Comparative Analysis of the NHANES (1999-2004) Public-Use and Restricted-Use Linked Mortality Files: 2010 Data Release

Suggested citation: Data Linkage Team. "Comparative analysis of the NHANES (1999-2004) public-use and restricted-use linked mortality files: 2010 public-use data release" National Center for Health Statistics. May 2010. Hyattsville, Maryland. (Available at the following address: http://www.cdc.gov/nchs/data_access/data_linkage/mortality/nhanes_99_04_linkage.htm)

Introduction

In 2009, NCHS completed a mortality linkage for National Health and Nutrition Examination Survey (NHANES) participants for the years 1999 to 2004[1], with mortality ascertained through December 31, 2006. Due to requirements to protect the confidentiality of the NHANES participants, restricted-use versions of the NHANES (1999-2004) Linked Mortality Files were made available only through the NCHS Research Data Center (RDC). To complement the restricted-use files and increase data access, NCHS has developed a plan to allow for a public-use release of linked mortality data.

In 2010, NCHS released public-use versions of the NHANES (1999-2004) Linked Mortality Files. The public-use data release includes the addition of perturbed data and was developed with the intent of eliminating re-identification risk to survey participants, maximizing the amount of mortality data included in the public-use release, while at the same time limiting the amount of synthetic data introduced to the data file.

This report describes a comparative analysis of the public-use and restricted-use NHANES (1999-2004) Linked Mortality Files. We used Cox proportional hazards models to compare the relative risk estimates for a standard set of socio-demographic covariates for all-cause as well as cause-specific mortality risk. NCHS is conducting this comparative analysis to demonstrate the comparability between the two versions of linked mortality files.

[1] Since 1999, the NHANES survey data have been released in two-year increments and the continuous NHANES Linked Mortality Files follow this same protocol.

Description of NHANES (1999-2004) Linked Mortality Data Resources

Mortality status for eligible NHANES (1999-2004) survey participants is ascertained primarily through probabilistic record matching with the National Death Index (NDI). For a complete description on the matching methodology please refer to http://www.cdc.gov/nchs/data/datalinkage/nh99+_mortality_matching_methodology_final.pdf .

The restricted-use files include detailed mortality information for all eligible survey participants including children. The restricted-use files include the following variables: survey respondent eligibility status, mortality status, age at death, age last known alive, date of death (month, day and year), underlying and multiple causes of death, date of birth, and NHANES interview and exam dates (month, day, and year).

Due to confidentiality protections, the public-use files include only eligible survey participants 18 years and older and a limited set of mortality variables. In addition, the public-use versions were subjected to data perturbation techniques to reduce the risk of respondent re-identification. Synthetic data were substituted for the actual date of death and underlying cause-of-death data for selected decedent records. Information regarding vital status was not perturbed. Variables provided on the public-use NHANES (1999-2004) Linked Mortality Files include: survey respondent eligibility status, mortality status, person months of follow-up from interview date, person months of follow-up from exam date, and 113 grouped recodes of underlying causes of death. In addition, three variables were created to indicate the presence of diabetes, hypertension, or hip fracture in the multiple cause-of-death codes, when these conditions are reported as contributing, rather than underlying causes of death.

Methods

Sample selection

To effectively compare the restricted-use and public-use data sets, we merged the public-use NHANES demographic file for each of the two-year increments from 1999 to 2004 with the accompanying public-use and restricted-use mortality files, respectively, to create the analytic samples. We restricted all analyses to those eligible for mortality follow-up, who were at least 25 years of age at the time of the NHANES interview, who were non-Hispanic white, non-Hispanic black, or Mexican American, with no missing values for education level, and with person months of follow-up greater than zero. Due largely to the sample restrictions of eligible adults 25 years and older, the final sample for the comparative analyses included 12,682 records for all cause mortality and 9,732 for cause-specific mortality. The cause-specific analyses excluded Mexican Americans because of their small numbers of cause-specific deaths.

Outcome measurement

We examined mortality in the public-use and restricted-use NHANES (1999-2004) Linked Mortality files using person-months of follow-up from the NHANES interview until death. Respondents who were not identified as deceased by the end of the follow-up period were assumed to be alive. For the public-use file, we used the person-months of follow-up variable from interview date that is provided on the linked mortality file. More information on the calculation of this variable can be found at http://www.cdc.gov/nchs/data/datalinkage/nh99+_mort_file_layout_public_2010.pdf . For the restricted-use files, person-months of follow-up was calculated using complete information on the month, day, and year of the NHANES interview and the month, day, and year of death or, for respondents assumed alive, until the end of the follow-up period (December 31, 2006).

In addition to all-cause mortality, we examined cause-specific mortality for selected causes with cause-of-death coding based on the International Classification of Diseases, Tenth Revision (ICD-10). We present cause-specific results based upon the Underlying Cause-of-Death Recoded 113 Groups for heart disease (55-68), cancer from all sites (20-44) and lung cancer (27). We conducted analyses for additional causes of death (data not shown), such as ischemic heart disease (59-61) and cerebrovascular diseases (70).

Covariates

In all models, we included a standard set of socio-demographic characteristics, which were reported at the time of the NHANES interview: age in continuous years and top-coded at age 85, sex, race/ethnicity (non-Hispanic black, non-Hispanic white, Mexican American), and educational attainment (less than high school, high school diploma or GED, more than high school).

Data Analysis

We used Cox proportional hazards models to compare the relative risk estimates for the covariates for all-cause as well as cause-specific mortality risk. All relative risk estimates were calculated with the survival procedure in Software for Survey Data Analysis (SUDAAN), version 10.0 to take into account the complex survey design of the NHANES.[1] Due to an insufficient number of deaths for Mexican Americans, the cause-specific mortality analyses are restricted to non-Hispanic whites and non-Hispanic blacks.

Results

Descriptive Results

Table 1 shows the unweighted sample counts (n) and weighted percentage distributions for the covariates used in the analyses. Note that these descriptive statistics for covariates do not differ between the public-use and restricted-use files because the only differences between the two files are associated with the variables taken from the mortality file. Briefly, the distributions of covariates are as expected: the average age of this sample is 49.2 years and two percent of respondents are aged 85 or above. Females outnumber males, and non-Hispanic whites make up 80.7 percent of the sample while non-Hispanic blacks (11.8 percent) and Mexican Americans (7.4 percent) account for considerably smaller proportions. Over 50% of the sample has more than a high school education.

The number and weighted percentage of persons, in our sample, who were identified as dying in each of the two files (n = 1,218; Percent = 5.8) is identical, since for the public-use file the vital status of individuals was not changed as a result of the perturbation

process. The public-use file includes perturbed information for date of death for selected decedents, which affects the calculation of months of follow-up. The mean months of follow-up were approximately 57 months (weighted) for both files. We examined cause-specific mortality percentage distributions and, overall, the distributions are quite similar when comparing the two files. For both files, the percentage of deaths attributed to heart disease and cancer is approximately 27% and 28%, respectively, while lung cancer accounted for approximately 9% of deaths and ischemic heart disease and cerebrovascular diseases each accounted for about 6% (data not shown).

All-Cause Mortality Model Results

Table 2.1 displays results from two Cox proportional hazards models of all-cause mortality: one estimated from the public-use file and one estimated from the restricted-use file. Recall that while vital status was not changed between the two files, there are differences in the duration of follow-up variables due to the perturbation of date of death for selected decedents in the public-use file. The results of both models are consistent. Age, sex, race/ethnicity and education are all related to the risk of adult mortality in the expected directions. For example, men, non-Hispanic blacks and persons with less than a high school education display higher risks of mortality compared to their respective counterpart subgroups. Moreover, the relative risks and 95% confidence intervals are nearly identical for results from the public-use and restricted-use files.

The results of the all-cause Cox proportional hazards models of adult mortality that are estimated separately by sex are shown in Table 2.2. For each sex, results from the public-use and restricted-use files are shown, respectively. The sex-specific models yield consistent results, when the public-use and restricted-use files are compared. Finally, Table 2.3 shows the results of separate proportional hazards models for non-Hispanic whites, non-Hispanic blacks, and Mexican Americans, respectively. Again, the public-use file produces results consistent with the restricted-use file. For example, in both the public-use and restricted use files, males exhibit higher mortality than females in each racial/ethnic group, and this association is not statistically significant for Mexican

Americans in either file. For non-Hispanic whites and blacks, persons with less than a high school education demonstrate higher mortality risks over the follow-up period.

Cause-Specific Mortality Model Results

As previously mentioned, cause-specific results are limited to those identified as non-Hispanic white or non-Hispanic black. We present the results of the Cox proportional hazards models for heart disease, cancer and lung cancer in tables 3.1 - 3.3. Overall, a comparison of the results for the public-use and restricted-use files for all of the specific causes examined yields similar results. For example, cancer mortality risk increases just over seven percent for each additional year of age and males experience about 78 percent higher risk in both the public-use data model and the restricted-use data model. However, for lung cancer mortality (Table 3.3) there are slight differences in the results for sex. Although the relative risk estimates are consistent between the two files, only the public-use file shows statistically significant results. For example, in the public-use file, the relative risk of lung cancer mortality for men compared to women is (RR=1.7, p = 0.05), while in the restricted-use file the RR=1.6, p= 0.06. The public-use file has more deaths attributed to lung cancer among men, resulting in a slightly greater risk among men.

Additional Analyses

We conducted additional analyses on the NHANES (1999-2004) participants that completed an examination in the Mobile Examination Center (MEC). The final sample for the comparative analyses of participants who completed an exam included 11, 346 records for all cause mortality and 8,665 for cause specific mortality due to the additional exclusion of Mexican American from the cause specific analyses. Overall there were 929 deaths among NHANES (1999-2004) participants who completed a MEC exam. Similar to the previous analyses the results from models using the public-use and the restricted-use files yielded similar results with only minor differences when comparing the actual coefficients and standards errors of the models (data not shown).

Discussion

This report describes analyses comparing results obtained from the public-use version and restricted-use version of the NHANES (1999-2004) Linked Mortality Files, with mortality follow-up through 2006. In the public-use version of the data files, a limited amount of information for decedents was perturbed. Also, the public-use files do not include detailed mortality or interview information, whereas the restricted-use version includes complete information on date of death; including month, day and year as well as date of interview and examination.

The comparative analysis finds that the two data files yield very similar descriptive and model results. This is particularly true when examining all-cause mortality. Because the perturbation process in the public-use file did not affect the vital status of any individuals in the file, differences in results between the two files when examining overall (all-cause) mortality arise because the public-use file has perturbed date of death information that is included in the calculated duration of follow-up variables provided on the public-use file. Differences in results for all-cause mortality between the public-use file and restricted-use file were very minor.

There are only 1,218 deaths in the NHANES (1999-2004) Linked Mortality Files and limited follow-up time, which limits analyses of cause-specific mortality. However, for the few causes of death examined, the comparative analysis across the public-use and restricted-use versions yielded only slight differences in model results. The perturbation process in the public-use version will impact the frequency distributions for cause-of-death and should be kept in mind when conducting cause-specific analyses of the public-use files. In addition, analysts should refer to the NHANES Analytic Guidelines and the on-line NHANES Tutorial when assessing the reliability of the estimated standard error for various subgroups of interest within the total NHANES population as well as for further details on other analytic issues. Both of these are available on the NHANES website.

Our findings should provide analysts with the confidence to use these public-use data files providing mortality follow-up for eligible adult NHANES 1999-2004 respondents. However, there are some analytic considerations that should be noted by all potential users. First, we estimated relative risks using SUDAAN 10.0 because it accounts for complex survey designs. Moreover, caution in using the public-use files is urged when examining the mortality patterns of small subgroups of the population, such as numerically small racial/ethnic minority groups, very old individuals, or young adults. This is particularly the case when cause-specific analyses of such numerically small demographic subgroups are performed. Caution is also urged when conducting analyses that allow participants to age into varying age strata over the follow-up period. The availability of more precise and detailed age and follow-up information on the restricted use file could lead to different samples being obtained in the various age strata. Researchers using the public-use data for such analyses are strongly encouraged to confirm their findings with the restricted-use data.

In sum, the 2010 release of a public-use version of the NHANES (1999-2004) Linked Mortality Files provide the public health, social science, demographic, and medical communities with a data set that is easily available, nationally representative, and rich in detail for both mortality covariates and specificity in outcomes. The public-use files are an important resource for researchers and policymakers in further understanding the adult mortality trends and patterns.

References

1. SUDAAN: Software for the Statistical Analysis of Correlated Data, 10.0. RTI International.

Table 1. Baseline sample characteristics, NHANES (1999-2004): n = 12,682

	Unweighted (n)	Weighted percentage or mean
Age in years (mean)	n/a	49.2
Age (grouped)		
25-44	4,713	44.9%
45-64	3,910	35.7
65-84	3,469	17.4
85+	590	2.0
Sex		
Male	6,061	48.0%
Female	6,621	52.0
Race/Ethnicity		
non-Hispanic white	7,148	80.7%
non-Hispanic black	2,584	11.8
Mexican American	2,950	7.4
Education level		
Less than high school	4,191	19.6%
High school/GED	3,005	26.5
More than high school	5,486	53.9

Table 2.1. Relative Risks for all-cause mortality: NHANES (1999-2004) linked mortality files, mortality follow-up through 2006 (n = 12,682)

	Public-use			Restricted-use		
	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI
Age in years	1.090	1.080	1.099	1.090	1.080	1.099
Sex (female)						
Male	1.704	1.472	1.973	1.703	1.471	1.972
Race/ethnicity (NHW)						
NHB	1.309	1.088	1.576	1.308	1.087	1.574
Mexican American	0.812	0.633	1.042	0.813	0.634	1.043
Education (More than high school)						
Less than high school	1.713	1.399	2.098	1.713	1.400	2.096
High school	1.372	1.111	1.694	1.372	1.111	1.695

Notes:
Relative Risks are estimated from a Cox proportional hazards model.
All models adjust for sample weights and the NHANES complex survey design using the SUDAAN software program (10.0).
NHW refers to non-Hispanic white; NHB refers to non-Hispanic black.
Values in parenthesis are reference categories.

Table 2.2. Relative Risks for all-cause mortality by sex: NHANES (1999-2004) linked mortality files, mortality follow-up through 2006 (n = 12,682)

	Men						Women					
	Public-use			Restricted-use			Public-use			Restricted-use		
	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI
Age in years	1.088	1.077	1.099	1.088	1.077	1.099	1.092	1.081	1.104	1.092	1.081	1.104
Race/ethnicity (NHW)												
NHB	1.303	0.956	1.775	1.303	0.957	1.774	1.324	0.999	1.754	1.320	0.997	1.747
Mexican American	0.640	0.413	0.990	0.641	0.414	0.993	1.096	0.739	1.627	1.097	0.741	1.624
Education (more than high school)												
Less than high school	1.807	1.348	2.424	1.808	1.349	2.423	1.617	1.195	2.189	1.615	1.195	2.184
High school	1.255	0.900	1.749	1.256	0.902	1.750	1.491	1.133	1.964	1.490	1.131	1.962

Notes:
Relative Risks are estimated from a Cox proportional hazards model.
All models adjust for sample weights and the NHANES complex survey design using the SUDAAN software program (10.0).
NHW refers to non-Hispanic white; NHB refers to non-Hispanic black.
Values in parenthesis are reference categories.

Table 2.3. Relative Risks for all-cause mortality by race/ethnicity: NHANES linked mortality files, mortality follow-up through 2006 (n = 12,682)

	non-Hispanic whites						non-Hispanic blacks					
	Public-use			Restricted-use			Public-use			Restricted-use		
	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI
Age in years	1.093	1.081	1.106	1.094	1.081	1.106	1.073	1.059	1.086	1.073	1.059	1.087
Sex (female)												
Male	1.745	1.464	2.081	1.744	1.463	2.079	1.696	1.226	2.346	1.696	1.227	2.346
Education (More than high school)												
Less than high school	1.794	1.421	2.265	1.794	1.423	2.263	1.484	1.056	2.084	1.481	1.055	2.078
High school	1.411	1.121	1.777	1.412	1.122	1.777	1.129	0.751	1.697	1.128	0.751	1.695

	Mexican Americans					
	Public-use			Restricted-use		
	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI
Age	1.083	1.064	1.102	1.083	1.064	1.102
Sex (female)						
Male	1.111	0.663	1.861	1.113	0.665	1.862
Education (More than high school)						
Less than high school	0.902	0.419	1.943	0.900	0.418	1.937
High school	0.531	0.222	1.270	0.531	0.222	1.268

Notes:
Relative Risks are estimated from a Cox proportional hazards model.
All models adjust for sample weights and the NHANES complex survey design using the SUDAAN software program (10.0).
Values in parenthesis are reference categories.

Table 3.1. Relative Risks for heart disease mortality: NHANES (1999-2004) linked mortality files, mortality follow-up through 2006, non-Hispanic whites and blacks only (n=9,732)

	Public-use			Restricted-use		
	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI
Age in years	1.109	1.085	1.134	1.107	1.084	1.131
Sex (female)						
Male	1.549	1.240	1.934	1.555	1.245	1.943
Race/ethnicity (NHW)						
NHB	1.112	0.790	1.564	1.177	0.847	1.635
Education (More than high school)						
Less than high school	2.096	1.479	2.970	2.144	1.519	3.026
High school	1.116	0.739	1.685	1.110	0.738	1.669

Notes:
Relative Risks are estimated from a Cox proportional hazards model.
All models use sample weights and take into account the NHANES complex survey design using the SUDAAN software program (10.0.).
NHW refers to non-Hispanic white; NHB refers to non-Hispanic black.
Values in parenthesis are reference categories.

Table 3.2. Relative Risks for all cancer mortality: NHANES (1999-2004) linked mortality files, mortality follow-up through 2006, non-Hispanic whites and blacks only (n=9,732)

	Public-use			Restricted-use		
	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI
Age in years	1.071	1.060	1.083	1.072	1.060	1.083
Sex (female)						
Male	1.787	1.330	2.402	1.776	1.326	2.380
Race/ethnicity (NHW)						
NHB	1.167	0.748	1.821	1.134	0.729	1.764
Education (More than high school)						
Less than high school	1.643	1.099	2.458	1.623	1.081	2.436
High school	1.454	0.998	2.119	1.449	0.993	2.112

Notes:
Relative Risks are estimated from a Cox proportional hazards model.
All models use sample weights and take into account the NHANES complex survey design using the SUDAAN software program (10.0).
NHW refers to non-Hispanic white; NHB refers to non-Hispanic black.
Values in parenthesis are reference categories.

Table 3.3. Relative Risks for lung cancer mortality: NHANES (1999-2004) linked mortality files, mortality follow-up through 2006, non-Hispanic whites and blacks only (n=9,732)

	Public-use			Restricted-use		
	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI	Relative Risk	Lower Bound 95% CI	Upper Bound 95% CI
Age in years	1.064	1.049	1.078	1.064	1.050	1.080
Sex (female)						
Male	1.683	1.011	2.803	1.609	0.972	2.664
Race/ethnicity (NHW)						
NHB	1.367	0.748	2.500	1.308	0.699	2.448
Education (More than high school)						
Less than high school	2.338	1.240	4.405	2.198	1.157	4.173
High school	1.606	0.778	3.311	1.597	0.772	3.303

Notes:
Relative Risks are estimated from a Cox proportional hazards model.
All models use sample weights and take into account the NHANES complex survey design using the SUDAAN software program (10.0).
NHW refers to non-Hispanic white; NHB refers to non-Hispanic black.
Values in parenthesis are reference categories.