

Analyzing Health Surveys for Cancer-Related Objectives

Barry I. Graubard, Edward L. Korn

Large-scale health surveys conducted by government agencies record information on a large number of health-related variables. We review the use of these data for performing analyses that address cancer-related objectives. After describing the conduct of a large-scale health survey (the third National Health and Nutrition Examination Survey [NHANES III]), we discuss some of the issues involved in analyzing data collected in such a survey. In particular, the use of sample weights in the analysis and the importance of accounting for the complex survey design when estimating standard errors are discussed. Six applications are then presented that involve the following: 1) estimating demographic factors associated with snuff use, 2) estimating the association of type of health insurance with the probability of receiving a digital rectal examination, 3) estimating the association of body iron stores with the probability of later developing cancer, 4) estimating the changing rates of mammography screening in the United States between 1987 and 1992, 5) evaluating smoking and alcohol consumption as risk factors for digestive cancer by use of a population-based, case-control study, and 6) evaluating a randomized community-intervention experiment to encourage smoking cessation. These applications use data from the National Health Interview Survey, the NHANES I Epidemiologic Followup Study, the 1986 National Mortality Followback Survey, and the Community Intervention Trial for Smoking Cessation. The availability of public-use data files is discussed for surveys sponsored by the U.S. government that collect health-related information. We demonstrate that statistical methods and computer software are available for analyzing public-use data files of surveys to address different types of cancer-related objectives. [J Natl Cancer Inst 1999;91:1005-16]

Health surveys provide a wealth of information about the incidence and prevalence of diseases, the occurrence of healthy and unhealthy behaviors, exposures to potential risk factors, dietary intake, physiologic measures of the population, and costs and utilization of health services. The large sample sizes of some health surveys, typically conducted by government agencies, enable one to study relatively small—but important—associations between variables, relatively rare events, and subpopulations of interest. Since appropriate statistical methods can make the results of an analysis of survey data representative of the population sampled, repeated surveys of the same population can be used to evaluate trends in the population. In addition, follow-up of individuals sampled in a baseline survey allows one to measure change at the individual level and to associate baseline risk factors with the development of diseases. Table 1 presents examples of the wide variety of cancer-related analyses that have been published using some selected health surveys.

In this review, we examine the use of data from large-scale health surveys to address cancer-related objectives. The health surveys that we consider are surveys that are representative of well-defined large populations, such as the U.S. population. In what follows, 1) we describe the design and conduct of such a survey, the third National Health and Nutrition Examination Survey (NHANES III), for those not familiar with this type of investigation; 2) we describe two aspects of survey data that can complicate an analysis, the sample weights associated with the observations and the clustering of the observations, and describe the statistical methods that are used to handle these complications; 3) we present six applications involving different types of cancer-related research questions and analyses using health survey data; 4) we discuss the availability of public use data files for U.S. government-sponsored surveys that collect health-related information; and 5) we describe some limitations of using survey data for analyses. The goal of this review is to increase awareness of the types of health survey resources that are available for cancer-related investigations and to encourage their increased utilization.

CONDUCTING A LARGE-SCALE HEALTH SURVEY: NHANES III

NHANES III is the seventh in a series of surveys that began in 1960 to examine the health of the U.S. population (40). Its goals are as follows [(40), p. 1]: “1) to estimate the national prevalence of selected diseases and risk factors; 2) to estimate national population reference distributions of selected health parameters; 3) to document and investigate reasons for secular trends in selected diseases and risk factors; 4) to contribute to an understanding of disease etiology; and 5) to investigate the natural history of selected diseases.” NHANES III sampled approximately 40 000 individuals during the period from 1988 through 1994 and cost approximately \$100 million dollars to conduct. The population sampled consisted of civilian, noninstitutionalized individuals, 2 months of age or older. In NHANES III, sampled individuals were first interviewed in their homes. To decide who to sample, if one had a list of all individuals in the United States, one could imagine taking a “simple random sample” from the list. This is equivalent to putting all of the names in a hat and pulling out 40 000 at random. However, this would not be a practical sampling design because then interviewers would have to travel to 40 000 locations dispersed across the country. In addition, sampled individuals were asked to go for a physical examination to a mobile examination center, which could be set up at only a limited number of sites. There-

Affiliations of authors: B. I. Graubard (Biostatistics Branch), E. L. Korn (Biometric Research Branch), National Cancer Institute, Bethesda, MD.

Correspondence to: Barry L. Graubard, Ph.D., National Institutes of Health, Executive Plaza South, Rm. 8024, Bethesda, MD 20892.

See “Note” following “References.”

Table 1. Examples of cancer-related analyses with the use of data from the National Health Interview Surveys (NHIS), National Health and Nutrition Examination Surveys (NHANES) (including the follow-up of the 1971–1975 survey), and the 1986 National Mortality Followback Survey

| Time trends of | | Reference No.* |
|--|---|----------------|
| Use of postmenopausal hormone replacement therapy | | (1) a |
| Cancer screening | | (2) bc, (3) bd |
| Smoking initiation | | (4) bef |
| <i>Cross-sectional associations with various outcomes</i> | | |
| Risk factors/subgroups | Outcome | |
| Demographic groups and veteran status | Smoking | (5) g, (6) bf |
| Demographic groups and diet | Exposure to environmental tobacco smoke | (7) h |
| Demographic groups | Urinary pesticide levels | (8) i |
| Demographic groups | Consumption of fruits and vegetables | (9) i |
| Demographic groups | Nutrition and cancer prevention knowledge, beliefs, and practices | (10) b |
| Demographic groups | Knowledge about indoor radon | (11) d |
| Demographic groups | Mammographic and Pap smear screening | (12) b |
| Type of health insurance | Cancer screening | (13) c |
| Demographic groups and cancer knowledge | Oral cancer examinations | (14) c |
| Sun exposure | Skin damage | (15) j |
| Demographic groups | Cancer prevalence | (16) b |
| <i>Longitudinal associations of risk factors and the development of cancer</i> | | |
| Risk factors | Cancer site | |
| Alcohol consumption, anthropometry, bowel function, dietary fat intake, family and pregnancy histories, and metabolic rate | Breast | (17–22) a |
| Iron intake | Colorectal | (23) a |
| Aspirin intake | Esophageal | (24) a |
| Dietary vitamins (A, E, and C) intake and occupation | Lung | (25) a, (26) a |
| Antigenic stimulation | Multiple myeloma | (27) a |
| Serum vitamin A | Prostate | (28) a |
| Adult stature, body iron stores, depression, physical activity, and serum cholesterol level | All | (29–33) a |
| <i>Case-control or proportional mortality analyses^k involving cancer deaths</i> | | |
| Risk factors | Cancer site | |
| Diet, smoking, alcohol consumption, and use of oral contraceptives | Adrenal | (34) l |
| Smokeless tobacco use | Digestive and oral cancer | (35) bl |
| Oral contraceptive use | Liver | (36) bl |
| Smoking and alcohol consumption | Nasopharyngeal | (37) l |
| Diet and alcohol consumption and tobacco use | Small intestine | (38) l |

*Footnotes to reference numbers designate the particular survey(s) used in the analysis: (a) NHANES I Epidemiologic Followup Study; (b) 1987 NHIS; (c) 1992 NHIS; (d) 1990 NHIS; (e) 1970, 1978, 1979, and 1980 NHIS; (f) 1988 NHIS; (g) Hispanic Health and Nutrition Examination Survey; (h) Third Health and Nutrition Examination Survey; (i) Second Health and Nutrition Examination Survey; (j) First Health and Nutrition Examination Survey; (k) Proportional mortality studies use deaths from other causes as “controls” (39); and (l) 1986 National Mortality Followback Survey.

fore, rather than using a simple random sample, NHANES III used a “multistage sampling design,” in which 81 counties (or sometimes two or more adjacent smaller counties) were first sampled, and then individuals within each sampled county were subsampled. The mobile examination centers needed to be positioned at only 89 locations, at which 300–600 individuals in neighboring areas were examined over a 4- to 6-week period.

The 81 selected counties (or areas) were not chosen as a simple random sample from a list of all such areas in the United States. Instead, to decrease the variability of parameter estimators based on data from the completed survey, large-population counties were sampled with a higher probability than small-population counties. Also, since one of the design considerations of NHANES III was to provide reliable estimates for the African-American and Mexican-American minority groups, counties with larger proportions of these minorities were included in the sample with higher probabilities. The sampling of certain subpopulations with higher probabilities than others is a hallmark of large surveys. NHANES III used “stratified sampling,” in which the units to be sampled were divided into a small number of groups (“strata”) and then sampled at different rates in the different strata.

For a sampled county, the second stage of sampling in NHANES III involved sampling area segments consisting of city or suburban blocks or other contiguous geographic areas contained within the county. Segments with larger minority populations were sampled with higher probability. The third stage of sampling involved listing all of the households within the sampled segments and then sampling them at a rate that depended on the segment characteristics, e.g., racial or ethnic composition. The fourth stage of sampling was to sample individuals within sampled households to be interviewed. The probabilities of individuals being chosen in this final stage of sampling were based on their sex, age, and race/ethnicity. Because of design considerations, only about one in five households sampled contributed sampled persons who were interviewed.

The NHANES III household interview consisted of an individual questionnaire for each sampled person, blood pressure measurements for persons aged 17 years and over, and a family questionnaire. The questionnaires contained questions about dietary intake and nutritional status, reproductive history and sexual behavior, use of vitamin and mineral supplements and medications, tobacco and alcohol use, physical activity, health care utilization and health insurance, and sociodemographic characteristics. Sampled persons who completed a household interview were invited to have a medical examination at a mobile examination center; transportation and a small cash payment were provided. The examinations included additional dietary and health interviews, body measurements, physical and dental examinations, venipuncture, urine collection, audiometry, x-rays, electrocardiograms, spirometry, oral glucose tolerance tests, ophthalmologic examinations, ultrasonography, bone density measurements, cognitive assessments, and allergy tests. Examination data were recorded, for the most part, directly into an automated data collection system.

ANALYSES ACCOMMODATING SAMPLE WEIGHTS AND SAMPLE CLUSTERING

Many of the statistical issues involved in analyzing survey data are the same as encountered when analyzing nonsurvey data. The two characteristics of survey data that most complicate

the analysis are the sample weights associated with the data and the fact that the data are clustered. In this section, we discuss these characteristics and methods for accommodating them in the analysis.

The data from each sampled individual are associated with a sample weight, which estimates the number of people in the population that he or she represents. To calculate these weights, consideration is taken of the differential probabilities that individuals were sampled, so that individuals sampled from a subgroup at a rate of one per 10 000 have larger sample weights than individuals sampled from a subgroup at a rate of one per 1000. Sample weights also adjust for the facts that not all sampled individuals participate in a survey ("nonresponse"), and inadvertently not all individuals may have had a chance to be sampled ("frame undercoverage"). These adjustments, which are based on statistical models, are thought to typically lessen bias due to nonresponse and frame undercoverage, but there are no guarantees that they accomplish this.

The classic approach to analyzing data with sample weights is to use (sample-)weighted estimators. Weighted estimation is equivalent to performing standard (unweighted) estimation on a new expanded dataset created from the original dataset of the sampled observations by duplicating each observation the number of times given by its sample weight. For example, an individual with a sample weight of 12 239 would have his or her data appearing 12 239 times in the expanded dataset. Of course, there are simple formulas for performing weighted estimation that do not require actually duplicating observations. In addition, one cannot calculate standard errors of parameter estimators using standard nonsurvey formulas applied to the expanded dataset, but other methods are required as described below.

The advantage of using weighted estimation over using unweighted estimation is that weighted estimators are estimating population parameters and not parameters that depend on the particular sample design used in the survey. In addition, unweighted estimation can sometimes give misleading results when estimating cause/effect relationships (41). A disadvantage of using weighted estimation is that parameter estimators can be very variable when the weights are very variable, especially when a few individuals have very large sample weights. Because of this potential, there has been some debate in the statistical literature concerning the role of sample weights in the analysis of survey data [e.g., (42,43)]. Our own approach is to use weighted estimation either when the analysis is primarily descriptive (as opposed to investigating cause/effect relationships) or when weighted estimation does not greatly increase the variability of estimators. Otherwise, we use unweighted estimation with variables that are used to construct the sample weights additionally included in the analysis model (44).

When estimating standard errors for parameter estimators by use of survey data, one needs to take into account the fact that the data are typically clustered. For example, in NHANES III, there are a few hundred observations in each sampled county. Since data from the same cluster tend to be more correlated than data from different clusters, clustering tends to make standard errors larger than would be obtained from a simple random sample with the same sample size (45). (This problem is not unique to survey data; analysts of animal litter data need to account for the correlation of observations within litters.) Fortunately, simple techniques have been developed to estimate standard errors that require only a designation for each indi-

vidual of the cluster at the first stage of sampling to which he or she belongs (46). [This is true provided that either the fraction of first-stage units sampled is relatively small or the inference required is for the assumed model underlying the generation of the data; see (47).] These first-stage sampling clusters are called "primary sampling units." Public-use data files for most health surveys contain the primary sampling unit designations of each sampled individual. An exception is some institutional surveys, e.g., the National Hospital Discharge Survey (48), where there are confidentiality concerns in releasing information about the hospitals (which are the primary sampling units). For these surveys, approximate methods for standard error estimation involving "variance curves" ("generalized variance functions") have been developed (46).

An alternative approach to estimating standard errors, which might be attractive to a statistician unfamiliar with survey methods, would be to model all of the stages of sampling with fixed and random effects [e.g., (49)]. However, we have shown on theoretical grounds that such modeling, besides being overly complex, does not automatically improve standard error estimation (50). The one instance when simple survey methods do require some modification is when the sample design is such that only a small number of primary sampling units are available for standard error estimation, e.g., 16 in the Hispanic Health and Nutrition Examination Survey (51). We have given recommendations for this special situation elsewhere (44).

Fuller discussions of the statistical issues involved in analyzing health surveys are given elsewhere (44,50,52-69).

APPLICATIONS

One can categorize analyses of health surveys for cancer-related objectives in various ways: by type of outcome (incidence of cancer, death from cancer, presence of a cancer risk factor, and utilization of a cancer-screening modality), by type of study design (cross-sectional, longitudinal, and case-control), and by type of surveys used in the analysis (single cross-sectional survey, single longitudinal survey, multiple cross-sectional surveys of the same population, and multiple cross-sectional surveys of different populations). The six applications presented in this section were chosen to provide interesting examples of all of these categories and to demonstrate the different ways survey data have been used to address cancer-related issues.

We utilized the computer software SUDAAN (70) and some of our own computer programs to perform the analyses. Other commercial software designed for survey analyses could equally well be used (71).

Factors Associated With Snuff Use, Derived From the 1987 National Health Interview Survey

Smokeless tobacco, i.e., snuff and chewing tobacco, have a number of adverse health effects (72). By use of data from the 1987 National Health Interview Survey (1987 NHIS), a model is developed to identify individuals at higher risk of using snuff based on their characteristics. Such an identification could be useful for targeting prevention initiatives. Table 2, A, contains the estimated proportions of men aged at least 18 years old who use snuff for each of the categories of a group of descriptive variables. We see, for example, that snuff use is low in the Northeast, in central cities, and among blacks and Hispanics. The percentages in Table 2, A, are sample-weighted estimates,

Table 2, A. Univariate descriptive analyses of the proportion of snuff use for men sampled in the 1987 National Health Interview Survey

| Variable | Sample size | Estimated population size, millions | Proportion, %* ± SE† |
|--|-------------|-------------------------------------|----------------------|
| Age, y | | | |
| 18–24 | 2143 | 10.9 | 6.79 ± 0.72 |
| 25–34 | 4026 | 18.7 | 3.59 ± 0.34 |
| 35–44 | 3401 | 14.7 | 2.45 ± 0.31 |
| 45–54 | 2082 | 9.7 | 1.23 ± 0.25 |
| 55–64 | 1870 | 8.3 | 1.82 ± 0.38 |
| 64–74 | 1621 | 6.8 | 1.79 ± 0.26 |
| ≥75 | 865 | 3.3 | 2.58 ± 0.63 |
| Race | | | |
| White | 12 791 | 58.2 | 3.56 ± 0.22 |
| Black | 1780 | 6.8 | 1.23 ± 0.40 |
| Hispanic | 1045 | 5.2 | 0.99 ± 0.37 |
| Other | 392 | 2.1 | 1.93 ± 0.96 |
| Region | | | |
| Northeast | 3201 | 14.9 | 1.21 ± 0.35 |
| Midwest | 4125 | 18.1 | 3.81 ± 0.37 |
| South | 5248 | 23.7 | 3.85 ± 0.38 |
| West | 3434 | 15.5 | 2.99 ± 0.38 |
| Metropolitan Statistical Area (MSA) category | | | |
| Central city MSA | 5387 | 22.5 | 1.56 ± 0.19 |
| MSA, but not in central city | 6794 | 33.2 | 2.76 ± 0.26 |
| Non-MSA | 3827 | 16.5 | 5.91 ± 0.58 |
| Education, y | | | |
| <12 | 3586 | 15.8 | 3.68 ± 0.38 |
| 12 | 5530 | 26.1 | 4.05 ± 0.32 |
| >12 | 6892 | 30.4 | 2.01 ± 0.21 |
| Income‡ | | | |
| ≤7999 | 1822 | 6.0 | 4.77 ± 0.66 |
| 8000–13 999 | 2101 | 8.1 | 4.31 ± 0.43 |
| 14 000–18 999 | 1777 | 7.3 | 3.66 ± 0.51 |
| 19 000–24 999 | 2185 | 9.7 | 3.57 ± 0.55 |
| 25 000–29 999 | 1677 | 7.5 | 2.87 ± 0.51 |
| 30 000–39 999 | 2584 | 12.5 | 2.81 ± 0.39 |
| 40 000–49 999 | 1703 | 8.9 | 2.52 ± 0.45 |
| ≥50 000 | 2159 | 12.1 | 1.68 ± 0.36 |
| Occupation | | | |
| White collar | 5915 | 26.9 | 1.95 ± 0.20 |
| Blue collar | 6372 | 29.9 | 3.11 ± 0.19 |
| Unemployed | 3721 | 15.4 | 2.99 ± 0.38 |
| Marital status | | | |
| Married | 9732 | 49.7 | 2.74 ± 0.20 |
| Unmarried | 6276 | 22.5 | 3.92 ± 0.34 |

Table 2, B. Logistic regression analysis for snuff use based on men sampled in the 1987 National Health Interview Survey (sample size = 16 008, estimated population size = 72.3 million)

| Variable | Coefficient§ ± SE | P (two-sided, Wald statistic) |
|------------------|---|-------------------------------|
| Intercept | -1.77 ± 0.84 | |
| Age, y | | —¶ |
| Age | -12.03 × 10 ⁻² ± 3.11 × 10 ⁻² | |
| Age ² | 11.09 × 10 ⁻⁴ ± 2.82 × 10 ⁻⁴ | |
| Race | | <.001 |
| White | 0# | |
| Black | -1.26 ± 0.33 | |
| Hispanic | -1.45 ± 0.39 | |
| Other | -0.67 ± 0.52 | |
| Region | | .013 |
| Northeast | 0# | |
| Midwest | 0.91 ± 0.31 | |
| South | 1.03 ± 0.31 | |
| West | 1.01 ± 0.33 | |

Table 2, B (continued). Logistic regression analysis for snuff use based on men sampled in the 1987 National Health Interview Survey (sample size = 16 008, estimated population size = 72.3 million)

| Variable | Coefficient§ ± SE | P (two-sided, Wald statistic) |
|--|---|-------------------------------|
| Metropolitan Statistical Area (MSA) category | | <.001 |
| Central city MSA | 0# | |
| MSA, but not in central city | 0.53 ± 0.15 | |
| Non-MSA | 1.02 ± 0.17 | |
| Education, y | | —¶ |
| (E0) <12 | 0# | |
| (E1) 12 | -0.70 ± 0.97 | |
| (E2) >12 | -0.52 ± 1.25 | |
| Income (\$1000)** | -1.09 × 10 ⁻² ± 0.50 × 10 ⁻² | .030 |
| Occupation | | .011 |
| (J0) White collar | 0# | |
| (J1) Blue collar | 0.40 ± 0.15 | |
| (J2) Unemployed | 0.09 ± 0.22 | |
| Marital status | | N.S.†† |
| Married | 0# | |
| Unmarried | -0.13 ± 0.12 | |
| Age × education | | <.001 |
| (Age × 12 y) | 6.30 × 10 ⁻² ± 4.60 × 10 ⁻² | |
| (Age × >12 y) | 3.64 × 10 ⁻² ± 6.15 × 10 ⁻² | |
| (Age ²) × (12 y) | -10.54 × 10 ⁻⁴ ± 4.82 × 10 ⁻⁴ | |
| (Age ²) × (>12 y) | -8.22 × 10 ⁻⁴ ± 6.46 × 10 ⁻⁴ | |

*Proportions are estimated with the use of the sample weights.

†Standard errors (SEs) account for the fact that the estimated proportions are weighted and that there is clustering in the sample design.

‡Annual family income, in dollars.

§Regression coefficients are estimated with the use of the sample weights.

||Standard errors (SEs) account for the fact that the estimated coefficients are weighted and that there is clustering in the sample design.

¶Dash (—) means not applicable because a higher-order interaction is in the model.

#Reference category.

**The mid-point of the dollar ranges specified on the questionnaires was used, e.g., 14 500 for 14 000–14 999. For the highest income category, ≥50 000, the value 55 000 was used.

††Not statistically significant.

so that they are representative of the U.S. civilian noninstitutionalized population in 1987.

For some applications, the univariate descriptive analyses given in Table 2, A, might be sufficient to address the relevant issues, for example, suggesting areas to target for a national advertising campaign to reduce snuff use. However, the results in this table have a major limitation: They do not provide estimates of the probabilities of snuff use for various combinations of levels of the independent variables. One could approach this by estimating proportions of snuff use cross-classified by more than one variable, for example, for each of the 16 (i.e., 4 × 4) cells defined by combinations of race-by-region categories. However, the sample sizes in these cells may become too small to yield reliable estimators of the proportions, especially if one cross-classifies by more than two variables.

Instead of cross-classifications, a logistic regression analysis can be used to model the probability of snuff use as a function of the levels of all the independent variables (Table 2, B). For example, by use of the estimated logistic regression coefficients in Table 2, B, one can estimate the probability of snuff use as 5.2% for a 40-year-old white individual who lives in the South in a non-Metropolitan Statistical Area, has 12 years of education

and a (family) income of \$15 000, who is a white-collar worker, and who has never been married. The interpretation of the individual regression coefficients in Table 2, B, is the same as that for a logistic regression in the nonsurvey setting. For example, the coefficient for a categorical independent variable not involved in an interaction is the logarithm of the estimated odds ratio (relative risk) of that category compared with the baseline category. For example, the estimated odds ratio associated with being unmarried is $0.88 = \exp(-.013)$. Note that this odds ratio is less than 1.0, which is the opposite of what is suggested by Table 2, A. However, the logistic regression result controls for possible imbalances in the other variables (e.g., age) between married and unmarried individuals.

Digital Rectal Examinations and Type of Health Insurance Coverage by Using the 1992 National Health Interview Survey

The American Cancer Society recommends annual digital rectal examinations for individuals aged 40 years or over for cancer screening (73). Of interest is the association of the probability that an individual has had an annual digital rectal examination with his or her type of health insurance; a full analysis including other types of cancer screening is given elsewhere (13). The third column of Table 3 shows the estimated proportions of digital rectal examinations cross-classified by type of health insurance, by use of the Cancer Control Supplement to the 1992 National Health Interview Survey (1992 NHIS) (74). In terms of a causal association between health insurance and the probability of examination, the observed proportions can be misleading because they do not control for patient characteristics. Thus, for example, the estimated proportion for individuals with

public health insurance may appear low because these individuals have a lower income, and income is positively associated with the probability of examinations.

Column 4 of Table 3 shows the predictive margin for the probability of digital rectal examination controlling for the demographic variables given in footnote § of the table. The method of predictive margins is a form of direct standardization by use of a regression model and has been developed for handling dichotomous outcomes analyzed with the use of nonlinear models like logistic regression (60,75). The predictive margin in the fourth column of Table 3 represents the probability of an examination for a hypothetical population with the same distribution of demographic characteristics as the 1992 NHIS, but where all of the individuals had each one of the health insurance types in turn. Controlling for individual characteristics in this way, we see that the estimated probability of an examination for the public health insurance group is actually higher than the other groups.

The predictive margin in the fifth column of Table 3 estimates the probability of an examination for a population with the same distribution of individual characteristics as the subgroup with no health insurance, but where all of the individuals had each one of the health insurance types in turn. This predictive margin is relevant if one was contemplating what would happen to the probabilities if the individuals with no insurance obtained different kinds of health insurance.

Body Iron Stores and Risk of Developing Cancer, Derived From the NHANES I Epidemiologic Followup Study

The first National Health and Nutrition Examination Survey (NHANES I) was conducted during the period from 1971 through 1975. Individuals sampled who were aged 25–74 years in this survey (or their proxies) have been periodically contacted concerning their health status as part of the NHANES I Epidemiologic Followup Study. This longitudinal study allows for estimating the association of risk factors measured at the baseline survey (and possibly updated with information from the follow-ups) with the development of various diseases. This type of longitudinal analysis could also be approached with a single cross-sectional survey by asking sampled individuals for a history of their risk factors. However, longitudinal surveys have the advantage that individuals sampled in the baseline survey—but who later die—provide important information, whereas obtaining information about dead individuals is problematic in a cross-sectional survey. In addition, individuals in a cross-sectional survey may not be able to recall accurately their risk factor exposures from the past, and certain risk factors that require immediate evaluation (such as blood chemistries) would not be available.

In this section, we estimate the association of transferrin saturation, a measure of body iron stores, with the development of cancer. This association was previously studied as part of a more general analysis (30). We restrict the analysis to the 1971–1974 cohort of NHANES I because the blood chemistry variables were measured only for that cohort. Individuals with cancer at the baseline survey are excluded from the analysis. Stevens et al. (30) expressed the theoretical concern that preclinical cancer might affect serum chemistry values and therefore restricted analysis to individuals who were alive and cancer free for at least 4 years after the baseline survey. For the same reason, in our

Table 3. Proportions and predictive margins for the probability of digital rectal examination as a function of type of health insurance plan, based on data from individuals between 40 and 64 years of age sampled in the 1992 National Health Interview Survey (sample size = 3657, estimated population size = 57.0 million)

| Health insurance* | Sample size | Proportion† ± SE‡ | Predictive margin§ ± SE (population = all) | Predictive margin§ ± SE (population = none)¶ |
|-------------------|-------------|-------------------|--|--|
| None | 532 | 0.13 ± 0.02 | 0.14 ± 0.02 | 0.13 ± 0.02 |
| FFS (large) | 1153 | 0.34 ± 0.02 | 0.33 ± 0.02 | 0.27 ± 0.02 |
| FFS (other) | 867 | 0.30 ± 0.02 | 0.29 ± 0.02 | 0.22 ± 0.02 |
| HMO/PPO | 813 | 0.37 ± 0.02 | 0.37 ± 0.02 | 0.35 ± 0.03 |
| Public | 292 | 0.30 ± 0.03 | 0.45 ± 0.07 | 0.35 ± 0.05 |

*None = no private or public health care coverage reported; FFS (large) = one of the 50 largest fee-for-service plans held privately or through employer; FFS (other) = fee-for-service plan held privately or through employer, but not one of the 50 largest; HMO/PPO = enrolled in a Health Maintenance Organization or Preferred Provider Organization; Public = Medicaid or other public assistance program, but not an HMO/PPO.

†Proportions are estimated with the use of the sample weights.

‡Standard errors (SEs) account for the fact that the estimated proportions are weighted and that there is clustering in the sample design.

§Predictive margins control for age, family income (<20 000, ≥20 000), sex, race (white, black, and Hispanic), education (<12 years, 12 years, and >12 years), marital status, and self-reported health status (fair/poor, good, and excellent/very good).

||Standardizing population is all of the target population of the 1992 National Health Interview Survey.

¶Standardizing population is subpopulation of individuals who belong to the health insurance = "None" group.

analyses, we remove the first 4 years of follow-up for all individuals.

Table 4 shows the results of proportional hazards regressions for the association of developing cancer with transferrin saturation, smoking, race, income (family), and type of census enumeration district in which the individual lived. The same as in a proportional-hazards regression used to analyze a randomized clinical trial, the regression coefficient for a variable is interpreted as the logarithm of the relative hazard associated with a change in one unit in that variable. So, for example, the interquartile range (75th percentile minus 25th percentile) of transferrin saturation for men in this population is 13.7, and the relative hazard associated with this difference is $1.22 = \exp(13.7 \times 14.29 \times 10^{-3})$. Therefore, transferrin saturation is positively associated with the risk of cancer for men; *see also* Fig. 1. For women, the association is not statistically significantly different from zero.

The analyses shown in Table 4 do not use the sample weights because, for NHANES I, the sample weights are very variable, which leads to very large standard errors of estimated weighted regression coefficients. Therefore, we use unweighted estimators but include as additional independent variables the variables that were used in constructing the sample weights (race, income, and enumeration district). The sample clustering is taken into account in the analyses in Table 4, by the use of the primary sampling unit designations of the individuals available on the public-use data files.

Beside using the clustering in the data, there are some other important methodological differences between the analyses shown in Table 4 and the typical analysis of a randomized clinical trial. (Since these differences are somewhat technical, some readers may wish to skip the rest of this paragraph.) (a) In a

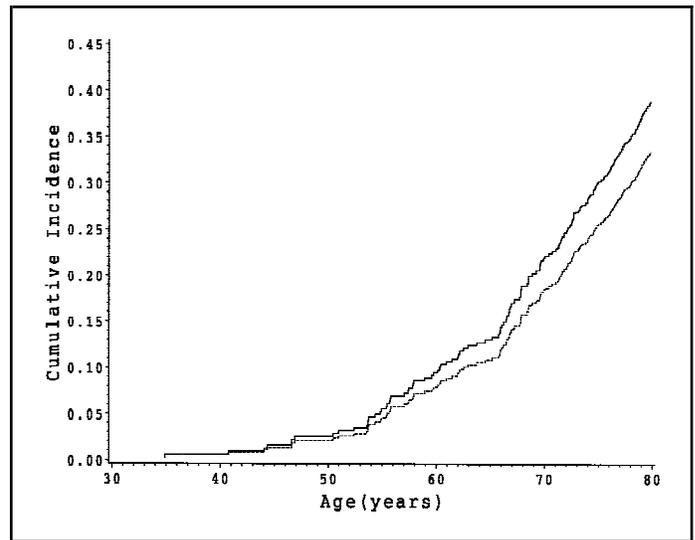


Fig 1. Predictive margins for cause-specific cumulative incidence of cancer for men for transferrin saturation at the 25th percentile (lower line) and the 75th percentile (upper line)

typical analysis of a randomized clinical trial, one would start the time axis at the time of randomization, which is usually close to the diagnosis of a certain stage of disease or the start of treatment. The time scale here is taken as age rather than time from the baseline survey because age is a more important determinant of the risk of cancer than time from the baseline survey for a healthy population (65). (b) Note that transferrin saturation is being measured only once for each individual, at the time of his or her baseline survey. The possibility that transferrin

Table 4. Unweighted regression coefficients for a proportional hazards regression of the incidence of developing cancer (as a function of age) on transferrin saturation and other covariates with data from the NHANES I Epidemiologic Followup Study; analysis stratified by 5-year birth cohorts

| Variable | Men (sample size = 3290, No. of cancer events = 232, estimated population size = 40.6 million) | | Women (sample size = 5270, No. of cancer events = 197, estimated population size = 45.9 million) | |
|------------------------|---|------|---|------|
| | Coefficient \pm SE* | P† | Coefficient \pm SE | P† |
| Transferrin saturation | $14.29 \times 10^{-3} \pm 4.36 \times 10^{-3}$ | .002 | $6.80 \times 10^{-3} \pm 6.82 \times 10^{-3}$ | N.S. |
| Smoking | | .002 | | N.S. |
| Never smoked | 0‡ | | 0‡ | |
| Former smoker | -0.22 ± 0.15 | | 0.03 ± 0.24 | |
| Current smoker | 0.12 ± 0.19 | | -0.05 ± 0.17 | |
| Unknown | 0.63 ± 0.19 | | -0.13 ± 0.31 | |
| Race | | | | |
| White | 0‡ | | 0‡ | |
| Nonwhite | 0.03 ± 0.18 | N.S. | -0.37 ± 0.22 | N.S. |
| Income, family | | N.S. | | .040 |
| <\$3000 | 0‡ | | 0‡ | |
| \$3000–\$6999 | 0.13 ± 0.18 | | -0.51 ± 0.22 | |
| \$7000–\$9999 | -0.13 ± 0.22 | | -0.07 ± 0.20 | |
| \$10 000–14 999 | 0.07 ± 0.24 | | -0.60 ± 0.24 | |
| \geq \$15 000 | -0.17 ± 0.25 | | -0.36 ± 0.32 | |
| Enumeration district | | | | |
| Nonpoverty | 0‡ | | 0‡ | |
| Poverty | 0.01 ± 0.15 | N.S. | 0.05 ± 0.12 | N.S. |

*Standard errors (SEs) account for the fact that there is clustering in the sample design.

†N.S. = Not statistically significant. P values are two-sided (Wald statistic).

‡Reference categories.

saturation means different things for individuals of different ages is controlled for in the analysis by stratifying on 5-year birth cohorts. (c) Individuals dying of other causes are considered to have their data censored at the time of death. Although there are other possibilities, this “cause-specific” analysis is most appropriate for understanding the biology of the association (76).

The analyses shown here are based on the 1987 follow-up of the NHANES I cohort; there was also a follow-up in 1992. Beside follow-ups involving contacts with the sampled individuals, the National Center for Health Statistics is also providing follow-up by use of the National Death Index (77) for this survey and for some of their other surveys, e.g., the second National Health and Nutrition Examination Survey. This type of follow-up provides less information than a personal contact but can be quite useful for studying associations with causes of death.

Changing Rates of Mammography Screening by Use of the National Health Interview Survey

Many of the annual National Health Interview Surveys include supplemental questionnaires that contain questions on cancer-related topics. The questionnaires in the 1987, 1990, and 1992 surveys contained similar questions about mammography screening (74,78,79), which we exploit to examine trends in mammography screening. We focus here on women more than 40 years of age and on whether or not they reported a screening mammographic examination in the last year. In particular, we examine how the percentages of these women reporting a screening mammographic examination are changing over time and whether the time trends are different depending on the educational levels of the women. Similar analyses with the 1987 and 1990 surveys have been previously performed (3).

Table 5 shows the estimated percentages of women reporting screening mammograms, cross-classified by level of education. The percentage of women reporting screening mammograms is increasing over the years from 1987 through 1992. It also appears that the increases are similar for the different educational levels. This could be formally checked by performing further analyses, which could also control for additional variables, such as race and whether or not the woman lived in an urban or rural area.

An interesting statistical consideration in analyzing these data is that, although the individuals sampled in each survey are different, the estimated percentages given in Table 5 are not statistically independent. This is because the same set of

sampled counties (primary sampling units) were used for the National Health Interview Survey in the years from 1985 through 1994. This is taken into account in the joint analysis of multiple years of the survey by use of the primary sampling unit sample design for any one year of the survey when estimating standard errors.

Smoking and Alcohol Consumption as Risk Factors for Digestive Cancer Using the 1986 National Mortality Followback Survey and the 1987 National Health Interview Survey

We used data from the 1987 National Health Interview Survey (1987 NHIS) and the 1986 National Mortality Followback Survey (1986 NMFS) to estimate digestive cancer death rates and their association with alcohol consumption and smoking. The 1986 NMFS is an example of a “followback” survey, in which individuals associated with sampled records are interviewed for further information. In this case, the records are death certificates, but other followback surveys use other types of records, e.g., birth certificates (80). In this section, we follow the general approach of Sterling et al. (35), who used the 1987 NHIS and the 1986 NMFS to examine the association of smokeless tobacco with oral and digestive cancer mortality. These surveys can be used together to estimate the death rates, where the numerators of the rates are estimated from the 1986 NMFS and the denominators are estimated from the 1987 NHIS. This is an example of a population-based, case-control study. Two advantages of a population-based, case-control study over a study that is not population based (e.g., hospital case patients with hospital control patients) are that many of the biases involved in choosing an appropriate control population are eliminated, and one can estimate death rates in addition to relative risks.

Because the information on deaths and alive individuals come from two different surveys, biases can arise because the modes of data collection of the surveys are different (mail response from an informant for the decedent for the 1986 NMFS and face-to-face interview for the 1987 NHIS) and the questions asked are different. In addition, the populations sampled differ in ways other than vital status. For example, the 1987 NHIS sampled only civilians and noninstitutionalized individuals, and the 1986 NMFS sampled deaths only for individuals aged 25 years or older. To make the populations sampled in the two surveys comparable, the analysis is restricted to civilians aged 25 years or older who, if they died, were institutionalized less than half of their last year of life.

Table 5. Estimated percentages of women over the age of 40 years reporting screening mammograms in the last year based on the 1987, 1990, and 1992 National Health Interview Surveys by years of education

| | Year of survey | | |
|--------------|--|---|--|
| | 1987 (sample size = 6449, estimated population size = 45.5 million), %* ± SE† | 1990 (sample size = 12 485 estimated population size = 48.8 million), %* ± SE† | 1992 (sample size = 3646, estimated population size = 49.7 million), %* ± SE† |
| Education, y | | | |
| <12 | 13.6% ± 0.9% | 23.6% ± 0.9% | 26.1% ± 1.6% |
| 12 | 23.4% ± 1.1% | 33.8% ± 0.8% | 39.7% ± 1.6% |
| >12 | 29.2% ± 1.3% | 42.0% ± 1.0% | 46.6% ± 1.9% |
| Overall | 21.9% ± 0.6% | 33.7% ± 0.5% | 38.6% ± 1.1% |

*Percents are estimated with the use of the sample weights.

†Standard errors (SEs) account for the fact that the estimated percents are weighted and that there is clustering in the sample design.

Table 6, A, displays a univariate descriptive analysis of the annual digestive cancer death rates cross-classified by sex, race, age, drinking, and smoking levels. The outcome variable is death due to digestive cancer in 1986, defined by an underlying cause of death with ICD-9th Revision codes 150–159 (81). (The 1986 NMFS public-use data files include a variable coded for this cause of death.) There were 785 such deaths sampled in the 1986 NMFS. Drinking or smoking information from 134 of the 785 digestive cancer deaths and from 1560 observations of the 19240 living individuals was partially or completely missing. Rather than just eliminating from the analyses these observations with missing drinking or smoking data, we used a “hot-deck” imputation to fill in their missing data with the data from randomly chosen individuals with nonmissing data who were similar to the individuals with the missing data [(82), p. 62–67]; each donor was chosen to be in the same age/race/sex category, and, if available, the same smoking or drinking category as the recipient. Imputation is a common technique in the analysis of surveys to minimize potential bias in measuring associations due to missing data. In the present application, in which the amount of missing drinking/smoking data is larger for the 1986 NMFS

than the 1987 NHIS, the imputation also serves to keep the rates in Table 6, A, from being biased low.

Beside taking account of sample design, there are two additional statistical subtleties in calculating the standard errors given in Table 6, A. The first is that, even though we are only interested in digestive cancer deaths, it is important not to just delete the other deaths from the dataset when performing the analysis. It can be shown that this will lead to standard errors being estimated incorrectly (59). Instead, all of the deaths should be kept in the analysis dataset and the subpopulation of interest (e.g., digestive cancer deaths) should be specified in the code of the computer software used for the analysis. The second subtlety concerns the imputation for the missing data. The standard errors shown in Table 6, A, treat the imputed data as if they were real data and thus are underestimates. Unfortunately, to our knowledge, at this time commercial computer software for surveys does not exist that can properly take account of the fact that some data have been imputed. However, calculations by use of our own computer programs show that the effect on the standard errors for this particular application are small (results not shown).

Table 6, A. Univariate descriptive analyses of digestive cancer death rates based on the 1986 National Mortality Followback Survey and the 1987 National Health Interview Survey

| Variable | Sample size | Estimate population size, millions | Annual rate* (per 10 ⁴) ± SE† |
|-----------|-------------|------------------------------------|---|
| Sex | | | |
| Male | 8325 | 70.4 | 7.26 ± 0.44 |
| Female | 11 700 | 78.8 | 6.48 ± 0.39 |
| Race | | | |
| White | 16 699 | 129.6 | 6.90 ± 0.33 |
| Nonwhite | 3326 | 19.6 | 6.51 ± 0.52 |
| Age, y | | | |
| 25–44 | 9782 | 76.0 | 0.44 ± 0.05 |
| 45–64 | 5554 | 44.7 | 6.15 ± 0.46 |
| 65–84 | 4305 | 26.6 | 23.18 ± 1.37 |
| ≥85 | 384 | 2.0 | 49.16 ± 7.69 |
| Drinking‡ | | | |
| 0 | 6477 | 47.0 | 4.46 ± 0.41 |
| 1–52 | 5232 | 38.6 | 8.92 ± 0.67 |
| 53–365 | 4787 | 36.7 | 5.61 ± 0.54 |
| ≥365 | 3529 | 26.9 | 9.76 ± 0.83 |
| Smoking§ | | | |
| 0–19 | 9353 | 68.6 | 6.42 ± 0.42 |
| 20–11 999 | 7667 | 58.1 | 5.03 ± 0.39 |
| ≥12 000 | 3005 | 22.6 | 12.80 ± 1.08 |

Table 6, B. Predictive margin|| (annual digestive cancer death rate per 10⁴ ± SE) for drinking, smoking, and drinking by smoking

| | Smoking, lifetime No. of packs of cigarettes smoked | | | Overall |
|---|---|--------------|--------------|--------------|
| | 0–19 | 20–11 999 | ≥12 000 | |
| Drinking, No. of alcoholic drinks consumed per year as an adult | | | | |
| 0 | 4.14 ± 0.44 | 2.16 ± 0.51 | 0.74 ± 0.34 | 2.78 ± 0.26 |
| 1–52 | 8.42 ± 1.05 | 10.87 ± 1.57 | 12.64 ± 1.84 | 9.95 ± 0.79 |
| 53–365 | 8.38 ± 1.37 | 10.21 ± 1.77 | 7.83 ± 1.64 | 8.57 ± 0.93 |
| ≥365 | 12.15 ± 2.42 | 10.38 ± 1.69 | 17.64 ± 2.48 | 12.58 ± 1.39 |
| Overall | 6.79 ± 0.50 | 6.52 ± 0.55 | 6.69 ± 0.59 | |

*Rates are estimated with the use of the sample weights.

†Standard errors (SEs) account for the fact that the estimated rates are weighted and that there is clustering in the sample design.

‡Number of alcoholic drinks consumed per year as an adult.

§Lifetime number of packs of cigarettes smoked.

||Predictive margin controls for sex, race (white versus nonwhite), and age category (25–44 years, 45–64 years, 65–84 years, and ≥71–85 years).

Since the rates in Table 6, A, are not age adjusted (except for the cross-classification by age), they can be misleading. For example, individuals in the three smoking categories have mean ages of 47, 43, and 57 years, respectively, so that the differences in the observed cancer death rates could easily be due to the age differences. A logistic regression analysis of digestive cancer deaths was performed with the independent variables being sex, race, age, smoking, and drinking (results not shown). By use of this regression, we calculated the predictive margin for drinking, smoking, and drinking by smoking (Table 6, B). When controlling for age, we see no association of smoking and digestive cancer death rates (last row of Table 6, B). There is, however, an association of drinking and digestive cancer mortality (last column of Table 6, B). The exact pattern of the association of drinking and the rates seems to depend on level of smoking; we have no plausible explanation for this pattern.

An early example of a population-based, case-control study using separate samples of deaths and living individuals is given by Haenszel et al. (83). They used a 10% sample of all deaths in the United States in 1958 and information collected from a supplement to the Current Population Survey in May 1958 to study the associations of smoking and residence (urban versus rural) with lung cancer mortality. Most population-based, case-control studies do not sample deaths but use information on all of the case patients in certain geographic areas (e.g., obtained from tumor registries) and a sample of control subjects from the same areas (e.g., by use of telephone surveys). For example, Brinton et al. (84) attempted to interview all of the women who were newly diagnosed with breast cancer during a fixed time period identified by tumor registries in three geographic areas and used telephone surveys of the same areas to ascertain the control population.

COMMIT, a Community Intervention Trial of Smoking Cessation

The Community Intervention Trial for Smoking Cessation (COMMIT) was an experiment in which one community from each of 11 matched community pairs was randomly assigned to a 4-year community level intervention to help smokers quit smoking (85,86). The other community from each matched pair served as a control for comparison purposes. The two communities within each pair were matched for geographic location, population size, general sociodemographic factors, and esti-

mated smoking prevalence rates. Although there were multiple objectives and analyses of COMMIT, we focus here on the effects of the intervention on the prevalence of adult cigarette smoking, one of the secondary analyses. These effects were assessed by performing telephone surveys in the 22 communities before and after the intervention, in 1988 and 1993, respectively (87). The telephone surveys used list-assisted random-digit dialing, in which blocks of 100 consecutive telephone numbers were first classified into two strata, depending on whether one or more numbers in the block were listed in residential telephone directories. Telephone numbers in blocks with residential numbers were sampled at a higher rate. The sample sizes and population sizes of the community surveys were varied but averaged about 4900 and 77 000, respectively. The baseline surveys were performed before the randomization.

Table 7 shows the estimated smoking prevalence for each of the communities before and after the intervention. Averaged over the 11 intervention communities, the smoking prevalence went down 2.9 percentage points, from 24.6% to 21.6%. However, the smoking prevalence also went down in the 11 comparison communities an average of 2.7 percentage points, from 25.1% to 22.5%. The intervention offered a 0.27 percentage point advantage in the lowering of the smoking prevalence over the comparison, with a 90% confidence interval for this advantage given by (-0.74 to 1.28). One can conclude that the intervention did not have an impact on smoking prevalence beyond the general decreasing time trends. This study reinforces the importance of having a control group; without such a group, one might have incorrectly assumed that the intervention was successful in lowering smoking prevalence.

AVAILABILITY OF PUBLIC-USE DATA FILES

If a survey is required of individuals living in a relatively small geographic area, e.g., of the control population associated with patient information obtained from a tumor registry, then investigators can hire a private organization to perform the survey for them. This is also a reasonable option for a more widely dispersed population if there exists a list of the individuals in the population to be sampled, e.g., physicians who are board certified in a certain specialty. Obtaining national estimates for the general population is a much larger undertaking. Fortunately, many U.S. government agencies make the data available from the surveys that they sponsor to investigators who are interested

Table 7. Smoking prevalence among adults (ages >18 years) for each community in years 1988, 1993, and the change from 1988 to 1993, expressed as percentage of adults smoking

| Community pair | Intervention communities | | | Comparison communities | | | Difference* |
|-----------------|--------------------------|------|--------|------------------------|------|--------|-------------|
| | 1988 | 1993 | Change | 1988 | 1993 | Change | |
| 1 | 26.1 | 19.4 | -6.7 | 24.9 | 19.4 | -5.5 | -1.20 |
| 2 | 32.0 | 29.8 | -2.2 | 28.1 | 24.8 | -3.3 | 1.06 |
| 3 | 22.4 | 21.8 | -0.6 | 26.2 | 23.7 | -2.5 | 1.90 |
| 4 | 26.3 | 23.1 | -3.2 | 29.1 | 26.1 | -3.0 | -0.24 |
| 5 | 26.5 | 21.1 | -5.4 | 28.8 | 26.2 | -2.6 | -2.80 |
| 6 | 22.0 | 18.6 | -3.4 | 19.5 | 17.0 | -2.6 | -0.84 |
| 7 | 24.8 | 22.4 | -2.3 | 24.9 | 20.0 | -4.9 | 2.56 |
| 8 | 26.5 | 24.2 | -2.3 | 25.5 | 23.3 | -2.2 | -0.12 |
| 9 | 22.8 | 19.7 | -3.2 | 25.7 | 26.0 | 0.4 | -3.51 |
| 10 | 21.1 | 19.9 | -1.1 | 18.3 | 16.2 | -2.1 | 0.93 |
| 11 | 20.0 | 18.2 | -1.9 | 25.5 | 24.4 | -1.2 | -0.74 |
| Community means | 24.6 | 21.6 | -2.9 | 25.1 | 22.5 | -2.7 | -0.27 |

*Difference in change in intervention community minus the change in comparison community.

in performing their own analyses. These data can be obtained in computer-readable form at many university and government libraries or computer centers and are also available directly from many government-agency websites.

The "Appendix" section contains a list of such surveys that collect health information on a national scale. Some of these surveys are focused on health information, whereas others collect health-related information peripherally. We have not tried to restrict this list to surveys collecting cancer-related information because any health-related variable could conceivably be of interest to a cancer-related investigation. We do not claim that this list of surveys is exhaustive and suggest contacting the individual agencies for further information about their surveys; *see also* the FEDSTATS website (www.fedstats.gov) and the website of the Inter-university Consortium for Political and Social Research, Institute for Social Research, University of Michigan (www.icpsr.umich.edu).

Besides the information collected in the surveys listed in the "Appendix" section, investigators also have the option of collaborating with one of the government sponsors to have certain required information collected in the future sample of one of the continuing surveys. This approach is useful only for long-term projects, since the time span for a national survey from idea for data collection to data availability can be years.

Although the focus of this review has been surveys, it is useful to note that health-related data are also available for some variables on essentially the whole population. For example, the National Center for Health Statistics makes available data files that have information on every birth and death occurring within the United States ("vital statistics"), and the Health Care Financing Administration has information available on all Medicare claims in the United States. Interested investigators should contact the relevant agencies for information about what resources are available.

LIMITATIONS OF USING SURVEY DATA FOR ANALYSES

National health surveys will generally not contain sufficient numbers of sampled individuals to estimate directly parameters for small geographic areas, e.g., county-specific rates. Although statistical methods, usually referred to as "small area estimation," have been developed to estimate indirectly such parameters (88), for many applications it will be necessary to perform a survey of the particular geographic area needed. This was the situation in the breast cancer population-based, case-control study mentioned previously, in which telephone surveys were performed in the geographic areas where the case patients were identified. National health surveys may also not contain sufficient numbers of individuals in special populations (e.g., Native Americans, the oldest-old) to perform analyses restricted to these populations. Some surveys oversample some special populations to be able to provide reliable inferences for them. (It is interesting to note that the use of sample-weighted estimators allows one to also make valid national estimates from these surveys despite the oversampling.)

It is sometimes possible to avoid the problem of small sample sizes sampled by health surveys in local geographic areas and subpopulations by the use of other data sources. For example, the Current Population Survey, which is designed to provide characteristics of the labor force, occasionally uses supplements that involve health issues. An example is given by the Tobacco Use Supplement that was added to 3 months of the Current

Population Survey during the period from 1992 through 1993 (89). (The Current Population Survey samples approximately 60 000 households in the United States each month, which is much larger than nearly all health surveys.) As mentioned previously, an additional data source is information on vital statistics or other records that are kept on all individuals. If the research question can be answered with these resources, then the problem of small numbers can be avoided.

The availability of public-use data files from health surveys and commercial software for performing survey analyses makes it easy for investigators to address their research questions by use of survey data. However, the ability to analyze survey data should not lead one to ignore the principles of good scientific method. In particular, given the large numbers of variables recorded in a typical survey, it is easy to examine a multitude of possible associations and discover some spurious ones. In addition, except in the situation of randomized assignment of an intervention (e.g., COMMIT described above), associations found between risk factors and outcomes could theoretically always be due to confounding variables and not be of a causal nature. Finally, as is well known, statistical significance is not the same as scientific importance. With large sample sizes of some surveys, it is possible that an association between two variables can be small, yet still be statistically significant (e.g., $P < .05$). Presentation of confidence intervals with estimates can help avoid misinterpretations. With these caveats in mind, we believe that there is tremendous potential in using health surveys as a resource of addressing cancer-related objectives.

APPENDIX

U.S. Government-sponsored surveys that collect health information that have public-use data files (with years of surveys)

- Agency for Health Care Policy and Research (<http://www.ahcpr.gov>):
 - Medical Expenditure Panel Survey (1996)
 - National Medical Care Expenditure Survey (1977)
 - National Medical Expenditure Survey (1987)
- Bureau of Labor Statistics (<http://www.bls.gov>):
 - Consumer Expenditure Survey (1980-)
 - National Longitudinal Surveys (four cohorts in 1966, 1967, 1968, and 1979, with follow-up)
- Bureau of the Census (<http://www.census.gov>):
 - Current Population Survey (1980-)
- Centers for Disease Control and Prevention (<http://www.cdc.gov>):
 - Behavioral Risk Factor Surveillance System (1984-)
 - Youth Risk Behavior Surveillance Survey (1990-)
- Department of Agriculture (<http://www.usda.gov>):
 - Continuing Survey of Food Intakes by Individuals (1985-1986, 1989-1991, and 1994-1996)
 - Diet and Health Knowledge Survey (1989-1991 and 1994-1996)
 - Nationwide Food Consumption Survey (periodic, starting in 1936)
- Health Care Financing Administration (<http://www.hcfa.gov>):
 - Medicare Current Beneficiary Survey (1991-, with follow-up)
- National Center for Health Statistics (<http://www.cdc.gov/nchswww>):
 - Hispanic Health and Nutrition Examination Survey (1982-1984)
 - National Ambulatory Medical Care Survey (1974-1981, 1985, and 1989-)
 - National Employer Health Survey (1994)
 - National Health Examination Survey (1959-1970)

National Health Interview Survey (1957–, with Longitudinal Study of Aging)

National Health and Nutrition Examination Surveys (periodic, starting in 1971, with follow-up of NHANES I)

National Home and Hospice Care Survey (1992–)

National Hospital Ambulatory Medical Care Survey (1992–)

National Hospital Discharge Survey (1965–)

National Maternal and Infant Health Survey (1988, with follow-up)

National Medical Care Utilization and Expenditure Survey (1980)

National Mortality Followback Survey (periodic, starting in 1961)

National Nursing Home Survey (periodic, starting in 1973)

National Natality Surveys (1963, 1964–1966, 1968–1969, 1972, and 1980)

National Survey of Ambulatory Surgery (1994–)

National Survey of Family Growth (periodic, starting in 1973)

National Survey of Personal Health Practices and Consequences (1979–1980)

National Institute on Aging/Center for Demographic Studies, Duke University (<http://cds.duke.edu>):

National Long Term Care Survey (1982, 1984, 1989, and 1994, with follow-up)

National Institute on Alcohol Abuse and Alcoholism (<http://www.niaaa.nih.gov>):

National Longitudinal Alcohol Epidemiologic Survey (1992)

National Institute of Dental and Craniofacial Research (<http://www.nidr.nih.gov>):

National Surveys of Oral Health (1985–1987)

National Institute on Drug Abuse (<http://www.nida.nih.gov>):

Monitoring the Future Study (1975–, with follow-up)

National Pregnancy and Health Survey (1992–1993)

National Institute of Mental Health/Department of Health Care Policy, Harvard Medical School (<http://www.hcp.med.harvard.edu>):

National Comorbidity Survey (1991)

Substance Abuse and Mental Health Services Administration (<http://www.samhsa.gov>):

National Household Survey on Drug Abuse (1971–)

REFERENCES

- (1) Brett KM, Madans JH. Use of postmenopausal hormone replacement therapy: estimates from a nationally representative cohort study. *Am J Epidemiol* 1997;145:536–45.
- (2) Anderson LM, May DS. Has the use of cervical, breast, and colorectal cancer screening increased in the United States? *Am J Public Health* 1995; 85:840–2.
- (3) Breen N, Kessler L. Changes in the use of screening mammography: evidence from the 1987 and 1990 National Health Interview Surveys. *Am J Public Health* 1994;84:62–7.
- (4) Gilpin EA, Lee L, Evans N, Pierce JP. Smoking initiation rates in adults and minors: United States, 1944–1988. *Am J Epidemiol* 1994;140:535–43.
- (5) Haynes SG, Harvey C, Montes H, Nickens H, Cohen BH. Patterns of cigarette smoking among Hispanics in the United States: results from HHANES 1982–84. *Am J Public Health* 1990;80 Suppl:47–53.
- (6) Klevens RM, Giovino GA, Peddicord JP, Nelson DE, Mowery P, Grummer-Strawn L. The association between veteran status and cigarette-smoking behaviors. *Am J Prev Med* 1995;11:245–50.
- (7) Pirkle JL, Flegal KM, Bernert JT, Brody DJ, Etzel RA, Maurer KR. Exposure of the US population to environmental tobacco smoke: the Third National Health and Nutrition Examination Survey, 1988 to 1991. *JAMA* 1996;275:1233–40.
- (8) Kutz FW, Cook BT, Carter-Pokras OD, Brody D, Murphy RS. Selected pesticide residues and metabolites in urine from a survey of the U.S. general population. *J Toxicol Environ Health* 1992;37:277–91.
- (9) Patterson BH, Block G, Rosenberger WF, Pee D, Kahle LL. Fruit and vegetables in the American diet: data from the NHANES II survey. *Am J Public Health* 1990;80:1443–9.
- (10) Cotugna N, Subar AF, Heimendinger J, Kahle L. Nutrition and cancer prevention knowledge, beliefs, attitudes, and practices: the 1987 National Health Interview Survey. *J Am Diet Assoc* 1992;92:963–8.
- (11) Ehemann CR, Ford E, Garbe P, Staehling N. Knowledge about indoor radon in the United States: 1990 National Health Interview Survey. *Arch Environ Health* 1996;51:245–7.
- (12) Calle EE, Flanders WD, Thun MJ, Martin LM. Demographic predictors of mammography and Pap smear screening in US women. *Am J Public Health* 1993;83:53–60.
- (13) Potosky AL, Breen N, Graubard BI, Parsons PE. The association between health care coverage and the use of cancer screening tests. Results from the 1992 National Health Interview Survey [published erratum appears in *Med Care* 1998;36:1470]. *Med Care* 1998;36:257–70.
- (14) Horowitz AM, Nourjah PA. Factors associated with having oral cancer examinations among US adults 40 years of age or older. *J Public Health Dent* 1996;56:331–5.
- (15) Engel A, Johnson ML, Haynes SG. Health effects of sunlight exposure in the United States. Results from the first National Health and Nutrition Examination Survey, 1971–1974. *Arch Dermatol* 1988;124:72–9.
- (16) Byrne J, Kessler LG, Devesa SS. The prevalence of cancer among adults in the United States: 1987. *Cancer* 1992;69:2154–9.
- (17) Schatzkin A, Jones DY, Hoover RN, Taylor PR, Brinton LA, Ziegler RG, et al. Alcohol consumption and breast cancer in the epidemiologic follow-up study of the first National Health and Nutrition Examination Survey. *N Engl J Med* 1987;316:1169–73.
- (18) Swanson CA, Jones DY, Schatzkin A, Brinton LA, Ziegler RG. Breast cancer risk assessed by anthropometry in the NHANES I epidemiological follow-up study. *Cancer Res* 1988;48:5363–7.
- (19) Micozzi MS, Carter CL, Albanes D, Taylor PR, Licitra LM. Bowel function and breast cancer in US women. *Am J Public Health* 1989;79:73–5.
- (20) Jones DY, Schatzkin A, Green SB, Block G, Brinton LA, Ziegler RG, et al. Dietary fat and breast cancer in the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study. *J Natl Cancer Inst* 1987;79:465–71.
- (21) Madigan MP, Ziegler RG, Benichou J, Byrne C, Hoover RN. Proportion of breast cancer cases in the United States explained by well-established risk factors. *J Natl Cancer Inst* 1995;87:1681–5.
- (22) Freni SC, Eberhardt MS, Turturro A, Hine RJ. Anthropometric measures and metabolic rate in association with risk of breast cancer (United States). *Cancer Causes Control* 1996;7:358–65.
- (23) Wurzelmann JI, Silver A, Schreinemachers DM, Sandler RS, Everson RB. Iron intake and the risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 1996;5:503–7.
- (24) Funkhouser EM, Sharp GB. Aspirin and reduced risk of esophageal carcinoma. *Cancer* 1995;76:1116–9.
- (25) Yong LC, Brown CC, Schatzkin A, Dresser CM, Slesinski MJ, Cox CS, et al. Intake of vitamins E, C, and A and risk of lung cancer. The NHANES I epidemiologic follow-up study. First National Health and Nutrition Examination Survey. *Am J Epidemiol* 1997;146:231–43.
- (26) Leigh JP. Occupations, cigarette smoking and lung cancer in the epidemiological follow-up to the NHANES I and the California Occupational Mortality Study. *Bull N Y Acad Med* 1996;73:370–97.
- (27) Bourguet CC, Logue EE. Antigenic stimulation and multiple myeloma. A prospective study. *Cancer* 1993;72:2148–54.
- (28) Reichman ME, Hayes RB, Ziegler RG, Schatzkin A, Taylor PR, Kahle LL, et al. Serum vitamin A and subsequent development of prostate cancer in the first National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. *Cancer Res* 1990;50:2311–5.
- (29) Albanes D, Jones DY, Schatzkin A, Micozzi MS, Taylor PR. Adult stature and risk of cancer. *Cancer Res* 1988;48:1658–62.
- (30) Stevens RG, Jones DY, Micozzi MS, Taylor PR. Body iron stores and the risk of cancer. *N Engl J Med* 1988;319:1047–52.
- (31) Zonderman AB, Costa PT Jr, McCrae RR. Depression as a risk for cancer morbidity and mortality in a nationally representative sample. *JAMA* 1989; 262:1191–5.
- (32) Albanes D, Blair A, Taylor PR. Physical activity and risk of cancer in the NHANES I population. *Am J Public Health* 1989;79:744–50.
- (33) Schatzkin A, Hoover RN, Taylor PR, Ziegler RG, Carter CL, et al. Serum cholesterol and cancer in the NHANES I epidemiologic follow-up study. National Health and Nutrition Examination Survey. *Lancet* 1987;2:298–301.
- (34) Hsing AW, Nam JM, Co Chien HT, McLaughlin JK, Fraumeni JF Jr. Risk

- factors for adrenal cancer: an exploratory study. *Int J Cancer* 1996;65:432–6.
- (35) Sterling TD, Rosenbaum WL, Weinkam JJ. Analysis of the relationship between smokeless tobacco and cancer based on data from the National Mortality Followback Survey. *J Clin Epidemiol* 1992;45:223–231.
- (36) Hsing AW, Hoover RN, McLaughlin JK, Co-Chien HT, Wacholder S, Blot WJ, et al. Oral contraceptives and primary liver cancer among young women. *Cancer Causes Control* 1992;3:43–8.
- (37) Nam JM, McLaughlin JK, Blot WJ. Cigarette smoking, alcohol, and nasopharyngeal carcinoma: a case-control study among U.S. whites. *J Natl Cancer Inst* 1992;84:619–22.
- (38) Chow WH, Linet MS, McLaughlin JK, Hsing AW, Chien HT, Blot WJ. Risk factors for small intestine cancer. *Cancer Causes Control* 1993;4:163–9.
- (39) Breslow NE, Day NE. Statistical methods in cancer research. Vol. II—the design and analysis of cohort studies. International Agency for Research on Cancer. IARC Sci Publ 1987:45–6.
- (40) National Center for Health Statistics. Plan and operation of the Third National Health and Nutrition Examination Survey, 1988–94. *Vital Health Stat Series 1*(32) 1994.
- (41) Korn EL, Graubard BI. Examples of differing weighted and unweighted estimates from a sample survey. *Am Stat* 1995;49:291–5.
- (42) Hoem JM. The issue of weights in panel surveys of individual behavior. In: Kasprzyk D, Duncan G, Kalton G, Singh MP, editors. Panel surveys. New York (NY): John Wiley & Sons; 1989. p. 539–65.
- (43) Kalton G. Modeling considerations: discussion from a survey sampling perspective. In: Kasprzyk D, Duncan G, Kalton G, Singh MP, editors. Panel surveys. New York (NY): John Wiley & Sons; 1989. p. 575–85.
- (44) Korn EL, Graubard BI. Analysis of large health surveys: accounting for the sampling design. *J Royal Stat Soc A* 1995;158:263–95.
- (45) Cochran WG. Sampling techniques. 3rd ed. New York (NY): John Wiley & Sons; 1977. p. 240–3.
- (46) Wolter KM. Introduction to variance estimation. New York (NY): Springer-Verlag; 1985.
- (47) Korn EL, Graubard BI. Variance estimation for superpopulation parameters. *Stat Sinica* 1998;8:1131–51.
- (48) Graves EJ. National Hospital Discharge Survey: annual summary, 1990. *Vital Health Stat Series 13* (112)1992.
- (49) Pfeiffermann D, LaVange L. Regression models for stratified multi-stage cluster samples. In: Skinner CJ, Holt D, Smith TM, editors. Analysis of complex surveys. New York (NY): John Wiley & Sons; 1989. p. 237–60.
- (50) Graubard BI, Korn EL. Modelling the sampling design in the analysis of health surveys. *Stat Methods Med Res* 1996;5:263–81.
- (51) Kovar MG, Johnson C. Design effects from the Mexican American portion of the Hispanic Health and Nutrition Examination Survey: a strategy for analysts. *American Statistical Association 1986 Proceedings of the Section on Survey Research Methods*; 1986. p. 396–9.
- (52) Aday LA. Designing and conducting health surveys. 2nd ed. San Francisco (CA): Jossey-Bass; 1996.
- (53) Botman SL, Jack SS. Combining National Health Interview Survey Datasets: issues and approaches. *Stat Med* 1995;14:669–77.
- (54) Brock DB, Beckett LA, Bienias JL. Sample surveys. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. Vol. 5. Chichester (U.K.): John Wiley & Sons; 1998.
- (55) Corder LS, Manton KG. National surveys and the health and functioning of the elderly: the effects of design and content. *J Am Stat Assoc* 1991;86:513–25.
- (56) Cox BG, Cohen SB. Methodological issues for health surveys. New York (NY): Marcel Dekker; 1985.
- (57) Delgado JL, Johnson CL, Roy I, Trevino FM. Hispanic Health and Nutrition Examination Survey: methodological considerations. *Am J Public Health* 1990;80 Suppl:6–10.
- (58) Graubard BI, Fears TR, Gail MH. Effects of cluster sampling on epidemiologic analysis in population-based case-control studies. *Biometrics* 1989;45:1053–71.
- (59) Graubard BI, Korn EL. Survey inference for subpopulations. *Am J Epidemiol* 1996;144:102–6.
- (60) Graubard BI, Korn EL. Predictive margins for survey data. *Biometrics*. In press 1999.
- (61) Ingram DD, Makuc DM. Statistical issues in analyzing the NHANES I Epidemiologic Followup Study. *Vital Health Stat Series 2* (121) 1994.
- (62) Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: accounting for the sampling design. *Am J Public Health* 1991;81:1166–3.
- (63) Korn EL, Graubard BI. Scatterplots with survey data. *Amer Stat* 1998;52:58–69.
- (64) Korn EL, Graubard BI. Analysis of health surveys. New York (NY): John Wiley & Sons. In press 1999.
- (65) Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up for a survey: choice of the time-scale. *Am J Epidemiol* 1997;145:72–80.
- (66) Landis JR, Lepkowski JM, Eklund SA, Stehouwer SA. A statistical methodology for analyzing data from a complex survey: the first National Health and Nutrition Examination Survey. *Vital Health Stat Series 2* (92) 1982.
- (67) McCarthy PJ. Replication: an approach to the analysis of data from complex surveys. *Vital Health Stat Series 2* (14) 1966.
- (68) Rust KF, Rao JN. Variance estimation for complex surveys using replication techniques. *Stat Methods Med Res* 1996;5:283–310.
- (69) Weinkam JJ, Rosenbaum WL, Sterling TD. Computation of relative risk based on simultaneous surveys: an alternative to cohort and case-control studies. *Am J Epidemiol* 1992;136:722–9.
- (70) Shah BV, Barnwell BG, Bieler GS. SUDAAN user's manual, release 7.5. Research Triangle Park (NC): Research Triangle Institute; 1997.
- (71) Cohen SB. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *Am Stat* 1997;51:285–92.
- (72) Consensus Conference. Health applications of smokeless tobacco use. *JAMA* 1986;255:1045–8.
- (73) American Cancer Society. Guidelines for the cancer-related checkup (80-1MM-Rev.2/93-No.2070-LE). Atlanta (GA): American Cancer Society; 1993.
- (74) Benson V, Marano MA. Current estimates from the National Health Interview Survey, 1992. *Vital Health Stat Series 10* (189) 1994.
- (75) Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics* 1982;38:613–21.
- (76) Korn EL, Dorey FJ. Applications of crude incidence curves. *Stat Med* 1992;11:813–29.
- (77) Patterson BH, Bilgrad R. Use of the National Death Index in cancer studies. *J Natl Cancer Inst* 1986;77:877–81.
- (78) Schoenborn CA, Marano M. Current estimates from the National Health Interview Survey, United States, 1987. *Vital Health Stat Series 10* (166) 1988.
- (79) Adams PF, Benson V. Current estimates from the National Health Interview Survey, 1990. *Vital Health Stat Series 10* (181) 1991.
- (80) Sanderson M, Scott C, Gonzalez JF. 1988 National Maternal and Infant Health Survey: methods and response characteristics. *Vital Health Stat Series 2* (125) 1998.
- (81) World Health Organization (WHO). Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death: based on the recommendations of the ninth Revision Conference, 1975, and adopted by the 29th World Health Assembly. Geneva (Switzerland): WHO; 1977.
- (82) Little RJ, Rubin DB. Statistical analysis with missing data. New York (NY): John Wiley & Sons; 1987.
- (83) Haenszel W, Loveland DB, Sirken MG. Lung-cancer mortality as related to residence and smoking histories. I. white males. *J Natl Cancer Inst* 1962;28:947–1001.
- (84) Brinton LA, Daling JR, Liff JM, Schoenberg JB, Malone KE, Stanford JL, et al. Oral contraceptives and breast cancer risk among younger women. *J Natl Cancer Inst* 1995;87:827–35.
- (85) Community Intervention Trial for Smoking Cessation (COMMIT): summary of design and intervention. COMMIT Research Group. *J Natl Cancer Inst* 1991;83:1620–8.
- (86) Gail MH, Byar DP, Pechacek TF, Corle DK. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT) [published erratum in *Controlled Clin Trials* 1992;14:253–4]. *Controlled Clin Trials* 1992;13:6–21.
- (87) Community Intervention Trial for Smoking Cessation (COMMIT): II. Changes in adult cigarette smoking prevalence. *Am J Public Health* 1995;85:193–200.
- (88) Ghosh M, Rao JN. Small area estimation: an appraisal. *Statistical Science* 1994;9:55–93.
- (89) Shopland DR, Hartman AM, Gibson JT, Mueller MD, Kessler LG, Lynn WR. Cigarette smoking among U.S. adults by state and region: estimates from the current population survey. *J Natl Cancer Inst* 1996;88:1748–58.

NOTE

Manuscript received September 28, 1998; revised February 10, 1999; accepted April 6, 1999.