American Journal of Epidemiology
Published by the Johns Hopkins Bloomberg School of Public Health 2008.

Vol. 168, No. 3
DOI: 10.1093/aje/kwn123
Advance Access publication May 23, 2008

Practice of Epidemiology

The Public-Use National Health Interview Survey Linked Mortality Files: Methods of Reidentification Risk Avoidance and Comparative Analysis

Kimberly Lochner[1], Robert A. Hummer[2], Stephanie Bartee[1], Gloria Wheatcroft[1], and Christine Cox[1]

[1] Office of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD.
[2] Population Research Center, Department of Sociology, University of Texas, Austin, TX.

Received for publication January 2, 2008; accepted for publication April 9, 2008.

The National Center for Health Statistics (NCHS) conducts mortality follow-up for its major population-based surveys. In 2004, NCHS updated the mortality follow-up for the 1986–2000 National Health Interview Survey (NHIS) years, which because of confidentiality protections was made available only through the NCHS Research Data Center. In 2007, NCHS released a public-use version of the NHIS Linked Mortality Files that includes a limited amount of perturbed information for decedents. The modification of the public-use version included conducting a reidentification risk scenario to determine records at risk for reidentification and then imputing values for either date or cause of death for a select sample of records. To demonstrate the comparability between the public-use and restricted-use versions of the linked mortality files, the authors estimated relative hazards for all-cause and cause-specific mortality risk using a Cox proportional hazards model. The pooled 1986–2000 NHIS Linked Mortality Files contain 1,576,171 records and 120,765 deaths. The sample for the comparative analyses included 897,232 records and 114,264 deaths. The comparative analyses show that the two data files yield very similar results for both all-cause and cause-specific mortality. Analytical considerations when examining cause-specific analyses of numerically small demographic subgroups are addressed.

confidentiality; epidemiologic methods; health surveys; longitudinal studies; mortality

Abbreviations: CI, confidence interval; HR, hazard ratio; ICD, *International Statistical Classification of Diseases, Injuries, and Causes of Death*; NCHS, National Center for Health Statistics; NHIS, National Health Interview Survey.

Federally sponsored health surveys are a critical source of information on public health in the United States. National health surveys provide rich information on risk factors such as smoking, height and weight, health status, and socioeconomic circumstances, but they often lack information on outcomes such as changes in health status over time or mortality risk. There is increasing demand for statistical agencies to incorporate information from additional sources in order to enhance the availability and quality of information on exposures and outcomes and to make such data files publicly available. However, government statistical agencies must balance the desire to provide publicly available, high-quality, and timely data with the maintenance of appropriate safeguards to ensure the confidentiality of individual responses.

The National Center for Health Statistics (NCHS) collects health information from survey participants under assurances of confidentiality as is mandated under section 308(d) of the Health Services Research and Evaluation and Health Statistics Act of 1974 (P.L. 93–353) (1). To ensure that identifiable information is not released, statistical disclosure limitation methods must be applied before the data file can be made publicly available. Standard

Correspondence to Dr. Kimberly Lochner, Office of Analysis and Epidemiology, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782 (e-mail: KLochner@cdc.gov).

TABLE 1. Selected variables on the NHIS* Linked Mortality Files

	Restricted use	Public use
Final mortality status	Yes	Yes
Death date	Yes (month, day, year)	Yes (quarter, year)
Underlying cause of death	Yes	Yes (grouped recode)
Contributing cause of death	Yes	Yes (diabetes, hypertension, hip fracture only)
Age at interview	Yes	Yes, top coded at >85 years
Age at death	Yes	No
Age when last presumed alive	Yes	No
Date of birth	Yes (month, day, year)	Yes (month, year)
Interview date	Yes (month, day, year)	Yes (quarter, year)

* NHIS, National Health Interview Survey.

approaches include limiting the amount of information that is released, for example, creating categorical variables from continuous data, as well as masking techniques that modify the data, for example, noise addition, swapping values, and imputation (2–4).

NCHS periodically conducts mortality follow-up studies through record linkage to the National Death Index (5) for its major population-based surveys. The resultant linked mortality files fill research gaps by providing data resources that contain high-quality sociodemographic, health, and mortality information for nationally representative US samples. However, the linking of records from different data sources can increase the chance that an individual's information may now be at risk of being disclosed. For this reason, the most recent update of the National Health Interview Survey (NHIS) Linked Mortality Files was made available only through the NCHS Research Data Center. Recognizing that this would limit researchers' use of previously highly utilized public-use data files, NCHS recently developed a public-use version of the files. To create the new public-use version, NCHS modified the NHIS Linked Mortality Files in two ways: limiting the amount of mortality information available and masking the data by imputing values for date or cause of death for cases determined to be at increased risk of reidentification.

This article discusses the NHIS Linked Mortality Files, the statistical disclosure avoidance techniques applied prior to releasing the public-use version, and the findings of a comparative analysis between the restricted-use and the public-use files to determine whether analyses using the pubic-use files can reproduce analyses using the restricted-use files. The reidentification risk and data-masking approaches discussed in this paper apply only to those employed for a public-use release of linked mortality files. This paper does not address the various disclosure avoidance techniques used in the release of other NCHS public-use data files.

MATERIALS AND METHODS

NHIS Linked Mortality Files

In 2004, NCHS completed a mortality follow-up study for the 1986–2000 NHIS years. Because of confidentiality protections for the NHIS participants, the NHIS Linked Mortality Files were made available only through the NCHS Research Data Center. We refer to this version of the data as ''restricted use.'' In 2007, NCHS developed a public-use version of the files.

The NHIS is a cross-sectional household interview survey of the civilian noninstitutionalized population of the United States. The NHIS collects data on a broad range of health topics and sociodemographic information. Sampling and interviewing are continuous throughout each year. Descriptions of the NHIS design have been published elsewhere (6, 7). The National Death Index maintains a national file of death record information, beginning with 1979 deaths, compiled from death certificate records collected from state vital statistics offices. More information on the National Death Index can be found at www.cdc.gov/nchs/ndi.htm (accessed December 11, 2007). The NHIS Linked Mortality Files include the NHIS years 1986–2000 and are based upon a probabilistic linkage of eligible adults (18 years or older) to the National Death Index, with mortality follow-up through December 31, 2002. A complete description of the methodology used to link NHIS records to the National Death Index can be found at www.cdc.gov/nchs/data/datalinkage/matching_methodology_nhis_final.pdf (accessed December 11, 2007) (8).

Table 1 lists key mortality variables included on the restricted-use files, as well as the reduced number of variables available on the public-use NHIS Linked Mortality Files. The public-use files replace the exact follow-up time with an approximate follow-up time and limit information on the exact cause of death to a grouped recode, which for some records has been imputed. For example, the restricted-use files include mortality status, exact date of death (month, day, year), and underlying and contributing cause-of-death codes for both the Ninth Revision and the Tenth Revision of the *International Statistical Classification of Diseases, Injuries, and Causes of Death* (ICD-9 and ICD-10, respectively) as reported on the death certificate. The public-use files include mortality status, quarter and year of death, a grouped recode of the underlying cause of death, and a variable indicating whether diabetes, hypertension, or hip fractures were reported in the contributing cause-of-death codes. In addition, the restricted-use files provide more

detail on the NHIS interview date and NHIS participant age than what is available on the NHIS public-use files (e.g., NHIS core or person files). For example, NCHS has made available on the restricted-use NHIS Linked Mortality Files the exact date of the NHIS interview (month, day, year), as well as detailed information on age in years at interview (not top coded), date of birth (month, day, year), and age at death. Such detail on interview date, age, and timing of death facilitates the creation of more detailed and specific follow-up times for mortality analyses. The public NHIS core or person files include information on interview quarter and year, age in years, and month and year of birth.

Reidentification risk simulation and data perturbation plan

We assessed reidentification risk by matching key NHIS public-use sociodemographic variables and mortality information for NHIS decedents to existing publicly available data sources. For each publicly available data source, we identified the unique records and then compared the unique records between files. We considered all NHIS decedent ''unique'' records, which were correctly matched to these public data sources, to be at risk for being reidentified.

After identifying the cases at risk for reidentification, we constructed a perturbation plan to modify the mortality data for those NHIS decedents at risk for being reidentified and to allow for the release of a NHIS linked mortality public-use file. All cases considered potentially ''reidentifiable'' were subject to data perturbation and were randomly assigned to have either the date of death or the underlying cause of death perturbed. To further reduce reidentification risk, we subjected an additional random sample of decedents to perturbation. Information regarding vital status was not perturbed.

Cases requiring the date of death perturbation had either the quarter or year randomly perturbed, and in some cases both fields were perturbed. For those cases requiring underlying cause-of-death perturbation, we implemented a hot-deck method, by replacing the original value with a value from a decedent with similar characteristics, and imputed the 113 grouped underlying cause-of-death recode (9). The perturbed cases are not identified on the public-use files.

Comparative analysis

Once we determined that the modifications made to the public-use version of the NHIS Linked Mortality Files would offer adequate protection of the NHIS participant's identity, we replicated the analyses conducted on the restricted-use files on the public-use files to demonstrate the comparability between the two versions of the linked mortality files. We used Cox proportional hazards models to compare the relative hazards for a standard set of sociodemographic covariates for all-cause and cause-specific mortality risk (10).

Analytical sample. To effectively compare the public-use and restricted-use data sets, we merged the public-use NHIS person-level file for each year 1986–2000 with the accompanying public-use and restricted-use mortality files, respectively, to create the two analytical 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 NHIS interview and non-Hispanic White, non-Hispanic Black, or Hispanic, with a sample weight greater than zero, and with no missing values for educational level, marital status, and cause of death.

Outcome measurement. We examined all-cause and cause-specific mortality in the public-use and restricted-use NHIS Linked Mortality Files using time from NHIS 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 files, duration of follow-up was constructed by use of NHIS interview year and year of death. Respondents who died in the same year as their NHIS interview were assigned ½ year of follow-up time. All other decedents were assigned to have a ½ year of follow-up during the year of their interview, a full year of follow-up for each year after their year of interview until the year prior to their death, and then another ½ year of follow-up during the year of their death. For respondents assumed alive, their follow-up time was calculated by assigning ½ year of follow-up during their NHIS interview year and a full year of follow-up for each year thereafter until the end of 2002. For the restricted-use files, duration of follow-up was calculated by use of complete information on the month, day, and year of the NHIS interview and the month, day, and year of death or, for respondents assumed alive, until the end of the follow-up period, December 31, 2002.

In addition to all-cause mortality, we examined 14 causes of death that are among the 10 leading causes of death in the United States and/or contribute to the most years of potential life lost (11): heart disease, ischemic heart disease, cancer (all sites), lung cancer, colorectal cancer, breast cancer (estimated for women only), prostate cancer, cerebrovascular diseases, diabetes, pneumonia and influenza, chronic liver diseases and cirrhosis, unintentional injuries, suicide, and homicide. Because deaths in this analysis span the transition from the ICD-9 guidelines to the ICD-10, the cause-specific death categories are based upon a recode of the underlying causes of death into 113 selected causes. This list of 113 selected causes was developed for the general analysis of ICD-10 mortality and codes all deaths occurring prior to 1999 coded under ICD-9 guidelines into comparable ICD-10 underlying cause-of-death groups. However, the cause-specific analyses presented in this paper do not control for the transition in coding rules between ICD-9 and ICD-10, because that transition does not affect the comparisons of interest in this paper (12).

Covariates. We included in all models a standard set of sociodemographic characteristics, which were observed at the time of NHIS interview: age in continuous years, sex, race/ethnicity (non-Hispanic Black, non-Hispanic White, Hispanic), educational attainment (less than high school, high school diploma, some college, college degree or more), marital status (widowed, divorced/separated, never married, married), and region of the country (South, Midwest, Northeast, West).

TABLE 2. Mortality characteristics of 897,232 adults aged 25 years or older in the 1986–2000 NHIS* Linked Mortality Files†

	Public use		Restricted use	
	Unweighted no.	Weighted %	Unweighted no.	Weighted %
Mean follow-up period (years)	9.1	8.7	9.1	8.6
Assigned vital status				
Dead	114,264	11.8	114,264	11.8
Alive	782,968	88.2	782,968	88.2
Cause-specific deaths‡				
Diseases of the heart	37,272	32.5	36,689	32.0
Ischemic heart disease	11,434	10.0	11,290	9.8
Cancer, all sites	30,220	26.6	30,197	26.5
Lung cancer	8,838	7.8	8,395	7.4
Colorectal cancer	3,044	2.6	3,094	2.7
Breast cancer§	2,421	4.3	2,372	4.2
Prostate cancer¶	1,762	3.0	1,786	3.0
Cerebrovascular diseases	7,802	6.8	7,855	6.8
Diabetes	3,361	2.9	3,384	2.9
Pneumonia/influenza	3,306	2.9	3,342	2.9
Chronic liver disease/cirrhosis	1,238	1.1	1,268	1.1
Unintentional injuries	3,242	2.9	3,294	2.9
Suicide	1,097	1.0	1,117	1.1
Homicide	410	0.3	425	0.4

* NHIS, National Health Interview Survey.
† Mortality follow-up was through December 31, 2002.
‡ Underlying cause-of-death codes are based upon the *International Statistical Classification of Diseases, Injuries, and Causes of Death,* Tenth Revision, recode into 113 selected causes. Weighted percentages for cause-specific deaths are based upon the sample of decedents.
§ Women only.
¶ Men only.

Data analysis. We used Cox proportional hazards models to compare the relative hazards in the public-use and restricted-use files among covariates for all-cause as well as cause-specific mortality risk. Because of an insufficient number of deaths in certain population subgroups, we restricted the cause-specific mortality analyses to non-Hispanic Whites and non-Hispanic Blacks and collapsed educational attainment into three categories. We calculated all hazard ratios and 95 percent confidence intervals with the survival procedure in Survey Data Analysis (SUDAAN), version 9.0.1, software (13) to take into account the complex survey design of the NHIS.

RESULTS

The public-use and restricted-use pooled 1986–2000 NHIS Linked Mortality Files each contain 1,576,171 records and 120,765 deaths. The final sample for the comparative analyses included 897,232 records and 114,264 deaths. The distribution for the covariates included in the models is the same for both sets of analyses using the public-use and restricted-use linked mortality files. The average age of this sample is 47.9 years, and fewer than 2 percent of respondents are aged 85 years or more. Females outnumber males (from 52.6 to 47.4 percent, respectively), and non-Hispanic Whites make up just over 80 percent of the sample while non-Hispanic Blacks (10.9 percent) and Hispanics (8.2 percent) account for considerably smaller proportions. A vast majority of the sample is married at the time of NHIS interview (69.0 percent), and the modal educational category is a high school degree or general equivalency diploma (a certificate representing the equivalent of a high-school diploma) (36.0 percent), with 20.4 percent having less than a high school education, 21.4 percent some college, and 22.1 percent at least a college degree. Over 35 percent of the sample resides in the South, while nearly 25 percent resides in the Midwest and 19 percent in the West.

Table 2 shows the comparative descriptive statistics for mortality outcome variables among the public-use and restricted-use files, respectively. The total number and percentage of persons who were identified in each of the two files as having died ($n = 114{,}264$; 11.8 percent) are identical. As mentioned above, this illustrates that the vital status

TABLE 3. All-cause mortality by sociodemographic characteristics for adults aged 25 years or older in the 1986–2000 NHIS* Linked Mortality Files†

	Public use		Restricted use	
	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval
Age in years	1.09	1.09, 1.09	1.09	1.09, 1.09
Sex				
Women	1.00		1.00	
Men	1.69	1.67, 1.71	1.69	1.67, 1.71
Race/ethnicity				
Non-Hispanic White	1.00		1.00	
Non-Hispanic Black	1.15	1.13, 1.18	1.15	1.13, 1.18
Hispanic	0.89	0.86, 0.92	0.89	0.87, 0.92
Marital status				
Married	1.00		1.00	
Widowed	1.23	1.21, 1.25	1.23	1.21, 1.25
Divorced/separated	1.40	1.36, 1.43	1.40	1.36, 1.43
Never married	1.48	1.44, 1.53	1.48	1.44, 1.53
Educational level				
Less than high school	1.68	1.64, 1.72	1.68	1.64, 1.72
High school/GED*	1.41	1.37, 1.44	1.41	1.37, 1.44
Some college	1.28	1.25, 1.31	1.28	1.25, 1.31
College degree or more	1.00		1.00	
Region				
Northeast	0.97	0.95, 1.00	0.98	0.95, 1.00
Midwest	0.99	0.96, 1.01	0.99	0.96, 1.01
South	1.05	1.03, 1.08	1.05	1.03, 1.08
West	1.00		1.00	

* NHIS, National Health Interview Survey ($n = 897{,}232$; deaths = 114,264); GED, general equivalency diploma (a certificate representing the equivalent of a high-school diploma).

† Mortality follow-up was through December 31, 2002.

‡ Estimated from a Cox proportional hazards model.

of individuals was not changed for anyone as a result of the perturbation process for the public-use file. However, there are some modest differences in the cause-of-death distributions when comparing the public-use and restricted-use files. For example, the number of deaths attributed to some of the more common causes of death, such as heart disease ($n = 37{,}272$) and lung cancer ($n = 8{,}838$), in the public-use file is greater than the number of deaths attributed to those causes in the restricted-use file ($n = 36{,}689$ and $n = 8{,}395$, respectively). Similarly, there are modest differences for some of the less common causes, such as unintentional and intentional injuries.

Table 3 presents 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. The results of both models are consistent with expectations, given the results from similar models that used an earlier version of this data set (14). Age is very strongly and positively related to the risk of adult mortality, and men, non-Hispanic Blacks, persons with less than a high school education, never married individuals, and those living in the South display higher risks of mortality compared with their respective counterpart subgroups. Moreover, hazard ratios and 95 percent confidence intervals are essentially identical when comparing the results from the public-use and restricted-use files. For example, in the restricted-use file, mortality from all causes was higher for men compared with women (hazard ratio (HR) = 1.69, 95 percent confidence interval (CI): 1.67, 1.71), and this result was replicated in the public-use data. For the other covariates, similar results were obtained using the two data files.

Models estimated separately for men and women are shown in table 4. The sex-specific models yield results that are consistent with previous research, and again the public-use and restricted-use files obtain nearly identical hazard ratios and 95 percent confidence intervals. For example, controlling for age and other sociodemographic factors, non-Hispanic Black men have an increased risk of mortality compared with non-Hispanic White men. The hazard ratio for the race covariate estimated from the restricted-use data was 1.16 (95 percent CI: 1.12, 1.20), which was replicated in the public-use data. The findings were similar for the models restricted to women. We also estimated separate proportional hazards models for non-Hispanic Whites, non-Hispanic Blacks, and Hispanics (table 5). Again, results are similar from the public-use and restricted-use files for each of the three racial/ethnic groups. For each group, covariates exhibit associations with all-cause mortality that are consistent with what one would expect from the literature (14). For example, males exhibit 60–70 percent higher mortality than do females in each racial/ethnic group, and persons with less than a high school education demonstrate higher mortality risks over the follow-up period in each racial/ethnic group compared with persons in the more highly educated groups. Given differences in the way that the duration of follow-up variable was calculated for the restricted-use and public-use versions of the NHIS Linked Mortality Files, the slight differences in model results for all-cause mortality can be accounted for by differences in the duration of follow-up variables.

Each cause-specific table compares the model results from the public-use version and the restricted-use version of the NHIS Linked Mortality Files. Because of space constraints, we present results for two of the 14 cause-specific analyses here, with the remaining 12 cause-specific analyses available in Web Appendix tables 1–12. (These supplementary tables are posted on the *Journal*'s website (http://aje.oxfordjournals.org/).) A comparison of the results for the public-use and restricted-use files for each of the 14 causes yields no substantive differences in conclusions and hazard ratios and confidence intervals that are very similar. However, there tends to be less agreement in the estimates for the less common causes of death when comparing results from the public-use data and restricted-use data models.

Table 6 presents cause-specific results for all-cancer mortality. The mortality risk increases just over 7 percent for each additional year of age in both the public-use data

TABLE 4. All-cause mortality by sociodemographic characteristics for men and women aged 25 years or older in the 1986–2000 NHIS* Linked Mortality Files†

	Men (deaths = 57,218)				Women (deaths = 57,046)			
	Public use		Restricted use		Public use		Restricted use	
	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval
Age in years	1.09	1.09, 1.09	1.09	1.09, 1.09	1.09	1.09, 1.09	1.09	1.09, 1.09
Race/ethnicity								
Non-Hispanic White	1.00		1.00		1.00		1.00	
Non-Hispanic Black	1.16	1.12, 1.20	1.16	1.12, 1.20	1.15	1.11, 1.18	1.15	1.11, 1.18
Hispanic	0.90	0.86, 0.95	0.90	0.86, 0.95	0.88	0.85, 0.92	0.89	0.85, 0.92
Marital status								
Married	1.00		1.00		1.00		1.00	
Widowed	1.18	1.14, 1.22	1.18	1.14, 1.22	1.25	1.22, 1.28	1.25	1.22, 1.28
Divorced/separated	1.45	1.40, 1.50	1.45	1.40, 1.50	1.35	1.30, 1.39	1.35	1.30, 1.39
Never married	1.57	1.51, 1.63	1.57	1.51, 1.64	1.38	1.33, 1.44	1.38	1.33, 1.44
Educational level								
Less than high school	1.76	1.71, 1.80	1.76	1.71, 1.80	1.55	1.50, 1.60	1.55	1.50, 1.60
High school/GED*	1.46	1.42, 1.51	1.46	1.42, 1.51	1.30	1.26, 1.35	1.30	1.26, 1.35
Some college	1.34	1.29, 1.38	1.34	1.29, 1.38	1.19	1.14, 1.23	1.19	1.14, 1.23
College degree or more	1.00		1.00		1.00		1.00	
Region								
Northeast	1.01	0.98, 1.04	1.01	0.98, 1.04	0.94	0.91, 0.98	0.94	0.91, 0.98
Midwest	0.99	0.96, 1.03	0.99	0.96, 1.03	0.98	0.94, 1.01	0.98	0.95, 1.01
South	1.10	1.06, 1.13	1.10	1.06, 1.13	1.01	0.97, 1.04	1.01	0.97, 1.04
West	1.00		1.00		1.00		1.00	

* NHIS, National Health Interview Survey (n = 897,232); GED, general equivalency diploma (a certificate representing the equivalent of a high-school diploma).

† Mortality follow-up was through December 31, 2002.

‡ Estimated from a Cox proportional hazards model.

model and the restricted-use data model. Hazard ratios and 95 percent confidence intervals vary slightly for the other covariates. Men experience higher cancer mortality risk than do women over the course of the follow-up period (public-use data HR = 1.57, 95 percent CI: 1.53, 1.62; restricted-use data HR = 1.59, 95 percent CI: 1.55, 1.63). Educational differences in overall cancer mortality risk favor those with more than a high school education in both the public-use and restricted-use data sets. In a comparison of those who attained more than a high school education with those who had less than a high school education, the risk estimates were essentially the same from restricted-use data (HR = 1.37, 95 percent CI: 1.32, 1.42) and public-use data (HR = 1.36, 95 percent CI: 1.31, 1.41).

Mortality from homicide, an example of an underlying cause that is far less common than all-cancer mortality, is shown in table 7. In our analytical samples, homicide accounts for only 0.3 percent of deaths. Both the public-use and restricted-use files show similar results, but there is more variation in point estimates and their associated standard errors than for all-cause or the more common cause-specific mortality outcomes. In the restricted-use files, homicide mortality is 2.7 times more likely for men than women, 3.9 times more likely for non-Hispanic Blacks compared with non-Hispanic Whites, and 2.3 times more likely for those with less than a high school education compared with those with more than a high school degree. The hazard ratios in the public-use files are 2.7, 4.0, and 2.4 for men, non-Hispanic Blacks, and those with less than a high school education, respectively.

DISCUSSION

The availability of nationally representative longitudinal mortality follow-up data that have high-quality information on risk factors and sociodemographic characteristics is critical for epidemiologic research. The updated mortality follow-up for the NHIS creates a prospective component to these cross-sectional data, and the 2007 public-use release of the NHIS Linked Mortality Files expands access to this rich data source. The modifications made to the public-use

TABLE 5. All-cause mortality by sociodemographic characteristics for non-Hispanic White, non-Hispanic Black, and Hispanic adults aged 25 years or older in the 1986–2000 NHIS* Linked Mortality Files†

	Non-Hispanic Whites (deaths = 91,426)				Non-Hispanic Blacks (deaths = 16,575)				Hispanics (deaths = 6,263)			
	Public use		Restricted use		Public use		Restricted use		Public use		Restricted use	
	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval
Age in years	1.09	1.09, 1.09	1.09	1.09, 1.09	1.07	1.07, 1.07	1.07	1.07, 1.07	1.07	1.07, 1.07	1.07	1.07, 1.07
Sex												
Women	1.00		1.00		1.00		1.00		1.00		1.00	
Men	1.70	1.68, 1.73	1.70	1.68, 1.73	1.66	1.60, 1.73	1.66	1.60, 1.73	1.61	1.52, 1.71	1.61	1.52, 1.71
Marital status												
Married	1.00		1.00		1.00		1.00		1.00		1.00	
Widowed	1.22	1.20, 1.24	1.22	1.20, 1.25	1.21	1.15, 1.28	1.21	1.15, 1.28	1.22	1.12, 1.33	1.22	1.12, 1.32
Divorced/separated	1.46	1.42, 1.50	1.46	1.42, 1.50	1.26	1.19, 1.33	1.26	1.19, 1.33	1.15	1.05, 1.27	1.15	1.04, 1.27
Never married	1.44	1.39, 1.49	1.44	1.39, 1.49	1.50	1.41, 1.59	1.50	1.41, 1.59	1.28	1.13, 1.44	1.28	1.13, 1.44
Educational level												
Less than high school	1.67	1.63, 1.71	1.67	1.63, 1.71	1.68	1.56, 1.81	1.67	1.55, 1.80	1.65	1.48, 1.84	1.65	1.48, 1.85
High school/GED*	1.39	1.36, 1.43	1.39	1.36, 1.43	1.42	1.31, 1.53	1.42	1.31, 1.53	1.31	1.16, 1.47	1.31	1.16, 1.48
Some college	1.27	1.23, 1.30	1.27	1.23, 1.30	1.27	1.17, 1.39	1.27	1.17, 1.39	1.24	1.10, 1.41	1.24	1.09, 1.41
College degree or more	1.00		1.00		1.00		1.00		1.00		1.00	
Region												
Northeast	0.98	0.95, 1.00	0.98	0.95, 1.01	0.99	0.90, 1.09	1.00	0.91, 1.10	0.96	0.88, 1.04	0.96	0.88, 1.04
Midwest	0.99	0.96, 1.01	0.99	0.96, 1.01	1.06	0.98, 1.15	1.07	0.98, 1.16	0.90	0.81, 1.00	0.90	0.81, 1.01
South	1.05	1.02, 1.08	1.05	1.02, 1.08	1.09	1.01, 1.18	1.09	1.01, 1.18	1.11	1.03, 1.18	1.11	1.04, 1.18
West	1.00		1.00		1.00		1.00		1.00		1.00	

* NHIS, National Health Interview Survey (n = 897,232); GED, general equivalency diploma (a certificate representing the equivalent of a high-school diploma).
† Mortality follow-up was through December 31, 2002.
‡ Estimated from a Cox proportional hazards model.

TABLE 6. Mortality from cancer by sociodemographic characteristics for non-Hispanic White and non-Hispanic Black adults aged 25 years or older in the 1986–2000 NHIS* Linked Mortality Files†

	Public use (deaths = 28,709)		Restricted use (deaths = 28,679)	
	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval
Age in years	1.07	1.07, 1.07	1.08	1.07, 1.08
Sex				
Women	1.00		1.00	
Men	1.57	1.53, 1.62	1.59	1.55, 1.63
Race/ethnicity				
Non-Hispanic White	1.00		1.00	
Non-Hispanic Black	1.17	1.13, 1.22	1.18	1.13, 1.22
Marital status				
Married	1.00		1.00	
Widowed	0.86	0.83, 0.90	0.87	0.84, 0.91
Divorced/separated	1.29	1.24, 1.35	1.29	1.23, 1.35
Never married	0.92	0.87, 0.98	0.94	0.89, 0.99
Educational level				
Less than high school	1.36	1.31, 1.41	1.37	1.32, 1.42
High school/GED*	1.24	1.20, 1.29	1.24	1.20, 1.29
More than high school	1.00		1.00	
Region				
Northeast	1.08	1.04, 1.13	1.08	1.04, 1.13
Midwest	1.04	1.00, 1.09	1.05	1.00, 1.09
South	1.09	1.05, 1.14	1.09	1.05, 1.14
West	1.00		1.00	

* NHIS, National Health Interview Survey (n = 802,387); GED, general equivalency diploma (a certificate representing the equivalent of a high-school diploma).
† Mortality follow-up was through December 31, 2002.
‡ Estimated from a Cox proportional hazards model.

TABLE 7. Mortality from homicide by sociodemographic characteristics for non-Hispanic White and non-Hispanic Black adults aged 25 years or older in the 1986–2000 NHIS* Linked Mortality Files†

	Public use (deaths = 320)		Restricted use (deaths = 331)	
	Hazard ratio‡	95% confidence interval	Hazard ratio‡	95% confidence interval
Age in years	0.98	0.97, 0.99	0.99	0.98, 1.00
Sex				
Women	1.00		1.00	
Men	2.70	2.13, 3.42	2.70	2.14, 3.40
Race/ethnicity				
Non-Hispanic White	1.00		1.00	
Non-Hispanic Black	4.01	3.01, 5.33	3.90	2.92, 5.20
Marital status				
Married	1.00		1.00	
Widowed	1.26	0.70, 2.29	1.50	0.88, 2.57
Divorced/separated	1.60	1.15, 2.21	1.62	1.15, 2.27
Never married	1.88	1.32, 2.68	1.89	1.33, 2.69
Educational level				
Less than high school	2.44	1.71, 3.50	2.31	1.63, 3.26
High school/GED*	1.65	1.22, 2.23	1.55	1.16, 2.07
More than high school	1.00		1.00	
Region				
Northeast	0.46	0.30, 0.70	0.46	0.30, 0.71
Midwest	0.82	0.55, 1.20	0.80	0.54, 1.18
South	1.07	0.76, 1.52	1.03	0.72, 1.47
West	1.00		1.00	

* NHIS, National Health Interview Survey (n = 802,387); GED, general equivalency diploma (a certificate representing the equivalent of a high-school diploma).
† Mortality follow-up was through December 31, 2002.
‡ Estimated from a Cox proportional hazards model.

file to allow its release include both limiting mortality information compared with that in the restricted-use file and perturbing data for a small, select number of records.

With the release of public-use linked mortality files, NCHS has intended to balance the data needs of the research community while protecting the confidentiality of survey participants. Thus, this article makes a pragmatic, but unique, contribution to both the providers and users of data by discussing the issues related to confidentiality protection, demonstrating that the masking procedures implemented to reduce reidentification risk resulted in a public-use file with many of the variables that data users will need, for example, information for the calculation of follow-up time and cause-of-death information and providing a comparative analysis of the restricted-use and public-use versions of the data.

The comparative analysis shows that the two data files yield similar descriptive and model results. This is particularly true when examining all-cause mortality. Because the perturbation process in the public-use files did not affect the vital status of any individuals in the file, the only differences in results between the two files when examining overall mortality arose because of less specificity in the time-to-death information, that is, follow-up time to the nearest year, in the public-use files. In the end, the differences that resulted from the comparisons of all-cause mortality between the public-use files and restricted-use files were minor, which is not surprising if we assume that the risk of mortality remains constant over each of the 1-year follow-up periods. Thus, the specification of deaths within 1-year follow-up intervals resulted in little, if any, lost information regarding basic mortality differences. The comparative analysis of cause-specific mortality across the public-use and restricted-use versions of the NHIS Linked Mortality Files also yielded only slight differences in model results, even for causes of death such as chronic liver disease and cirrhosis, homicide, unintentional injuries, and suicide, each of which represents fewer than 3

percent of all US adult deaths. The frequency distributions that were shown for cause of death for the public-use and restricted-use versions of the NHIS Linked Mortality Files demonstrated that the perturbation process in the public-use version had a minor impact on the number of persons identified as having died from each cause, as well as the overall distribution of deaths. This should be kept in mind when conducting cause-specific analyses of the public-use files. Nevertheless, the relative hazards and 95 percent confidence intervals in the cause-specific models that we have estimated demonstrate that such differences in the identification of causes of death for some cases result in only very slight changes in the comparative results. Overall, we did not reach any different conclusions when using the public-use file in comparison with the restricted-use file.

However, there are some analytical considerations that should be noted by all potential users. For the public-use files, length of follow-up time was calculated for this article using only the year of interview and the year of death or, for those assumed alive, the end of 2002 as the endpoint of follow-up. Using only year information resulted in 32 distinct follow-up times. The resulting tied failure times could cause bias in model estimates if not handled correctly. We used the statistical software package SUDAAN, version 9.0.1, because it estimates Cox proportional hazard models for sample surveys and uses Efron's likelihood for tied failure times as the default (15). We conducted additional analyses using the available information on quarter of interview and quarter of death to calculate length of follow-up time in the public-use file, which yielded 256 distinct failure times, and found no substantive or significant differences in the results compared with those presented in this article. Moreover, caution in using the public-use files is urged for researchers requiring more detail on timing of death or age or 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.

In sum, our findings should provide analysts with the information needed to use data from the public-use NHIS Linked Mortality Files that provide mortality follow-up for eligible NHIS respondents. The new public-use version of the NHIS Linked Mortality Files provides the public health, social science, demographic, and medical communities with a data set that is readily available, very large, nationally representative, and rich in detail for both baseline 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 that characterize our diverse society. More information on NCHS's data linkage activities and access to the public-use linked mortality data files for the National Health Interview Survey, the Third National Health and Nutrition Examination Survey, and the Second Longitudinal Study of Aging can be found at the NCHS data linkage website: www.cdc.gov/nchs/r&d/nchs_datalinkage/data_linkage_activities.htm.

ACKNOWLEDGMENTS

Participation by R. A. H. in the preparation of this article was supported by the National Institute of Child Health and Human Development (grant 1 R01 053696).

Conflict of interest: none declared.

REFERENCES

1. National Center for Health Statistics staff manual on confidentiality. Hyattsville, MD: National Center for Health Statistics, 1997.
2. Fienberg SE, Willenborg LCRJ. Introduction to the special issue: disclosure limitation methods for protecting the confidentiality of statistical data. J Off Stat 1998;14:337–45.
3. Domingo-Ferrer J, Torra V. Disclosure protection methods and information loss for microdata. In: Doyle P, Lane J, Theeuwes J, et al, eds. Confidentiality, disclosure and data access. Amsterdam, Netherlands: North-Holland, 2001:91–110.
4. Winkler WE. Masking and re-identification methods for public-use microdata: overview and research problems. In: Domingo-Ferrer J, Torra V, eds. Privacy in statistical databases. Berlin, Germany: Springer, 2004:231–46.
5. The National Death Index. Hyattsville, MD: Division of Vital Statistics, National Center for Health Statistics, 2007. (http://www.cdc.gov/nchs/ndi.htm).
6. Massey JT, Moore TF, Parsons VL, et al. Design and estimation for the National Health Interview Survey, 1985–94. Hyattsville, MD: National Center for Health Statistics, 1989. (Vital and health statistics, series 2: data evaluation and methods research, no. 110).
7. Botman SL, Moore TF, Moriarity CL, et al. Design and estimation for the National Health Interview Survey, 1995–2004. Hyattsville, MD: National Center for Health Statistics, 2000. (Vital and health statistics, series 2: data evaluation and methods research, no. 130).
8. The 1986–2000 National Health Interview Survey Linked Mortality Files: matching methodology. Hyattsville, MD: Office of Analysis and Epidemiology, National Center for Health Statistics, 2005. (http://www.cdc.gov/nchs/data/datalinkage/matching_methodology_nhis_final.pdf).
9. Lessler JT, Kalsbeek WD. Nonsampling errors in surveys. New York, NY: John Wiley & Sons, Inc, 1992.
10. Cox DR. Regression models and life tables (with discussion). J R Stat Soc (B) 1972;34:187–220.
11. Health, United States, 2006. Hyattsville, MD: National Center for Health Statistics, 2006.
12. Anderson RN, Minino AM, Hoyert DL, et al. Comparability of cause of death between ICD-9 and ICD-10: preliminary estimates. Hyattsville, MD: National Center for Health Statistics, 2001. (http://www.cdc.gov/nchs/data/nvsr/nvsr49/nvsr49_02.pdf).
13. SUDAAN: software for the statistical analysis of correlated data, 9.01. Research Triangle Park, NC: RTI International, 2005.
14. Rogers R, Hummer RA, Nam CB. Living and dying in the U.S.A. San Diego, CA: Academic Press, 2000.
15. Hertz-Picciotto I, Rockhill B. Validity and efficiency of approximation methods for tied survival times in Cox regression. Biometrics 1997;53:1151–6.