Guide to Weighting and Variance Estimation in the 1986-2000 Public Release NHIS Linked Mortality Files

I. Introduction

NCHS has conducted a mortality follow-up for 15 years of the National Health Interview Survey (NHIS) by linking eligible adult NHIS participants to the National Death Index (NDI). This linkage of the NHIS survey participants with the NDI provides the opportunity to conduct studies designed to investigate the association of a variety of health factors with mortality, using the richness of the NHIS core and supplement questionnaires. Basic documentation for these data files can found at the NCHS Data Linkage Activities website.

Before attempting any analytical work, the data user should thoroughly review the documentation for the *NHIS Linked Mortality Files*, in particular the Matching methodology report and the Analytic guidelines. Attention should be directed to the criteria for participant eligibility and the criteria for assigning vital status. Furthermore, the reader is directed to Korn and Graubard (1999) for general discussion of the types of analytical issues specific to complex surveys like the NHIS.

The *NHIS Linked Mortality Files* are available for each NHIS survey year from 1986 through 2000, which covers survey data from two different NHIS sample design periods, years 1986 to 1994, NCHS (1989), and years 1995 to 2000, NCHS (2000). The *eligible adult* participants from these surveys have been linked to the NDI for mortality follow-up through December 31, 2002. Those adults classified as eligible have their original NHIS sampling weight adjusted to compensate for those ineligible.

This paper provides general guidelines for standard survey design-based analyses for the *NHIS Linked Mortality Files*. Section II discusses the new sample weights created to account for ineligible status due to insufficient identifying information for linkage. Section III discusses new design variables created to provide appropriate variance estimates when using any single or combined sets of NHIS years from 1986 to 2000.

II. Weighting Issues and Adjustments

For the 1986 - 1996 NHIS data, the unweighted percentage of adults ineligible for matching due to insufficient identification data was in the range of 1.5% to 3% per sample year. Starting in 1997, due to changes in the NHIS questionnaire design and administration, the percentage of NHIS adults who were ineligible for NDI matching increased to approximately 10%. Ignoring these ineligibles in the *NHIS Linked Mortality Files* may lead to biased mortality estimates, certainly for estimates of totals and quite likely for many statistics based on ratios, e.g., mortality rates. A complete discussion of weighting adjustments is beyond the scope of this paper. We direct the reader to Korn

and Graubard (1999) section (4.2) for more information on methods of weighting adjustment.

For the *NHIS Linked Mortality Files* the eligible sample from the person file was treated as a subsample from the original NHIS samples and the original postratification adjustment method was used to inflate the sampling weights. The tacit assumption is that the adjustment cells used will mitigate estimation bias due to using only the eligible sample. The reader is referred to NCHS (1989) (chapter 4) and NCHS (2000) (chapter 3) for a discussion of poststratification methods specific to the 1985-1994 and 1995-2004 NHIS designs.

In the *NHIS Linked Mortality Files* an eligibility-adjusted sample weight, *WGT_NEW,* is included for all individuals who are also in the annual NHIS person files. Here, *WGT_NEW* = 0 for all ineligible individuals[1]. For the NHIS person-level database associated with each survey year the adjustment weightings have the following control-total relationship: If *A* represents all NHIS participants age 18 and older and *E* represents all NHIS participants age 18 and older *and* eligible for mortality linkage then

$$\Sigma_A \, WTFA = \Sigma_E \, WGT\_NEW = \text{NHIS target population of adults 18 and older.}$$

For any response variable, *x*, original or linked, the assumption will be made that the weighted sum of the sample variable, $\Sigma_E \, x \, WGT\_NEW$, will be an unbiased estimator for the corresponding population total for characteristic *x*, and traditional design-based sampling methods can be applied. That is, a traditional method that is used with original annual data can also be used with the extended linked data. Users should note that the eligibility adjusted weight is specific to the person-level NHIS files and eligibility adjusted weights specific to the sample adult files, sample child files or any NHIS supplemental files have NOT been created.

III. Variance estimation for combining multiple *NHIS Linked Mortality Files*

The *NHIS Linked Mortality Files* cover the NHIS survey years 1986 to 2000, and analysts are likely to need to pool several years of the files to implement many types of analyses, such as survival analyses. For the 1986-2000 *NHIS Linked Mortality Files,* the 1986-1994 and 1995-2000 components are based on two independent NHIS designs. Each design has a complex-sampling structure resulting in sampled households being clustered geographically over time, and households/persons being selected with differential sampling rates. Furthermore, in years 1986 and 1996 samples were drastically reduced by 50% and 38% respectively, and in 1997 new NCHS data confidentiality policies regarding the dissemination of micro-level geographical information placed limitations on the release of design information.

[1]Starting in 1997, some military adults with children are included in the original NHIS, but with the adults receiving weight of 0. Although these records may be classified eligible in the NHIS Linked Mortality files, they receive *WGT_NEW* = 0 based on sampling year status.

To serve the vast majority of data users of the *NHIS Linked Mortality Files*, two variance estimation variables have been created in the NHIS Linked Mortality Files. These variables *STR_POOL* and *PSU_POOL,* are consistent with existing NHIS public use file variance estimation variables and represent strata and primary sampling units (PSUs) defined over the 15 years of data.

III.i Guidelines for the Data Analyst

1. The *NHIS Linked Mortality Files* and associated NHIS participants can be treated as one large NHIS survey, but with the interviews over a 15 year period.
2. The final weight will be *WGTL* = (*WGT_NEW* / k) where k is the number of annual files used, k = 1, 2 , …, 15. For example, if all 15 years of the NHIS are included in an analysis, an appropriate statement in the program would be
    - WGTL = WGT_NEW/15;

3. The design structure can be represented as strata labeled *STR_POOL* and PSUs labeled *PSU_POOL*
4. The analysis can be run on any software that is capable of handling a "two sample PSU per Stratum" complex design.
5. Users should note that these revised weights are specific to the person-level NHIS files and not for use with the sample adult files, sample child files or any NHIS supplemental files.

III.ii Discussion of conceptual *NHIS Linked Mortality Files* survey design

As an example of using the above rules with the SUDAAN software, the design structure can be specified with the following statements:

```
PROC ... DESIGN = WR ;
NEST STR_POOL PSU_POOL ;
WEIGHT WGTL ;
```

The structure provided by *STR_POOL* and *PSU_POOL* is consistent with the public use file variance estimation structures described at the NHIS "Methods" webpage, http://www.cdc.gov/nchs/about/major/nhis/methods.htm.The periods 1986-94, 1995-96, and 1997-2000 are treated as being statistically independent from each other, while the data within each of the periods are not treated as being statistically independent. The user is strongly encouraged to read the discussion on *subsetted data analysis* in the guidance materials at the NHIS Methods webpage.

The suggested weight, *WGTL,* is simply the average over the eligible-adjusted weights, *WGT_NEW,* of the combined files. Thus, the target population becomes the average population total over the same specified years. This is the traditional approach used when combining years of NHIS data. If for analytical reasons a different approach to

weighting is desired, then the user will need to re-weight the data. This specialized topic is beyond the scope of these guidelines. It should be noted that since the suggested combined-year weight *WGTL* is a simple scale-factor adjustment of *WGT_NEW*, computations of means, proportions, regression coefficients and their standard errors will be identical using either weight, *WGTL* or *WGT_NEW*. Depending upon the application, the user may not need to calculate *WGTL* for pooled year analyses and instead may use *WGT_NEW*.

In general, the rules stated above are targeted for analyses covering geographic areas which are dispersed somewhat broadly over the nation and are sampled in many PSUs. This conceptual-design is not intended for tracking small geographic areas, e.g., counties, over time. Methods and design structures for such analyses are beyond the scope of these guidelines.

NCHS (1989), ``Design and Estimation for the National Health Interview Survey'', 1985-1994, *Vital and Health Statistics*, Series 2, NO. 110

NCHS (2000), ``Design and Estimation for the National Health Interview Survey'', 1995-2004, *Vital and Health Statistics*, Series 2, NO. 130

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

Last updated: September 5, 2007