Epidemiologic Reviews
Vol. 20, No. 1
Printed in U.S.A.

Population-based Cohort Studies

Moyses Szklo

INTRODUCTION

In population-based cohort studies, a sample, or even the entirety, of a defined population is selected for longitudinal assessment of exposure-outcome relations. Because of their typically high cost and logistic complexities, population-based cohort studies generally evaluate multiple hypotheses, some defined a priori as well as some suggested in the course of the study either based on interim analyses or on advances in the field, particularly when data are repeatedly collected on cohort members. Perhaps the most often cited justification for conducting a population-based study is its external validity—that is, the applicability of its results to a defined population.

The study of a cohort which is representative of a defined population offers three additional advantages. 1) It allows the estimation of distributions and prevalence rates of relevant variables in the reference population (information on risk factors is used for the calculation of population attributable risks) (1). 2) Risk factor distributions measured at baseline in a study involving periodic examinations of the cohort can be compared with distributions in future cross-sectional samples so as to assess risk factor trends over time (2). 3) A representative sample is the ideal setting in which to carry out unbiased evaluations of relations, not only of confounders to exposures and outcomes, but also among any other variables of interest, even those which were not specified in the original study hypotheses.

This review will briefly summarize some key features of population-based cohort studies, and then the main rationale for conducting these studies, that is external validity, will be discussed.

DEFINITION OF THE POPULATION

Sampling frames for population-based cohort studies include any well-defined population. Such populations encompass those that are defined by geographic boundaries (e.g., the Atherosclerosis Risk in Communities (ARIC) study cohort (2)) as well as those defined by other criteria, such as membership in health maintenance organizations (e.g., studies conducted among Kaiser Permanente enrollees (3)) or occupation (e.g., studies of occupational cohorts). For space reasons, however, this review will focus mostly on geographically-defined cohort studies (e.g., the Framingham study).

SAMPLING AND RECRUITMENT

Although it is sometimes possible to study whole populations, particularly when record linkage on a population basis can be done (e.g., the Washington County, Maryland, studies linking total population census to cancer registry data (4)), most population-based studies cohorts are based on probability samples of the reference populations. The exact sampling approach, however, varies from study to study, or even among sites in the same study. For example, in one of the sites of the ARIC study, Washington County, two sources were used for comprising the population frame, a private population-wide census and a list of individuals possessing driver's licenses provided by the local motor vehicle administration office (2). In other sites, the ARIC cohorts were identified by means of area sampling (Forsyth County, North Carolina) or ascertainment of persons eligible for jury duty, that is, with driver's licenses, voter registration cards, or identification cards (selected suburbs of Minneapolis, Minnesota).

Other strategies for selection of a representative cohort include telephone or door to door recruitment, such as done in the Coronary Artery Risk Development in Young Adults (CARDIA) study (5); and for studies involving elderly cohorts, the Medicare eligibility lists of the Health Care Financing Administration (HCFA) have often been used as a sampling frame (6). As discussed later in this review, attempts at

Received for publication December 22, 1997, and accepted for publication April 20, 1998.

Abbreviations: ARIC, Atherosclerosis Risk in Communities; CARDIA, Coronary Artery Risk Development in Young Adults; CHS, Cardiovascular Health Study; HCFA, Health Care Financing Administration; NHANES I, first National Health and Nutrition Examination Survey; NHEFS, National Health and Nutrition Examination Epidemiologic Follow-up Study.

From the Department of Epidemiology, School of Hygiene and Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205. (Reprint requests to Dr. Szklo at this address.)

including representative cohorts through random sampling are often marred by poor response rates at baseline and losses during follow-up.

It is important to point out that the extent to which a cohort sample is representative of the total reference population depends on the completeness of the population frame available to the investigator. For example, if a substantial proportion of the reference population does not have a drivers' license, and if those who are licensed differ from those who are not licensed, use of motor vehicle administration listings, such as done in one of the ARIC field centers (2), will not result in a representative cohort sample. To cite another example, when selecting the cohort by random digit dialing, the strategy often relies on the unwarranted assumption that those who have telephones in their homes comprise a representative sample of the total reference population (7). An additional important consideration is whether sampling fractions varied for different subgroups of the reference population, making it necessary to use sample weights when estimating average population parameters based on the cohort.

Maximizing cohort recruitment

The selection of a random sample is merely the first step toward achieving cohort participation. The importance of keeping good relations with the community was recognized by the original Framingham investigators, who established an "executive committee" broadly representative of the various groups in the community (8). Such committees provide a mechanism to involve cohort participants in the discussions of the public health, logistics, and ethical aspects of the study. Translating key results into lay people's language, which can then be communicated to cohort participants (for example, by means of mass media approaches or newsletters) is yet another strategy to improve initial participation and minimize losses to follow-up. In studies where a large portion of the cohort participants is expected to develop health problems leading to participants' inability to come to the study clinics, visits to study sites can be replaced with home visits using simplified procedures.

Another important strategy relates to securing the support of health care providers in the study communities, particularly in view of the need to refer cohort members with values of key health variables outside normal bounds (e.g., systolic blood pressure levels ≥140 mmHg) to these providers. In this context, investigators and staff should underscore to the health care providers that health care is not provided to cohort participants by the study and that the data are being collected for research purposes only. In the Framingham study, the joint charges of a "professional committee" of physicians and dentists and the "executive committee" formed by community representatives aptly summarize the main issues related to community participation in population-based studies: 1) to assist in planning a program which would be acceptable to the community, 2) to interpret the goals of the study in a way that would be understandable to the community, and 3) to bring community leaders into active participation in the organizational aspects of the study (8).

DATA COLLECTION AND FOLLOW-UP

In population-based cohort studies, the best strategy to avoid medical surveillance bias when obtaining data on outcomes (9) is to carry out systematic, standardized, and periodic data collection procedures on all, or a sample of, cohort members regardless of exposure status.

As for cohort studies in general, data collection in population-based studies relies both on linkage with available databases (e.g., hospital records, death certificates, cancer registries) and on procedures designed especially for ascertaining study variables. Examples of the former are the Washington County studies taking advantage of community-wide data collected as part of previous or ongoing studies, two population censuses carried out in 1963 and 1975, and a cancer registry (4). Serum and data banks are, for example, available for population-based studies done in 1974 ("CLUE I") and 1989 ("CLUE II") which included close to 26,000 and 33,000 individuals, respectively. Data from these sources have been cross-linked as well as linked to death certificate data and data from other population-based studies, thus allowing prospective and nested case-control examinations of numerous hypotheses (10–12).

As an example of an historic cohort study using available data from the Washington County censuses of 1963 and 1975 and county-wide mortality data, Helsing et al. (13) assessed the relation of widowhood to survival between the census years. The population-based nature of this inexpensive study lent considerable credence to its principal finding that widowed men, but not women, when compared with their married controls, had a shortened survival. Study participants were not followed up individually, but because a 5 percent sample of households included in the 1963 census was contacted 8 years later to ascertain their vital status and residence, it was possible to adjust the population figures in the study of Helsing et al. (13) for losses. This lead Comstock et al. to remark that "while the need for. . .individual follow-up is often cited as a necessary evil of prospective studies, it is not necessarily so" (4, p. 1027).

In other population-based cohort studies, data collection procedures have to be exclusively designed for the study purposes. For example, in the ARIC study, data on putative and hypothesized risk factors are obtained for all cohort members in study clinics at 3-year intervals (2). In addition, data collection procedures are carried out between clinic visits, including annual telephone interviews and ongoing death certificate and hospital record searches.

When cohort studies involve periodic contacts with study participants, it is possible to change the information collected from contact to contact, thus making it easier to test new hypotheses. For example, in the ARIC study, cognitive function tests were applied in follow-up visits, but not at baseline (14). For some variables, it may not be necessary to collect information in each contact; this decision depends, among other considerations, on their relative importance and rate of change over time. For "stable" variables, such as educational level in middle aged or elderly subjects, baseline information may be sufficient. Data needed to assess the comparability between individuals who are subsequently lost to follow-up, and those who are not, must also be obtained in every contact. In cohort studies involving both home and clinic visits, the protocol should specify key variables to be collected in the home, lest the cohort participant fail to come to the study clinic; this consideration is particularly important in studies of elderly cohorts (6).

The need for high quality, standardized data collection procedures makes quality assurance and quality control key activities in population-based cohort studies, particularly those including multiple sites (2, 5, 6) and, thus, multiple data collectors.

CASE-COHORT DESIGN AND NESTED CASE-CONTROL STUDIES

Both the population-based and nonpopulation-based cohort study designs are ideal settings in which to test hypotheses by means of nested case-control analyses, as they meet the major assumption inherent in case-control studies with reasonable certainty, that cases and controls originate from the same reference population (4). Nested case-control studies are especially appropriate when biologic specimens (such as blood) are collected and stored for all cohort participants at baseline; the measurement of analytes in only those who develop the outcome on follow-up (cases) and a sample of noncases (controls) permits a cost-effective evaluation of associations free of temporal bias (11). The nested approach is also useful to avoid recall bias vis-a-vis traditional case-control studies; this is illustrated by the impact of collecting information on tanning ability after the occurrence of melanoma in cohort participants of the Nurses' Health Study. Although the use of information collected in a nested case-control mode (i.e., before the development of melanoma) suggested a protective effect associated with a lower ability to tan (relative odds = 0.7), the postdiagnosis data suggested the opposite (relative odds = 1.6), likely reflecting a memory bias limited to cases (15).

The nested case-control strategy as an alternative to examining the cohort data in a "prospective" mode had been recognized by the Framingham authors decades ago (8). Because of its historic importance, it is worthwhile to quote these investigators ipsis literis: "At the end of 5 years a portion of the base population, which was normal at the time of the initial examination with respect to the diseases studied, will have passed the borderline into definite abnormality, and a few will have died. At that point it will be possible to study the differences, as of the time of the initial examination, between those who remained essentially normal, and those who became abnormal (or diseased)" (8, p. 285).

The comparison of cases occurring in the cohort with a sample of the whole cohort, often referred to as a case-cohort study design, has frequently been used when large sample sizes are involved. The use of a sample to estimate total population parameters in a cohort study seems to have been used for the first time by Comstock and Lundin (16) in a study evaluating the relations of maternal smoking to the risks of neonatal and postneonatal deaths. The study involved identifying deaths in these categories occurring in Washington County over the 10 years prior to 1963, as well as a systematic 3 percent sample of live births occurring in the same period. After estimating the total denominator by dividing the live births by the reciprocal of the sampling fraction for the categories denoting presence or absence of maternal smoking in the index pregnancy, the authors tested their hypothesis using the traditional "prospective" mode, that is, comparing rates of neonatal and postneonatal deaths between mothers who smoked in the index pregnancy and those who did not. However, particularly for postneonatal deaths, which would have also been included in the sampling frame of live births, the comparison of the odds of maternal smoking between cases and the live birth sample would have been fully akin to a case-cohort study in that the odds ratio of maternal smoking would have estimated the risk ratio obtained in the "prospective" analysis (17).

Following this pioneering effort by Comstock and Lundin, both the case-cohort and the nested case-control approaches have become often used strategies in the context of cohort studies, especially over the last decade or so (18, 19). As mentioned previously, the cohort sample used as a control group in the case-cohort design allows the estimation of risk factor distributions and prevalence rates needed for population attributable risk estimates, thus making this design preferable to the nested case-control (noncase) approach, particularly when the outcome of interest is not rare. An additional advantage of the case-cohort study is that the relative odds of exposure using this approach provides a direct estimation of the rate or risk ratio (7). This not only does not require the rarity assumption, but may, in addition, be a more readily interpretable measure of association than the relative odds estimated when using the nested case-control strategy.

EXTERNAL VALIDITY

Concern about external validity was recognized as early as 1950 when Dawber et al., commenting on the Framingham study design, stated that its narrow geographic coverage "clearly limits the generality of conclusions which can be reached" (8, p. 281). Because external validity is regarded as the main advantage of a population-based cohort study, it is relevant to discuss the extent to which it can be achieved in real life. Representativeness of the cohort depends on eligibility criteria for inclusion, initial response of the sample, and the stability of the cohort on follow-up. In addition, it is a function of the variability of the factors under study which, in observational studies, comprise a sine qua non condition for detecting an association.

An exemplar of a population-based cohort study: the National Health and Nutrition Examination Epidemiologic Follow-up Study (NHEFS)

NHEFS was the first attempt ever to follow-up a probability sample of the total US population for ascertainment of health outcomes, thus representing a setting which optimizes external validity (20). It originated as a joint project between the National Center for Health Statistics and the National Institute on Aging with support from other National Institutes of Health and Public Health Service agencies, and was designed to follow-up the probability sample included in the first National Health and Nutrition Examination Survey (NHANES I) (20). Of the approximately 21,000 noninstitutionalized persons aged 25–74 years included in NHANES I, conducted in 1971–1975, the 14,407 (70 percent) who were medically examined were included in the first NHEFS follow-up contact conducted in 1982–1984. Whereas the NHANES I included physical examinations, laboratory tests, and interviews, NHEFS was primarily designed to assess outcomes on the basis of self-reports, hospital and nursing home record review, and search for death certificates using the National Death Index.

Initial nonresponse. Initial nonresponse of individuals chosen for inclusion in a population-based cohort study obviously limits the external validity of the cohort sample. Table 1 shows that nonresponse rates in NHANES I were approximately 30 percent. More importantly, the variation according to age, gender, race, and income underscores the notion that, as nonresponse in a cohort study is rarely random across segments of the reference population, those included in the baseline examination may not be a representative sample of the population with regard to variables of interest. Furthermore, the level of representativeness may vary as a function of interactions between these variables. Thus, as seen in table 1, nonresponse in NHANES I increased with age in whites, but not obviously so in blacks. In addition, opposite patterns of nonresponse according to income between races

TABLE 1. Percent nonresponse in the first National Health and Nutrition Examination Survey (NHANES I) by sex, race, age, and income*

Characteristic	Total nonresponse	White		Black	
		Male	Female	Male	Female
Age (years)					
25–34	26.7	29.7	24.3	32.1	28.2
35–44	27.4	27.6	26.7	35.3	26.1
45–54	29.6	27.5	30.0	31.2	35.2
55–64	32.5	31.0	35.1	29.5	27.2
65 and older	35.3	32.0	39.8	24.9	33.6
Income (annual)					
<$7,000	28.6	28.4	29.7	26.0	27.4
$7,000–14,999	27.6	26.2	28.0	30.1	29.5
≥$15,000	25.8	25.8	25.4	32.0	24.3
Total	30.5	30.0	31.1	29.6	29.8

* From Cohen et al. (20).

and genders were suggested; nonresponse decreased with higher incomes in white males and females, and possibly also in black females. However, in black males, the reverse seemed to be true. When age-, race-, and gender-adjusted odds ratios of nonresponse were examined by selected health-related variables, statistically significantly ($p < 0.001$) higher odds were seen for those reporting arthritis (1.45) or vision trouble (1.64).

Follow-up rates. Notwithstanding nonresponse at baseline, follow-up of the NHEFS cohort was attempted for all 14,407 subjects sampled for inclusion in the NHANES I (figure 1). As a result, close to 85 percent of the original NHANES I sample was interviewed. However, complete measurements of pulse, blood pressure, and weight could only be achieved for about 70 percent of the sample. In addition, models including age, gender, race, and an interaction term for race and gender indicated markedly higher odds of loss to follow-up for younger compared with older subjects; for example, the odds ratio for those aged less than 35 years compared with those aged 55 years and older was over 4.0. Age-adjusted odds ratios were also heterogeneous by race and gender as follows: white males, 1.0; white females, 1.3; black males, 3.2; and black females, 2.3. Health-related variables significantly associated with higher odds of loss to follow-up included current smoking (odds ratio = 1.9) and systolic blood pressure of 160 mmHg or higher (odds ratio = 0.8).

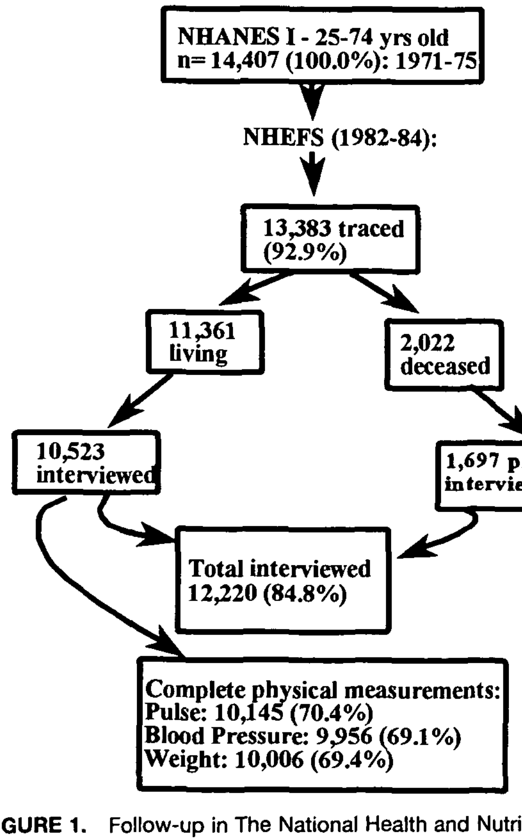

FIGURE 1. Follow-up in The National Health and Nutrition Examination Epidemiologic Follow-up Study (NHEFS) (20). NHANES I, first National Health and Nutrition Examination Survey.

Other studies

Other population-based studies have typically used more restrictive inclusion criteria than those used in the NHEFS, and thus provide examples of the influences of both eligibility criteria and response rates on external validity.

In the Cardiovascular Health Study (CHS), adults aged 65 years and older were sampled from the HCFA Medicare eligibility listings in four US communities (21). Recruited cohort members had to be able to complete an in-depth home interview and an examination lasting approximately 4 hours to be conducted in especially set-up clinics. In addition, the eligibility criteria included the expectation of returning for the 3-year follow-up visit. As a result, 9.6 percent of the subjects were found to be ineligible. Among the eligible persons, 38.6 percent refused to participate in the study. As for the NHEFS, participants did not seem to be comparable to nonparticipants in that they were more likely to be younger, had a higher average educational level, were more often married, and, most importantly given the age range of the cohort, were less likely to report limitations in their physical activities. Those who had quit smoking were more likely to agree to participate, a finding in agreement with the first NHEFS follow-up contact. On the other hand, in this elderly cohort, presence of high blood pressure was less common in those who enrolled, whereas in the NHEFS the odds of participation was 20 percent higher when high blood pressure was present (20). As in any other cohort study, the ability to generalize results from the CHS is a function of both study eligibility criteria (healthy elderly) and the characteristics related to participation in the study.

In the ARIC study, the recruitment of the cohort aged 45–64 years involved four major sequential operational steps—sampling, enumeration of households, home interview, and examination at study clinics—thus allowing the authors to compare eligible individuals who completed both the home interviews and clinic procedures (respondents) with those who participated in home interviews but not clinic examinations (nonrespondents) (22). Of the age-eligible individuals (aged 45–64 years), 75 percent completed the home interviews and 60 percent underwent clinic procedures. Nonresponse was higher in blacks (58 percent in men and 51 percent in women) than in whites (approximately 33 percent in both sexes). As expected, white respondents were more likely to report excellent health and were less likely to have been hospitalized in the past year; however, no such differences were seen in blacks. The authors aptly underscore the effect that low response rates may have on estimates of prevalence in community-based studies, even when random selection of the cohort is attempted; for example, in the ARIC communities, point prevalence rates of current smoking were found to be 38 percent and 25 percent for black male and female respondents, respectively, yet the community rates were estimated at 45 percent and 29 percent, respectively. The availability of home interview data for smoking in the ARIC study allows estimating the impact of nonresponse on population-wide rates; however, for variables for which measurement requires laboratory procedures, and which thus could not be included in the home interviews, such as total plasma cholesterol levels, the magnitude of bias cannot be easily estimated. This makes it difficult to generalize baseline cohort mean levels to the total reference populations.

Another illustration of the impact of eligibility and response rates on generalizability of findings is given by the CARDIA study, designed to assess the influence of lifestyle and health behaviors on cardiovascular risk factors in individuals aged 18–30 years living in four urban areas (5). Besides age, race, and residence requirements, eligibility criteria included freedom from long-term diseases or disabilities, such as deafness, blindness, and mental retardation, that may interfere with the examination. In addition to exclusion on the basis of ineligibility, approximately 35 percent of the 2,213 eligible study subjects did not complete the first examination. More importantly, marked differences were found between attendees and nonattendees, with the former more likely than the latter to be white (49 versus 38 percent), older (55 versus 46 percent), and have completed high school (60 versus 48 percent). These differences may have compromised one of the stated goals of recruitment of the CARDIA study, that is, to obtain samples as representative as possible of the underlying populations (5). In some population-based cohort studies, failure to achieve initial representativeness may impact negatively on the use of the baseline examination of the cohort as a way to provide a cross-sectional profile of the reference population with which later survey results can be compared for purposes of monitoring of risk factors and other health-related variables.

It is interesting that for many population-based studies, once cohorts are recruited, follow-up rates among survivors appear to be high. As discussed previously, the first NHEFS follow-up interview had a response rate of about 85 percent (20). In the ARIC study, after an initial overall enrollment of only approximately 60 percent of cohorts selected from total populations aged 45–64 years in four US communities, return rates after 9 years remain high for a 4-hour in-clinic examination (about 80 percent) (The ARIC Investigators: report to the ARIC Policy Board, November 1997, unpublished). For follow-up through annual telephone interviews, a response rate of close to 100 percent is being achieved even 12 years after the study onset, thus underscoring the possibility of a high response rate when using strategies representing a small burden to cohort participants. Follow-up rates from both the CHS and the ARIC studies suggest that cohort stability may be more important than initial representativeness, thus providing a rationale for nonpopulation-based cohort studies, such as the American Cancer Society studies of volunteers (23).

Impact of representativeness of the cohort on external validity

The debate pertaining to the ability to generalize results from a given population-based cohort study is not new. In the landmark Framingham study, the random sample forming the cohort had to be "enriched" by volunteers to compensate for its relatively low (69 percent) response rate (figure 2) (24). Framingham investigators acknowledged that external validity of study findings could suffer from the limited response of those randomly selected, particularly in view of the differences between participants and nonparticipants, with the latter including a higher proportion of single men, excessive alcohol drinkers, and those with a low educational background. Nevertheless, they aptly argued that the determination of absolute measures of disease frequency (incidence rates) "had only limited applicability since the population of this New England town was in itself not totally representative of the United States" (24, p. 22). They further stressed that "random sampling is not essential if the purpose. . .is to compare subgroups of the population determined by

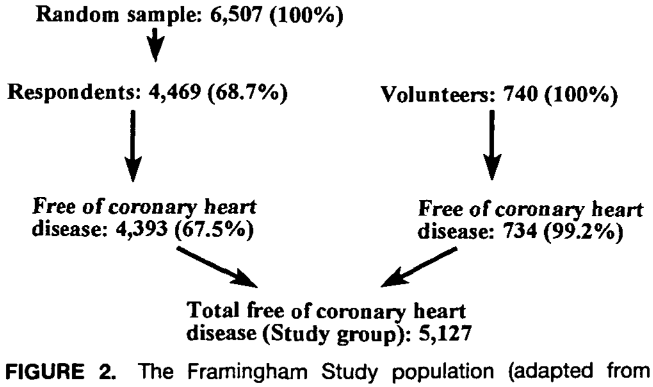
Random sample: 6,507 (100%)
Respondents: 4,469 (68.7%)
Volunteers: 740 (100%)
Free of coronary heart disease: 4,393 (67.5%)
Free of coronary heart disease: 734 (99.2%)
Total free of coronary heart disease (Study group): 5,127

FIGURE 2. The Framingham Study population (adapted from Dawber (24)).

specific characteristics" (24, p. 22). (They instead emphasized that "the primary concern should be that the population contain sufficient numbers of subjects with these characteristics to enable comparison" (24, p. 22), thus underscoring one of the key problems in epidemiologic studies discussed later in this paper, i.e., that for true associations to be detected in observational studies, sufficient variability must exist in both the exposure and the outcome of interest.)

A similar reasoning was offered by Keys, who stated with regard to the Seven Countries study that, "whether the men. . .are representative of some larger population in regard to particular characteristics at entry is not the issue. We are interested in relationships" (25, p. 12). These points underscore the importance of recognizing the difference between the external validity of an absolute measure of disease frequency from the external validity of measures of association. Whereas the former's applicability is often limited to the reference population and clearly depends on whether the cohort is a representative sample of this population, the latter may be valid even if the cohort is not a random sample. The explanation underlying the belief that it is possible to obtain unbiased and generalizable results from "unrepresentative" cohorts when assessing exposure-outcome relations is, as pointed out by Schlesselman (26), that as long as the groups under comparison are equally biased, relative risks (or odds ratios) are estimated accurately. Thus, in a cohort study, even if incidence rates (or odds) of a disease in exposed and unexposed individuals are not externally valid, it is possible to obtain unbiased estimates of the relative risk or the odds ratio.

Generalizing beyond the study population

As pointed out almost 20 years ago, "the ultimate test of the validity and generalizability of scientific observations is the ability, based on them, to predict phenomena and circumstances independent of those used to record the original findings" (27, p. 256). An assessment of the predictive values of cardiovascular risk factors beyond specific study populations was first attempted by investigators participating in the so-called Pooling Project (27): in this project, 10-year follow-up data of five cohort studies of major coronary events (nonfatal and fatal myocardial infarcts, and sudden deaths from coronary heart disease)—two of which were population-based (Framingham and Tecumseh)—were pooled and then applied to two cohorts originally considered but eventually not included in the main project pool (Pool 5).[1] For both cohorts not included in Pool 5, the observed relative and absolute excesses were remarkably similar to those predicted on the basis of Pool 5 results (table 2), leading the authors to state that for the studies included in Pool 5, results are "widely generalizable to US middle-aged white men" (27, p. 256).

The most thorough effort to examine comparability among cohort studies comes from Chambless et al. (28). More specifically, these authors discussed the appropriateness of using logistic regression function scores estimated from a given study population to predict risk in other populations. Similar conceptual frameworks to assess limitations in external validity have been proposed by Chambless et al. (28) and Szklo (29) and can be summarized as follows:

1. If the circumstances of the exposure of interest, such as duration, level, route, and time of exposure vary over time or, for the same time interval, among populations or subgroups of the same population, these differences may limit the applicability of results. For instance, this would be the case if the chemical composition of a drug, such as estrogen replacement therapy, varied from population to population or over time in the same population. A related mechanism pertains to the timing of the studies. Particularly in the assessment of newly introduced exposures with long latent periods, studies carried out at a time when minimum latency has not yet gone by will tend to underestimate the strength of the association with the relevant outcome, thus compromising their external validity.

2. Results are not generalizable when there are differences in susceptibility to the risk factor of interest between the study population and the populations to which one wishes to apply the study findings. For example, in an early study of the effect of maternal smoking on neonatal mortality, Comstock and Lundin (16) found a strong interaction between maternal smoking and father's educational level, leading these investigators to suggest that the generalizability of their findings to other populations would be affected by their socioeconomic characteristics. Similarly, results obtained from a subgroup of the population may not be generalizable to another group. Thus, estimates of the relative effects of risk factors such as smoking

[1] As the studies had been selected for inclusion in Pool 5 on the basis of comparability of their findings, it is not surprising that results were fairly homogeneous for two of the three major cardiovascular risk factors. Thus, rate ratio comparing the fifth quintile to the combined first and second quintiles of diastolic blood pressure had a range of 1.8–2.5, and that from comparing smokers of one pack/day or more with nonsmokers, of 2.2–3.3. The only exception was total cholesterol, for which the Tecumseh study rate ratio, based on the fifth versus the combined first-second quintiles, was markedly higher (4.9), and that of the Chicago Peoples Gas Company lower (1.5) than the rate ratios of the other three studies (2.5, 2.7, and 2.7). The combined Pool 5 rate ratios and 95 percent confidence limits were, for diastolic blood pressure, 1.9 (1.6–2.4), for total cholesterol, 2.4 (1.9–2.9), and for smoking, 2.5 (2.1–3.1).

TABLE 2. Predicted and observed risk of a first major coronary event for men aged 40-59 years at baseline of two cohort studies not included in Pool 5, Pooling Project, 10-year follow-up*

Quintile of predicted risk†	Risk/1,000 men (corrected for duration of follow-up)			
	Minnesota business and professional men (*n* = 280)		Minnesota-based railroad workers (*n* = 2,422)	
	Predicted	Observed	Predicted	Observed
I	37.6	53.6	17.0	16.5
II	57.0	71.4	31.0	12.4
III	80.9	125.0	44.7	31.0
IV	113.0	107.1	65.5	68.2
V	199.1	214.3	116.1	102.9
All quintiles	97.5	114.3	54.9	46.2
Ratio of quintile V to quintile I	5.3	4.0	6.8	6.3
Difference between quintile V and quintile I	161.5	160.7	99.1	86.4

* From the Pooling Project Research Group (27).

† Defined by diastolic blood pressure, serum cholesterol, smoking, and age using a multiple logistic function.

on cardiovascular outcomes obtained from a relatively young study population cannot be generalized to populations with a high proportion of elderly individuals, in view of the known decrease with age of relative risks associated with these risk factors (24).

3. When modeling relative risks or odds ratios from data obtained from a given population, results may not be applicable to populations with different levels of exposure, as the relations may not be linear across all exposure levels. Thus, studies done in populations with high levels of radiation exposure may not be generalizable to populations with exposure to low levels (30).

4. The distribution of confounding variables unaccounted for may be different between the study population and populations to which one wishes to generalize results.

5. The definitions of exposure and outcome variables may not be applicable to other populations or even different groups of the same population. In addition, construct validity of a given variable may differ among populations or groups within a population. For example, although educational level is often used as a marker of socioeconomic status, there is evidence that it is differentially related to income in whites and blacks in the United States (31).

6. Another concern is that associations between risk factors and clinical outcomes may be influenced by the prevalence of the underlying subclinical process (e.g., atherosclerosis). In the Seven Countries study, for example, the association between smoking and clinical coronary heart disease was much stronger for the western European cohorts than for the former Yugoslavian cohort (25), possibly because the thrombogenic effect of smoking leading to clinical events was more likely to occur where subclinical atherosclerosis was more common (i.e., western Europe).

7. Finally, as for most diseases, risk factor-outcome associations are time dependent. Thus, the length of follow-up (and other methodological aspects) of a given study needs to be taken into account when generalizing results to other populations.

To illustrate these problems, Chambless et al. (28) reviewed several cohort studies of coronary heart disease, including international studies. Their results pertaining to the major cardiovascular risk factors underscore the problems influencing external validity of population-based studies (table 3). Using the above list of factors affecting interpopulation comparability as a framework, Chambless et al. noted that there was variation from study to study in factors influencing the odds ratio estimates, including the definition of the event "coronary heart disease", the length of follow-up (4–24 years), the potential confounders adjusted for to arrive at the odds ratio estimates, and the range of values for the variables of interest. In addition, there may have also been differences among the study populations in the extent of underlying atherosclerosis at the beginning of the study and in time since initiation of exposure to the cardiovascular risk factors considered by the authors. For smoking, the extraordinarily wide range of odds ratios did not seem to result from outlying values, as even when the maximum was re-

TABLE 3. Longitudinal studies of incidence of coronary heart disease events in men*

Risk factor	Unit	Odds ratio range†			Population	
		Maximum	Minimum	Ratio	Maximum	Minimum
Age (years)	10 years	2.66	1.40	1.90	Northern Europe (25)	Honolulu (32)
Cholesterol (mg/dl)	60 mg/dl	2.03	1.28	1.59	Sweden (37)	Southern Europe (25)
Systolic blood pressure (mmHg)	40 mmHg	2.51	1.57	1.60	British Regional Heart Study (38)	Southern Europe (25)
Smoking	1 pack/day or 25 cigarettes/day vs. nonsmoker	13.3	1.03	12.91	Sweden (37)	Southern Europe (25)

* Adapted from Chambless et al. (28). Number of studies: age = 11; cholesterol = 14; systolic blood pressure = 14; and smoking = 11.

† Except for smoking, the minimum and maximum values do not appear to be outliers. Median odds ratios: age = 1.80; cholesterol = 1.56; systolic blood pressure = 2.14; and smoking = 2.03 (for smoking, the second highest odds ratio is 8.0, and the third highest is 2.92).

placed with the third highest odds ratio, the ratio of odds ratios remained high (about 3). As recognized by the authors, the definition of "smoking" varied widely and, thus, the estimates are not directly comparable across studies. To provide for some degree of comparability, the smoking category was defined by Chambless et al. (28) in such broad terms (≥1 pack/day or ≥25 cigarettes/day) that it is possible that the differences among studies illustrated in table 3 reflect differences in exposure level within the open-ended "exposed" category, rather than true heterogeneity of effects.

Unlike the investigators involved in the Pooling Project (27) and the Framingham study (24), Chambless et al. concluded that, in spite of the fact that the risk factors were consistently found to increase risk of coronary heart disease across studies, the use of a measure of risk from a given population to infer risk in another population "may be more misleading than advantageous" (28, p. 394). On the other hand, with the exception of smoking, and notwithstanding the fact that the analyses of Chambless et al. included studies from countries around the world, the ratios of maximum to minimum odds ratios were found to be less than 2. When only the four US based studies (Pooling Project (27), Honolulu Heart Program (32), Western Collaborative Heart Study (33), and Framingham (34)) were considered, these ranges were similar for age, total cholesterol and blood pressure, and much narrower for smoking. (Using the same units shown in table 3, the ratios of maximum to minimum odds ratios for the US studies were: age, 1.57; total cholesterol, 1.32; systolic blood pressure, 1.51; and smoking, 1.62).

Variability of the factors under study

Another important limitation in generalizing results of a cohort study is the need for sufficient variability of exposure (and outcome) levels—a sine qua non condition for detecting associations in observational studies. For example, the fact that variability in salt consumption is greater among than within countries (35) may explain why the relation between salt intake and hypertension is more clearly seen in between-country data than in studies carried out within populations. Concern about the natural variability of the exposures under study was already expressed by those who masterminded the Framingham study (8), as the plan was to select a population which would vary in the characteristics believed to be related to coronary heart disease.

The need for a within-cohort, in addition to between-cohort, exposure gradient was also recognized by the investigators of the Seven Countries study; for example, the sampling approach to select the US cohort of railroad employees took account not only of the size of the community, but also attempted to provide occupational subsamples which would permit comparing sedentary and physically active individuals (25).

Single versus multiple populations

Population-based cohort studies may be conducted in single populations, such as the Framingham (24) and the Honolulu Heart studies (36), or use cohorts selected from several defined populations, such as the ARIC (2), CARDIA (5), and the CHS (6) studies. The decision as to whether single or multiple reference populations should be included is determined, not only by power considerations, but also by whether heterogeneity in population characteristics is relevant to the hypotheses of interest, as was the case of the Seven Countries study (25).

When, in multicenter studies, a sampling approach taking into consideration within-population variability in factors of interest is not used, it may be difficult to discern the effect of a given variable which is homogeneous within the populations from that of the geographic location. In the ARIC study, for example, the vast majority of African-Americans were sampled from one of the four target populations (Jackson, Mississippi); in two other target populations, cohort samples were almost exclusively whites, whereas in the fourth, only about 12 percent were African-American (22). This sampling strategy, although advantageous in many respects, made it difficult to discern the effect of "race" from that of geographic location, as the vast majority of African-American cohort members were Jackson residents. The inability to distinguish racial background from the Jackson general community characteristics (e.g., diet, access to medical care) in the ARIC study underscores the need to achieve both within- and between-population variations in the factors of interest so as to be able to assess their independent effects.

CONCLUSIONS

Both population-based and nonpopulation-based cohort studies have made landmark contributions to the elucidation of risk factor-outcome associations. Among the advantages of population-based cohort studies is that health-related parameters for the reference population can be measured, thus allowing the estimation of population attributable risk figures needed for planning purposes. On the other hand, follow-up studies of volunteers or other nonpopulation-based samples are often more efficient than population-based cohort studies. In the American Cancer Society's large prospective studies (23), for example, volunteers for the American Cancer Society not only recruited cohort participants, but also served as their contacts for follow-up retention purposes. Thus, in these studies, cost-effectiveness resulting from the ease of tracing cohort participants may well have outweighed the fact that they were not population-based. Ultimately, the decision as to whether to conduct a population-based study depends on whether the level of participation and retention of the cohort necessary to make it "representative" of the reference population is sufficient to outweigh the superior cost-effectiveness which often characterizes nonpopulation-based cohort studies. If participation rates at baseline and retention rates on follow-up are not simultaneously affected by exposure levels and outcome, it is expected that "compensating bias" would ensue, thus rendering risk factor-outcome association inferences valid when using either type of strategy.

REFERENCES

1. Levin M. The occurrence of lung cancer in man. Acta Unio Contra Cancrum 1953;9:531–41.
2. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. Am J Epidemiol 1989;129:687–702.
3. Sternfeld B, Stang P, Sidney S. Relationship of migraine headaches to experience of chest pain and subsequent risk for myocardial infarction. Neurology 1995;45:2135–42.
4. Comstock GW, Bush TL, Helzlsouer KJ, et al. The Washington County Training Center: an exemplar of public health research in the field. Am J Epidemiol 1991;134:1023–9.
5. Friedman GD, Cutter GR, Donahue RP, et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. J Clin Epidemiol 1988;41:1105–16.
6. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1991;1: 263–76.
7. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven, 1998.
8. Dawber TR, Meadors GF, Moore FE Jr. Epidemiologic approaches to heart disease: the Framingham Study. Am J Public Health 1951;41:279–86.
9. Sackett DL. Bias in analytic research. J Chronic Dis 1979;32: 51–63.
10. Comstock GW, Burke AE, Hoffman SC, et al. Serum concentrations of α-tocopherol, β-carotene, and retinol preceding the diagnosis of rheumatoid arthritis and systemic lupus erythematosus. Ann Rheum Dis 1997;56:323–5.
11. Rothman N, Cantor KP, Blair A, et al. A nested case-control study of non-Hodgkin lymphoma and serum organochlorine residues. Lancet 1997;350:240–4.
12. Diez-Roux AV, Nieto FJ, Comstock GW, et al. The relationship of active and passive smoking to carotid atherosclerosis 12–14 years later. Prev Med 1995;24:48–55.
13. Helsing KJ, Szklo M. Mortality after bereavement. Am J Epidemiol 1981;114:41–52.
14. Szklo M, Cerhan J, Diez-Roux AV, et al. Estrogen replacement therapy and cognitive functioning in the Atherosclerosis Risk in Communities (ARIC) study. Am J Epidemiol 1996; 144:1048–57.
15. Weinstock MA, Colditz GA, Willett WC, et al. Recall (report) bias and reliability in the retrospective assessment of melanoma risk. Am J Epidemiol 1991;133:240–5.
16. Comstock GW, Lundin FE Jr. Parental smoking and perinatal mortality. Am J Obstet Gynecol 1967;98:708–18.
17. Wacholder S, Silverman DT, McLaughlin JK, et al. Selection of controls in case-control studies. III. Design options. Am J Epidemiol 1992;135:1042–50.
18. Liao D, Cai J, Rosamond WD, et al. Cardiac autonomic function and incident coronary heart disease: a population-based case-cohort study: the ARIC study. Am J Epidemiol 1997;145:696–706.
19. van der Klauw MM, Stricker BH, Herings RM, et al. A population based case-cohort study of drug-induced anaphylaxis. Br J Clin Pharmacol 1993;35:400–8.
20. Cohen BB, Barbano HE, Cox CS, et al. Plan and operation of the NHANES I Epidemiologic Followup Study: 1982–84. Vital Health Stat [1] 1987;(22):1–142.
21. Tell GS, Fried LP, Hermanson B, et al. Recruitment of adults 65 years and older as participants in the Cardiovascular Health Study. Ann Epidemiol 1993;3:358–66.
22. Jackson R, Chambless LE, Yang K, et al. Differences between respondents and nonrespondents in a multicenter community-based study vary by gender and ethnicity. J Clin Epidemiol 1996;49:1441–6.
23. Thun MJ, Heath CW Jr. Changes in mortality from smoking in two American Cancer Society prospective studies since 1959. Prev Med 1997;26:422–6.
24. Dawber TR, ed. The Framingham study: the epidemiology of atherosclerotic disease. Cambridge, MA: Harvard University Press, 1980.
25. Keys AB. Seven countries: a multivariate analysis of death and coronary heart disease. Cambridge, MA: Harvard University Press, 1980.
26. Schlesselman JJ. Case-control studies: design, conduct, analysis. New York, NY: Oxford University Press, 1982:128–9.
27. Relationship of blood pressure, serum cholesterol, smoking habit, relative weight and ECG abnormalities to incidence of major coronary events: final report of the Pooling Project. The Pooling Project Research Group. J Chronic Dis 1978;31:201–306.
28. Chambless LE, Dobson AJ, Patterson CC, et al. On the use of a logistic risk score in predicting risk of coronary heart disease. Stat Med 1990;9:385–96.
29. Szklo M. Evaluation of epidemiologic information. In: Gordis L, ed. Epidemiology and health risk assessment. New York, NY: Oxford University Press, 1988:268–72.
30. Land CE. Influence of theoretical and experimental radiobiology on the epidemiology of radiation carcinogenesis. In: Gordis L, ed. Epidemiology and health risk assessment. New York, NY: Oxford University Press, 1988:234–47.
31. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health 1997;18:341–78.
32. Yano K, Reed DM, McGee DL. Ten-year incidence of coronary heart disease in the Honolulu Heart Program: relationship to biologic and lifestyle characteristics. Am J Epidemiol 1984; 119:653–66.
33. Rosenman RH, Brand RJ, Sholtz RI, et al. Multivariate prediction of coronary heart disease during 8.5 year follow-up in the Western Collaborative Group Study. Am J Cardiol 1976; 37:903–10.
34. McGee D. Section 28: The probability of developing certain cardiovascular diseases in eight years at specified values of some characteristics. In: Kannel WB, Gordon T, eds. The Framingham study: an epidemiological investigation of cardiovascular disease. Washington, DC: US GPO, 1973. (DHEW publication no. (NIH) 74–618).
35. Szklo M. Epidemiologic patterns of blood pressure in children. Epidemiol Rev 1979;1:143–69.
36. Trombold JC, Moellering RC Jr, Kagan A. Epidemiological aspects of coronary heart disease and cerebrovascular disease: the Honolulu Heart Program. Hawaii Med J 1966;25:231–4.
37. Wilhelmsen L, Wedel H, Tibblin G. Multivariate analysis of risk factors for coronary heart disease. Circulation 1973;48:950–8.
38. Shaper AG, Pocock SJ, Walker M, et al. Risk factors for ischaemic heart disease: the prospective phase of the British Regional Heart Study. J Epidemiol Community Health 1985; 39:197–209.