Comparative Analysis of the 2004 NNHS Public-Use and Restricted-Use Linked Mortality File: 2010 Data Release

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

Introduction

In 2009, NCHS completed a mortality linkage for the 2004 National Nursing Home Survey (NNHS) [1] residents, with mortality ascertained through December 31, 2006. Due to requirements to protect the confidentiality of the NNHS participants, a restricted-use version of the 2004 NNHS Linked Mortality File was made available only through the NCHS Research Data Center (RDC). To complement the restricted-use file and increase data access, NCHS has developed a plan to allow for a public-use release of linked mortality data.

In 2010, NCHS released a public-use version of the 2004 NNHS Linked Mortality File. 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 2004 NNHS Linked Mortality Files. We used Cox proportional hazards models to compare the relative risk estimates for all-cause mortality at three different follow-up time periods. NCHS is conducting this comparative analysis to demonstrate the comparability between the two versions of linked mortality files.

[1] The 2004 National Nursing Home Survey (NNHS) is one in a continuing series of nationally representative sample surveys of United States nursing homes, their services, their staff, and their residents. The NNHS was first conducted in 1973-1974 and repeated in 1977, 1985, 1995, 1997, 1999, and most recently in 2004.

Description of 2004 NNHS Linked Mortality Data Resources

Mortality status for eligible 2004 NNHS 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/matching_methodology_04nnhs_final.pdf .

The restricted-use file includes detailed mortality information for all eligible survey participants. The restricted-use file includes 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 the 2004 NNHS resident admission and interview date (month, day, and year).

Due to confidentiality protections, the public-use file includes only eligible survey participants 18 years and older and a limited set of mortality variables. In addition, the public-use version was 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 2004 NNHS Linked Mortality File includes: survey respondent eligibility status, mortality status, person days of follow-up from admission 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 2004 NNHS resident file with the accompanying public-use and restricted-use mortality file, respectively, to create the analytic sample. We restricted all analyses to those eligible for mortality follow-up, who were at least 25 years of age at the time of

admission, who were non-Hispanic white, non-Hispanic black, and with person days of follow-up greater than zero. In addition, to minimize bias due to those nursing home residents surviving to the interview date, we limited the analytic sample to those residents who were admitted in 2004. Since date of admission is not available on the public use file, residents were selected if their age at interview was equal to their age at admission, as a proxy for admission in 2004. The final sample for the comparative analyses included 3,571 records.

Outcome measurement

We examined mortality in the public-use and restricted-use 2004 NNHS Linked Mortality files using person days of follow-up from the admission date until death. We used three different follow-up time periods to compare the two mortality files. We assessed the relative risks for residents who died within 6 months post admission, within one year post admission and until the end of follow-up, December 31, 2006.

For the public-use file, we used the person days of follow-up variable from the admission 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/nnhs04_mort_file_layout.pdf. For the restricted-use file, person days of follow-up was calculated using complete information on the month, day, and year of the admission date and the month, day, and year of the end of the follow-up period. For the six-month post-admission analysis, residents who were not identified as deceased by six months post admission were assumed to be alive. For the one-year post admission analysis, residents who were not identified as deceased by one-year post admission were assumed to be alive. For the analysis that spanned the entire follow-up period, through December 31, 2006, residents who were not identified as deceased by the end of the follow-up period were assumed to be alive.

Data Analysis

We used Cox proportional hazards models to examine the relative risk of age at admission (in continuous years), sex, and race/ethnicity (non-Hispanic white, non-

Hispanic black) on all-cause mortality for the three different follow-up time periods. 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 NNHS.[1]

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 average age of this sample admitted in 2004 is 79 years. Females outnumber males, and non-Hispanic whites make up 87 percent of the sample.

The public-use file includes perturbed information for date of death for selected decedents, which affects the calculation of days of follow-up. Although this leads to slight differences between the public-use and restricted-use files in the number of decedents who died within six months and one-year from admission in 2004, the percentages dying within these two follow-up time periods is similar, about 16% in the six month post admission analysis and about 28% in the one year post admission analysis. For the analysis that spanned the entire follow-up period of the linked mortality file (December 31, 2006), the number of survey respondents dying in the two files remains identical (n=1,759) and the mean days of follow-up were approximately 629.5 days (weighted) for both files.

Although specific causes of death were not analyzed in this report, the cause-specific mortality percentage distributions are quite similar when comparing the two files through the maximum follow-up period (2006). For both files, the percentage of deaths attributed to heart disease and cancer is approximately 27% and 13%, respectively, while lung cancer accounted for approximately 3% of deaths, ischemic heart disease and

Alzheimer's each accounted for about 6%, and cerebrovascular diseases accounted for about 8% (data not shown).

All-Cause Mortality Model Results

Table 2 displays results from the Cox proportional hazards models for the three different follow-up time periods for 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. For all three follow-up time periods, the model results between the public-use and restricted-use files are consistent, with relative risks and 95% confidence intervals that are nearly identical. For example, the relative risk and 95% confidence interval of dying within 1 year post admission for men is 1.33 (1.12, 1.58) compared to women, using either the public use or restricted use dataset.

Discussion

This report describes analyses comparing results obtained from the public-use version and restricted-use version of the 2004 NNHS Linked Mortality File. In the public-use version of the data file, a limited amount of information for decedents was perturbed. Further, there is less detail on mortality information in the public-use version, compared to the restricted-use file, where no information has been perturbed and there is complete information on date of death; including month, day and year as well as date of admission.

The comparative analyses examined all-cause mortality at three different follow-up time periods in order to illustrate differences in results between the two files that may 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. For all three follow-up time periods, 6 months post admission, one year post admission, and through the end of the study follow-up period, the comparative analyses found that the two data files yield similar results.

However, there are some analytic considerations that should be noted by all potential users. First, analysts should refer to the Sample Design, Data Collection and Estimation Procedures documentation for the 2004 National Nursing Home Survey (http://www.cdc.gov/nchs/data/nnhsd/2004NNHS_DesignCollectionEstimates_072706tags.pdf). In particular, sample sizes should be assessed when working with the 2004 NNHS Linked Mortality file and caution in using the public-use file is urged when examining the mortality patterns of small subgroups of the population, such as numerically small racial/ethnic minority groups and cause-specific deaths or when conducting analyses that allow participants to age into varying age strata over the follow-up period.

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 file. For the nursing home population, cause-specific information from multiple cause-of-death codes may be needed, but is not available on the public-use file, with the exception of flags indicating the presence of diabetes, hypertension, or hip fracture reported as a contributing rather than an underlying cause-of-death. Multiple cause of death information is available on the restricted-use file.

Residents who were admitted prior to 2004 may introduce survivor bias into a survival analysis because only living residents were eligible to be selected for the 2004 NNHS. In order to minimize bias due to those nursing home residents surviving to the interview date, we limited the analytic sample to those residents who were admitted in 2004, but analysts should consider appropriate methodologies to account for such potential bias[2]. Finally, the complex survey design of the 2004 NNHS should be taken into account. We estimated relative risks using SUDAAN 10.0.

The 2010 release of a public-use version of the 2004 NNHS Linked Mortality File is an important resource for researchers and policymakers in further understanding the adult mortality trends and patterns in a nursing home population and our findings should

provide analysts with the confidence to use the public-use data file providing mortality follow-up for eligible 2004 NNHS residents.

References

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

2. Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: Wiley.

Table 1. Baseline sample characteristics for survey respondents admitted to a nursing home in 2004, NNHS 2004: n = 3,571

	Unweighted (n)	Weighted percentage or mean
Age in years, mean	n/a	79.3
Age at admission (grouped)		
Under 65	454	12.4%
65-74	471	12.8
75-84	1,269	35.5
85+	1,377	39.3
Sex		
Male	1,223	34.2%
Female	2,348	65.8
Race/Ethnicity		
non-Hispanic white	3,165	86.7%
non-Hispanic black	406	13.3

Table 2. Relative Risks for all-cause mortality for 2004 NNHS residents admitted in 2004: (n =3,571)

	Six Month Post-admission						One Year Post-admission					
	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	1.02	1.01	1.03	1.02	1.01	1.03	1.02	1.01	1.03	1.02	1.01	1.03
Sex (female)												
Male	1.34	1.05	1.70	1.33	1.05	1.69	1.33	1.12	1.58	1.33	1.12	1.58
Race/ethnicity (NHW)												
NHB	0.91	0.65	1.26	0.93	0.67	1.29	0.83	0.65	1.06	0.86	0.67	1.09

	Follow-up through 12/31/2006					
	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	1.03	1.02	1.03	1.03	1.02	1.03
Sex (female)						
Male	1.31	1.16	1.48	1.31	1.16	1.48
Race/ethnicity (NHW)						
NHB	0.95	0.79	1.14	0.95	0.80	1.14

Notes:
Relative Risks are estimated from a Cox proportional hazards model.
All models adjust for sample weights and the NNHS 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.