease development has been detected by other means as well. These additional sources include records of hospitalizations and of local physicians' office visits, and information about deaths from death certificates, coroner's reports, and reports of relatives. The diagnosis of any disease studied has been made according to strict criteria so as to include only definite cases in the diseased group.

Maintaining a continuing program of biennial examinations for a few thousand persons has involved a major investment in the operation of the study clinic. A staff of physicians, nurses, laboratory technicians, receptionists, clerical personnel, and others have been necessary for the smooth operation of the clinic and to assure the collection of complete and accurate data. Epidemiologically oriented physicians and statisticians located both on-site and at the National Heart and Lung Institute headquarters in Bethesda, Maryland have carried out the research analyses of data and the preparation of scientific papers.

The study findings have emerged in a large series of reports over the years since 1951 and can only be summarized briefly here. Several representative papers are listed in the references under the first authors, Dawber, Kannel, Gordon, and Friedman.

The study has been able to confirm in great detail that the atherosclerotic diseases do not strike persons at random as they age, but that highly susceptible individuals can be identified in advance of any definite clinical manifestations. Indications of susceptibility, or "risk factors," that have been found in the Framingham Study and other epidemiologic investigations include male sex, advancing age, high serum lipid concentrations, high blood pressure, cigarette smoking, diabetes mellitus (or even milder degrees of carbohydrate intolerance), obesity, low vital capacity, and certain electrocardiographic abnormalities. Other risk factors that have been emphasized more by other studies include certain psychosocial factors, family history of coronary heart disease, and physical inactivity.

The detailed information and large population available at Framingham have permitted more intensive investigation of the unique role of each risk factor. For example, it was found that obesity is not related equally to all manifestations of coronary heart disease. Although it does appear to predispose to angina pectoris and to sudden unexpected death, it is not related to myocardial infarction per se. Also, sufficient numbers of cases emerged to permit the study of interrelationships of several risk factors. One important finding was that persons with combinations of risk factors (for example hypertensive male smokers with high serum lipid levels) are at especially high risk of developing coronary heart disease.

As the study population ages, more emphasis can be placed on the diseases of the elderly such as stroke. Furthermore, the wide scope of information collected in Framingham has permitted the epidemiologic study of other nonatherosclerotic diseases as well, for example, rheumatic heart disease, gout, and gallbladder disease. In addition, several studies of epidemiologic methods have been carried out there.

At present the major research efforts in the epidemiology of coronary heart disease are being switched more and more from observational studies, of which Framingham has been one of the most important, to experimental trials attempting actually to lower the risk of disease. The predictive value of serum lipids, blood pressure, and cigarette smoking have been repeatedly demonstrated. Many feel that it is now necessary to prove that actively changing these characteristics by diet, drugs, and other means will safely lower risk and prevent or postpone atherosclerotic disease before widespread measures are applied to the general public or to high-risk individuals. Thus, at the time of this writing the National Institutes of Health is initiating a large-scale Multiple Risk Factor

Intervention Trial which will be a controlled experiment (see Chap. 9) to evaluate active preventive measures, involving the collaboration of several medical centers in the United States.

While it is generally accepted, then, that enough has been learned about factors predisposing to coronary heart disease to justify serious attempts at prevention, this does not mean that observational epidemiologic studies and other efforts to identify causal factors are no longer needed. There are many individuals developing the disease who by present criteria are at low risk. Conversely, many persons in the apparent high risk groups remain free of clinical coronary heart disease. Thus, our power to predict coronary heart disease is limited, and further studies are needed to identify pertinent risk factors.

Example 2: Mortality in Radiologists—Does Radiation Shorten Their Lives?

As the use of man-made sources of ionizing radiation has increased, so has the concern that these may be producing a variety of adverse effects on life and health (MacMahon, 1967; Whittenberger, 1967). While intense acute exposures have clearly proved to be quite harmful or even fatal, the evidence is less obvious regarding the consequences of chronic exposure to relatively low levels of radiation. Experimental animals subjected to chronic exposure have died sooner than expected, but findings in animals are not always applicable to man.

The effects on man's life-span are clearly a matter requiring epidemiologic study. Laboratory investigations of radiation effects on animals, cells, and other biological or biochemical systems, however important and illuminating, do not answer the basic question, *Does exposure to mild and moderate levels of radiation actually shorten human lives?*

Since the intentional exposure of human beings to radiation for the sole purpose of answering this question is ethically unthinkable, one problem for the epidemiologist is to locate human groups who have been or are being exposed for other reasons, so that their mortality experience may be investigated. Groups already studied for a relationship between ionizing radiation and overall mortality or cancers of various types include uranium miners, residents of Hiroshima and Nagasaki who survived the atom bomb, patients receiving radiation therapy for noncancerous conditions such as enlargement of the thymus gland or ankylosing spondylitis, and children exposed in utero to diagnostic x-rays of their mothers' abdomen and pelvis.

Radiologists have also been studied for possible life-shortening effects. Since the findings of some of the earlier studies of radiologists were inconclusive, either because of small numbers of subjects or because of questionable comparison groups and measures of outcome, Seltser and Sartwell (1965) undertook a study of all members of an organization of radiologists compared to members of other medical specialty societies.

The Radiological Society of North America was the radiologists' organization studied. Founded in 1915, it existed during some of the early years of radiology when many radiologists were much less concerned and self-protective about radiation exposure than they have been more recently. (Some of the old-time radiologists even placed their own hand next to the patient routinely, so that its image on the x-ray photograph would help in judging the exposure time.) It was hypothesized in advance that the radiologists were the high-exposure, *high-risk* medical specialty group. The American College of Physicians has been composed largely of internists and was studied as a probable *intermediate-risk* group, since some physicians in this group have fluoroscoped patients to aid in diagnosis. The hypothesized *low-risk* specialty society was the American Academy of Ophthalmology and Otolaryngology, whose membership would contain only a few persons exposed routinely to radiation.

This investigation is described here as an example of a *retrospective* cohort study, contrasting greatly with the Framingham Study in scope and expense. In this study, all the events to be studied had already taken place and the required data were already recorded.

Because the data were already recorded does not mean that preparing them for analysis was an easy task. Several years of work were required to extract the necessary information from the files of the specialty societies and the American Medical Association's Directory Department. All specialists studied were traced from the time of joining their societies in or after 1915 until the end of 1958, and the time and place of death for all deceased members were noted. The cause of death was determined for over 99 percent of the deceased subjects by obtaining death certificates or reviewing other death records. The study was limited to men.

The end point of this study was, of course, mortality. The data were analyzed in terms of person-years of observation. Each physician was considered to have contributed one-half person-year of observation during the year he joined—a convenient approximation which represents the average—plus a full person-year for each subsequent calendar year survived through 1958. Subjects dying before the end of 1958 were credited with one-half year during the year they died, again a convenient approximation. All told, there were 16,339 physician specialists studied, of whom 3,521 were radiologists. Person-years of observation totaled 232,708, of which the radiologists contributed 48,895.

Mortality rates were summarized for three age groups, 35–49 years, 50–64 years, and 65–79 years as well as for the total group. Similarly, mortality experience was looked at in three separate time periods, 1935–1944, 1945–1954, and 1955–1958.

As hypothesized, the death rate was highest among radiologists, intermediate in internists, and lowest in ophthalmologists and otolaryngologists. The differences were larger in the earlier time periods than in later ones and more apparent in older than in younger men. In fact, after 1944, radiologists in the 35–49-year group showed no increase in mortality over the other specialists of the same age.

The authors interpreted these age and time relationships as being consistent with a cumulative harmful effect of x-ray exposure becoming manifest in later life, and a decreasing or disappearing effect in more recent years due to improvements in equipment, techniques, and safety measures.

It was of interest that the radiologists' death rates were similar to those of all U.S. white males. Since physicians are, on the average, of higher socioeconomic status and probably receive better medical care, they would be expected to show a lower mortality rate than all males. This illustrates the importance of selecting appropriate comparison groups when special cohorts, such as radiologists or other occupational groups, are followed up. Comparison with all men would have revealed no mortality difference. The more appropriate comparison, with other medical specialists, *did* reveal a difference.

Putting the age-specific death rates into one cross-sectional analysis of life expectancy starting at age 40 (see Chap. 5, p. 57) was another way of looking at the data. This revealed a similar relationship to medical specialty. The median age at death for 40-year-olds starting in the three successive time periods, 1935–1944, 1945–1954, and 1955–1958, respectively, were radiologists—71.4, 72.0, and 73.5 years; internists—73.4, 74.8, and 76.0 years; and otolaryngologists and ophthalmologists—76.2, 76.0, and 76.4 years.

Recognizing the limitations of death-certificate diagnoses, the investigators noted that the causes of death for each medical specialist group would probably have been recorded with reasonably equal accuracy. They compared the rates for major causes such as cardiovascular disease and cancer. The mortality ratios for major causes in radiologists as compared to ophthalmologists and otolaryngologists were relatively close to the overall ratio of 1.4 for all deaths.

Leukemia showed a higher mortality ratio—2.5, based on 19 *observed* leukemia deaths in the radiologists as compared to the 7.7 *expected* if the eye and ear group's mortality rates had applied to the radiologists. This is consistent with the results of other studies showing that radiation increases the risk of developing leukemia. It was pointed out, though, that the approximate 11 excess deaths from leukemia (19 observed minus 7.7 expected) constituted only a small fraction of the 228 total excess deaths. Thus, the higher death rate in radiologists appeared to be largely a nonspecific across-the-board increase.

In evaluating the findings, the investigators considered other possible sources of the mortality differences among the specialties, such as place of residence and initial self-selection of a medical specialty on the basis of health. The additional information available suggested that these factors did not account for the relatively shorter life expectancy of radiologists and that occupational exposure to ionizing radiation was the most likely explanation.

The investigators stressed, rightfully, that their findings were enhanced by the fact that they had predicted the outcome in advance. This deserves special emphasis because of the fact that epidemiologists and other scientists can be trapped by the so-called post hoc, or after-the-fact, explanation. Given a set of findings or measurements, the human mind is usually ingenious enough to produce a reasonable theory or explanation as to why they occurred. This is accomplished with special ease in fields like medicine or psychology which deal with systems of great complexity. Quite plausible explanations can be brought forth to explain diametrically opposite observations, and almost any result can be made to appear consistent with someone's pet theory. A much better test of a theory is whether it will predict specific outcomes of a study *in advance*.

This is not meant to detract from the importance of exploring data in order to develop new hypotheses or theories for further study. However, once such hypotheses are arrived at, they sooner or later will have to be tested to see whether they *predict* study outcomes.

Role of Incidence Studies

It should be clear from the description of the Framingham Study why prospective incidence studies of general populations are infrequently carried out. They are difficult and expensive, and require the initial willingness to make a long-term commitment and the continuing patience on the part of both the sponsoring agencies and the study personnel. Yet the investment may well prove its worth in the depth and variety of information that such a study can produce.

The need for either a long-term follow-up or a very large study population or both, rests fundamentally on the fact that most diseases studied in this manner have surprisingly low incidence rates. Coronary heart disease is the leading cause of death in the United States, and coronary atherosclerosis is well known to be common in middle-aged men at autopsy. Yet, the incidence of new *clinically identified* cases of coronary heart disease in middle-aged men is only about 1 percent per year. Similarly, although hypertension is a highly *prevalent* condition in U.S. adults, many hypertensives seem to have drifted gradually into their present state, making it difficult both to define and to find *new* cases in a population for an incidence study.

Retrospective incidence studies, of course, can be accomplished relatively quickly if suitable cohorts can be identified and if adequate data about them are available. Yet many diseases of interest are so rare that case-control studies currently represent the only practical epidemiologic approach to studying them.

It now appears that technological changes will increase the feasibility of cohort studies in the future. Storage of medical and demographic information in computer data banks is becoming an accepted approach to improving the efficiency and quality of medical care. A by-product will be the increased availability of information about a variety of cohorts that can be studied both retrospectively and prospectively. On-going efforts in the area of "record-linkage" (i.e., the combination of a variety of records about each person, such as birth, physical examination, illness, and death records) will increase the number of different relationships that can be studied—relationships between a variety of initial characteristics and a variety of disease outcomes.

REFERENCES

Dawber, T. R., and W. B. Kannel. 1962. Atherosclerosis and you: Pathogenetic implications from epidemiologic observations. *J. Am. Geriat.*, **10**:805–821.

Dawber, T. R., W. B. Kannel, and L. P. Lyell. 1963. An approach to longitudinal studies in a community: The Framingham study. *Ann. N.Y. Acad. Sci.*, **107**:539–556.

Friedman, G. D., W. B. Kannel, and T. R. Dawber. 1966. The epidemiology of gallbladder disease: Observations in the Framingham study. *J. Chron. Dis.*, **19**:273–292.

Friedman, G. D., W. B. Kannel, T. R. Dawber, and P. M. McNamara. 1967. An evaluation of follow-up methods in the Framingham heart study. *Am. J. Public Health*, **57**:1015–1024.

Gordon, T., and W. B. Kannel. 1972. Predisposition to atherosclerosis in the head, heart, and legs. *J. Am. Med. Assoc.*, **221**:661–666.

Kannel, W. B., An epidemiologic study of cerebrovascular disease, in *Cerebral Vascular Diseases, 5th Conference*, edited by C. H. Millikan et al. (New York: Grune and Stratton, 1966), pp. 53–66.

Kannel, W. B., W. P. Castelli, and P. M. McNamara. 1967. The coronary profile: 12-year follow-up in the Framingham study. *J. Occup. Med.*, **9**:611–619.

Kannel, W. B., T. R. Dawber, A. Kagan, N. Revotskie, and J. Stokes. 1961. Factors of risk in the development of coronary heart disease: six-year follow-up experience: The Framingham study. *Ann. Intern. Med.*, **55**:33–50.

Kannel, W. B., E. J. LeBauer, T. R. Dawber, and P. M. McNamara. 1967. Relation of body weight to development of coronary heart disease: The Framingham study. *Circulation*, **35**:734–744.

MacMahon, B., Cancer. Chap. 24 in *Preventive Medicine*, edited by D. W. Clark and B. MacMahon, (Boston: Little, Brown, 1967), pp. 423–426.

Whittenberger, J. L., The physical and chemical environment. Chap. 34 in *Preventive Medicine*, edited by D. W. Clark and B. MacMahon, (Boston: Little, Brown, 1967), pp. 630–638.

Seltser, R., and P. E. Sartwell. 1965. The influence of occupational exposure to radiation on the mortality of American radiologists and other medical specialists. *Am. J. Epidemiology*, **81**:2–22.

Chapter 9

Experimental Studies

Experimental studies resemble incidence studies in that they require follow-up of the subjects to determine outcome. However, the essential distinguishing feature of experiments is that they involve some *action* or *manipulation* or *intervention* on the part of the investigators; that is, something is done to at least some of the study subjects. This contrasts with incidence and other observational studies, where the investigators take no action, but only observe.

Experiments are believed to be the best test of a cause-and-effect relationship. Something is done to an *experimental group* and the observed outcome is presumed to be the effect of that action, provided that the same outcome did not occur in an equivalent *control group* that was not acted upon. A cause-and-effect relationship can also be demonstrated by *removing* or *reducing* the alleged causal factor in the experimental group and showing a disappearance or reduction in the effect, while no change is observed in the control group.

The latter approach is especially relevant to epidemiologic experiments in preventive medicine (Hutchison, 1967). If a factor is removed or reduced and the disease incidence declines as a result, the factor is, for practical purposes, a causal one.

Although great value is placed on experimental evidence, experimental studies are often exceedingly difficult to carry out. In addition, they raise some ethical issues which must be considered.

Ethical Problems

In observational studies, the investigator's chief ethical problem, aside from the need for objectivity and conscientious work, is to maintain the confidentiality of his, records about each person studied. Harm might come to an individual if some of his characteristics, recorded in confidence for medical or scientific purposes, were made available to others, or were communicated to the individual, himself, in an inappropriate manner. In the main, though, the observational epidemiologist is a passive observer of nature with few ethical problems.

The experimentalist's ethical position is quite different, since he takes it upon himself to do something to people. He must have good reason to believe that what he proposes to do has an excellent chance of helping them. On the other hand, he must also have ample doubt about the value of what is to be done, compared to doing nothing or doing what had been done in the past. Otherwise he could not, in good conscience, subject the control group to no action or to the traditional action.

Thus, medical experiments can only be carried out in a situation of uncertainty. Unfortunately, some potential investigators are so convinced as to the benefits of a treatment or preventive measure, that they are unwilling to carry out a controlled experimental test of its effects. Their *feeling* of certainty, even if based on inadequate evidence, makes them reluctant to withhold the treatment from a control group. Similarly, the unreasonable skeptic, convinced of the value of either the traditional treatment or doing nothing, may be unwilling to try new methods on an experimental basis. Both types of "believers" should realize that the failure to carry out a controlled experiment, when it is needed and feasible, is also unethical (Hill, 1971).

Sensitivity to the ethical aspects of human experimentation has resulted in the formation of committees in universities and other research institutions to review and approve all proposed studies of human subjects. It is now commonly believed that whenever possible, the potential subject should share in the decision as to whether he or she should participate in the study. This decision should be made with adequate understanding of the potential risks and benefits involved. Accordingly, informed consent is generally required from experimental subjects or from appropriate relatives or guardians.

How Experiments Are Carried Out

Experimental epidemiology is concerned primarily with testing the efficacy of measures to *prevent* disease. The preventive measure to be tested is applied to a group of persons. The incidence of the disease or disease-related outcome, such as disability, is measured in this experimental, or treated, group.

In order for the experiment to be informative, it must be controlled; that is, the outcome must be compared to some standard to determine whether any benefit has resulted. The standard may be the outcome in another similar group who do not receive the preventive measure. This control group may, instead, receive either no preventive measure or whatever is currently being applied.

Experiments may involve comparisons among several groups. For example, different amounts or dosages of the treatment may be tested. Or, there may be two or more aspects or elements in a preventive program. In this case, each experimental group may receive a different element or combination of elements. Experiments may even be designed in a more complex fashion so that each group receives a variety of treatments in sequence, possibly including periods of time with no treatment (Smart, 1970).

Randomized Control Groups The traditional and most accepted means of defining the treated and control groups is to identify one large group of all study subjects and then divide them randomly into two or more groups. If only chance determines who gets into one group or another, then the usual tests of statistical significance can be applied, to see whether chance could have produced the observed outcome. Random assignment to groups should be done *after* the subjects are shown to be qualified and willing to participate. This will minimize subsequent losses from one or more groups.

If it is crucial that the treated and control groups be equivalent with regard to certain characteristics that might affect the outcome, the entire study population can be divided, or stratified, into subgroups and each subgroup can then be randomly divided into treated and control subjects. For example, stratification into age subgroups can be accomplished to assure that the treated and control groups have similar age distributions.

If after randomization has taken place, the experimenter would like to be sure that some nonstratified crucial characteristic is similar in the treated and control groups, he should examine the distribution of this characteristic in the two groups. If crucial characteristics differ appreciably, then the experimenter had bad luck in the randomization process. Randomization may have to be repeated, or if not possible, the results of the experiment will have to be analyzed in a way that takes into account the differences in these important characteristics. Appropriate analytic methods are discussed in Chap. 11.

Nonrandom Control Groups Randomized control groups are not always available for epidemiologic experiments. The reason may be economic. Funds may not be adequate for careful follow-up of both a treated and control group of adequate size. Or, the extra assurance that can be provided by this more ideal method may be judged to be not worth the cost involved. Also, there may not be enough subjects available for the two groups.

Even if there are enough subjects and enough money, randomization into subgroups may be impossible or may fail in actual practice. Randomization is impossible if the preventive measure can be applied only to the entire population, as when something is added to the water supply of a total community. Or, learning of the preventive measure through conversations with members of the treated group or through publicity campaigns, the control group may adopt the preventive measure to almost the same extent as does the treated group.

If randomized control groups are not used, alternative standards of comparison are available. A comparison group may be selected from persons known to be similar to the experimental group with respect to several pertinent characteristics such as age, sex, occupation, and social class. Or, if the preventive program is applied to an entire community, a similar untreated community may be used as a control.

Another approach is to have the experimental group serve as its own control. That is, a before-after comparison is made, in which there is a baseline period of observation on the experimental group before any preventive program is applied. The disease experience during this period can be compared with what happens after the program is put into effect.

Even when a separate comparison group is used, a baseline observation period is helpful. If systematic differences between the groups are noted during the baseline period, these can be taken into account in comparing the groups after the preventive measure is applied.

Possible biases or underlying group differences should always be searched for when nonrandom control groups are used. Having a group serve as its own control seems especially attractive, since this appears to eliminate virtually all group differences. However, the control and experimental observations are made during different time periods. Thus, there is the real danger that with the passage of time, other things have happened to the study group leading to the appearance of benefit from the preventive measure when none exists, or conversely, masking true benefits. Rapid changes in diagnostic and treatment methods or even in ways of life are the order of the day; these may result in real or apparent changes in disease incidence that have nothing to do with preventive methods being tested.

Subject Cooperation Many preventive measures require the cooperation or active participation of the study subjects. Experimental evaluations of these measures must take into account the failure of many subjects to cooperate. Even after initially agreeing to participate, persons drop out of the study for a variety of reasons. Also, in the treated group there will be those who take none or only part of the treatment. Similarly, in the control group there may be some who openly or surreptitiously obtain the treatment on their own.

Study of outcomes should not be limited to the cooperators in each group since they represent a self-selected subgroup, often characterized by higher educational level, higher socioeconomic status, more concern about health and better health habits. Furthermore, if the preventive measure is eventually adopted, it will be applied in the "real world," which also has its full share of noncooperators.

Thus, the most important comparison to be made is of the *entire* study group versus the *entire* control group. This will provide the best estimate of the overall benefit to be obtained from the preventive measure if it is put into practice.

Blind Experiments If possible, experimental subjects should be kept unaware of whether they are treated or control subjects. Then, their own prejudices or enthusiasms will not result in behavior that promotes or inhibits the recognition of disease outcomes. Often, however, the nature of the treatment makes it impossible to keep the subjects "blind" to their assignment to treated or control groups.

More important is that the *assessment* of outcome be blind. Whenever possible, the physicians or others who determine whether the disease outcome has occurred should be unaware of whether the individual is a treated or control subject. The use of objective tests and criteria for diagnosis will help prevent any bias in favor of the treated or control group.

Even when experiments are designed to be blind, the subjects or their evaluators often become aware of their status. If drugs are involved in the treatment, characteristic side effects may reveal their identity. Also, unbeknown to the investigator, medical personnel involved in the care of the subjects may have access to the code or other information which identifies treated and control groups.

Thus, blind experiments are often desired but less often achieved. As for any type of study, careful evaluation of methods and results for possible bias is necessary.

The term "double-blind" is frequently encountered. Some authors use it to refer to experiments where both the assignment to treatment or control group and the assessment of results are blind. Others use it to refer to experiments in which neither the patient nor the physician knows whether the patient is in the experimental or control group.

Sample Size Considerations and Sequential Analysis Statistical methods are available for determining in *advance* how large the treatment and control groups must be, to obtain answers of the desired precision (Ipsen and Feigl, 1970). In general, the more subjects, the greater assurance that the results of the experiment are accurate and not subject to chance variation.

The desirability of having large numbers of subjects is counterbalanced by practical considerations of cost and difficulty. Ethics also enter into decisions about sample size. The more subjects included, the more who will have received the inferior treatment, if either the experimental or control regimen proves to be better.

Sometimes subjects are brought into an experiment over a relatively long period of time rather than all at once. The results for the subjects who started early may be available before the experiment is completed as planned. It is tempting to peek at early results for a few subjects and end the experiment if a difference between experimental and control groups is apparent. Unfortunately, these preliminary findings will not have the accuracy that was originally planned and agreed upon for the experiment. Stopping the experiment at this point may seem economically or ethically justified, but unless the differences noted are striking and compelling, the investigators may later regret reaching a conclusion on the basis of incomplete data. On the other hand, treatment-control differences may be much greater than originally expected, and therefore accurately demonstrable on a small number of subjects. The investigators would certainly not wish to continue the experiment, if they could be sure that this were the case.

Sequential analysis is a relatively new statistical method which allows an experiment to be ended as soon as an answer of the desired precision is obtained. The result of the comparison of each pair of subjects, one treated and one control, is looked at as soon as it becomes available and is added to all previous results. A criterion for deciding in favor of either the experimental or control treatment is specified in advance with the desired degree of accuracy. The comparison of a relatively small number of pairs may show sufficient differences to permit the decision to be reached. If not, the results for each additional pair are added as soon as they become available until the decision criterion is met, or until it becomes apparent that there is no appreciable difference. As soon as any conclusion is reached, the experiment is stopped. The use of sequential analysis in medical experiments is described further by Armitage (1960) and Smart (1970).

Example 1. Controlled Field Trials of Poliomyelitis Vaccine

The first poliomyelitis vaccine that was widely used in the United States was the injectable vaccine containing inactivated virus, developed by Dr. Jonas Salk. By 1953, evidence had accumulated that this vaccine could be safely administered to man and that it stimulated the production of antibody that protected against the three known types of poliomyelitis virus. What was needed next was an experimental trial of the vaccine to demonstrate whether it was safe and effective when put into general use.

A large-scale cooperative field trial was undertaken in 1954, coordinated by the Poliomyelitis Vaccine Evaluation Center at the University of Michigan (Francis et al., 1955). Through the cooperation of state and local health authorities, over 200 areas participated. These were selected partly because they had experienced higher than average poliomyelitis incidence rates in previous years.

The initial plan was to inoculate school children in the second grade and observe the first- and third-graders as a control group. Although this would not permit a blind assessment of outcome, many states had agreed to participate on this basis, and this procedure was carried out in 127 counties or towns in 33 states (called "observed areas"). Eleven states were willing to cooperate in a blind experiment with a randomized control group. In the 84 counties and towns in this latter group (called "placebo areas"), participating children in the first through third grades would all receive a series of three injections, but half would receive the vaccine and half would receive an inactive *placebo*, or *dummy*.

All children in the first through third grades of the participating schools were first identified by means of a "registration form" on which was also recorded birth date, sex, race, and previous history of poliomyelitis or disability. Each child was to give a "participation request" form to his parents. This form described the observed or placebo study and provided space for the parent to sign a request that his child participate in the study. A vaccination record form was used to record all inoculations given to each participant.

Unique identification of each child on all the forms, plus cross-checking and editing of the information was carried out to ensure a high degree of accuracy. In this study there were 200,745 vaccinated and 201,229 receiving placebo among the 1,829,916 first- to third-grade children in the placebo areas, and 221,998 vaccinated second-graders and 725,173 first- and third-grade controls among the 1,080,680 first- to third-graders in the observed areas.

The vaccination phase took place between April 26, 1954 and June 15, 1954. Participating children in each classroom received vaccine or placebo from numbered vials in such a way that all three injections would be of the same material. In the placebo areas, there were vaccinated and placebo children in virtually every class. The vial code numbers could be interpreted as representing vaccine or placebo only at the Evaluation Center. Pre- and post-inoculation blood specimens were obtained from a sample of children to assess antibody response.

During follow-up, through the rest of the year, uniform procedures were instituted to detect and investigate all suspected cases of poliomyelitis among first- through third-grade children, regardless of their participation or vaccination status. The Evaluation Center was notified of *all* suspected cases plus *all* deaths from any cause. Each local health department arranged for the complete investigation of each case. The data collected included (1) a complete clinical report including history, physical examination, and spinal fluid findings; (2) laboratory specimens, including stool and blood samples for viral and antibody studies; (3) examinations by a physical therapist to classify the patient according to physical disability; and (4) autopsies, when obtainable for fatal cases.

Checking systems plus a good deal of correspondence with physicians and other persons involved were required to make certain that the data collected were complete. By December 31, 1954 290 case records of the total of 1,103 reported were still incomplete. A campaign of telegrams, telephone calls, letters, and field visits reduced the number of incomplete reports to 78 by the end of January, but the last delinquent report was not received until March 9, 1955.

Criteria were drawn up for interpreting the laboratory and clinical findings, and on the basis of these, the investigated cases were classified as either "not polio," "doubtful polio," "nonparalytic polio," or "paralytic polio." Paralytic cases were further divided into spinal, bulbar, bulbospinal, and fatal. These decisions were all made without knowledge of the vaccination status of the children.

The experiment clearly established the benefits of the vaccines. In the placebo areas the incidence of poliomyelitis was less than half as great in those who were vaccinated (28 per 100,000) as in those who were given placebo (71 per 100,000). Similarly, in the observed areas the incidence was 25 per 100,000 in the vaccinated second-graders and 54 per 100,000 in the first- and third-grade controls. These differences were highly significant statistically. The protection appeared to be only against paralytic poliomyelitis, since there were no appreciable differences between vaccinated and controls in the incidence of nonparalytic disease.

Supporting evidence for the vaccine's effectiveness was obtained from the antibody studies. Furthermore, cases occurring among the vaccinated tended to occur in children who received vaccine which was independently judged less effective, on the basis of antigenic response. Other detailed analyses revealed that the vaccine conferred greater protection against more severe forms of paralysis and that older children appeared to benefit more than younger ones.

No ill effects of the vaccine could be demonstrated. School absenteeism for 6 weeks after the inoculations did not differ significantly among the vaccinated, placebo, and noninoculated populations. Nor was there any difference in the occurrence of rashes or other allergic manifestations, which were very rare despite the presence of small amounts of penicillin in the vaccine and placebo. Other symptoms and illnesses at the time of the injection series were quite unusual and occurred no more often in the vaccinated than in the placebo group. The minute quantities of kidney protein in the vaccine caused some concern about possible side effects on the kidney, but none could be demonstrated in the study, nor could any deaths be reasonably attributed to the vaccine.

This study represents a major achievement in experimental epidemiology. The low incidence of poliomyelitis required that a very large population be studied to provide adequate cases to reliably demonstrate the vaccine's effectiveness. Coordinating a large-scale field trial of this nature is a difficult undertaking. This summary has emphasized study design and data collection efforts, but major problems of a logistical nature should not be forgotten. For example, hundreds of thousands of children all over the country had to be supplied with the right vaccines at the right times, and thousands of blood specimens had to be drawn and transported to 28 different laboratories.

Example 2. Fluoride and Tooth Decay

Experimental studies to test the effects of adding fluorides to community water supplies were begun around 1945. The expectation that raising the fluoride concentration of drinking water to one part per million would safely lower the incidence of tooth decay was based on a number of previous observational studies. These studies had demonstrated that ingestion of water containing large amounts of fluorides during the years of tooth enamel calcification resulted in discoloration and even pitting of the teeth. However, these "mottled" teeth appeared to be quite resistant to decay. Comparisons of dental status in communities with differing fluoride concentrations in their drinking water showed that where the level was about one part per million, the decay rates were relatively low and no disfiguring mottling of the enamel was apparent.

On the basis of these findings the water supply of certain low-fluoride communities was treated on an experimental basis to bring the fluoride concentration up to the desired one-part-per-million concentration. Since randomized control groups could not be obtained for these studies, the experiment was controlled by concurrently measuring dental health status in similar but untreated low-fluoride communities. Furthermore, the dental health of children in the treated communities was assessed before the addition of fluoride, to provide a before-after comparison. Still another comparison was made of each treated community with Aurora, Illinois, where the naturally occurring fluoride concentration in water was 1.2 parts per million and relatively little tooth decay was observed. One of these investigations, the Newburgh-Kingston Caries-Fluorine Study (Dean, 1956, Hilleboe, 1956, Schlesinger et al., 1956, Ast et al., 1956) will be described here.

The cities studied, Newburgh and Kingston, New York are located on the Hudson River about 35 miles apart. Each had a population of about 30,000. Newburgh agreed to serve as the treated community, and beginning May 2, 1945, sodium fluoride was added to its drinking water to raise the fluoride content from about 0.1 part per million to 1.0–1.2 parts per million. Kingston agreed to serve as the control community, and its water supply with a fluoride concentration of about 0.1 part per million was left unchanged.

During the year prior to adding fluoride, baseline dental examinations were carried out on the public and parochial school children, ages 6–12, in both communities. Baseline pediatric examinations were performed on smaller samples. Kingston and Newburgh children were, at first, similar regarding both general health and the prevalence of tooth decay.

Periodic assessments of both dental and other health measures were made subsequently. Although the caries experience in Kingston children remained relatively stable, a continuing improvement was noted in Newburgh.

A final evaluation was carried out after the experiment had gone on for 10 years. Over 2,000 children, ages 6–16 were given dental examinations in each community. They were selected by taking every second school child who was present on the day of the examination. Although the clinical dental examinations were not conducted in a blind fashion, x-rays were taken and were randomized at the state health department so that the interpreters would not know whether they were reading Kingston or Newburgh films.

The data analysis was carried out for separate age groups. The Newburgh subjects, ages 6–9, had used fluoridated water all their lives. The older age groups had been exposed to fluoridation starting at later periods in their dental development, and thus might be expected to show less benefit.

The efficacy of fluoridated water in preventing dental decay was clearly shown in this experiment. One of the indexes of the prevalence of tooth decay was the number of decayed, missing, or filled (DMF) permanent teeth per 100 erupted permanent teeth. For the 6–9-year-olds, this measure was 23.1 in Kingston and 10.0 in Newburgh, a relative reduction of 57 percent of the Kingston rate. The reduction in Newburgh was present in all age groups but was relatively less in older children. Thus the DMF rates in 16-year-olds were 58.9 in Kingston and 34.8 in Newburgh, a relative reduction of 41 percent of the Kingston rate. The Kingston-Newburgh differences were found in both the clinical and x-ray examinations.

Dental-caries prevalence rates in Newburgh and other communities with experimental water fluoridation programs were reduced to levels very similar to those noted in Aurora, Illinois. Thus, artifically fluoridated water was also shown to have the same benefit as observed for the naturally occurring fluoride.

Adverse effects of fluoridation were also looked for. There were no instances of disfiguring dental fluorosis or mottling. About 18 percent of the Newburgh children were found to have questionable or mild fluorosis when examined by an expert trained in detecting the effects of fluoride. The mild changes noted would have been hardly noticeable to the average dentist. On the other hand, 19 percent of children in Kingston had nonfluoride opacities or circular patches in the enamel which would have been obvious even to the untrained eye. These were found in only 8 percent of Newburgh children.

The medical examinations, x-ray estimates of bone maturation, measures of growth and development, eye and ear tests, blood counts, and quantitative studies of urinary excretion of albumin, red blood cells, and casts, all revealed no significant differences between Kingston and Newburgh children. Vital statistics data showed no consistent differences between the two communities in cancer and cardiovascular-renal death rates or in infant mortality, maternal mortality, or stillbirth rates.

These community studies present rather convincing evidence of the benefits of water fluoridation. They illustrate how well-designed preventive medical experiments can be carried out even when randomized control groups are not available.

Example 3: Evaluating the Periodic Multiphasic Health Checkup

An experiment to evaluate the long-term effects of periodic multiphasic health checkups is currently in progress at the Kaiser-Permanente Medical Care Program in northern California. Although the results are only beginning to appear at the time of this writing, this experiment is described to introduce the reader to studies of preventive medical services that go beyond the prevention of single diseases.

It is widely accepted in the United States that annual physical examinations are an important means of maintaining good health. The rationale for annual checkups is that the physician may detect early or asymptomatic disease and initiate treatment before serious consequences develop.

Because of this belief, many persons request and expect annual checkups as part of the medical-care services they receive. Providing checkups to large numbers of patients can consume a substantial proportion of a physician's time—time that might also be used to provide more care of the sick. Because of the growing awareness in this country of the high costs and limitations of physician time and medical care resources, efforts to simplify the checkup are being developed and evaluated. Along these lines, paramedical personnel and automated instruments are being used to assist in examinations in order to save physician time.

Yet the basic question still remains as to just how much overall benefit periodic checkups actually offer. While common sense supports the value of early disease detection and treatment, physicians must also conclude that at least some aspects of checkups (such as listening to the heart and lungs of a young healthy patient every year, year after year) are almost always a waste of time.

The available scientific data on this question are surprisingly limited. A few studies have shown reductions in mortality and in other unfavorable outcomes in groups who received periodic health examinations. However, the comparison groups have not been randomly selected but have been superficially similar populations not receiving examinations. Persons who receive examinations have been shown to be like volunteers and other "cooperators" in that they tend to be more educated, more health-conscious, less prone to smoke cigarettes, and so on. Thus, serious questions can be raised about the comparability of the examined and nonexamined populations in these earlier studies.

In the Kaiser-Permanente experiment, the control group is quite comparable to the examined, or "study," group. Both groups of over 5,000 subjects were selected on the basis of having certain digits in their medical record numbers, a systematic sampling method that is equivalent to random sampling, since these numbers are assigned in sequence with no relationship to any personal characteristics. These two samples were drawn from a large pool of Kaiser Foundation Health Plan members living in Oakland, Berkeley, and San Francisco, California and aged 35–54 when the study started in 1964. To minimize losses to follow-up, another selection criterion for potential study subjects was that they must have been Health Plan members for at least 2 years, since persons quitting the Plan tend to do so soon after joining.

Each study-group subject has been telephoned and urged to have a multiphasic health checkup every year. Control-group subjects have not been urged or reminded to have these checkups, but, of course, they are entitled to receive this service if they so choose. On the average, 20 to 24 percent of the control group have sought this service each year, and during the first 7 years of the study, the average number of examinations received per subject was 1.34, with 47 percent of control members having received none. In contrast, 60 to 70 percent of the urged study group have been examined annually, and the average number of examinations per subject in 7 years was 3.54, with only 17 percent of study group having had no examinations. Thus the urging has resulted in a considerably larger "dosage" of multiphasic checkups for the study group.

Follow-up of the two groups has consisted of a number of components to measure the development of morbidity, mortality, and disability and to assess the utilization and costs of all medical-care services. Hospitalizations and outpatient visits are tabulated, and the names of all persons lost to follow-up are sent to the state health department for a check against death certificate lists to see if they have died. A questionnaire survey is sent to both groups at approximately 2-year intervals to learn of the development of disability and other pertinent problems.

Whenever possible, assessment of various outcomes is made in such a way as to avoid bias in favor of study or control group. For example, even though submitting subjects' recent addresses would help the state health department search for deaths, this is not done, since the annual telephone contact with the study group leads to more accurate and up-to-date information about addresses than is available for the control group.

As mentioned, this study is still in progress. Results in the first 7 years show that the checkup program has had an impact on the discovery and diagnosis of a variety of conditions. The older men in the study group, those aged 45–54 when the experiment started, showed some benefit from these examinations in the form of less disability and time lost from work than was experienced by the older control group men. There also appeared to be some reduction in the study group of mortality from conditions that would be expected to be influenced by early detection and therapy, such as hypertension and its complications. Economically, the added costs of the examinations were more than made up for by the greater earning power of the examined group due to their diminished disability and mortality (Cutler et al., 1973, Ramcharan et al., 1973, Dales et al., 1973, Collen et al., 1973).

REFERENCES

Armitage, P., *Sequential Medical Trials.* (Springfield, Ill.: Charles C Thomas, 1960).

Ast, D. B., D. J. Smith, B. Wachs, and K. T. Cantwell. 1956. Newburg-Kingston caries-fluorine study XIV. Combined clinical and roentgenographic dental findings after 10 years of fluoride experience. *J. Am. Dent. Assoc.*, **52**:314–325.

Collen, M. F., L. G. Dales, G. D. Friedman, C. D. Flagle, R. Feldman, and A. B. Siegelaub. 1973. Multiphasic Checkup Evaluation Study: 4. Preliminary cost benefit analysis for middle aged men. *Preventive Medicine*, **2**:236–246.

Cutler, J. L., S. Ramcharan, R. Feldman, A. B. Siegelaub, B. Campbell, G. D. Friedman, L. G. Dales, and M. F. Collen. 1973. Multiphasic Checkup Evaluation Study: 1. Methods and population. *Preventive Medicine*, **2**:197–206.

Dales, L. G., G. D. Friedman, S. Ramcharan, A. B. Siegelaub, B. A. Campbell, R. Feldman, and M. F. Collen. 1973. Multiphasic

Checkup Evaluation Study: 3. Outpatient clinic utilization, hospitalization and mortality experience after seven years. *Preventive Medicine*, **2**:221–235.

Dean, H. T. 1956. Fluorine in the control of dental caries. *J. Am. Dent. Assoc.*, **52**:1–8.

Francis, T., Jr., R. F. Korns, R. B. Voight, M. Boisen, F. M. Hemphill, J. A. Napier, and E. Tolchinsky. May 1955. An evaluation of the 1954 poliomyelitis vaccine trials: Summary report. *Am. J. Public Health*, **45**:(No. 5, Part 2)1–63.

Hill, A. B., *Principles of Medical Statistics*, 9th ed. (London: Oxford University Press, 1971), Chap. 20.

Hilleboe, H. E. 1956. History of the Newburgh-Kingston caries-fluorine study. *J. Am. Dent. Assoc.*, **52**:291–295.

Hutchison, G. B., Evaluation of preventive measures, in *Preventive Medicine*, edited by D. W. Clark and B. MacMahon (Boston: Little, Brown, 1967), pp. 39–54.

Ipsen, J., and P. Feigl, *Bancroft's Introduction to Biostatistics*, 2d ed., (New York: Harper and Row, 1970), pp. 180–184.

Ramcharan, S., J. L. Cutler, R. Feldman, A. B. Siegelaub, B. Campbell, G. D. Friedman, L. G. Dales, and M. F. Collen. 1973. Multiphasic Checkup Evaluation Study: 2. Disability and chronic disease after seven years of multiphasic health checkups. *Preventive Medicine*, **2**:207–220.

Schlesinger, E. R., D. E. Overton, H. C. Chase, and K. T. Cantwell. 1956. Newburgh-Kingston caries-fluorine study XIII. Pediatric findings after ten years. *J. Am. Dent. Assoc.*, **52**:296–306.

Smart, J. V., *Elements of Medical Statistics*, 2d ed. (London: Staples Press, 1970) Chaps. 5, 8, 10–12.