# Statistical Analysis Plan

## Study design and background

From 1993 to 2013, the NHSS has been conducted for five times (every five years) which is a cross-sectional survey. The latest one was conducted in 2013, which was the fifth one. The 5th Health Service Survey of Shanghai was the extension of 5th NHSS. This survey was organized and conducted by the Shanghai Municipal Commission Health and Family Planning. The sampling method and quality assurance measures used in Shanghai survey were consistent with the national sampling approach and principle. The 5th Health Service Survey of Shanghai cover all of the 17 districts in Shanghai, and a three-stage, stratified, random sampling method was adopted. First stage, 100 towns/townships were selected randomly from all these 17 districts. Second stage, one thousand villages/communities were sampled randomly from these selected towns/townships. Third stage, about 12,000 households were identified randomly.

A face-to-face interview approach using structured household questionnaire, which was developed by the National Commission Health and Family Planning of China, was conducted for each household. The questionnaire contained the general information of household, the demographic characteristics of residents, the relationship of household members, self-reported illness and injury, outpatient and inpatient information.

To ensure the quality of this survey, some quality assurance measures were applied during the process of data collection. Logic errors would be checked among the data by survey constitutors. If there were logic errors, the investigators would contact the household members and verify the relevant information. The accuracy of data information were assessed by revisit-approach. The investigators revisited 5% of the sampled households and collected ten key questions to check the consistency of the information recorded. The consistency rates between these two visits was near to 99 %.

## Objective

The aim of our study is to estimate the association between an adult's own chronic condition status and the chronic condition status of other household members.

## Data management

Those who meet the following conditions will be excluded.

1. Residents less than 18 years old
2. Only one person in a household

Because the NHSS of Shanghai conducted well, there is little data missing in the data set. If one resident missed some important variables, he/she will be excluded from analyses.

## Outcomes

We choose five chronic conditions with high prevalence in this survey: hypertension, diabetes, ischemic heart disease (IHD), cerebrovascular disease (CVD), and obesity. The definition of these chronic conditions is based on the question in the questionnaire,

the disease coding list of the NHSS, and Body mass index (BMI).

**Hypertension:** Hypertension for every resident is indicated based on the question in the questionnaire "Have you ever been told by a doctor that you have hypertension". If one choose "YES", the hypertension status of the participant is "1". If the one choose "NO", the hypertension status of the participant is "0".

**Diabetes:** Diabetes for every resident is indicated based on the question in the questionnaire "Have you ever been told by a doctor that you have diabetes". If one choose "YES", the diabetes status of the participant is "1". If the one choose "NO", the diabetes status of the participant is "0".

**IHD:** The questionnaire will record all of the chronic diseases for every resident, and these chronic diseases will been encoded in accordance with the disease coding list of the NHSS. IHD included angina pectoris (061), myocardial infarction (062), and other ischemic heart disease (063). If one choose any of these three codes, the IHD status of the participant is "1". If the one did not choose any of these three codes, the IHD status of the participant is "0".

**CVD:** This chronic condition will be indicated according to the disease coding list of the NHSS: cerebrovascular disease (067). If one choose this code, the CVD status of the participant is "1". If the one did not choose this code, the CVD status of the participant is "0".

**Obesity:** Body mass index (BMI) will be calculated for every resident by height and weight. Obesity was indicated by the World Health Organization (WHO) International BMI categories (BMI $\geq$ 30 kg/m$^2$). If the BMI of the one is equal or over 30 kg/m$^2$, the obesity status of the participant is "1". If the BMI of the one is lower than 30 kg/m$^2$, the obesity status of the participant is "0".

**Any chronic conditions:** If any of these five chronic conditions is "Yes" for a participant, his or her status of "Any chronic conditions" is "Yes", otherwise "Any chronic conditions" is "No"

## Covariates

Socio-demographic characteristics added to our models as covariates include age (continuously specified in years), education status (illiteracy/primary, secondary or college), health insurance status (yes or no), marriage status (married, unmarried, divorced or widowed), smoking (yes or no), and drinking (yes or no). Except the analyses of spouse subsample, gender (male or female) and is also included as a covariate.

## Subsamples

To make a comprehensive estimation for the association, we conduct the analyses in three subsamples: all adult household members, adult children, and wives.

**Subsample 1 all adult household members:** We can estimate the general association between an adult's own chronic condition status and the chronic condition status of other household members based on this subsample. The chronic condition status of participant is treated as outcome, and the household situation of chronic condition for each participant is treated as exposure factor. If any other residents (excluding self)

have the given chronic conditions, the exposure for the one is "YES".

**Subsample 2 adult children:** We only include adult children (age > 18 years old) in this subsample, and those participants are excluded if the disease information of parents were not available. The chronic conditions status of their parents are considered as exposure, and the chronic condition status of adult children is treated as outcome. If any of the one's parents have the given chronic conditions, the exposure for the one is "YES". The results of this subsample might show the effect of genetic factors and common living environment in the relationships that we are interested in.

**Subsample 3 wives:** we include married women in the analyses. We define the chronic condition of wife as the outcome and the chronic condition of husband as the exposure. The results of this subsample might show the effect of common living environment without genetic ties in the relationships that we are interested in.

## Statistical analysis

### Descriptive analyses

Descriptive statistics are summarized for the covariates of three subsamples, respectively. Mean (standard deviation, SD) is calculated for continuous variables, and counts (percentages) are calculated for categorical variables.

### Generalized Estimating Equations

The generalized estimating equations (GEE) model with logit link will be used to find out the relationship between one's chronic conditions and the others with chronic conditions living in the same household. We consider two-level hierarchical structure of the model (individual within household), and choose the exchangeable working correlation matrix for GEE model. The GEE model is performed using "GENMOD" procedure in SAS software. The analyses model is based on individual's data without considering sampling weight, because there is no relevant information.

The adjusted GEE models will include age, gender, health insurance status, education status, drinking and smoking (gender is excluded in the third subsample). The chronic condition status of participants is included in models as independent variable. And the exposure (status of other members) is included in models as dependent variable. The final conclusion will be based on the results of the adjusted models.

The odds ratio (OR) and 95% confidence interval (CI) are estimated by GEE model to indicate the association between any chronic condition or each given chronic condition of individual with the same condition of household member (e.g., the association between the hypertensive status of a participants and that same condition in his or her other household members). And the association of different chronic condition will also be assessed (e.g., the association between hypertension in a participant and diabetes in other household members).

### Sensitivity analysis

We will also estimate the unadjusted odds ratio and 95% confidence interval as sensitivity analysis. And the unadjusted results are required by the STROBE checklist.

### Subgroup analysis

We plan to perform two subgroup analysis. One is by gender (male or female) and the

other is by education status (illiteracy/primary, secondary or collage). Illiteracy/primary means the education years were from 0 to 5 years, secondary means the education years were from 6 to 12 years, and collage means the education years of residents were higher than 12 years including undergraduate and graduate degrees. The reason why we perform this two subgroup analysis is this two factors may have a significant influence on living habits which is the cause of chronic conditions.

**Other information**

All data management and statistical analyses will be performed using SAS software (version 9.4; SAS Institute Inc., Cary, NC). All reported p values are two-sided and p value $< 0.05$ is regarded as statistically significant. This study will be reported according to the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) guidelines.