ANALYTIC GUIDELINES

Using the datasets

The NYC HANES data consist of the following 6 datasets (specific information on each dataset can be found under Data and Data Documentation):

- **SPfile (Study Participant File)**
- **CAPI (Computer Assisted Personal Interview)**
- **ACASI (Audio Computer-Assisted Self-Interview)**
- **CIDI (Composite International Diagnostic Interview)**
- **EXAM**
- **LABS.**

The Study Participant (SP) File, SPfile, contains variables necessary for all analyses, therefore, when using the other datasets, they should be merged to this file. SP_ID is the unique identifier used to merge all datasets. Merging information from multiple NYC HANES datasets using SP_ID ensures that the appropriate information for each SP is linked correctly. (SAS datasets must be sorted by SP_ID prior to merging.) Please note that NYC HANES datasets may not have the same number of records for each component because some participants did not complete each component. For an example of how to use and merge the datasets, refer to the sample program *Sample Program – Merging Datasets.sas.*

Due to the complex survey design of NYC HANES, all analyses should utilize software (e.g. SUDAAN or STATA) that adjusts for clustering in the sampling design. Clustering, which may lead to correlation between observations, will impact the variance of any estimates produced from the survey data. Analytic approaches that do not adjust for clustering will underestimate standard errors of estimates and associated significance levels. The sample programs provided use SAS-Callable SUDAAN.

Weighting Methodology and Variables

To ensure that the 1,999 NYC HANES participants are representative of the adult New Yorkers ages 20+ years old (N=5,827,719), weights were created to:

1) Account for the probability of selection into the survey - the chance that a participant will be selected into the survey based on the probability selections for the segment, household and person(s) within each household

2) Account for non-response to the survey - adjustment that attempts to reduce the bias that occurs when characteristics of survey non-respondents differ from those of respondents.

3) Poststratify to population demographics - ensures that the weighted survey totals by select demographic subpopulations match those of accepted US figures (e.g., 2000 Census)

It is important for data users to utilize the sample weights that have been provided in order to produce unbiased citywide estimates. Three types of weights have been constructed for the NYC HANES sample:

- **WTSF1CH** - 'Clinic + Home' weight used for interview & exam components captured *either at the clinic or home exam*, e.g., CAPI, EXAM.
    - All 1,999 SPs have this weight.

- **WTSF1C** - 'Clinic only' weight used for interview & exam components that were *only* captured at the clinic, e.g., ACASI, CIDI, non-fasting measures from LABS.
    - 1,861 of all 1,999 SPs completed at least one of these components and have this weight.

- **WTSF1F** - 'Fasting weight' used only for fasting subsample, e.g., fasting glucose and triglycerides.
    - 1,350 of all 1,999 SPs were included in this subsample and have this weight.

Each of the three weights represents the total adult population ages 20+ years old in New York City and each sums to 5,827,719. The determination of the most appropriate weight for analysis depends upon the variables and components used. When an analysis involves more than one variable/component, the analyst should use a "least common denominator" approach to determine which weight to use ('Clinic + Home' is greatest, then 'Clinic only', then 'Fasting'). For example, if the analyst is using one variable which is collected at the clinic only, and another variable which is collected both at the clinic and in the home, then the analyst will need to subset to 'clinic only' cases and will need to use the 'Clinic only' weight.

Please see the sample programs for examples of when and how to use the weights.

Subsample Adjusted Weights

Analysts can consider adjusting the survey weight to account for missing data that often occurs due to non-response when looking at a specific subset. In many cases, there are missing data due to component or item non-response. A few examples include:

- SPs did not complete CIDI and therefore do not have data on depression or anxiety.
- SPs may not have provided blood or may not have results for all laboratory tests.
- SPs did not self-report a history of hypertension or did not have blood pressure measurements, both of which are used in determining hypertension outcomes.

SPs who are missing data on the primary outcome of interest (e.g., depression, hepatitis C, hypertension) would not be included in the main analysis or cross tabulations with other variables. The weighted sample size would be less than 5,827,719, which is the total population of adult New Yorkers ages 20+ years old represented in the survey.

To account for missing data that occurred due to component or item non-response, analysts can consider adjusting the weight. The assumption underlying the adjusted weight is that there is some type of selection bias that has occurred. The creation of the adjusted weight accounts for another stage of selection and additional non-response based upon age, gender, and race/ethnicity. Adjustment of the weight ensures that the subsample one analyzes is reflective of the target population of 5,827,719 New York City adult residents.

Analysts can create a subset or subsample excluding those with missing data and then create appropriate adjustment weights. The following steps give specifications for performing a weighting adjustment for analysis of items or components with high non-response. The procedure involves creating a subset or subsample excluding those with relevant missing data and then computing appropriate adjustment weights. The steps below assume that response to the item within each age-gender-race/ethnicity cell is random. If response is thought to be related to another variable, e.g. health status or borough, then that variable should be used in addition to or instead of age, gender, and race/ethnicity.

Weights for subsamples defined as the respondents to a survey item or component may be obtained as follows:

1. Determine whether the Clinic+Home, Clinic, or Fasting weight is appropriate, depending on the component/item of interest.

2. Subset file to include only the cases with responses to the item of interest.

3. Within each category of age-gender-race/ethnicity, sum the clinic, clinic+home, or fasting weights of the respondents. Call this Sum B.

4. For the Clinic+Home, Clinic, or Fasting weights, use the relevant Control Totals which is the sum for the appropriate age-gender-race/ethnicity category. Call this Sum A.
    - Clinic+Home Control Totals
    - Clinic Control Totals
    - Fasting Control Totals.

5. For each SP in the cell, adjust the weight by the factor A/B to create the new adjusted weight:

- New weight = Old weight * (A/B).

- New weight should sum to 5,827,719, the total population of adult New Yorkers ages 20+ years old represented in the survey.

Please see the following sample programs for examples of how to create adjustment weights:

1) Sample program - Clinic + Home Weight Adjusted.sas
2) Sample program - Clinic Only Weight Adjusted.sas
3) Sample program - Fasting Weight Adjusted.sas.

Reporting Guidelines

Analysts are advised to consider the reliability of estimates that they are reporting. The relative standard error (RSE) of an estimate can be used to determine if an estimate is unstable. The RSE is equal to the standard error divided by the percent, and is often expressed as a percentage. Typically, an estimate with an RSE $\geq$ 30% is considered unreliable and notated that it should be interpreted with caution.

For further details on measures of reliability, including minimum sample size requirements, and on other issues outlined above key to analyzing NYC HANES data, please refer to the NHANES Analytic Guidelines website (http://www.cdc.gov/nchs/about/major/nhanes/nhanes2003-2004/analytical_guidelines.htm) which includes the following reports:

- The National Health and Nutrition Examination Survey (NHANES) Analytic and Reporting Guidelines: http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_guidelines_dec_2005.pdf.

- Previous versions that serve as supplemental guidelines include:

    1. NHANES Analytic Guidelines June 2004 Version: http://www.cdc.gov/nchs/data/nhanes/nhanes_general_guidelines_june_04.pdf.
    2. NHANES 1999-2000 Addendum to the NHANES III Analytic Guidelines: http://www.cdc.gov/nchs/data/nhanes/guidelines1.pdf.
    3. Analytic and Reporting Guidelines: The Third National Health and Nutrition Examination Survey, NHANES III (1988-94): http://www.cdc.gov/nchs/data/nhanes/nhanes3/nh3gui.pdf.

Age Adjustment

Analysts can refer to the following articles for information on when and how to age

adjust estimates:

- Age Adjustment Using the 2000 Projected U.S. Population, http://www.cdc.gov/nchs/data/statnt/statnt20.pdf.

- Age Standardization of Death Rates: Implementation of the Year 2000 Standard, http://www.cdc.gov/nchs/data/nvsr/nvsr47/nvs47_03.pdf.