Vital and Health Statistics

Statistical Issues in Analyzing the NHANES I Epidemiologic Followup Study

Series 2:
Data Evaluation and Methods Research
No. 121

This report presents alternative strategies for analysis of data from the NHANES I Epidemiologic Followup Study (NHEFS) using Cox proportional hazards and person-time logistic regression models Analytic issues related to the complex survey design of the NHANES I and the variable length of followup of NHEFS participants are discussed

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control and Prevention
National Center for Health Statistics

Hyattsville, Maryland
May 1994
DHHS Publication No. (PHS) 94-1395

National Center for Health Statistics

Manning Feinleib, M D , Dr P H , *Director*

Jack R Anderson, *Deputy Director*

Jacob J Feldman, Ph D , *Associate Director for Analysis, Epidemiology, and Health Promotion*

Gail F Fisher, Ph D , *Associate Director for Planning and Extramural Programs*

Peter L Hurley, *Associate Director for Vital and Health Statistics Systems*

Robert A Israel, *Associate Director for International Statistics*

Stephen E Nieberding, *Associate Director for Management*

Charles J Rothwell, *Associate Director for Data Processing and Services*

Monroe G Sirken, Ph D , *Associate Director for Research and Methodology*

Division of Health and Utilization Analysis

Thomas A Hodgson, Ph D , *Acting Director*

John L Kiely, Ph D , *Chief, Infant and Child Health Studies Branch*

Diane M Makuc, Dr P H , *Chief, Analytical Studies Branch*

Cooperation of the U S Bureau of the Census

Under the legislation establishing the National Health Survey the Public Health Service is authorized to use insofar as possible the services or facilities of other Federal State or private agencies

In accordance with specifications established by the National Center for Health Statistics the U S Bureau of the Census under a contractual arrangement participated in planning the survey and collecting the data

Contents

Appendixes

Text tables

Symbols

- - -	Data not available
	Category not applicable
–	Quantity zero
0 0	Quantity more than zero but less than 0 05
Z	Quantity more than zero but less than 500 where numbers are rounded to thousands
*	Figure does not meet standard of reliability or precision
#	Figure suppressed to comply with confidentiality requirements

Statistical Issues in Analyzing the NHANES I Epidemiologic Followup Study

by Deborah D Ingram, Ph D , and Diane M Makuc, Dr P H , Division of Health and Utilization Analysis

Introduction

This report is concerned with statistical issues faced by analysts of the Epidemiologic Followup to the first National Health and Nutrition Examination Survey (NHEFS) The NHEFS is a longitudinal study that uses as its baseline those persons 25–74 years of age who were examined during the first National Health and Examination Survey (NHANES I) The NHEFS is composed of a series of followup surveys and was designed to examine the relationship of baseline clinical, nutritional, and behavioral factors assessed during 1971–75 to subsequent morbidity, mortality, functional impairment, and institutionalization (1)

Most analysts of the NHEFS are interested in assessing the relationship between a set of risk factors measured at baseline and some outcome event, usually death or disease incidence Analysis of data from the NHEFS is not straightforward because the analyst must consider differential lengths of followup as well as the complex survey design

This report uses simulated data and NHEFS data to compare three models for analyzing data from the NHEFS, namely, the Cox proportional hazards model, the person-time logistic regression model, and the cumulative logistic regression model The Cox model is commonly used to analyze data from epidemiologic followup studies because it takes into account differential followup time Statistical methods and software to incorporate the complex survey design in the Cox model have recently been developed (2,3) The cumulative logistic regression model (generally referred to simply as the logistic regression model) is also used to analyze data from followup studies However, the logistic model is not entirely appropriate for use with the NHEFS data because it does not take into account differential length of followup The person-time logistic model is a modification of the cumulative logistic model and can incorporate the differential followup times (4)

This report also examines the effect of incorporating different aspects of the complex survey design in the analysis of NHEFS data The effect of the survey design on regression coefficients and their standard errors from the Cox proportional hazards model is assessed by performing analyses under four different options

- Ignoring all aspects of the complex survey design
- Incorporating only the stratification and clustering
- Incorporating only the sample weights
- Incorporating both the stratification and clustering and the sample weights

Additional approaches considered are

- Trimming the sample weights to reduce their variability
- Stratifying the analysis on variables used in the sample design
- Including variables used in the design as covariables in the model

This report also addresses several other statistical issues that arise in the analysis of the NHEFS data

- Calculation of followup time for incidence and mortality studies
- Development of sample weights for analyses of the NHEFS that include all of the 100 NHANES I sampling locations
- Description of "pseudo-stratum" and "pseudo-primary sampling unit (PSU) codes" for variance estimation

This report provides a practical guide, including SAS and SUDAAN code, for using the Cox and person-time logistic regression models to analyze the NHEFS data under four analysis options

The authors gratefully acknowledge the help of Jennifer Madans, Christine Cox, and Lester R Curtin of the National Center for Health Statistics in their review of this report, and the help of Barry Graubard of the National Cancer Institute, Frank Potter of the Research Triangle Institute, and Van Parsons and Meena Khare, both of the National Center for Health Statistics, in consultations about weighted analyses

The NHANES I Epidemiologic Followup Study was initiated jointly by the National Institute on Aging and the National Center for Health Statistics and has been developed and funded by the National Center for Health Statistics, the National Institute on Aging, the National Cancer Institute, the National Center for Chronic Disease Prevention and Health Promotion, the National Heart, Lung, and Blood Institute, the National Institute of Mental Health, the National Institute on Alcohol Abuse and Alcoholism, the National Institute of Allergy and Infectious Diseases, the National Institute of Neurological and Communicative Disorders and Stroke, the National Institute of Arthritis and Musculoskeletal and Skin Diseases, the National Institute of Child Health and Human Development, and the National Institute of Diabetes and Digestive and Kidney Diseases

Description of the study

Baseline design

NHANES I, which took place during 1971-75, provided the baseline sample for the NHEFS NHANES I collected data on a multistage, national probability sample of the U S civilian noninstitutionalized population 1–74 years of age, excluding persons in Alaska, Hawaii, and reservation lands of American Indians (5–8) Details of the plan, complex sample design, response, and operation have been published (5–8) Aspects of the design of NHANES most pertinent to the analysis of the NHEFS are described in this section

NHANES I was conducted at 100 locations across the United States and consisted of 6 nationally representative samples that were not mutually exclusive (table A) During 1971–74, the survey included persons 1–74 years of age from locations 1–65 During 1974–75, the survey included persons 25–74 years of age from locations 66–100 Locations 1–35 (data collected during 1971–72) also composed a nationally representative sample to produce early national estimates for the nutrition portion of the survey

NHANES I included a home interview, medical examination, and laboratory procedures for all participants As a result of the varied design features of NHANES I, not all study subjects received the same questions or examinations For example, only persons in locations 1–65 received the nutrition questionnaires A random sample of approximately 20 percent of those 25–74 years of age in locations 1–65 received a more detailed medical examination The subsamples of persons from locations 1–35 and locations 1–65 receiving the detailed medical examination were each nationally representative samples (1–35 detail and 1–65 detail) All persons in locations 66–100 received the detailed medical examination (66–100 detail) The combined 1–65 detail sample and the 66–100 sample also form a nationally representative sample (1–100 detail)

The complex survey design of the NHANES I involved several stages of selection In hierarchical order, these stages were primary sampling units (PSU's), enumeration districts, segments (cluster of households), households within clusters, and persons within households Each PSU was either a standard metropolitan statistical area (SMSA), a single county, or a group of two or three contiguous counties The approximately 1,900 PSU's were collapsed into 40 superstrata For the 1971–74 period of the survey (locations 1–65), 15 of the superstrata were selected with certainty (10 in locations 1–35 and 5 in locations 36–65) Each of the certainty strata contained one PSU that consisted of a single large metropolitan area with a population of more than 2 million The 25 remaining superstrata, referred to as the noncertainty strata, contained multiple PSU's One PSU was selected from each of the 25 noncertainty strata for the first 35 locations, and a second PSU was selected from each of the noncertainty strata for locations 36–65 Thus, the first-stage sample of 65 PSU's included 15 large metropolitan certainty areas and 50 paired selections (2 x 25) from the noncertainty areas For the 1974–75 period of the survey (locations 66–100), only 5 of the 15 superstrata (consisting of a single large SMSA) were drawn into the sample with certainty The other 10 of these superstrata were collapsed into 5 groups of 2 PSU's each, from which only 1 PSU was selected One PSU was selected from each of the 25 noncertainty strata Thus, for the augmentation stage of the survey, 10 of the 35 PSU's were large metropolitan areas and 25 were noncertainty areas Clusters of sample persons were selected from the 100 PSU's

Followup design

The baseline sample for the NHEFS is a national probability sample consisting of the 14,407 participants in NHANES I who were 25–74 years of age at the time of the baseline examination Thus, the NHEFS sample is a composite of the 11,348 persons aged 25–74 years from locations 1–65 of NHANES I and the 3,059 persons from locations 66–100 of NHANES I The number of NHEFS sample persons from each of the NHANES I samples is shown in table A

The NHEFS consists of an ongoing series of followup surveys (1, 9–11) The first wave of followup was conducted during 1982–84 and included vital status ascertainment, a personal interview with each participant or a proxy, and collection of health care facility records and death certificates Of the 14,407 study persons in the 1982–84 NHEFS, 93 percent (13,383 persons) were successfully traced (table B) Interviews were completed for 93 percent (10,523 persons) of those traced alive, and proxy interviews were completed for 84 percent (1,697 decedents) of decedents who were traced (9)

The second wave of followup was conducted during 1986 for members of the NHEFS cohort who were 55–74 years of age at the time of their baseline examination All persons in this age group not known to be deceased at the 1982–84 NHEFS, including those who were not traced, were in the second wave At the end of the 1986 survey period, 95 percent

Table A Number of examined persons, by sample location, type of examination, years of data collection, and age of target population NHANES I and NHANES I Epidemiologic Followup Study

			Number of examined persons	
Sample location and examination	*Years of data collection*	*Age of target population*	*Total*	*Persons 25–74 years of age (NHEFS sample)*
NHANES I samples				
Locations 1–35 nutrition	1971–72	1–74	10 127	5 500
Locations 1–35 detail[1]	1971–72	25–74	1 892	1 892
Locations 1–65 nutrition	1971–74	1–74	20 749	11 348
Locations 1–65 detail[1]	1971–74	25–74	3,854	3,854
Locations 66–100 detail	1974–75	25–74	3,059	3 059
Locations 1–100 detail[2]	1971–75	25–74	6 913	6 913
NHEFS sample				
Locations 1–100[3]	1971–75	25–74	14 407	14,407

[1] Detail sample is a subsample of nutrition sample.

[2] Locations 1–100 detail sample is a combination of locations 1–65 detail sample and locations 66–100 detail sample.

[3] NHEFS sample is composed of persons 25–74 years of age from the locations 1–65 nutrition sample and the locations 66–100 detail sample.

NOTES: NHANES I is National Health and Nutrition Examination Survey I. NHEFS is NHANES I Epidemiologic Followup Study.

Table B Number and percent distribution of respondents by status at followup, according to followup wave: NHANES I Epidemiologic Followup Study

		Status at followup		
Followup wave	*All respondents*	*Alive*	*Deceased*	*Lost to followup*
		Number		
1982–84	14 407	11 361	2 022	1 024
1986	3 980	3 132	635	213
1987	11 750	10 463	555	732
		Percent distribution[1]		
1982–84	100.0	78.9	14.0	7.1
1986	100.0	78.7	16.0	5.4
1987	100.0	89.0	4.7	6.2

[1] May not add to 100 percent because of rounding.

NOTE: NHANES I is the National Health and Nutrition Examination Survey I.

(3,767 persons) of the 3,980 subjects in the 1986 Followup cohort had been successfully traced. Interviews were completed for 97 percent of those traced alive, and proxy interviews were completed for 91 percent of decedents (10). The 1986 NHEFS collected information on changes in the health and functional status of participants since their 1982–84 followup. The 1986 NHEFS consisted of vital status ascertainment, a telephone interview with each participant or a proxy, and collection of health care facility records and death certificates.

The third wave of followup took place in 1987. An attempt was made to recontact the entire surviving NHEFS cohort, including persons who had not been traced or interviewed in the first and second waves of followup. At the end of the 1987 survey period, 94 percent (11,018 persons) of the 11,750 subjects in the 1987 Followup cohort had been successfully traced. Interviews were completed for 91 percent (9,526 persons) of subjects traced alive and for 85 percent (472 decedents) of decedents (11). The 1987 NHEFS collected information on changes in the health and functional status of the entire surviving NHEFS cohort since the last contact. The design and data collection procedures of the 1987 NHEFS were very similar to those used in the two previous followups, in that subjects were traced, subject and proxy interviews were conducted, and health care facility records and death certificates were collected.

Sample weights

The final NHANES I sample weight for each individual is the product of the basic sample weight, a nonresponse adjustment factor, and a poststratification adjustment factor. The basic sample weight is the reciprocal probability of selection for an individual and reflects the oversampling of subgroups in locations 1–65 (8).

Oversampling—Elderly persons (65–74 years), women of childbearing age (25–44 years), and persons residing in poverty areas were oversampled in locations 1–65. No oversampling of subgroups occurred in locations 66–100. The oversampling in locations 1–65 is illustrated by the following sampling rates: 1:4 for men 20–44 years, 1:2 for women 20–44 years, 1:4 for persons 45–64 years, and 1:1 for persons 65–74 years. Persons receiving the detailed medical examination were randomly selected from the 1–65-locations sample using different sampling rates. Initially, poverty areas were oversampled at a rate of 8:1, later this ratio was changed to 2:1.

Nonresponse adjustment—The response rate in NHANES I was high for the home interview (about 99 percent for persons 25–74 years of age) but lower for the medical examination (about 70 percent for those 25–74 years of age). Nonresponse adjustment factors were computed within five annual income groups (less than $3,000, $3,000–$6,999, $7,000–$9,999, $10,000–$14,999, and $15,000 or more) within

Table C Sample weight percentiles by sample location and type of examination NHANES I Epidemiologic Followup Study

Sample locations and examination	*Number of respondents*	*Sample weight percentile*					
		0	*5*	*50*	*95*	*98*	*100*
Locations 1–65 nutrition	11 348	471	1 055	6 314	26 491	31 737	100 990
Locations 1–65 detail[1]	3 854	1 616	4 200	21 803	66,374	84 111	178 994
Locations 66–100 detail	3 059	10 411	16 561	33 326	60,185	74 421	166 038
Locations 1–100, detail[2]	6,913	1 004	3 351	12,021	39 427	49 472	121,040
Locations 1–100 all persons[3]	14,407	442	1,010	5,867	18,263	22 209	68 027

[1] Detail sample is a subsample of the nutrition sample

[2] Locations 1–100 detail sample is a combination of the locations 1–65 detail and 66–100 detail samples

[3] NHANES I Epidemiologic Followup Study sample is comprised of locations 1–65 nutrition sample and locations 66–100 detail sample

NOTES NHANES I is National Health and Nutrition Examination Survey I NHEFS is NHANES I Epidemiologic Followup Study

each location (8) The factor is the ratio of the sum of basic sample weights for all sample persons to the sum of basic sample weights for all responding sample persons within the same group For current NCHS surveys, nonresponse adjustment factors are truncated at 2 However, in NHANES I some of the nonresponse-adjustment factors were between 2 and 3

Poststratification adjustment—A poststratification adjustment procedure was employed to ensure agreement between final sample estimates of the population and independent age–race–sex-specific controls prepared by the U S Bureau of the Census

As a result of the oversampling at baseline and of the nonresponse adjustment, the NHANES I sample weights are highly variable and skewed to the right (table C) For example, the sample weights for the 3,854 persons in the NHEFS from the 1–65 locations detailed sample range from 1,616 to 178,994, so that the ratio of the largest weight to the smallest is nearly 111 1 For the 66–100-locations sample, which had no oversampling, the ratio of the largest weight to the smallest is only 16 1 The weights for all 14,407 persons in the NHEFS (from locations 1–100) range from 442 to 68,027, so that the largest weight is 154 times the smallest weight The 98th percentile weights are considerably smaller than the maximum sample weights For example, the 98th percentile weight for the total NHEFS sample is 22,209 compared with the maximum weight of 68,027

An individual respondent with a large sample weight may have a large and possibly undesirable influence on estimates, particularly if the individual has an unusual value for the variable of interest In addition, some relatively small groups of individuals have large weights because of the oversampling and can strongly influence estimates For example, in locations 1–65, only 17 percent (8 out of 47) of the black males aged 65–74 years with 9 or more years of education lived outside the oversampled poverty areas at baseline However, this 17 percent accounts for 53 percent of the weights for this group

A different set of sample weights was needed for each of the NHANES I samples so that each sample could be used to obtain national estimates Originally, sample weights were calculated only for the six NHANES I samples shown in table A No sample weights were calculated for the entire NHANES I sample (all persons in locations 1–100) Thus, another set of sample weights for use with all 14,407 participants in the NHEFS was developed as described in a later section

The NHANES I sample weights are used for analyses of the NHEFS data They have not been adjusted for the nonresponse and loss to followup in the different NHEFS followups

Models for analyzing study data

This section contains a description and comparison of three regression models that can be used to examine the relationship between a set of risk factors and some outcome event The three regression models presented are the Cox proportional hazards regression model, the cumulative logistic regression model, and the person-time logistic regression model Simulated data sets and data sets from the NHEFS are used to compare parameter estimates from the three models In the analyses presented in this section, it is assumed that the data are from a simple random sample

In the comparison of the three models that follows, it is assumed that longitudinal, rather than cross-sectional, analyses are of interest The NHEFS was designed for longitudinal analysis, not for cross-sectional analysis The primary problem with using the NHEFS for cross-sectional analyses is that, for any given wave of followup, the NHEFS sample is not a nationally representative sample because some subjects have been lost to followup and some were traced but not interviewed In addition, the NHEFS sample at the followup waves does not reflect changes in the structure of the population resulting from migration that has occurred since the baseline sample was drawn in 1971–75 For these reasons, the estimation of prevalence rates from NHEFS data is especially problematic

Lengths of followup for subjects in the NHEFS are highly variable because of the staggered entry times (1971–75) and the staggered followup interview times and because deaths and censoring have occurred throughout the study period Thus, the Cox model is the preferred model for analyzing data from the NHEFS because it takes into account differential followup time and does not require assumptions about the survival time distribution The Cox model has been used in the vast majority of published analyses of NHEFS data

The cumulative logistic regression model is not entirely appropriate for analyzing the data from NHEFS because it does not take into account differential followup time Nevertheless, some researchers choose to use the cumulative logistic regression model when analyzing data from the NHEFS either because they prefer this model or because calculation of length of followup is problematic for the outcome event being studied In this section we demonstrate that the cumulative regression model can produce seriously biased estimates as a result of its failure to take into account differential followup time Researchers who wish to use a logistic model may find the person-time logistic regression model useful because it takes into account differential followup time

The Cox proportional hazards model

The Cox proportional hazards model assesses the relationship between a set of risk factors and some outcome event, usually death or disease incidence (12) The model measures the relative risk of death or disease in (infinitesimally) small time intervals under the assumption that the relative risk is constant over the followup period The model utilizes both covariates (risk factors) measured on each individual and the time each outcome event occurs The parameter estimates depend on the rank ordering of the event times rather than on the exact time an outcome event occurs

The cumulative logistic regression model

The cumulative logistic regression model (generally referred to simply as the logistic regression model) also assesses the relationship between a set of risk factors and some specified outcome event However, the metric it uses to measure this association differs from that used by the Cox model The cumulative logistic regression model measures the relative odds of death or disease after a fixed duration of followup The model is analogous to a multiple regression model with a dichotomous dependent variable Unlike the Cox model, which takes into account differential followup time, the logistic regression model assumes that all individuals are followed for the same length of time Thus, for studies with considerable differences in followup times, such as the NHEFS, the logistic model may produce biased parameter estimates

When length of followup varies, two approaches to the cumulative logistic regression model can be taken The first approach includes all individuals in the analysis, regardless of their followup time This approach assumes that length of followup has little effect on the parameter estimates This assumption is not strictly valid in mortality and morbidity studies because the likelihood of observing an event increases with the length of time an individual is followed This approach is used in the NHEFS analyses that follow because it is the approach that is most commonly used An alternative approach is to "stop" the study after some specified period of time With this second approach, any survivor (or decedent from a cause other than the one of interest) whose followup time was less than the specified length of time is excluded from the analysis In addition, decedents whose death occurred after the specified length of time are included in the analysis as survivors

The person-time logistic regression model

The person-time logistic regression model is a modification of the cumulative logistic regression model that takes into account differential followup time (4) The former may be a reasonable alternative to the latter for those researchers who prefer to work with a logistic model The person-time logistic model may also prove useful if the exact time of death or disease occurrence is not known, but the survival data can be grouped in time intervals

The person-time logistic regression model involves expressing the dependent variable as the number of outcome events per person-time unit of followup rather than per person, as is the case for the cumulative logistic regression model For the person-time logistic model, the followup period is divided into equal-length intervals (for example weeks, months, or years), the number of persons at risk and outcome events in each interval are counted, and these counts are aggregated over all of the intervals Each individual contributes his or her status in each interval followed (for example, alive = 0, and dead = 1, noncase = 0, and case = 1) to the numerator and the total number of intervals followed to the denominator An individual who dies or is lost in an interval does not contribute information for subsequent intervals

To illustrate, suppose that the time interval chosen is 1 month and that an individual is followed for 10 years, or 120 months For the person-time logistic analysis, this individual contributes information for all 120 intervals If this individual is alive at the end of followup, his or her vital status for each of the 120 intervals is "alive " If a second individual is followed for 5 years, or 60 months, he or she contributes information for only 60 intervals If this individual dies in the last month of followup, the vital status for 59 of the intervals is "alive," and for the last interval the vital status is "dead "

A basic assumption of the person-time logistic model is that the probability of death for an individual in any time interval is independent of the number of time intervals already survived (13) In other words, the assumption is exponential survival time The survival time distributions of the NHEFS samples we have examined have been reasonably well approximated by the exponential distribution

Comparison of the three models

Simulation data sets

A previous simulation study examined the effects of the following quantities on the parameter estimates from the three models disease incidence, risk factor strength of association, length of followup, proportion censored, nonproportional hazards, and sample size (4) This section contains a brief review of the findings of this study, discussing the effect of disease incidence, length of followup, and nonexponentiality

For simplicity, the models for the simulations contained only one dichotomous variable that designated group membership (Group 1 and Group 2) The length of followup was set at 10 years for most of the simulations, approximately the length of time between the baseline examination and the first wave of followup in the NHEFS In the analyses presented here, the proportion dead at 10 years in Group 1 was fixed at 10 percent and 40 percent and the relative risk of death for Group 2 (compared with Group 1) was fixed at 2 0

To ensure proportionality of the hazards over the followup period, the survival times were generated according to the Weibull distribution The shape parameter for the Weibull distribution (c) was set to 1 2 for most simulations This value was estimated from the NHEFS data

Simulations were performed to evaluate the effect of censoring For these simulations the expected proportion censored by 10 years was fixed at 50 percent Censoring times were generated according to the Weibull distribution

One of the assumptions of the person-time logistic model is exponential survival time When the Weibull shape parameter, c, is equal to 1, the Weibull distribution reduces to the exponential distribution Thus, to assess the effect of nonexponentiality on the person-time logistic estimates, simulations were performed with $c = 1\ 0$, $c = 0\ 5$, and $c = 2\ 0$

Simulation results

The simulations showed that parameter estimates from the person-time logistic regression model closely resembled those from the Cox model when the survival time distribution was close to exponential ($c = 1\ 2$) The estimates from the person-time logistic regression model remained similar to the Cox

Table D Simulation results showing the effect of length of followup and proportion dead on estimates from three alternative regression models

	Proportion dead		*Cox model*		*Person-time logistic model*		*Cumulative logisitic model*	
Length of followup	*Group 1*	*Group 2*	*Beta*	*Standard error*	*Beta*	*Standard error*	*Beta*	*Standard error*
	Percent							
5 years	4	9	0 70	0 25	0 70	0 26	0 72	0 27
10 years	10	19	0 70	0 18	0 69	0 18	0 75	0 19
15 years	16	29	0 69	0 14	0 68	0 14	0 78	0 16
5 years	20	36	0 68	0 12	0 67	0 12	0 80	0 15
10 years	40	64	0 68	0 09	0 66	0 09	0 97	0 13
15 years	56	81	0 69	0 08	0 65	0 08	1 19	0 15

NOTES From simulations with two groups of size=500 hazard ratio=2 0 c=1 and no censoring Estimates are based on a mean over 100 replicates

Table E Simulation results showing the effect of censoring on estimates from three alternative regression models

Proportion dead		*Proportion censored*		*Cox model*		*Person time logistic model*		*Cumulative logistic model*	
Group 1	*Group 2*	*Group 1*	*Group 2*	*Beta*	*Standard error*	*Beta*	*Standard error*	*Beta*	*Standard error*
		Percent							
10	19	None	None	0 70	0 18	0 69	0 18	0 75	0 19
10	19	50	50	0 69	0 21	0 69	0 21	0 74	0 23
10	19	50	None	0 69	0 20	0 72	0 20	0 37	0 21
10	19	None	50	0 69	0 19	0 65	0 19	1 12	0 21
40	64	None	None	0 68	0 09	0 66	0 09	0 97	0 13
40	64	50	50	0 69	0 10	0 67	0 10	0 99	0 17
40	64	50	None	0 70	0 10	0 71	0 10	0 58	0 15
40	64	None	50	0 67	0 10	0 62	0 10	1 37	0 15

NOTES Results are from simulations with two groups of size=500 followup period=10 years hazard ratio=2 0 c=1 2 Censoring times generated by Weibull distribution with c=1 2 and 50 percent censored expected by 10 years Estimates are based on a mean over 100 replicates

estimates as the length of the followup period increased from 5 to 15 years and as the proportion dead increased (table D) The person-time logistic estimates also were similar to the Cox estimates when there was censoring (table E) The estimates from the two models closely resembled each other as long as censoring occurred at the same rate in Group 1 and Group 2 Even when censoring occurred at different rates in the two groups, the estimates from the two models were fairly similar as long as the survival time distribution was close to exponential

The person-time logistic regression coefficients differed substantially from the Cox regression coefficients when the survival time distribution was not close to exponential and censoring occurred at different rates in the two groups (table F) The person-time logistic regression coefficients could be larger or smaller than those from the Cox model depending on the value of *c* and on which group had the censoring However, although the effect of nonexponentiality was substantial when there was unequal censoring, it had only a moderate effect as long as censoring occurred at the same rate in the groups

Parameter estimates from the cumulative logistic model were similar to those from the Cox and person-time logistic regression models when the proportion dead was small (table D) However, as the proportion dead increased (the relative risk of death remained constant at 2 0), the cumulative logistic estimates also increased, thus becoming increasingly disparate from the Cox estimates, which did not change For example, when the proportion dead in the two groups increased from 10 percent and 19 percent to 40 percent and 64 percent, the cumulative logistic regression coefficient increased from 0 75 to 0 97 (table D) The corresponding regression coefficients for the Cox model were 0 70 and 0 68 The parameter estimates from the cumulative logistic regression model also increased as the length of followup increased (table D) As can be seen, this increase occurred because the proportion dead increases with time

The simulations also showed that the cumulative logistic regression model, unlike the Cox model, was quite sensitive to unequal censoring rates across the groups (table E) When more censoring occurred in the group with the smaller proportion dead (Group 1), the cumulative logistic regression coeffi-

Table F Simulation results showing the effect of nonexponentiality on estimates from the person-time logistic regression model

						Person-time logistic model exponentiality parameter[1]					
Proportion dead		*Proportion censored*		*Cox model*		*c=0 5*		*c=1 0*		*c=2 0*	
Group 1	*Group 2*	*Group 1*	*Group 2*	*Beta*	*Standrd error*	*Beta*	*Standard error*	*Beta*	*Standard error*	*Beta*	*Standard error*
		Percent									
10	19	None	None	0 70	0 18	0 71	0 18	0 70	0 18	0 68	0 18
10	19	50	50	0 69	0 21	0 71	0 21	0 69	0 21	0 67	0 21
10	19	50	None	0 69	0 20	0 59	0 20	0 69	0 20	0 79	0 20
10	19	None	50	0 69	0 19	0 83	0 19	0 69	0 19	0 56	0 19
40	64	None	None	0 68	0 09	0 78	0 09	0 69	0 09	0 60	0 09
40	64	50	50	0 69	0 10	0 79	0 10	0 69	0 10	0 61	0 10
40	64	50	None	0 70	0 10	0 66	0 10	0 70	0 10	0 73	0 10
40	64	None	50	0 67	0 10	0 90	0 10	0 68	0 10	0 49	0 10

[1] The survival time distribution is exponential when *c*=1 0

NOTES Results are from simulations with two groups of size=500 followup period=10 years and hazard ratio=2 0 Censoring times generated by Weibull distribution and 50-percent censored expected by 10 years Estimates are based on a mean over 100 replicates cient was substantially smaller than the Cox regression coefficient When more censoring occurred in the group with the larger proportion dead (Group 2), the cumulative logistic regression coefficient was substantially larger than the Cox regression coefficient The effect of unequal censoring was more pronounced as the proportion dead increased

Analyses of NHEFS data are more complex than the simulation models and typically include numerous risk factors with differing strengths of association and involve unequal censoring across groups To gain more understanding about the performance of the three models, we compared the models using two data examples from the NHEFS

NHEFS data sets

The first data example involved the effects of age, race (white, black), sex, systolic blood pressure (SBP), and smoking (all variables measured at baseline) on subsequent mortality among persons 50–74 years of age (all 100 sampling locations) In the models, age and SBP were treated as continuous variables, and cigarette smoking was categorical (current, former, or never) Because vital statistics data suggest that excess mortality among black persons diminishes with increasing age, an age-by-race interaction term was included in the model Vital status information and followup time came from the 1987 followup There were 6,400 subjects in the analysis, of whom 2,675 (42 percent) had died The proportion dead ranged from 7 percent among white females 50–54 years of age to 77 percent among black males 70–74 years of age Length of followup among survivors ranged from 6 to 16 years with a mean of 14 years The time interval used for the person-time logistic model was 1 month For this example, the survival time distribution was nearly exponential

The results for the first NHEFS data example are shown in table G The Cox and person-time logistic models yielded similar results in terms of both the regression coefficients and their standard errors The regression coefficients from the cumulative logistic regression model were larger than those from the Cox or person-time logistic regression models with correspondingly larger standard errors Five of the seven regression coefficients for the cumulative logistic model were outside 95-percent confidence limits for the Cox model coefficients However, given the strength of the relationship between the risk factors and death, χ^2 tests for the cumulative logistic model coefficients were similar to those for the Cox model coefficients

The second data example involved race-specific analysis of all-cause mortality as a function of serum albumin levels (in tertiles less than 4 2, 4 2–4 4, and greater than 4 4 gm/dl), adjusting for age, educational attainment (less than 12 years, 12 years or more), systolic blood pressure, cigarette smoking (current, former and never), history of diabetes, and total serum cholesterol (less than 200, 200–239, or greater than or equal to 240 mg/dl) All variables were measured at baseline Vital status and followup time came from the 1987 wave of followup The analysis included 2,291 white males and 437 black males 45–74 years of age from locations 1–65 of NHANES I, of whom 1,120 white males and 260 black males died Length of followup among survivors for all cause mortality ranged from 10 to 16 years with a mean of 15 years The time interval used for the person-time logistic model was 1 month In this data example, the survival time distribution was nearly exponential This analysis does not incorporate the complex survey design

The results from the analysis of the association between serum albumin and all cause mortality are shown in table H The Cox and person-time logistic regression models yielded similar regression coefficients The absolute value of the person-time logistic regression coefficients tended to be somewhat smaller than the Cox estimates The standard errors from the two models were essentially identical The regression coefficients from the cumulative logistic regression model were not as similar to the Cox coefficients as the person-time logistic coefficients were For example, the coefficient for diabetes for white males is 0 806 for the Cox model, 0 699 for the person-time logistic regression model, and 1 531 for the cumulative logistic regression model Five of the nine coefficients for the cumulative logistic regression model were

Table G Results from three alternative regression models relating death and selected baseline risk factors among persons 50–74 years of age

	Cox model			*Person time logistic model*			*Cumulative logistic model*		
Risk factor	*Beta*	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2
Sex (male)	0 569	0 043	**176 0	0 526	0 043	**149 9	0 808	0 063	**166 5
Race (black)	0 325	0 073	**19 9	0 337	0 073	**21 5	0 504	0 094	**28 5
Age (years)[1]	0 095	0 004	**644 1	0 090	0 004	**587 3	0 131	0 005	**656 5
Age by race	–0 024	0 008	**8 6	–0 026	0 008	**9 5	–0 029	0 012	*6 1
SBP (mmHG)[2]	0 007	0 001	**80 0	0 006	0 001	**70 2	0 011	0 001	**89 5
Current smoker	0 504	0 047	**114 5	0 463	0 047	**96 2	0 760	0 072	**110 8
Former smoker	0 092	0 055	2 8	0 078	0 055	2 0	0 133	0 080	2 8

* 01<p ≤ 05

**p ≤ 01

[1] Age is centered at 60 years

[2] SBP is systolic blood pressure

NOTES Results are based on analysis of persons 50–74 years of age at baseline from locations 1–100 Vital status data are from the 1987 followup wave of the NHANES I Epidemiologic Followup Study NHANES I is National Health and Nutrition Examination Survey I

Table H Results from three alternative regression models relating death, serum albumin, and selected baseline risk factors for males 45–74 years of age

	Cox model			*Person-time logistic model*			*Cumulative logistic model*		
Race and risk factor	*Beta*	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2
White									
Albumin (4 2–4 4)	–0 214	0 073	**8 7	–0 186	0 073	*6 6	–0 354	0 123	**8 2
Albumin (>4 4)	–0 321	0 080	**16 1	–0 288	0 080	**12 9	–0 515	0 131	**15 6
Age (years)	0 083	0 005	**318 7	0 076	0 005	**276 0	0 116	0 007	**300 7
Education (< 12 years)	0 213	0 067	**10 3	0 193	0 067	**8 4	0 320	0 101	**10 0
Diabetes history (yes)	0 806	0 103	**61 0	0 699	0 103	**45 8	1 531	0 248	**38 2
SBP (mmHg)[1]	0 005	0 001	**17 0	0 005	0 001	**14 9	0 010	0 002	**21 8
Smoking (yes)	0 445	0 063	**50 6	0 396	0 062	**40 1	0 582	0 104	**31 3
Cholesterol (200–239)	–0 159	0 076	*4 4	–0 147	0 076	3 8	–0 271	0 123	*4 9
Cholesterol (≥240)	0 064	0 076	0 7	0 053	0 076	0 5	0 062	0 126	0 2
Black									
Albumin (4 2–4 4)	–0 181	0 145	1 6	–0 150	0 145	1 1	–0 173	0 265	0 4
Albumin (>4 4)	–0 400	0 174	*5 3	–0 335	0 173	3 7	–0 528	0 283	3 5
Age (years)	0 057	0 009	**37 7	0 050	0 009	**30 5	0 082	0 014	**35 5
Education (<12 years)	0 341	0 225	2 3	0 344	0 225	2 3	0 577	0 311	3 4
Diabetes history (yes)	0 624	0 218	**8 2	0 489	0 217	*5 1	1 209	0 535	*5 1
SBP (mmHg)[1]	0 006	0 002	*6 1	0 005	0 002	*5 0	0 012	0 004	**8 7
Smoking (yes)	0 278	0 129	*4 6	0 240	0 129	3 5	0 660	0 231	**8 1
Cholesterol (200–239)	–0 120	0 155	0 6	–0 109	0 155	0 5	–0 192	0 271	0 5
Cholesterol (≥240)	–0 100	0 159	0 4	–0 090	0 159	0 3	–0 053	0 278	0 0

* 01<*p*≤ 05

***p*≤ 01

[1] SBP is systolic blood pressure

NOTES Results are based on analysis of white and black males 45–74 years of age at baseline from locations 1–65 Vital status data are from the 1987 followup wave of the NHANES I Epidemiologic Followup Study NHANES I is the National Health and Nutrition Examination Survey I

outside 95-percent confidence limits for the Cox model coefficients for both white males and black males In contrast to the person-time logistic regression coefficients, the absolute value of the cumulative logistic regression coefficients tended to be larger than the Cox coefficients The standard errors from the cumulative logistic regression model were consistently larger than those from the Cox model

For white males, conclusions derived from χ^2 test statistics were generally similar for the three models For black males, those with the highest albumin levels had a significantly reduced risk of death according to the Cox model, but this relationship did not reach statistical significance according to the other two models For this variable, the person-time logistic model yielded a weaker regression coefficient than the Cox model, whereas the cumulative logistic regression model yielded a stronger regression coefficient but also had a larger standard error

Summary

Three regression models that can be used to analyze data from the NHEFS have been presented here Lengths of followup for subjects in the NHEFS are highly variable For morbidity and mortality analyses, it is important to take into account the length of time each subject was actually followed The Cox and person-time logistic regression models take into account differential followup time, but the cumulative logistic regression model does not, assuming instead that individuals are followed for the same length of time

The simulations demonstrated that parameter estimates from the Cox and person-time logistic regression models are nearly identical to each other as long as deaths are exponentially distributed The simulations also showed that parameter estimates from the cumulative logistic regression model can differ substantially from those from the Cox or person-time logistic regression models unless the disease is rare, the risk factor association is moderate, and censoring occurs at the same rate across subgroups

When censoring occurs at the same rate across subgroups, the parameter estimates from the three models are not affected by the censoring However, when censoring occurs at different rates, the estimates from the Cox model are unaffected, the estimates from the person-time logistic regression model are slightly affected, and the estimates from the cumulative logistic regression model are seriously biased

The two data examples further illustrated the similarities and dissimilarities among estimates from the three models In the first example, the regression coefficients from the cumulative logistic regression model were larger than those from the Cox and person-time logistic regression models (outside 95-percent confidence limits for the Cox model coefficients), and the standard errors also were larger Thus, the cumulative logistic regression model overestimated the strength of the association between the risk factors and death and produced wider confidence intervals for these estimates However, the associations in this analysis were so strong that the conclusions regarding statistical significance from the three models were similar

In the all cause mortality and serum albumin example, the person-time logistic model tended to underestimate slightly the strength of the association between the risk factors and death, and the cumulative logistic regression model again overestimated the strength of the associations The standard errors from the person-time logistic regression model were almost identical to those from the Cox model, those from the cumulative logistic regression model were larger Both the person-time logistic and cumulative logistic regression models failed to detect one of the main effects found by the Cox model

Although the disparities observed in the estimates from the Cox and cumulative logistic regression models are partly the result of the different metrics being estimated (the Cox model estimates relative risk, and the cumulative logistic regression model estimates relative odds), these disparities also are partly the result of bias in the cumulative logistic regression estimates resulting from the differential followup times The disparities observed in the parameter estimates from the Cox and person-time logistic regression models may also reflect the different metrics being estimated (the person-time logistic regression model estimates relative odds in such a way as to approximate relative risk) However, it seems more likely that the disparities reflect the effect of slight nonexponential survival on the estimates The simulation results showed that, when *c* is greater than 1 0, the regression coefficients for the person-time logistic regression model are smaller than those for the Cox model, and, for these data examples, *c* equals 1 3

In conclusion

- The Cox model is the preferred model for analyzing data from the NHEFS because it takes into account differential followup time and does not require that survival time be exponentially distributed
- The person-time logistic regression model produces parameter estimates similar to those obtained from the Cox model as long as the survival time distribution is reasonably approximated by the exponential distribution The person-time logistic regression model takes into account differential followup time
- The cumulative logistic regression model is not entirely appropriate for use with NHEFS data because it does not take into account differential followup When length of followup varies, the Cox model utilizes more information than the cumulative logistic regression model and should provide a more accurate and powerful assessment of the relationship between risk factors and the event of interest
- The cumulative logistic regression model produces parameter estimates similar to those obtained from the Cox and person-time logistic regression models when the disease is rare, the risk factor strength is moderate, length of followup is short, and censoring occurs at the same rate across subgroups
- The cumulative logistic regression model tends to overestimate the strength of the association between the risk factors and the outcome event and to produce wider confidence intervals
- The cumulative logistic regression model can produce seriously biased estimates, especially when censoring rates differ across subgroups

Incorporating the complex survey design of the study

The scientific literature contains many publications based on data from the NHEFS Most of these publications make no mention of the complex survey design of the NHEFS Others mention the oversampling of subgroups when describing the NHEFS, but not when describing the statistical methods used for the analysis Because the NHEFS is a complex survey, some discussion of the design in the context of analysis is needed In this section, several issues regarding the complex survey design that need to be considered when analyzing data from the NHEFS are summarized In addition, the effect of incorporating the survey design in Cox models is examined, using data examples from the NHEFS

Classical sampling theory advocates incorporating the survey design in the analysis of data from a complex survey (14–20) The observations from a complex survey are not independent and identically distributed (IID) because the survey typically involves stratification, clustering, and unequal probabilities of selection Given that a basic underlying assumption of standard statistical methodology is that the observations are IID, failure to incorporate the survey design in an analysis may result in biased parameter estimates and underestimation of the standard errors and thus may produce misleading results Therefore, the argument that one need not consider the complex survey design when studying associations between variables is not correct Korn and Graubard present an extreme example that clearly illustrates how misleading results can be obtained if the survey design is not taken into account in a regression analysis (21) The literature on complex surveys has placed greater emphasis on estimation of descriptive parameters such as means and totals than on estimation of parameters for more analytic uses of surveys, such as regression coefficients However, more research is becoming available on modeling data from a complex survey (2,15–17,19,20) Binder has derived a design-based procedure to estimate regression coefficients and standard errors for the Cox model (2) Computer software for logistic regression and Cox regression models is accessible to many analysts, but software that incorporates a complex survey design into such analyses has been less accessible Lack of easily accessible information and software to incorporate a complex survey design in regression analyses has led many analysts to ignore the survey design

Two broad approaches to the analysis of complex survey data have been identified the aggregated approach and the disaggregated approach (19) The aggregated approach involves defining a model without regard to the sampling design and then using procedures that take into account the design to make inferences from the model The disaggregated approach involves defining a model that includes variables used in the survey design (such as strata or clusters) in addition to the variables of analytic interest The disaggregated approach may allow for different regression models for subgroups defined by strata or clusters, for example In general, the aggregated approach to complex survey analysis is taken in this report However, survey variables that were used to define oversampled subgroups may be included in models or used to stratify analyses, this is similar in concept to the disaggregated approach

There are two aspects of the survey design that must be considered when analyzing data from the NHEFS

- Stratification and clustering
- Sample weights

Stratification generally reduces the variance of the estimates, and clustering generally increases the variance A sample weight indicates the number of individuals in the target population that the sample person represents Sample weights are functions of the probability of selection and nonresponse and poststratification adjustments The use of sample weights in an analysis generally reduces bias but results in larger estimated variances for the parameters When the sample weights are extremely variable, as in the NHEFS, the use of the sample weights may result in overestimation of variances

Previous studies

Three previous studies have examined the effect of incorporating the survey design when analyzing data from NHANES I and the NHEFS (8,21,22) One study recommends incorporating the entire survey design in analyses of NHANES I, but the other two studies point out the disadvantage of using the highly variable NHANES I sample weights in analyses of the NHEFS The three studies are summarized briefly in this section

Landis et al calculated means and regression coefficients for three data examples under three options (8)

1 Simple random sampling, incorporating neither the stratification and clustering nor the sample weights
2 Incorporating only the sample weights
3 Incorporating both the stratification and clustering and the sample weights

This study did not consider the effect of stratification and clustering without the sample weights Standard statistical software was used for the analyses under options 1 and 2 The approach used to estimate the weighted standard errors under option 2 did not show the full influence of the sample weights on the variance-covariance structure of the parameters and resulted in standard errors that were too small The analyses under option 3 were performed using specialized software that incorporated the sample weights and the complex design in the calculation of both the parameter estimates and their variances

Landis et al found that the parameter estimates and standard errors obtained under options 1 and 2 were similar, but that the standard errors obtained under option 3 were considerably larger, suggesting that the increases in the standard errors obtained under option 3 were the result of the stratification and clustering Landis et al recommended performing initial analyses of NHANES I data ignoring the survey design (because it is simpler and cheaper) and performing final analyses using the survey design The conclusions and recommendations of Landis et al may need to be modified in view of later work that calculates the standard errors under option 2, taking into account the variability of the sample weights, and examines the effect of the stratification and clustering independent of the sample weights

In the second study, Makuc and Kleinman compared proportions, Kaplan-Meier estimates, and Cox proportional hazards estimates from analyses of the relationship between educational attainment and mortality among persons 65–74 years of age, using NHEFS data under four analysis options (22)

1 Simple random sampling
2 Incorporating the stratification and clustering but ignoring the sample weights
3 Incorporating sample weights only
4 Incorporating both the stratification and clustering and the sample weights

Standard statistical software was used to obtain all estimates under options 1 and 3 The standard errors obtained under option 3 did not adequately reflect the influence of the sample weights on the variances as in the study by Landis et al A jackknife procedure was used to obtain estimates of the standard errors under options 2 and 4

Makuc and Kleinman found that the stratification and clustering had relatively little effect on the estimates of standard errors whereas the sample weights had a larger effect The highly skewed sample weights caused a relatively small number of observations to strongly influence the findings and to inflate the estimates of standard errors, suggesting that it might be appropriate to ignore the survey design This study controlled for the oversampling of the elderly by limiting the analysis to persons 65–74 years of age This study also concluded that there was a need for further work using additional variables and population subgroups

In the third study, Korn and Graubard presented a general discussion of the use of the survey design in epidemiologic analyses and used as an illustration a re-analysis of data from the NHEFS (21) The data example compared mean total iron-binding capacity for respondents who developed cancer and for those who did not, adjusted by linear regression for selected risk factors under five options

1 Simple random sampling
2 Sample weights only
3 Stratification and clustering only
4 Stratification and clustering and sample weights
5 Stratification and clustering and unweighted analysis, adjusted for many of the variables used to define the sample weights

This analysis ignored the differential survival times of the respondents Standard statistical software was used to obtain the parameter estimates and their standard errors under option 1 Specialized software (SURREGR) that uses Taylor series linearization variance estimation was used to calculate all estimates under the other four options (23) For option 2 (sample weights only), all individuals were assigned to the same stratum, and each individual was assigned to a unique PSU Variance estimates obtained using this approach ignore the impact of the survey design's stratification and clustering, but do reflect the variability of the sample weights

Korn and Graubard found that the stratification and clustering had a relatively small effect on the estimates, whereas the sample weights had a larger effect The results under option 5 were very similar to those obtained under option 1 (simple random sampling) Korn and Graubard concluded that it was preferable not to use the sample weights because of their extreme variability and instead incorporated variables used to define the oversampled subgroups as covariates in the model

Empirical results

The effect of incorporating the survey design on parameter estimates from Cox models was assessed by performing analyses under the following four options using the two NHEFS data examples from the previous section

1 Simple random sampling (SRS)
2 Stratification and clustering only
3 Sample weights only
4 Stratification and clustering and sample weights

A fifth approach, in which variables used to define the sample weights were included in the model, also was assessed using the second data example The effect of stratifying the analyses and the effect of trimming extreme sample weights also were examined for the second data example

Note that the stratification and clustering affect only the variance estimates, whereas the sample weights affect both the regression coefficients and the variance estimates Thus, the regression coefficients for options 1 and 2 (the unweighted analyses) are identical, and the regression coefficients for options 3 and 4 (the weighted analyses) are identical

Standard statistical software was used to estimate the regression coefficients and their standard errors under option 1 (PROC PHGLM in Version 5 of SAS) (24) The SUDAAN procedure SURVIVAL was used to obtain estimates under options 2, 3, and 4 (2,3,25) SUDAAN uses a first-order Taylor series linearization approach to variance estimation (25–27)

To incorporate the stratification and clustering while ignoring the sample weights (option 2), stratum and PSU variables were used with a dummy sample weight of 1 for each individual To incorporate only the sample weights (option 3), sample weights were used with dummy stratum and PSU codes (All individuals were assigned to the same stratum and each individual was assigned to a unique PSU) This approach ignores the stratification and clustering effect while accounting for the variability of the sample weights in the variance estimation To incorporate both the stratification and clustering and the sample weights (option 4), stratum and PSU variables and sample weights were used in the analysis

For the fifth approach, the stratification and clustering were incorporated but not the sample weights (as in option 2) A variable indicating residence in a poverty area was added to the model to account for the oversampling in poverty areas Other variables used to define the sample weights were already in the model (age and sex)

Additional analyses, stratified by age as well as by race, were performed Analyses often are stratified by age, race, and sex because risk factor associations differ across age–race–sex groups In analyses of the NHEFS data, stratification may also be an effective technique to account for the oversampling of the subgroups Stratification is particularly important when risk factor variables differ by variables used in oversampling, such as age

Tests of the regression coefficients were obtained using χ^2 tests for option 1 and Satterthwaite adjusted χ^2 tests for options 2, 3, 4, and 5 (25,28,29) A detailed description of how to execute Cox and person-time logistic models under the four options, as well as SAS and SUDAAN code, is provided in appendix I

As was discussed in the section "Description of the study," the NHANES I sample weights are highly variable and skewed to the right because of the oversampling of subgroups in locations 1–65 and the untruncated nonresponse adjustments Use of sample weights in an analysis results in larger estimated variances of the parameters When the weights are highly variable and skewed to the right, the variance estimates that result from a weighted analysis may be inflated Weight trimming is a procedure that reduces the size and number of extreme sample weights (30–33) Weight trimming may reduce the variance estimates but may also introduce bias into the regression coefficients Two weight-trimming procedures, the inspection procedure and the estimated mean square error (MSE) procedure, were applied to the second NHEFS data example to assess the possible benefit of trimming the sample weights A description of these procedures is provided in appendix II Weights were trimmed within 24 groups based on age (25–44, 45–64, and 65–74 years), race (black, other than black), sex, and poverty residence (yes, no), because the sample weight distributions were different across these groups as a result of oversampling

NHEFS data example 1

The results for the first data example, involving the relationship between selected risk factors and subsequent mortality, are shown in table J The unweighted regression coefficients were within the 95-percent confidence intervals for the weighted regression coefficients The standard errors obtained under option 2 (stratification and clustering only) were slightly smaller than those obtained under option 1 (SRS), with the exception of the smoking variables The standard errors obtained from the weighted analyses (options 3 and 4) also were quite similar to each other Paralleling the unweighted analyses, the standard errors obtained under option 4 were slightly smaller than those obtained under option 3, with the exception of the smoking variables Thus, the effect of the stratification and clustering in this data example was minimal

Comparison of the weighted and unweighted standard errors shows that the weighted standard errors were consider-

Table J Results from Cox regression models relating death and selected baseline risk factors for persons 50–74 years of age, by analysis option

	Unweighted analyses					*Weighted analyses*				
	Option 1			*Option 2*		*Option 3*			*Option 4*	
Risk factor	*Beta*	*Standard error*	χ^2	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2	*Standard error*	χ^2
Sex (male)	0 569	0 043	**176 0	0 041	**193 2	0 560	0 062	**82 7	0 061	**85 6
Race (black)	0 325	0 073	**19 9	0 061	**28 3	0 268	0 109	*6 0	0 099	**7 4
Age (years)[1]	0 095	0 004	**644 1	0 004	**672 3	0 097	0 005	**411 6	0 004	**520 3
Age by race	–0 024	0 008	**8 6	0 008	**8 8	–0 022	0 013	2 9	0 012	3 3
SBP (mmHg)[2]	0 007	0 001	**80 0	0 001	**75 9	0 007	0 001	**36 5	0 001	**35 3
Current smoker	0 504	0 047	**114 5	0 050	**100 2	0 607	0 068	**79 9	0 074	**67 6
Former smoker	0 092	0 055	2 8	0 063	2 1	0 116	0 077	2 2	0 080	2 1

* 01<p≤ 05

**p≤ 01

[1] Age is centered at 60 years

[2] SBP is systolic blood pressure

NOTES Results are based on analysis of persons 50–74 years of age at baseline from locations 1–100 Vital status data are from the 1987 followup wave of the NHANES I Epidemiologic Followup Study NHANES I is the National Health and Nutrition Examination Survey I ably larger (generally by about 50 percent) than the unweighted standard errors As previously discussed, because of the variability of the sample weights, the weighted standard errors may be excessively large Note that, given the minimal effect of the stratification and clustering, the effect of the survey design was almost entirely due to the sample weights

χ^2 test statistics of the regression coefficients tended to be smaller for the weighted analyses than for the unweighted analyses The conclusions derived from the χ^2 tests generally were similar for the unweighted and weighted analyses, although the age-by-race interaction achieved statistical significance only in the unweighted analyses The age-by-race interaction failed to reach statistical significance in the weighted analysis because the weighted standard error was about 50 percent larger than the unweighted standard error

NHEFS data example 2

The results for white males from the data example involving the relationship between serum albumin and death are given in table K In the age-stratified analyses for white males 45–64 and 65–74 years of age, all of the unweighted regression coefficients were within 95-percent confidence intervals for the weighted coefficients (table K) The unweighted standard errors (options 1 and 2) were similar to each other, and the weighted standard errors (options 3 and 4) were similar to each other, indicating that the stratification and clustering had little impact on the standard errors The weighted standard errors from the age-stratified analyses were about 10–20 percent larger than the unweighted standard errors The regression coefficients show that the associations between

Table K Results from Cox regression models relating death, serum albumin, and selected baseline risk factors for white males 45–74 years of age, by analysis option

	Unweighted analyses					*Weighted analyses*							
	Option 1			*Option 2*		*Option 3*			*Option 4*		*Option 5*		
Age group and risk factor	*Beta*	*Standard error*	χ^2	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2
45–64 years													
Albumin (4 2–4 4)	–0 436	0 152	**8 3	0 167	**6 8	–0 462	0 183	*6 3	0 184	*6 3	–0 425	0 168	*6 4
Albumin (>4 4)	–0 480	0 157	**9 4	0 160	**9 0	–0 475	0 185	*6 6	0 152	**9 8	–0 467	0 162	**8 3
Age (years)	0 078	0 012	**45 2	0 009	**79 9	0 075	0 014	**28 8	0 012	**38 2	0 078	0 009	**79 0
Education (< 12 years)	0 465	0 126	**13 6	0 120	**14 9	0 635	0 156	**16 6	0 150	**17 9	0 432	0 117	**13 5
Diabetes history (yes)	1 251	0 205	**37 1	0 157	**63 7	1 070	0 260	**16 9	0 223	**23 2	1 251	0 155	**65 4
SBP (mmHg)[1]	0 011	0 003	**16 3	0 003	**17 3	0 014	0 004	**14 4	0 003	**15 6	0 011	0 003	**17 1
Smoking (yes)	0 427	0 121	**12 6	0 114	**14 1	0 460	0 145	**10 0	0 136	**11 5	0 422	0 114	**13 8
Cholesterol (200–239)	–0 193	0 157	1 5	0 152	1 6	–0 167	0 188	0 8	0 173	0 9	–0 195	0 150	1 7
Cholesterol (≥240)	0 080	0 154	0 3	0 147	0 3	0 006	0 183	0 0	0 177	0 0	0 082	0 144	0 3
Poverty segment											0 159	0 128	1 5
65–74 years													
Albumin (4 2–4 4)	–0 151	0 083	3 3	0 089	2 9	–0 252	0 098	*6 6	0 123	*4 2	–0 159	0 088	3 2
Albumin (>4 4)	–0 267	0 093	**8 2	0 087	**9 6	–0 322	0 110	**8 6	0 115	**7 8	–0 268	0 087	**9 4
Age (years)	0 078	0 012	**40 4	0 014	**32 7	0 084	0 015	**31 0	0 017	**23 4	0 077	0 014	**31 6
Education (< 12 years)	0 117	0 078	2 2	0 057	*4 2	0 144	0 090	2 5	0 074	3 8	0 089	0 058	2 4
Diabetes history (yes)	0 680	0 120	**32 1	0 132	**26 7	0 712	0 135	**28 0	0 140	**25 8	0 704	0 127	**30 6
SBP (mmHg)[1]	0 004	0 001	**7 5	0 001	**9 2	0 004	0 002	3 7	0 002	*5 4	0 004	0 001	**8 8
Smoking (yes)	0 444	0 073	**36 6	0 076	**34 6	0 372	0 094	**15 8	0 083	**20 2	0 439	0 076	**33 7
Cholesterol (200–239)	–0 142	0 087	2 7	0 108	1 7	–0 077	0 105	0 5	0 113	0 5	–0 146	0 107	1 9
Cholesterol (≥240)	0 057	0 088	0 4	0 073	0 6	0 019	0 109	0 0	0 099	0 0	0 056	0 071	0 6
Poverty segment											0 139	0 073	3 6
45–74 years of age													
Albumin (4 2–4 4)	–0 214	0 073	**8 7	0 079	**7 3	–0 357	0 111	**10 3	0 124	**8 3	–0 219	0 079	**7 8
Albumin (>4 4)	–0 321	0 080	**16 1	0 079	**16 7	–0 401	0 118	**11 5	0 111	**12 9	–0 319	0 080	**15 8
Age (years)	0 083	0 005	**318 7	0 005	**339 5	0 078	0 006	**149 1	0 006	**187 5	0 083	0 004	**351 7
Education (<12 years)	0 213	0 067	**10 3	0 051	**17 2	0 432	0 104	**17 3	0 099	**18 8	0 183	0 051	**12 9
Diabetes history (yes)	0 806	0 103	**61 0	0 104	**59 7	0 876	0 163	**28 7	0 140	**39 0	0 828	0 102	**66 6
SBP (mmHg)[1]	0 005	0 001	**17 0	0 001	**21 6	0 008	0 002	**15 0	0 002	**20 2	0 005	0 001	**20 9
Smoking (yes)	0 445	0 063	**50 6	0 062	**51 0	0 426	0 098	**19 0	0 090	**22 2	0 439	0 062	**50 5
Cholesterol (200–239)	–0 159	0 076	4 4	0 083	3 7	–0 151	0 117	1 7	0 102	2 2	–0 163	0 083	*3 9
Cholesterol (≥240)	0 064	0 076	0 7	0 070	0 8	–0 010	0 115	0 0	0 108	0 0	0 065	0 069	0 9
Poverty segment											0 154	0 064	*5 8

* 01<p≤ 05

**p≤ 01

[1] SBP is systolic blood pressure

NOTES Results are based on analysis of white males aged 45–74 years at baseline from locations 1–65 Vital status data are from the 1987 followup wave of the NHANES I Epidemiologic Followup Study NHANES I is National Health and Nutrition Examination Survey I albumin and death and between education and death were stronger among white males 45–64 years of age than among white males 65–74 years of age Given that risk factor associations differ by age, the age-stratified analysis presented here was the most appropriate one for this example

Results from analyses for all white men 45–74 years of age illustrate the impact of the sample weights when risk factor associations differ for elderly men 65–74 years of age who were oversampled and for middle-aged men 45–64 years of age who were not (table K) The unweighted education coefficient for men 45–74 years of age was half as large as the weighted coefficient (0 213 compared with 0 432) and outside the 95-percent confidence limit for the weighted regression coefficient The weighted analysis places more emphasis on middle-aged men (because of their larger sample weights) than does the unweighted analysis Thus, the weighted education coefficient for men 45–74 years of age was substantially larger than the unweighted coefficient because of the stronger association between education and death among middle-aged men and because of the greater emphasis on middle-aged men in the weighted analysis

Inclusion of the design variable designating residence in a poverty area did not reduce the effect of the sample weights, the coefficients and standard errors obtained from the option 5 analyses were similar to those obtained from the unweighted analyses (options 1 and 2)

χ^2 tests of the regression coefficients generally were smaller for the weighted analyses than for the unweighted analyses However conclusions concerning statistical significance of the regression coefficients, derived from the χ^2 test statistics, were similar for the weighted and unweighted analyses

Trimming the sample weights at selected percentiles (98th, 95th, 90th, and 80th) had a small to moderate effect on the regression coefficients or their standard errors in the analysis of white males 45–64 years of age (table L) The MSE's for the set of variables in the model were minimized when the weights were trimmed at the 90th percentile Trimming the weights at the 80th and 95th percentile produced similar MSE's

For black males, the unweighted and weighted regression coefficients tended to be quite different (table M) For example, the weighted coefficients for albumin were about 2 5 times the unweighted coefficients However, except for the cholesterol variable (greater than or equal to 240 mg/dl), the unweighted coefficients were within the 95-percent confidence intervals for the weighted coefficients (The confidence intervals for the weighted coefficients were quite wide because of the large standard errors) The unweighted standard errors (options 1 and 2) were similar to each other, likewise, the weighted standard errors (options 3 and 4) were similar to each other The weighted standard errors were considerably larger (more than twice as large) than the unweighted standard errors Thus, the effect of the stratification and clustering was small, whereas the effect of the sample weights was large The effect of the sample weights was more extreme in these analyses of black males than it was in the analyses of white males

Table L Results showing the effect of weight trimming on estimates from Cox regression models relating death, serum albumin, and selected risk factors for white males 45–64 years of age by trimming percentile and risk factor

Risk factor and trimming percentile	*B*	*SE*	χ^2	*Estimated mean square error*
Albumin (4 2–4 4)				
No trimming	–0 462	0 184	*6 3	0 0337
98th percentile	–0 431	0 180	*5 7	0 0335
95th percentile	–0 435	0 181	*5 8	0 0335
90th percentile	–0 441	0 182	*5 9	0 0335
80th percentile	–0 440	0 183	*5 8	0 0340
No weights	–0 436	0 152	**8 8	0 0238
Albumin (greater than 4 4)				
No trimming	–0 475	0 152	**9 8	0 0231
98th percentile	–0 468	0 152	**9 4	0 0232
95th percentile	–0 466	0 155	**9 1	0 0240
90th percentile	–0 475	0 157	**9 2	0 0245
80th percentile	–0 482	0 158	**9 3	0 0250
No weights	–0 480	0 157	**10 7	0 0246
Education (< 12 years)				
No trimming	0 635	0 150	**17 9	0 0226
98th percentile	0 589	0 140	**17 7	0 0218
95th percentile	0 582	0 139	**17 5	0 0221
90th percentile	0 583	0 138	**17 8	0 0218
80th percentile	0 573	0 137	**17 4	0 0227
No weights	0 465	0 126	**13 4	0 0450

* 01<$p \le$ 05

**$p \le$ 01

NOTES Results are based on analysis of white males 45–64 years of age at baseline from locations 1–65 Vital status from the 1987 followup wave of the NHANES I Epidemiologic Followup Study NHANES I is National Health and Examination Survey I

Age-stratified analyses could not be performed for black males because of small numbers The coefficients and standard errors obtained under option 5 were similar to those obtained from the unweighted analyses (options 1 and 2) Thus, inclusion of the design variable designating residence in a poverty area did not reduce the effect of the sample weights in this analysis of black males

The main conclusions concerning statistical significance of the albumin regression coefficients were similar for the unweighted and weighted analyses However, age, smoking, and systolic blood pressure were statistically significant in the unweighted analysis, but failed to reach statistical significance in the weighted analyses Age failed to reach statistical significance in the weighted analysis because the weighted standard error was much larger than the unweighted standard error, and the weighted regression coefficient for age was somewhat smaller than the unweighted regression coefficients As age is a known risk factor in mortality analyses, its failure to achieve statistical significance in the weighted analyses is disturbing and puts in doubt the credibility of the weighted analysis This result is another indication that using the highly variable and skewed sample weights when analyzing NHEFS data results in standard errors that are too large

Trimming the sample weights had a large impact on the regression coefficients and their standard errors in the analyses of black males (table N) Trimming the sample weights at the

Table M Results from Cox regression models relating death, serum albumin, and selected baseline risk factors for black males 45–74 years of age, by analysis option

	Unweighted analyses					*Weighted analyses*							
	Option 1			*Option 2*		*Option 3*			*Option 4*		*Option 5*		
Risk factor	*Beta*	*Standard error*	χ^2	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2	*Standard error*	χ^2	*Beta*	*Standard error*	χ^2
Albumin (4 2–4 4)	–0 181	0 145	1 6	0 136	1 8	–0 450	0 315	2 0	0 305	2 2	–0 179	0 137	1 7
Albumin (> 4 4)	–0 400	0 174	*5 3	0 158	*6 4	–1 056	0 383	**7 6	0 376	**7 9	–0 400	0 156	*6 6
Age (years)	0 057	0 009	**37 7	0 008	**56 1	0 035	0 021	2 8	0 020	2 9	0 057	0 008	**55 6
Education (< 12 years)	0 341	0 225	2 3	0 191	3 2	0 909	0 460	*3 9	0 482	3 6	0 333	0 192	3 0
Diabetes history (yes)	0 624	0 218	**8 2	0 243	*6 6	0 843	0 316	**7 1	0 295	**8 2	0 616	0 247	*6 3
SBP (mmHg)[1]	0 006	0 002	*6 1	0 002	**7 3	0 006	0 005	2 0	0 005	2 0	0 006	0 002	**7 2
Smoking (yes)	0 278	0 129	*4 6	0 149	3 5	0 031	0 276	0 0	0 255	0 0	0 282	0 149	3 6
Cholesterol (200–239)	–0 120	0 155	0 6	0 152	0 6	0 536	0 327	2 7	0 330	2 6	–0 115	0 151	0 6
Cholesterol (≥240)	–0 100	0 159	0 4	0 136	0 5	0 531	0 299	3 2	0 328	2 6	–0 083	0 136	0 4
Poverty segment											0 169	0 150	1 3

* 01<p≤ 05

**p≤ 01

[1] SBP is sytolic blood pressure

NOTES Results are based on analysis of black males 45–74 years of age at baseline from locations 1–65 Vital status data are from the 1987 followup wave of the NHANES I Epidemiologic Followup Study NHANES I is the National Health and Nutrition Examination Survey I

98th percentile produced a marked change in most of the regression coefficients (they became more similar to the unweighted coefficients) and their standard errors (they became smaller) Further trimming, at the 95th and 90th percentiles, resulted only in small additional changes in the estimates More extreme trimming, at the 80th percentile, again resulted in larger changes in the estimates

Examination of the MSE's for the variables in the model when the weights were trimmed at the 98th, 95th, 90th, and 80th percentiles showed that trimming the weights at the 90th percentile minimized the MSE's for the set of variables Trimming the weights at the 95th percentile produced nearly equivalent results to trimming at the 90th percentile Trimming the weights at the 80th percentile resulted in larger MSE's Thus, in this analysis of black males, it appears to be beneficial to trim the sample weights at the 90th percentile

The effect of weight trimming on the regression coefficient and standard error for age was another indication of the importance of weight trimming in this analysis When the sample weights were trimmed at the 98th percentile, both the regression coefficient for age and its standard error changed considerably Further trimming had no effect on the regression coefficient but reduced the standard error slightly, thus reducing the MSE When the weights were trimmed, age reached statistical significance, with a p-value similar to that from the unweighted analyses

Summary

Classical sampling theory advocates the use of the complex survey design when analyzing data from the NHEFS Recent methodological and software developments have made it possible to incorporate the survey design in Cox models Incorporating the survey design when analyzing the two NHEFS data examples resulted in changes in the regression coefficients (sometimes large) and substantially larger variance

Table N Results showing the effect of weight trimming on estimates from Cox regression models relating death, serum albumin, and selected risk factors for black males 45–74 years of age, by trimming percentile and risk factors

Risk factor and trimming percentile	*Beta*	*Standard error*	χ^2	*Estimated mean square error*
Albumin (4 2–4 4)				
No trimming	–0 450	0 305	2 2	0 09287
98th percentile	–0 259	0 238	1 2	0 09343
95th percentile	–0 272	0 237	1 3	0 08816
90th percentile	–0 266	0 234	1 3	0 08866
80th percentile	–0 239	0 220	1 2	0 09332
No weights	–0 181	0 136	1 8	0 09087
Albumin (>4 4)				
No trimming	–1 056	0 376	**7 9	0 14174
98th percentile	–0 833	0 329	*6 4	0 15787
95th percentile	–0 843	0 328	*6 6	0 15336
90th percentile	–0 829	0 330	*6 3	0 16042
80th percentile	–0 725	0 327	*4 9	0 21598
No weights	–0 400	0 158	*6 4	0 18520
Age (years)				
No trimming	0 035	0 020	2 9	0 00041
98th percentile	0 050	0 013	**15 7	0 00040
95th percentile	0 050	0 012	**16 2	0 00039
90th percentile	0 050	0 012	**16 5	0 00038
80th percentile	0 052	0 012	**17 5	0 00044
No weights	0 057	0 008	**56 1	0 00057
Education (<12 years)				
No trimming	0 909	0 482	3 6	0 23221
98th percentile	0 739	0 439	2 8	0 22130
95th percentile	0 744	0 438	2 9	0 21959
90th percentile	0 737	0 435	2 9	0 21871
80th percentile	0 719	0 417	3 0	0 20999
No weights	0 341	0 191	3 2	0 35914

* 01≥p≤ 05

**p≥ 01

NOTES Results are based on analysis of black males 45–74 years of age at baseline from locations 1–65 Vital status data are from the 1987 followup wave of the NHANES I Epidemiologic Followup Study NHANES I is the National Health and Nutrition Examination Survey I estimates In both data examples, some risk factors that reached statistical significance in the unweighted analyses failed to reach statistical significance in the weighted analyses However, in both of the data examples, as well as in numerous other NHEFS analyses we have performed, the overall conclusions generally did not change when the sample weights were used

The stratification and clustering of the NHEFS had little effect on the standard errors of the regression coefficients in these analyses The effect of clustering in the NHANES I may be small because in most cases only one person was sampled from each household in the survey (8) Further, when analyses are performed for subdomains, the effect of the stratification and clustering is reduced

Most of the effect of the survey design was due to the sample weights The data example illustrated that the weights could have a large impact on both the regression coefficients and their standard errors Although the use of sample weights does increase variance estimates, the variability and skewness of the weights for NHEFS resulted in excessive increases in the variance estimates The impact of the sample weights on the age variable in the analysis of mortality among black males illustrated this When the sample weights were used, this known risk factor for mortality did not reach statistical significance because of a reduced regression coefficient and a substantially increased standard error This was a disturbing result and showed that the weights should not be used without care Trimming the weights moderately (98th percentile) reduced the standard error for age so that age again reached statistical significance

Several techniques were evaluated for use in reducing the effect of the sample weights on the parameter estimates The first approach examined was stratification Stratifying by age (one of the oversampling variables) was found to reduce the effect of the sample weights on the regression coefficients from the Cox models somewhat and to substantially reduce the standard errors Analyses are frequently stratified by age, race, and sex for epidemiologic reasons The data example showed the importance of stratification in the presence of interactions involving variables used in the oversampling Given that women 25–44 years of age and all persons 65–74 years of age were oversampled in locations 1–65, stratifying analyses by sex and age may be desirable when analyzing data from the NHEFS Unfortunately, small numbers in subgroups may prohibit stratification, as they did for black males in the second data example

Another approach to reduce the effect of the sample weights on parameter estimates is to include variables in the model that were used in the calculation of the sample weights (age, sex, residence in a poverty area, or family income) In the second data example, the inclusion of residence in a poverty area in the model did not effectively reduce the impact of the sample weights on the regression coefficients and their standard errors When including design variables in the model, the possibility of multicollinearity must be considered For example, family income and education probably should not be included in the same model because of their correlation

A third approach to reduce the effect of the sample weights is weight trimming Trimming the sample weights in the analysis of white males had little effect on the parameter estimates Trimming the weights in the analysis of black males had substantial effects both on the regression coefficients and their standard errors Trimming moderately (98th percentile) changed the regression coefficient for age and reduced its standard error so that this risk factor reached statistical significance Clearly, in this analysis of black males, the weights should be trimmed if they are used The effect of weight trimming should be evaluated for each analysis, in some situations it may be beneficial, but in others it may not

One analytical strategy is to perform preliminary analyses under option 1 (ignoring all aspects of the complex survey design) The final analyses also can be carried out under option 4 (and, if desired, options 2 and 3) to assess the effect of the stratification and clustering and the sample weights on the regression coefficients and their standard errors Some previously published studies have taken this approach, using cumulative logistic or person-time logistic models to compare unweighted and weighted results (34–39) In all instances, the authors concluded that the results from the unweighted and weighted analyses were similar, and presented the unweighted analyses The reasons given for presenting the unweighted results rather than the weighted results have been the highly variable sample weights that increase the variance estimates substantially and the small effect of clustering

In summary, given that the effect of incorporating the survey design is almost entirely due to the sample weights and that this effect can be excessive because the weights are highly variable and skewed, it is not clear that the weighted analysis is the most appropriate one However, the survey design must be considered when analyzing data from the NHEFS, it is inappropriate to assume a priori that the survey design should be ignored For each analysis, it is necessary to examine carefully the impact of the survey design Special care should be taken to check the data for outliers, because an outlier in conjunction with an extreme sample weight can have a substantial impact on the analysis Differences in risk factor associations by variables used to oversample groups should also be checked because unweighted results can be seriously biased in their presence if a stratified model is not employed Because of the oversampling of the elderly in NHEFS and the fact that many risk factor associations differ by age, age stratification in analyses of NHEFS data is often appropriate Techniques such as stratification, inclusion of design variables in the model, and weight trimming should be tried to reduce the effect of the sample weights on the estimates The decision whether to present unweighted or weighted results should be made for each analysis individually

Other statistical issues

Calculation of followup time

Analysis of data from a followup study typically focuses on the occurrence of some specified event, usually death or disease onset The longer an individual is under observation, the more likely it is that the event of interest will be observed Thus, it is essential in a followup study to take into account differences in the length of time individuals are observed Individuals in the NHEFS had different lengths of followup because they had baseline examinations at different times (1971–75), had followup interviews or were lost to followup at different times (Wave 1 1982–84, Wave 2 1986, and Wave 3 1987), died at some point during the study period, or were hospitalized for a condition at any time during the study period

Length of followup is calculated as the time between the individual's date of entry into the study and the last date the individual was known to be at risk for the event For the NHEFS, the date of entry is the date of the individual's NHANES I examination The date that the individual was last at risk depends on the endpoint of interest Determining this date is straightforward for mortality analyses, but more complicated for incidence analyses

For mortality analyses, the date the individual was last at risk is the date of death for decedents and the last date known alive for others Note that the last date known alive may or may not be the date of the last followup interview because some individuals were traced alive but not interviewed As an example of the calculation of length of followup in a mortality analysis, suppose an individual participated in NHANES I in 1973, was interviewed in 1983 (for the first wave of followup), was not interviewed in the 1986 or 1987 waves of followup, but was known to be alive in 1987 The last date this individual was known to be at risk of dying is 1987, not 1983 (the date of the last interview) Thus, for a mortality analysis, this individual's followup time would be the date last known alive minus the NHANES I examination date (1987 – 1973 = 14 years)

For incidence analyses, a number of subtleties may be involved in determining the date an individual was last at risk Information from one of the interviews, hospital (or other health care facility) records, or a death certificate may be used to identify incident cases Thus, for cases, the last date at risk could be the date of the last interview at which information on the event of interest was collected, a date earlier than the interview that was calculated from information obtained at the interview, the date of a hospitalization, or the date of death For noncases, the date last at risk could be the date of the last interview at which information about the event was collected or the date of death

To illustrate the calculation of followup time for an incidence analysis, consider an example in which incident cases are identified using information collected at each followup interview, the date of the interview is assumed to be the incident date, and the baseline examination was in 1973 If an individual participated in all waves of followup and did not experience the event of interest, then followup time would be the date of the last interview minus the date of the NHANES I examination (1987 – 1973 = 14 years) If an individual was interviewed only in 1983 (for the first wave of followup) and was not a case but was known to be alive in 1987, followup time would still be the date of last interview minus the NHANES I exam date (1983 – 1973 = 10 years) It would be inappropriate to calculate this individual's followup time using the last date known alive, as was done for the mortality analysis, because the individual could have become an incident case after the last interview Similarly, if an individual was a noncase when interviewed in 1983 and subsequently died, the last date the individual was at risk would be 1983, and length of followup would again be the date of last interview minus the NHANES I examination date (1983 – 1973 = 10 years)

Another subtlety that must be considered when calculating followup time is that proxy interviews were conducted for most decedents at the followup wave after their death The analyst must be careful not to use the date of this proxy interview when calculating followup time Even if information from the proxy interview is used to determine whether the decedent was an incident case, the date of death or some date prior to the date of death must be used when calculating followup time

Sample weights

Six sets of sample weights were originally calculated for NHANES I, one set for each of the six NHANES I samples shown in table A The six sets of sample weights can be used in the analysis of the corresponding samples from the NHEFS Note that the sample weights have not been adjusted for nonresponse and loss to followup in the different waves of followup The weights are available on the NHANES I and NHEFS data tapes Further information about these six sets of sample weights is given in the documentation for the data tapes as well as in the reports that describe the details of NHANES I (5–7,9–11)

No sample weights were calculated for the entire NHANES I sample (all persons from locations 1–100) Therefore, another set of sample weights had to be developed for use with all 14,407 persons in the NHEFS The NHEFS is a combination of two NHANES I samples sample persons 25–74 years of age from locations 1–65 and all sample persons from locations 66–100 Each of these samples has a set of sample weights that sum to the national population 25–74 years of age at the midpoint of the corresponding data collection period—1971–74 and 1974–75, respectively The weights corresponding to these two samples cannot be used directly in an analysis including all NHEFS sample persons for several reasons First, if the two sets of weights were used directly, the sample weights would sum to about twice the national population of persons 25–74 years of age Second, those persons in locations 66–100 composed about 21 percent of the total sample, but they would receive 51 percent of the weight in an analysis using the two sets of weights directly Further, because of the oversampling of the elderly, persons residing in poverty areas, and women of childbearing age in locations 1–65, the proportion of the total sample from locations 66–100 varies substantially among subgroups—from 5 6 percent of black persons 65–74 years of age to 31 7 percent of persons other than black 45–64 years of age In a weighted analysis that used the two sets of weights directly, the 5 6 percent of black persons 65–74 years of age from locations 66–100 would receive more than 50 percent of the weight for this subgroup Therefore, approximate sample weights were calculated for analyses involving the entire NHEFS sample

Approximate sample weights for the entire NHEFS sample were calculated using the NHANES I sample weights for all sample persons in locations 1–65 (wt165) and all sample persons in locations 66–100 (wt66100) To do this, the NHEFS sample was divided into 12 age–race–sex groups based on 3 age groups (25–44, 45–64, and 65–74 years) and 2 race groups (black and other than black) The new sample weights were defined so that

- The contribution of the 11,348 persons from locations 1–65 and the 3,059 persons from locations 66–100 was proportional to their contribution to the total sample within the 12 age–race–sex groups
- The new weights summed to about the national population at the midpoint of the data collection period 197–75

The formula used to calculate the new weights (wt1100) within age (i = 25–44, 45–64, and 65–74), race (j = black, other than black), and sex (k = male, female) groups was as follows

$$\text{wt1100} = \begin{cases} \text{wt165} \bullet \text{adj}_{ijk} & \text{for persons in locations 1–65} \\ \text{wt66100} \bullet (1 - \text{adj}_{ijk}), & \text{for persons in locations 66–100} \end{cases}$$

where

$adj_{ijk} = n_{ijk}/(n_{ijk}+m_{ijk})$

n_{ijk}= sample size for locations 1–65 in age group i, race group j, and sex group k

m_{ijk}= sample size for locations 66–100 in age group i, race group j, and sex group k

The new sample weights are not available on the NHEFS public-use data tapes They can be calculated using the SAS code provided in appendix III

Stratum and primary sampling unit codes for variance estimation

For purposes of variance estimation, stratum and PSU codes were provided on the NHANES I and NHEFS data tapes There are two problems with the original stratum and PSU codes on these tapes

- The use of segments as PSU's for the certainty strata makes variance estimation inefficient because of the large number of segments per stratum
- For the 1–35 and 66–100 location samples, noncertainty strata have to be grouped in order to have a minimum of two PSU's per stratum

To remedy these problems, revised stratum and PSU codes were derived and are available on the 1987 NHEFS data tapes (40) The revised codes are referred to as pseudo-stratum and pseudo–PSU codes to reflect the fact that they are modifications of the strata and PSU's used in the survey design

One set of pseudo-stratum and pseudo–PSU codes was derived for use in the analysis of data from the 1–65 and 1–100 location samples For these 2 samples, the segments within each of the 10 certainty strata were combined (by random assignment) into 3 groups resulting in the formation of 3 PSU's per stratum The PSU's within the 25 noncertainty strata (2 for the 1–65 sample and 3 for the 1–100 sample) were assigned a code of 1, 2, or 3 as follows a code of 1 if the PSU was from location 1–35, a code of 2 if the PSU was from locations 36–65, and a code of 3 if the PSU was from locations 66–100 Thus, under the new coding scheme, the 1–65 location sample has 10 certainty strata, each with 3 PSU's, and 25 noncertainty strata, each with 2 PSU's The 1–100 location sample has 35 pseudo-strata with 3 pseudo–PSU's

A second set of pseudo-stratum and pseudo–PSU codes was derived for use with data from the 1–35 and 66–100 location samples For these 2 samples, the segments within the 10 certainty strata were grouped (by random assignment) into 3 groups resulting in the formation of 3 PSU's per stratum The 25 noncertainty strata (each having only 1 PSU) were grouped into 12 strata using the collapsed-strata technique (12) Eleven of these strata were formed by grouping 2 strata together, 1 was formed by grouping 3 strata together Thus, under the new coding scheme, there are 22 pseudo-strata, 11 with 2 pseudo–PSU's and 11 with 3 pseudo–PSU's

References

1 Madans JH, Kleinman JC, Cox CS, et al 10 years after NHANES I report of initial followup, 1982–1984 Public Health Rep 101(5) 1986

2 Binder DA Fitting Cox's proportional hazards model for survey data Biometrika 79(1) 139–47 1992

3 Shah BV, Barnwell BG, Hunt PN, LaVange LM SUDAAN user's manual release 5 50 with addendum for SUDAAN changes from 5 50 to 6 30 Research Triangle Park, North Carolina Research Triangle Institute 1992

4 Ingram DD, Kleinman JC Empirical comparisons of proportional hazards and logistic regression models Stat Med 8(5) 525–38 1989

5 Miller HW Plan and operation of the Health and Nutrition Examination Survey, United States, 1971–73 National Center for Health Statistics Vital Health Stat 1(10a) 1973

6 National Center for Health Statistics Plan and operation of the Health and Nutrition Examination Survey, United States, 1971–73 National Center for Health Statistics Vital Health Stat 1(10b) 1977

7 Engel A, Murphy RS, Maurer K, Collins E Plan and operation of the NHANES I Augmentation Survey of adults 25–74 years, United States, 1974–75 National Center for Health Statistics Vital Health Stat 1(14) 1978

8 Landis JR, Lepkowski JM, Eklunnd SA, Stehouwer SA A statistical methodology for analyzing data from a complex survey the first National Health and Nutrition Examination Survey National Center for Health Statistics Vital Health Stat 2(92) 1982

9 Cohen BB, Barbano HE, Cox CS, et al Plan and operation of the NHANES I Epidemiologic Followup Study, 1982–84 National Center for Health Statistics Vital Health Stat 1(22) 1987

10 Finucane FF, Freid VM, Madans JH, et al Plan and operation of the NHANES I Epidemiologic Followup Study, 1986 National Center for Health Statistics Vital Health Stat 1(25) 1990

11 Cox CS, Rothwell ST, Madans JH, et al Plan and operation of the NHANES I Epidemiologic Followup Study, 1987 National Center for Health Statistics Vital Health Stat 1(27) 1992

12 Kalbfleisch JD, Prentice RL The statistical analysis of failure time data New York John Wiley & Sons, Inc 1980

13 Anderson S, Auquier A, Hauck WW, et al Statistical methods for comparative studies New York John Wiley & Sons, Inc 1980

14 Hansen MH, Madow WG, Tepping BJ An evaluation of model-dependent and probability-sampling inferences in sample surveys JASA 78(384) 776–807 1983

15 Holt D, Smith TMF, Winter PD Regression analyses of data from complex surveys J R Statist Soc, Part A 143 474–87 1980

16 Dumouche WH, Duncan GJ Using sample survey weights in multiple regression analyses of stratified samples JASA 78(3983) 535–543 1983

17 Roberts G, Rao JNK, Kumar S Logistic regression analysis of sample survey data Biometrika 74(1) 1–12 1987

18 Nathan G Inferences based on data from complex survey designs In Krishnaiah PR and Rao CR eds Handbook of statistics, Vol 6 New York Elsevier Science Publishers B V, 247–66 1988

19 Skinner CJ, Holt D, Smith TMF, eds Analysis of complex surveys New York John Wiley & Sons, Inc 1989

20 Sarndal CE, Swenson B, Wretnan J Model assisted survey sampling New York Springer-Verlag 1992

21 Korn EL, Graubard BI Epidemiologic studies utilizing surveys accounting for the sample design Am J Public Health 81 1166–73 1991

22 Makuc DM, Kleinman JC Survival analysis using complex survey data examples from the National Health and Nutrition Examination Survey Epidemiologic Followup Study Presented at the annual meetings of the American Statistical Association, Chicago 1986

23 Shah BV SURREGR standard errors of regression coefficients from sample survey data Research Triangle Park, North Carolina Research Triangle Institute 1982

24 SAS Institute Inc SAS supplemental library user's guide, version 5 edition Cary, North Carolina SAS Institute Inc 1986

25 Shah BV, Folsom RE, LaVange LM, et al Statistical methods and mathematical algorithms used in SUDAAN Research Triangle Park, North Carolina Research Triangle Institute 1993

26 Binder DA On the variance of asymptotically normal estimators from complex surveys Survey Methodology 7(2) 157–70 1981

27 Woodruff RS A simple method for approximating the variance of a complicated estimate JASA 66(334) 411–4 1971

28 Satterthwaite FE An approximate distribution of estimates of variance components Biometrics 2 110–4 1946

29 Thomas DR, Rao JNK Small-sample comparisons of level and power for sample goodness-of-fit statistics under cluster sampling JASA 82(398) 630–6 1987

30 Cox BG, McGrath DS An examination of the effect of sample weight truncation on the mean square error of survey estimates Presented at the Biometric Society ENAR meeting Richmond, Virginia March 1981

31 Potter FJ Survey of procedures to control extreme sample weights In Proceedings of the Section on Survey Research Methods, American Statistical Association American Statistical Association, 453–8 1988

32 Potter FJ A study of procedures to identify and trim extreme sample weights In Proceedings of the Section on Survey Research Methods, American Statistical Association American Statistical Association, 225–30 1990

33 Potter FJ The effect of weight trimming on nonlinear survey estimates Presented at the annual meetings of the American Statistical Association San Francisco, California 1993

34 Kleinman JC, Donahue RP, Harris MI, et al Mortality among diabetics in a national sample Am J Epidemiol 128(2) 289–401 1988

35 Feldman JJ, Makuc DM, Kleinman JC, Cornoni-Huntley J National trends in educational differentials in mortality Am J Epidemiol 129(5) 919–33 1989

36 Havlik RJ, LaCroix AZ, Kleinman JC, et al Antihypertensive drug therapy and survival by treatment status in a national survey Hypertension 13(suppl I) I28–I32 1989

37 Garg R, Madans JH, Kleinman JC Regional variation in ischemic heart disease incidence J Clin Epidemiol 45(2) 149–56 1992

38 Gillum RF, Makuc DM Serum albumin, coronary heart disease, and death Am Heart J 123(2) 507–13 1992

39 Gillum RF, Ingram DD, Makuc DM White blood cell count, coronary heart disease, and death the NHANES I Epidemiologic Followup Study Am Heart J 125(3) 855–63 1993

40 Rowland M, Parsons V, Makuc D Simplified design structure for NHANES I variance estimation In Proceedings of the Section on Survey Research Methods, American Statistical Association American Statistical Association 773–6 1988

41 SAS Institute Inc SAS technical report P–217, SAS/STAT software the PHREG procedure, version 6 Cary, North Carolina SAS Institute Inc 1991

42 SAS Institute Inc SAS/STAT user's guide, version 6, fourth edition, volume II Cary, North Carolina SAS Institute Inc 1990

Appendixes

Contents

Appendix tables

Appendix I
Using the Cox and person-time logistic regression models

This appendix is a practical guide to performing Cox and person-time logistic regressions under four analysis options

1 Ignore all aspects of the complex survey design
2 Incorporate only the stratification and clustering
3 Incorporate only the sample weights
4 Incorporate both the stratification and clustering and the sample weights

Standard statistical software can be used to perform the analyses under option 1, and an approach using SAS is given To obtain correct variance estimates, specialized software must be used to perform the analyses under options 2, 3, and 4 An approach using SUDAAN is given

This appendix also includes discussions about choice of a time interval for the person-time logistic regression model and about how to check the exponential assumption of this model

Definition of variables used

The following variables are used in the algorithms for performing Cox or person-time logistic regressions

VS = a dichotomous variable representing the outcome event of interest, usually death or disease incidence VS is coded 0 if the individual was censored and 1 if the individual had an event

VSN = a recode of VS created when the data are arranged for a person-time logistic analysis For PROC LOGISTIC in SAS Version 6, VSN should be coded 2 if the individual was censored and 1 if the individual had an event For SUDAAN PROC LOGISTIC, VSN should be coded 0 if the individual was censored and 1 if the individual had an event

FU = followup time, that is, the total length of time the individual was at risk for the outcome event of interest When the date functions in SAS are used for this calculation, the followup time is in days See "Calculation of followup time "

FUT = number of time intervals the individual was followed This variable, for use in a person-time logistic analysis, is derived from FU by dividing by the number of days in a time interval FUT is calculated so that if the individual is followed for part of an interval, the interval counts as a whole interval

NFUT = a recode of FUT created when the data are arranged for a person-time logistic analysis

SAMPWT = sample weight

FUTSMPWT = product of number of time intervals the individual was followed and the individual's sample weight

NOWT = dummy sample weight of 1

STRATUM = stratum code

NOSTRAT = dummy stratum code of 1

PSU = primary sampling unit (PSU) code

ID = unique identification code for each individual

Performing Cox regressions

Analysis under option 1

For option 1, the SAS procedure PHREG (SAS Version 6) can be used to perform the Cox regression analysis (41) The outcome variable should be coded 0 for a censored individual and 1 for an individual who had an event

SAS code—

```
PROC PHREG DATA = COX,

MODEL FU * outcome variable(1) = variables in model,
```

Analyses under options 2, 3, and 4

For options 2, 3, and 4, use PROC SURVIVAL in SUDAAN (25) Either a first-order SAS Version 5 data set or a sequential file (ASCII for PC SUDAAN) can be used as input for the SURVIVAL procedure The outcome variable should be coded 0 for a censored individual and 1 for an individual who had an event The coding of categorical variables that are in the model is different in SUDAAN than it is in SAS In SUDAAN, categorical variables must have positive nonzero values, and the largest value is the reference value For example, dichotomous variables (coded 0–1 in SAS) must be coded 1–2, where the 2 represents the reference group (that was previously coded 0)

The recommended design for the NHEFS is "WR," which means "with replacement " The NEST statement is used to specify the stratum and PSU variables and the WEIGHT statement is used to specify a weight variable Choice of appropriate stratum and PSU codes and sample weights for different NHEFS subsamples was discussed in the section "Other statistical issues " Dummy stratum and PSU codes are used for option 3 and dummy sample weights are used for option 2 Thus, the variables on the NEST and WEIGHT statements differ for the three options As a check, SUDAAN can be run with the design "SRS" (simple random sample) and no NEST statement Regression coefficients obtained from this analysis should be identical to those obtained from the SAS PHREG analysis, the standard errors obtained will be similar but not identical

The data set used as input for SUDAAN must be sorted by the stratum and PSU variables Hence, the data set is in a different sort order for option 3 than it is for options 2 and 4 When analyzing subgroups of the NHEFS sample, use the SUBPOPN statement to select the subgroups from the total sample rather than performing the analyses using subfiles When the SUBPOPN statement is used, SUDAAN is able to use the full design information to calculate the variances If subfiles are used, the variance estimates will be incorrect if there is not at least one person from the subgroup in each PSU within a stratum

Following are specific details for options 2, 3, and 4

Option 2—The regression coefficients under option 2 will be identical to those obtained under option 1 The estimates of the standard errors will be different The sample weights are ignored by using a dummy sample weight of 1 for each individual

Option 3—The stratification and clustering is ignored by assigning all individuals to the same stratum and having each individual represent a unique PSU Thus, a dummy stratum code of 1 is used for all individuals and dummy PSU codes are used so that each individual has a unique PSU code (for example, the ID's)

Option 4—The regression coefficients obtained under option 4 will be identical to those obtained under option 3 The standard error estimates will be different

SUDAAN code for options 2, 3, and 4

```
PROC SURVIVAL DATA=SASFIL COXxx
                              /* xx=24 for options 2,4 */
                              /* xx=3 for option 3     */
DESIGN= WR
FILETYPE=SAS
EST_NO=number of observations in analysis file,
*******************************************,
/*Use the appropriate NEST and WEIGHT statements for the
  option being performed                               */
/* NEST and WEIGHT statements for option 2             */
/*   NEST STRATUM PSU,                                 */
/*   WEIGHT NOWT,                                      */
/* NEST and WEIGHT statements for option 3             */
/*   NEST NOSTRAT ID,                                  */
/*   WEIGHT FUWT,                                      */
/* NEST and WEIGHT statements for option 4             */
/*   NEST STRATUM PSU,                                 */
/*   WEIGHT FUWT,                                      */
*******************************************,
EVENT VS,
SUBGROUP categorical variables in model,
LEVELS number of categories in each categorical variable,
SUBPOPN domain variables and ranges,
MODEL FU=variables in model,
SETENV DECWIDTH=5 LINESIZE=132,
TEST SATADCHI WALDCHI WALDF,
PRINT BETA SE_BETA T_BETA P_BETA DEFT/
  TEST=ALL STYLE=NCHS,
TITLE ' Cox regression",
```

Performing person-time logistic regressions

To perform a person-time logistic regression analysis, a time interval must be chosen, the data set must be arranged appropriately, and the exponential assumption must be checked These three aspects are discussed in this section, followed by a description of performing person-time logistic analyses under the four analysis options

Choice of time interval

The first step in performing a person-time logistic regression analysis is to choose the time interval We have done extensive work using both simulated data sets and the NHEFS data to examine the best choice for the time interval as well as how sensitive the parameter estimates are to this choice This work has shown that the parameter estimates are not very sensitive to choice of the time interval as long as it is short enough so that the probability of an event occurring in any given interval is small The simulation results show that the

Table I Simulation results showing the effect of the time interval on estimates from the person-time logistic regression model

Model and time interval	*Beta*	*Standard error*
Cox model	0 68	0 9
Person-time logistic model		
1 week	0 66	0 9
1 month	0 56	0 9
6 months	0 68	0 9
1 year	0 71	0 9
2 years	0 75	0 1

estimates from the models with 1-week, 1-month, 6-month, and 1-year time intervals are all quite similar to the Cox model estimates (table I) The estimates from the 2-year time interval model are somewhat larger than the Cox estimates but not entirely dissimilar We also examined the effect of choice of time interval when there is censoring and found similar results (data not shown) Analyses of NHEFS data using 1-month, 6-months, and 1-year time intervals also produced parameter estimates both similar to each other and to the Cox estimates (data not shown)

One consideration in choosing a time interval is computer time It is desirable to choose a time interval as long as possible because a longer interval means less computer time will be needed This is primarily an issue when the sample weights are used Differences in the amount of computer time used are negligible when no weights are involved Convention also may influence the choice of the time interval For example, a 1-year interval is commonly used in epidemiologic analyses whereas a 2-year interval is not For our analyses, we generally use a 1-month time interval

Checking the exponential distribution assumption

Before performing a person-time logistic regression analysis, it is advisable to check the assumption that the survival times are exponentially distributed, if they are not, the model may produce biased parameter estimates

The survival distribution function, *S(t)*, for the Weibull distribution is

$$S(t_i)_1 = \exp[-(t_i/b)^c]$$

where t_i is the *i*th followup time and *b* and *c* are unknown parameters The exponential distribution is a special case of the Weibull distribution with $c = 1$

Using the natural logarithmic function, the relationship between the survival function and time, *t*, can be expressed as a straight line

$$\begin{aligned} \log_e(-\log_e(S(t_i)) &= c\log_e(t_i) - c\log_e(b), \\ &= a + c\log_e(t_i) \end{aligned}$$

where $$a = c\log_e(b) \qquad (1)$$

By fitting the line in equation 1, it is possible to check (a) whether a straight line fits, in which case the survival distribution is in the Weibull family of distributions and thus the hazards are proportional, and (b) whether the slope, *c*, of the line is close to 1 For the NHEFS data sets we have studied, *c* has ranged from 1 16 to 1 26, and person-time logistic parameter estimates calculated have been close to those obtained from the Cox model To fit the line in equation 1, the survival distribution function, *S(t)*, must be estimated $S(t_i)$ can be estimated from the data using the formula

$$S(t_i) = p_0 p_1 \quad p_i, \qquad (2)$$

where

p_0 = 1 and $p_i = (N_i - d_i)/N_i$,
d_i = number of individuals who died in the *i*th interval,
N_i = number of individuals at risk at the beginning of the interval

The following SAS code can be used to estimate $S(t_i)$ and fit the line in equation 1

ATRISK = the number of individuals at risk at time *i*, it is initialized at the sample size,

ST = cumulative survival distribution function $p_0 p_1 \quad p_i$ It is initialized at 1,

NI = number of individuals who died or were censored at time *i*,

CASES = number of individuals who became a case at time *i*,

```
PROC SORT, BY FU,
PROC MEANS N SUM NOPRINT,
    VAR STATUS,
    BY FU,
    OUTPUT OUT=EXPTEST N=NI SUM=CASES,
DATA EXPTEST, SET EXPTEST,
    RETAIN ATRISK sample size ST 1,
    P=(ATRISK-CASES)/ATRISK,
    ST=ST*P,
    LST=LOG(-LOG(ST)),
    LFU=LOG(FU),
    ATRISK=ATRISK-NI, /*Reset number at risk at end of
                            time i*/
PROC REG,
MODEL LST=LFU,
```

The coefficient of LFU is the estimate of *c*

Arrangement of the data set

To perform a person-time logistic regression analysis, create a data set that has one observation for each individual who does not have an event during the followup period and two observations for each individual who does have an event during the followup period (one for the time intervals in which no event occurs and one for the interval in which the event occurs) Recode the VS variable so that it is 0 for the time intervals in which no event occurs and 1 for the time interval in which an event occurs Add a count variable to each record to represent the number of time intervals the individual is followed To illustrate, consider a data set with two individuals, one who is followed for 120 months and does not have an event, and one who has an event in the 60th month of followup Thus, for the individual who did not have an event, there will be one observation in the analysis data set This observation will have a count variable with a value of 120, representing the number of time intervals the individual was followed, and VSN = no event occurred For the individual who did have an event in the 60th month of followup, there will be two observations in the analysis data set The first observation will have a count variable with a value of 59 and VSN = no event occurred, the second will have a count variable with a value of 1 and VSN = event occurred SAS code for creating this data set is provided in this section (SAS Version 6 for SAS PROC LOGISTIC and SAS Version 5 for SUDAAN)

```
DATA PTL, SET ORIGINAL,

/*Calculate the number of time intervals an individual  */

/*was followed using the CEIL function so that if the   */

/*individual was followed for part of an interval, the  */

/*interval is counted                                    */

FUT=CEIL (FU/number of days in chosen time interval),

/*Create one observation for each individual who did not
  have an event                                          */
```

Code for SAS Version 5—

```
IF VS=0 THEN DO,
   VSN=0,
   NFUT=FUT,
   FUTSMPWT=NFUT*SAMPWT,
   OUTPUT,
END,
```

Code for SAS Version 6—

```
IF VS=0 THEN DO,
    VSN=2,
    NFUT=FUT,
    OUTPUT,
END,
/*Create two observations for each individual          */
/*who had an event during the following period,        */
/*one observation for the time intervals during        */
/*which no event occurred and one observation          */
/*for the time interval in which the event occurred    */
```

Code for SAS Version 5—

```
   IF VS=1 THEN DO,
      IF FUT > 1 THEN DO,
         VSN=0,
         NFUT=FUT-1,
         FUTSMPWT=NFUT*SAMPWT,
         OUTPUT, END,
      VSN=1,
      NFUT=1,
         FUTSMPWT=NFUT*SAMPWT,
         OUTPUT,
      END,
```

Code for SAS Version 6—

```
IF VS=1 THEN DO,
   IF FUT > 1 THEN DO,
      VSN=2,
      NFUT=FUT-1,
      OUTPUT, END,
   VSN=1,
   NFUT=1,
   OUTPUT,
END,
```

Analysis under option 1

For option 1, the SAS procedure LOGISTIC (SAS Version 6) can be used to perform the person-time logistic regression (42) The WEIGHT statement is needed to represent the count of the time intervals each individual was followed For SAS PROC LOGISTIC, the outcome variable should be coded 2 for an individual who does not have an event and 1 for an individual who does have an event The data set must be arranged as already described

SAS 6 code—

```
PROC LOGISTIC DATA=PTL,

MODEL VSN=variables in model,

WEIGHT NFUT,
```

Analysis under options 2, 3, and 4

For options 2, 3, and 4, use PROC LOGISTIC in SUDAAN to perform the person-time logistic regression analysis (25) Either a first-order SAS Version 5 data set or a sequential file (ASCII for PC SUDAAN) can be used as input for the LOGISTIC procedure The data set should be arranged as already described The outcome variable should be coded 0 for an individual who does not have an event and 1 for an individual who does have an event The coding of categorical variables in the model is different in SUDAAN than it is in SAS In SUDAAN, categorical variables must have positive nonzero values For example, a dichotomous variable (coded 0–1 in SAS) must be coded 1–2 in SUDAAN where 2 represents the reference group (coded 0 in SAS)

The recommended design for the NHEFS is "WR," which means "with replacement " The NEST statement is used to specify the stratum and PSU variables, and the WEIGHT statement is used to specify the weight variable Note that the weight variable for a person-time logistic regression using the data arrangement already described is the product of the sample weight and the number of time intervals of followup Choice of appropriate stratum and PSU codes and sample weights for different NHEFS subsamples was discussed in the section "Other statistical issues" Dummy stratum and PSU codes are used for option 3 and dummy sample weights are used for option 2 Thus, the variables in the NEST and WEIGHT statements differ for the three options

As a check, SUDAAN can be run with a design of "SRS" (simple random sampling) and no NEST statement The regression coefficients from the SRS analysis should be identical to those from the SAS LOGISTIC analysis The standard errors from the two analyses should be similar but not identical

The data set must be sorted by the stratum and PSU variables Thus, the data set is in a different sort order for option 3 than it is for options 2 and 4

The design effect calculated in SUDAAN when the data are arranged as previously described will not be correct because the weight variable represents both the sample weight and the number of time intervals of followup

When analyzing subgroups of the NHEFS sample, use the SUBPOPN statement to select the subgroups from the total sample rather than performing the analyses using subfiles When the SUBPOPN statement is used, SUDAAN is able to use the full design information to calculate the variances If subfiles are used, the variance estimates will be incorrect if there is not at least one person from the subgroup in each PSU within a stratum

Following are the specific details for options 2, 3, and 4

Option 2—The regression coefficients obtained under option 2 will be identical to those obtained under option 1 The estimates of the standard errors will be different The sample weights are ignored by using a dummy sample weight of 1 for each individual Thus, the weight variable for this option, which is the product of the sample weight and the intervals of followup, is just the count of the time intervals of followup

Option 3—For this analysis, all individuals are assigned to the same stratum, and each individual represents a unique PSU Thus, a dummy stratum code of 1 is used for all individuals and a set of dummy PSU codes (for example, the ID's) such that each individual has a unique code is used The weight variable for this person-time logistic regression analysis is the product of the individual's sample weight and the count of the time intervals the individual was followed

Option 4—The regression coefficients obtained under option 4 will be identical to those obtained under option 3 The standard error estimates will be different The weight variable for this analysis is the product of the individual's sample weight and the count of the time intervals the individual was followed

SUDAAN code for options 2, 3, and 4—

```
PROC LOGISTIC DATA=SASLIB PTLxx
                    /* xx=24 for options 2,4              */
                    /* xx=3 for option 3                  */
DESIGN=WR
FILETYPE=SAS,
*************************************,

/*Use the appropriate NEST and WEIGHT statements for the
option being performed,                                    */

/*NEST and WEIGHT statements for option 2—                 */
/*  NEST STRATUM PSU,                                      */
/*  WEIGHT NFUT,                                           */

/*NEST and WEIGHT statements for option 3—                 */
/*  NEST NOSTRAT ID,                                       */
/*  WEIGHT FUTSMPWT,                                       */

/*NEST and WEIGHT statements for option 4—                 */
/*  NEST STRATUM PSU,                                      */
/*  WEIGHT FUTSMPWT,                                       */
**********************************************,

  SUBGROUP categorical variables in model,
  LEVELS levels of each categorical variable,
  SUBPOPN domain variables and ranges,
  MODEL VSN=variables in model,
  SETENV DECWIDTH=5 LINESIZE=132,
  TEST SATADCHI WALDCHI WALDF,
  PRINT BETA SEBETA T_BETA
     P_BETA DEFT/TEST=ALL
     STYLE=NCHS,
  TITLE "Person-time logistic regression",
```

Appendix II
Weight trimming

Extreme variation in sample weights can result in excessively large variance estimates and loss of power Weight trimming, also called weight truncation, is a technique that can be used to reduce the size and number of extreme sample weights Weight trimming involves identifying extreme sample weights, reducing them to some specified maximum, and distributing the trimmed portion of these weights so that the adjusted weights sum to the same total as the original weights

The goal of weight trimming is to reduce the mean square error of parameter estimates The mean square error is the sum of the squared bias of an estimate and the variance of the estimate An optimal trimming point reduces the variance of the estimate enough to offset the bias that is introduced by trimming the sample weights

When the sample weight distribution differs across subgroups, weight trimming should be done within the subgroups In other words, if weights are to be trimmed at the 98th percentile, they should be trimmed at the 98th percentile for each subgroup rather than at the 98th percentile for the entire sample In the second data example (analysis of location 1–65 sample), in the section "Incorporating the complex survey design," we trimmed weights within eight age–race–poverty residence subgroups (table II)

Numerous trimming procedures are available (30–33) In this report we used the inspection procedure and the estimated mean square error (MSE) procedure using regression coefficients (31,33) Both procedures are described briefly in this section

Inspection procedure

The inspection approach is simple, but it is subjective and does not assess the effect of the trimming on the MSE of variables This procedure generally involves examining the mean, variance, coefficient of variation, and selected percentiles of the sample weight distribution to identify a logical trimming point

Estimated mean square error procedure using regression coefficients

The MSE procedure using regression coefficients involves calculating the mean square error for each of the variables in the model using weights trimmed at t different levels For each variable, the t-estimated MSE's are ranked The trimming level with the smallest average rank across the variables minimizes the MSE for the set of variables In the second data example, there were nine variables in the model, and we considered four different trimming levels (98th, 95th, 90th, and 80th percentiles) For each of the nine variables, the MSE was estimated using the four trimming levels and was assigned a rank of one to four Finally, for each trimming level, the nine

Table II Sample weight percentile for males 45–74 years of age in locations 1–65 by poverty residence, race, and age NHANES I Epidemiologic Followup Study

		Sample weight percentile				
Poverty residence, race, and age	*N*	*100*	*98*	*95*	*90*	*Minimum*
Nonpoverty residence						
White males						
45–64 years	681	90 940	46 743	38,568	30 854	4,546
65–74 years	787	21 866	12 456	9 420	7 512	1 013
Black males						
45–64 years	39	45 127	45 127	36,368	36 043	4 263
65–74 years	46	8 625	8,625	7,878	6 335	889
Poverty residence						
White males						
45–64 years	471	69,503	15,140	12,124	10 831	2 142
65–74 years	574	10,675	3,092	2,749	2 419	498
Black males						
45–64 years	175	59 809	10 217	9 610	8,863	969
65–74 years	248	4 553	2,518	2 286	2,054	471

NOTE NHANES I is the National Health and Nutrition Examination Survey I ranks were averaged, and the trimming level with the smallest average rank was chosen as the optimal trimming level

An approximate formula for the MSE of the estimate when the weights are trimmed at the jth percentile is

$$\mathrm{MSE}(X_j) = \mathrm{Var}\ (X_j) + (X_j - X_{100})^2$$

where

X_j = the parameter estimate obtained when the weights are trimmed at the jth percentile

X_{100} = the parameter estimate obtained when the weights are not trimmed

Appendix III
SAS code for computing sample weights for locations 1–100

As described in the section 'Other statistical issues," when analyzing the entire NHEFS sample (n = 14,407), a new set of sample weights must be calculated The SAS code in this section can be used to calculate these sample weights Two input files are needed NHEFS vital and tracing file and any NHANES I file (except 4091, 4140, and 4171)

Variables used in the algorithm—

SEQNO =	HANES I sequence number
WT165 =	HANES I sample weight for all persons from locations 1–65
WT66100 =	HANES I sample weight for locations 66–100
WT1100 =	new sample weight for all persons from locations 1–100

SAS code—

```
*Input NHEFS vital status file,

DATA NHEFSVTS,

INFILE IN1,

INPUT SEQNO 1-5

  AGE 25–26

  SEX 32

  RACE 33,

PROC SORT,

BY SEQNO,

*Input NHANES I sample weights,

DATA NHANESI,

INFILE IN2,

INPUT SEQNO 1–5

  WT165 176–181

  WT66100 182–187,

PROC SORT, BY SEQNO,

DATA COMBINE,

MERGE NHEFSVTS(IN=A) NHANESI,

BY SEQNO,

IF A,

IF AGE<45 THEN AGEC=25,

ELSE IF 45<= AGE<=64 THEN AGEC=45,

ELSE IF AGE>=65 THEN AGEC=65,

IF RACE=3 THEN RACEC=2, ELSE RACEC=1,

IF AGEC=25 AND SEX=1 AND RACEC=1 THEN ADJ=1255/
 1804,

ELSE IF AGEC=25 AND SEX=2 AND RACEC=1 THEN
 ADJ=2879/3661,

ELSE IF AGEC=45 AND SEX=1 AND RACEC=1 THEN
 ADJ=1152/1661,

ELSE IF AGEC=45 AND SEX=2 AND RACEC=1 THEN
 ADJ=1263/1875,

ELSE IF AGEC=65 AND SEX=1 AND RACEC=1 THEN
 ADJ=1361/1523

ELSE IF AGEC=65 AND SEX=2 AND RACEC=1 THEN
 ADJ=1503/1684

ELSE IF AGEC=25 AND SEX=1 AND RACEC=2 THEN
 ADJ=203/251,

ELSE IF AGEC=25 AND SEX=2 AND RACEC=2 THEN
 ADJ=658/734,

ELSE IF AGEC=45 AND SEX=1 AND RACEC=2 THEN
 ADJ=214/259,

ELSE IF AGEC=45 AND SEX=2 AND RACEC=2 THEN
 ADJ=250/309

ELSE IF AGEC=65 AND SEX=1 AND RACEC=2 THEN
 ADJ=294/313,

ELSE IF AGEC=65 AND SEX=2 AND RACEC=2 THEN
 ADJ=316/333,

IF WT66100 NE   THEN WT1100=ROUND(WT66100 *
 (1-ADJ),1),

ELSE WT1100=ROUND(WT165 * ADJ,1),
```

*U S Government Printing Office 1994 — 301-019/00005

Vital and Health Statistics series descriptions

SERIES 1 **Programs and Collection Procedures**—These reports describe the data collection programs of the National Center for Health Statistics They include descriptions of the methods used to collect and process the data definitions, and other material necessary for understanding the data

SERIES 2 **Data Evaluation and Methods Research**—These reports are studies of new statistical methods and include analytical techniques objective evaluations of reliability of collected data, and contributions to statistical theory These studies also include experimental tests of new survey methods and comparisons of U S methodology with those of other countries

SERIES 3 **Analytical and Epidemiological Studies**—These reports present analytical or interpretive studies based on vital and health statistics These reports carry the analyses further than the expository types of reports in the other series

SERIES 4 **Documents and Committee Reports**—These are final reports of major committees concerned with vital and health statistics and documents such as recommended model vital registration laws and revised birth and death certificates

SERIES 5 **International Vital and Health Statistics Reports**—These reports are analytical or descriptive reports that compare U S vital and health statistics with those of other countries or present other international data of relevance to the health statistics system of the United States

SERIES 6 **Cognition and Survey Measurement**—These reports are from the National Laboratory for Collaborative Research in Cognition and Survey Measurement They use methods of cognitive science to design evaluate and test survey instruments

SERIES 10 **Data From the National Health Interview Survey**—These reports contain statistics on illness, unintentional injuries disability, use of hospital medical and other health services and a wide range of special current health topics covering many aspects of health behaviors, health status and health care utilization They are based on data collected in a continuing national household interview survey

SERIES 11 **Data From the National Health Examination Survey, the National Health and Nutrition Examination Surveys, and the Hispanic Health and Nutrition Examination Survey**—Data from direct examination, testing and measurement on representative samples of the civilian noninstitutionalized population provide the basis for (1) medically defined total prevalence of specific diseases or conditions in the United States and the distributions of the population with respect to physical physiological and psychological characteristics and (2) analyses of trends and relationships among various measurements and between survey periods

SERIES 12 **Data From the Institutionalized Population Surveys**—Discontinued in 1975 Reports from these surveys are included in Series 13

SERIES 13 **Data From the National Health Care Survey**—These reports contain statistics on health resources and the public's use of health care resources including ambulatory, hospital, and long-term care services based on data collected directly from health care providers and provider records

SERIES 14 **Data on Health Resources Manpower and Facilities**—Discontinued in 1990 Reports on the numbers, geographic distribution, and characteristics of health resources are now included in Series 13

SERIES 15 **Data From Special Surveys**—These reports contain statistics on health and health-related topics collected in special surveys that are not part of the continuing data systems of the National Center for Health Statistics

SERIES 16 **Compilations of Advance Data From Vital and Health Statistics**—Advance Data Reports provide early release of information from the National Center for Health Statistics health and demographic surveys They are compiled in the order in which they are published Some of these releases may be followed by detailed reports in Series 10–13

SERIES 20 **Data on Mortality**—These reports contain statistics on mortality that are not included in regular annual, or monthly reports Special analyses by cause of death, age, other demographic variables, and geographic and trend analyses are included

SERIES 21 **Data on Natality, Marriage, and Divorce**—These reports contain statistics on natality, marriage, and divorce that are not included in regular annual or monthly reports Special analyses by health and demographic variables and geographic and trend analyses are included

SERIES 22 **Data From the National Mortality and Natality Surveys**—Discontinued in 1975 Reports from these sample surveys based on vital records are now published in Series 20 or 21

SERIES 23 **Data From the National Survey of Family Growth**—These reports contain statistics on factors that affect birth rates, including contraception, infertility, cohabitation, marriage divorce, and remarriage adoption, use of medical care for family planning and infertility and related maternal and infant health topics These statistics are based on national surveys of childbearing age

SERIES 24 **Compilations of Data on Natality, Mortality, Marriage, Divorce, and Induced Terminations of Pregnancy**—These include advance reports of births deaths, marriages, and divorces based on final data from the National Vital Statistics System that were published as supplements to the *Monthly Vital Statistics Report* (MVSR) These reports provide highlights and summaries of detailed data subsequently published in *Vital Statistics of the United States* Other supplements to the MVSR published here provide selected findings based on final data from the National Vital Statistics System and may be followed by detailed reports in Series 20 or 21

For answers to questions about this report or for a list of reports published in these series contact

Data Dissemination Branch
National Center for Health Statistics
Centers for Disease Control and Prevention
Public Health Service
6525 Belcrest Road Room 1064
Hyattsville MD 20782

(301) 436–8500

DEPARTMENT OF
HEALTH & HUMAN SERVICES

Public Health Service
Centers for Disease Control and Prevention
National Center for Health Statistics
6525 Belcrest Road
Hyattsville, Maryland 20782

DHHS Publication No (PHS) 94-1395, Series 2, No 121