

**HCHS/SOL Analysis Methods - Visit 3**

**November 2025**

**Version 1.1**

**Prepared by the**

**HCHS/SOL Coordinating Center**

Collaborative Studies Coordinating Center

UNC Department of Biostatistics

Jianwen Cai

Beibo Zhao

Daniela Sotres-Alvarez

Alex Akushevich

Alvaro Clemente Quijano Angarita

Wenyi Xie

Franklyn Gonzalez

**This document is CONFIDENTIAL and for EXCLUSIVE use by HCHS/SOL investigators and NHLBI-NIH. Its purpose is to illustrate methods not to report results. Please send questions, suggestions, and comments to fgonzale@email.unc.edu**

[Foreword 4](#_Toc211703330)

[Note to Users 4](#_Toc211703331)

[Additional Documentations 4](#_Toc211703332)

[List of Updates 5](#_Toc211703333)

[Version 1.1 (November 2025) 5](#_Toc211703334)

[List of Abbreviations 6](#_Toc211703335)

[1. Introduction 7](#_Toc211703336)

[1.1. Inferential Framework 7](#_Toc211703337)

[1.2. Modelling Approaches 8](#_Toc211703338)

[2. Cross-Sectional Analysis at Visit 3 9](#_Toc211703339)

[2.1. Visit 3 Sampling Weights 9](#_Toc211703340)

[2.2. Comparison of Estimates for Baseline Characteristics 11](#_Toc211703341)

[3. Longitudinal Analysis: Introduction 18](#_Toc211703342)

[3.1. Methods for Addressing Missing Visits 18](#_Toc211703343)

[3.1.1. Multiple Imputation 19](#_Toc211703344)

[3.1.2. Inverse Probability Weighting 20](#_Toc211703345)

[3.1.3. MI vs. IPW 21](#_Toc211703346)

[3.2. Visit 1 Sample vs. Visit 3 Sample 22](#_Toc211703347)

[3.3. Recommendations for Marginal Approaches 25](#_Toc211703348)

[Continuous Outcomes 25](#_Toc211703349)

[Binary Outcomes 25](#_Toc211703350)

[3.3.1. General Procedure for GEE with MI 25](#_Toc211703351)

[3.3.2. General Procedure for GEE with GLM-based IPW 26](#_Toc211703352)

[3.4. Analytic Dataset 27](#_Toc211703353)

[3.5. Data Management: wide-format and long-format 30](#_Toc211703354)

[4. Longitudinal Analysis of Continuous Outcomes 31](#_Toc211703355)

[4.1. Illustrative Analytic Example 31](#_Toc211703356)

[4.1.1. Model Specification and Covariates 31](#_Toc211703357)

[4.1.2. Implementation of MI 31](#_Toc211703358)

[4.1.3. Implementation of GLM-based IPW 32](#_Toc211703359)

[4.1.4. Choice and Specification of MI and IPW Models 33](#_Toc211703360)

[4.2. SAS/SUDAAN 34](#_Toc211703361)

[4.2.1. MI 34](#_Toc211703362)

[4.2.2. GLM-based IPW 36](#_Toc211703363)

[4.2.3. Application to Visit 1 Sample 39](#_Toc211703364)

[4.2.3.1. Complex-Survey GEE with MI 39](#_Toc211703365)

[4.2.3.2. Model-Based GEE with MI 40](#_Toc211703366)

[4.2.3.3. Complex-Survey GEE with GLM-Based IPW 43](#_Toc211703367)

[4.2.3.4. Model-Based GEE with GLM-based IPW 43](#_Toc211703368)

[4.2.4. Application to Visit 3 Sample 44](#_Toc211703369)

[4.2.4.1. Complex Survey GEE with GLM-based IPW 44](#_Toc211703370)

[4.2.4.2. Model-Based GEE with GLM-based IPW 44](#_Toc211703371)

[4.3. Stata 45](#_Toc211703372)

[4.3.1. Application to Visit 1 Sample 45](#_Toc211703373)

[4.3.1.1. Model-Based GEE with MI 45](#_Toc211703374)

[4.4. R 49](#_Toc211703375)

[4.4.1. Application to Visit 1 Sample 49](#_Toc211703376)

[4.4.1.1. Model-Based GEE with MI 49](#_Toc211703377)

[5. Longitudinal Analysis of Binary Outcomes 55](#_Toc211703378)

[References 55](#_Toc211703379)

Foreword

## Note to Users

* This document is for illustration purposes for longitudinal data analysis based on data from the first three HCHS/SOL clinic visits (Baseline/Visit 1, Visit 2, Visit 3).
* Because the HCHS/SOL cohort was selected through a stratified multi-stage area probability sample design (Lavange et al., 2010), the study design specifications are accounted for in all the analysis presented.
* For cross-sectional analysis based on Visit 3 data only, please refer to *HCHS/SOL Analysis Methods at Baseline* and use Visit 3 sampling weights.
* For longitudinal analysis using only two visits, for example, Visit 1 and Visit 3 or Visit 2 and Visit 3, please refer to *HCHS/SOL Analysis Methods – Visit 2* and use Visit 3 sampling weights.
* The document is not intended for direct citation.
* Statistical program outputs used in the examples throughout this document have been modified and/or formatted for presentation and clarity.
* Case sensitivity: In R and Stata, variable names as well as commands are case-sensitive.

## Additional Documentations

* HCHS/SOL Analysis Methods at Baseline <https://sites.cscc.unc.edu/hchs/node/405>
* HCHS/SOL Analysis Methods - Visit 2

<https://sites.cscc.unc.edu/hchs/node/6113>

* HCHS/SOL Baseline Physical Activity Data Overview, Methods & Guidelines

<https://sites9.cscc.unc.edu/hchs/node/415>

* SAS (Version 9.4) <https://support.sas.com/documentation/onlinedoc/stat/>
* STATA (Version 18)

<https://www.stata.com/features/documentation/>

* R (Version 4.4.1)

[https://www.r-project.org/l](https://www.statmodel.com/html_ug.shtml)

## List of Updates

### Version 1.1 (November 2025)

* Inv files version update (v1 inv5; v2 inv3; v3 inv2)

## List of Abbreviations

|  |  |
| --- | --- |
| BG | Block Group |
| CC | Coordinating Center |
| CART | Classification and Regression Tree |
| FCS | Fully Conditional Specification |
| GEE | Generalized Estimating Equation |
| GLM | Generalized Linear Model |
| HH | Household |
| IPW | Inverse Probability Weighting |
| MAR | Missing at Random |
| MCAR | Missing Completely at Random |
| MICE | Multiple Imputation by Chained Equations |
| MI | Multiple Imputation |
| MNAR | Missing Not at Random |
| PSU | Primary Sampling Unit |
| SRS | Simple Random Sampling |
| SSU | Secondary Sampling Unit |
| SUB | Subject |

1. Introduction

In the HCHS/SOL, data are collected longitudinally, with participants invited to in-person clinic visits to obtain measurements of interest such as anthropometry and biospecimens. This document contains two general parts. The first part (Chapter 2) describes the calculation of Visit 3 sampling weights. For how to conduct **cross-sectional analysis** for HCHS/SOL data involving Visit 3 data only, please refer to *HCHS/SOL Analysis Methods at Baseline* and use Visit 3 sampling weights. The second part (Chapters 3 and 4) provides guidelines on **longitudinal analysis** with repeated measures for HCHS/SOL data involving more than two clinic visits, focusing on modeling a continuous outcome over time. For how to conduct **longitudinal analysis** for HCHS/SOL datainvolving only two clinic visits, for example, Visit 1 and Visit 3 data only or Visit 2 and Visit 3 data only, focusing on modelling the difference, rate of change, incident event odds ratio, or incidence rate, please refer to *HCHS/SOL Analysis Methods - Visit 2* and use Visit 3 sampling weights.

Because the HCHS/SOL cohort was selected through a stratified multi-stage area probability sample design (Lavange et al., 2010), the study design specifications are accounted for in all the presented analysis. Sample codes and results using readily available software (e.g., SAS, Stata, R) are provided.

## Inferential Framework

In all our analysis, we adopt the following perspective: observations are assumed to be sampled from a fixed finite population using a pre-specified sampling design, with the variation in the sample resulting from the randomness from sampling, instead of distributional assumption about the data-generating process (Sterba, 2009). The values of variables of interest are treated as fixed in this finite population, and their inference considers the distribution of the estimator over repeated samples by using the same sampling design. For valid inference under this perspective, the sampling design (stratification, clustering and sampling weights) needs to be accounted for during the point and variance estimation of finite-population parameters. Analytic techniques that properly incorporate these features are referred to throughout this document as **design-based complex-survey procedures**. For certain model structures, particularly those involving **longitudinal or highly clustered data**, such complex survey procedures either do not yet exist or have not been fully implemented in standard statistical software.

In contrast to complex-survey procedures, **model-based procedures** assume that the observed data arise from an underlying stochastic (superpopulation) model that specifies the probability distribution of the measurements given model parameters. Inferences are drawn based on the likelihood function or estimating equations derived from that model, rather than from the randomization distribution of the sampling design. When applied appropriately, model-based approaches can provide consistent and unbiased estimates of finite-population parameters if they incorporate sampling weights to adjust for unequal selection probabilities and use robust variance estimators to account for intra-cluster correlation and model misspecification.

The **Coordinating Center (CC)** conducted **simulation studies** to evaluate both complex-survey methods (when available) and **model-based** alternatives as tools for obtaining finite-population estimates. The simulation results, which will be communicated in a separate document, show that **both** can yield valid and efficient finite-population inferences for HCHS/SOL data. Therefore, in this document, we present the use of **both** complex-survey **and model-based procedures** for analyzing HCHS/SOL data, highlighting approaches that have been empirically shown to provide appropriate inference for the target finite population.

## Modelling Approaches

Two statistical modelling approaches are commonly adopted to analyze longitudinal data with repeated measures, the **marginal approach** modeling the population-averaged longitudinal trend and the **conditional approach** modeling the subject-specific longitudinal trend. The marginal approach describes the linear relationship of a transformed mean response with the covariates without specifying the correlation structure for the responses within clusters. The coefficients (betas) of covariates have the interpretation of population-averaged effects; hence they are useful when one is interested in the covariate effects on the response but describing the amount of correlation of responses within clusters is not of particular interest. The conditional approach incorporates random effects to capture between-subject heterogeneity in response trend. The random effects are usually assumed to follow some parametric distribution. The coefficients of the covariates in the model (betas) represent subject-specific effects, quantifying how changes in covariates within a person affect individual responses conditioning on the random effects. By explicitly modeling the within-cluster correlation structure through random effects, this approach provides insights into how the responses within a person are correlated. The interpretation of covariate effects is specific to each subject rather than averaged across the population. The choice between the conditional and marginal approaches depends on whether or not the correlation of the responses within clusters is of interest. When the response variable is continuous and the link function is identity function, the beta coefficients in the marginal model are the same as the fixed effects in the conditional model.

**Generalized Estimating Equation (GEE)** is a marginal approach for longitudinal analysis with repeated measures (Liang & Zeger, 1986). GEE estimates the relationship of a mean response with the covariates through a quasi-likelihood function and accounts for the non-independence of units within clusters (e.g., repeated observations within participants) through the specification of a working correlation structure. GEE can provide asymptotically unbiased coefficient estimates, which are interpreted as population-averaged effects. The variance of the coefficients can be estimated using a cluster-robust variance estimator (also known as the sandwich estimator), which is robust against misspecification of the working correlation structure. Investigators can use this marginal approach when their primary interest lies in understanding the effects of change in covariates within a person/cluster on the response, rather than quantifying the correlation between responses within clusters.

1. Cross-Sectional Analysis at Visit 3

In this chapter, we describe the calculation of Visit 3 sampling weights. We also present estimates for baseline characteristics based on Visit 1 sample using Visit 1 sampling weights and based on Visit 3 sample using Visit 3 sampling weights. We expect the estimates to be similar because both are estimating the same population parameters.

For how to conduct **cross-sectional analysis** for HCHS/SOL data involving Visit 3 data only, please refer to *HCHS/SOL Analysis Methods at Baseline* and use Visit 3 sampling weights.

## Visit 3 Sampling Weights

The HCHS/SOL cohort at baseline was selected through a stratified multi-stage probability sampling design. Briefly, at the 1st stage, the **Primary Sampling Units (PSUs)** were the census **Block Groups (BGs)** and were selected with **Simple Random Sampling (SRS)** at each field center, stratified by cross-classification of 2000 Census high/low socioeconomic status and high/low Hispanic/Latino concentration. At the 2nd stage, the **Secondary Sampling Units (SSUs)** were the **Households (HHs)** and were selected with SRS in each of the sampled PSUs, stratified by having or not Hispanic/Latino surname from postal addresses purchased from Genesys. Households with Hispanic/Latino surname were over-sampled. Lastly, at the 3rd stage, **Subjects (SUBs)**, i.e., study participants, were selected in each of the eligible sampled SSUs. Participants aged 45-74 years were over-sampled. Therefore, participants were nested within household clusters, which were further nested within block group clusters with unequal probabilities of selection of BGs, HHs, and SUBs at their respective levels by this sampling design. The product of the reciprocals of the probabilities of being selected at each stage was used to calculate the basesampling weight for each participant in the cohort, which remains the same through all subsequent visits. These base weights were then adjusted for differential non-response at both the household and subject-level at baseline, forming the Visit 1 non-response adjusted sampling weights. Non-response adjustment factors were defined as the reciprocal of an estimate of the probability that a sample household agrees to be screened and to participate in the study, and the probability that a person selected into the sample agrees to participate and completes the clinic exam.

Visit 3 data collection initially began in January 2020. Due to the COVID-19 pandemic, it was paused in March 2020. To navigate the challenges posed by the pandemic, the HCHS/SOL Steering Committee decided to split Visit 3 visit into two parts: phone interview and in-person exam. The phone interviews were initiated in May 2020 and the in-person exam was resumed during the first quarter of 2021. Consequently, for Visit 3, there are two definitions of participation: (1) In-person participation only (including home visits) (N=9,090, i.e., excluding those who had phone interviews only); and (2) All participation (including phone-only interviews) (N=9,864). Of the 7,179 participants who started with phone interviews during the COVID pandemic, 6,405 (89%) later completed an in-person visit, while 774 (11%) had phone interviews only. The variables PARTICIPANT\_EXAMONLY\_V3 and PARTICIPANT\_ALL\_V3 are the indicator variables for Visit 3 participation based on the “Exam Only” definition and the “All” definition, respectively.

As with any complex survey design, the Visit 3 sampling weights account for non-response under both definitions. The non-response probability at Visit 3 is estimated using a **Classification and Regression Tree (CART)** analysis that allows an estimation of non-response profiles using all data collected at either baseline or over the course of follow-up. The idea is to form strata based on factors associated with the probability of returning for Visit 3 examination. To identify these factors, the R package 'rpart' was used to implement the CART. The advantage of the CART is that it takes interactions among factors into consideration and provides estimates for the cutpoints of continuous variables. The baseline factors considered include the following categorical variables: Hispanic/Latino Background, Age, Sex, PSU Strata, Education, Income, Health Insurance, Mental Health Status, Physical Health Status, Alcohol Use, Cigarette Use, Diabetes Status, Employment Status, Physical Activity, Prevalent Hypertension, Prevalent MI, Prevalent Stroke, Born in Mainland US, and Years Lived in US at the baseline, and AFU refusal; and the following continuous variables: Height, Weight, Body Mass Index (BMI), Cardiac Risk Ratio, eGFR, Triglycerides, HDL, LDL, Glucose, Creatinine, Urine Creatinine, Urine Micro albumin, Albumin/Creatinine Ratio, Cystatin C, and Insulin at baseline, and Log-Distance between V1 address and the last AFU address before V3 (referred to as Mobility Score hereafter).

The CART identified several factors associated with the probability of returning for Visit 3. For the "Exam Only" definition, these factors include AFU refusal, Mobility Score with a cutpoint of 3.94, Age group, Sex, PSU Strata, Cystatin C with cutpoints of 0.795, 1.09, and 1.2, and Income. For the "All" definition, the same factors were identified, except for Income. The CART divided the participants into groups, referred to as CART groups, based on identified factors (used cutpoints for continuous variables). The CART groups were further stratified by Cigarette Use. When forming the final strata for Visit 3 non-response adjustment, we imposed a minimum of 90 participants per stratum to ensure stability and reliability. If a stratum had less than 90 participants, it was combined with an adjacent tree branch that was grown from the same parent branch until sufficient number of participants was reached to form a stratum. Visit 3 response rates were then calculated within each of these strata.

Consistent with the approach used for overall sampling weights at baseline and Visit 2, the derivation of the overall sampling weight at Visit 3 follows the following procedure: (1) calculate Visit 3 non-response adjusted sampling weights by multiplying the Visit 1 non-response adjusted sampling weights by the inverse of the Visit 3 response rates, calculated for each stratum that is formed from the CART analysis described above; (2) trim extreme weights to control variability of the response rates; (3) calibrate to the age, sex and Hispanic/Latino background distributions from the 2010 US Census for the four study centers based on participants' Visit 1 age; (4) normalize to the overall sample.

The two definitions of participation at Visit 3 each have their corresponding overall sampling weights: WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 for the "Exam Only" definition, and WEIGHT\_NORM\_OVERALL\_ALL\_V3 for the "All" definition. Investigators using data from clinic/home exams or biospecimens should use the "Exam Only" dataset with 9,090 participants and the "Exam Only" sampling weights. However, if they are interested only in measures collected through phone interviews, they can use the larger dataset with 9,864 participants and the "All" sampling weights.

## Comparison of Estimates for Baseline Characteristics

The sampling weights released for Visit 1 and Visit 3 data are both designed for inferences in the HCHS/SOL target population. We compared estimates for some baseline characteristics using Visit 1 sampling weights (WEIGHT\_FINAL\_NORM\_OVERALL) with data from Visit 1 to two scenarios of those using Visit 3 sampling weights with data from Visit 3: (1) using WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 for Visit 3 participation based on the "Exam Only" definition (**Output 2.2‑1**), and (2) using WEIGHT\_NORM\_OVERALL\_ALL\_V3 for Visit 3 participation based on the "All" definition (**Output 2.2‑2**).

To compare the results, we examined the difference in estimated percentages or means, defined as (value\_v3 - value\_v1), and the relative difference, defined as the difference divided by value\_v1. Comparing the results, we note that most of these estimates have the absolute value of the difference less than 2.7% for percentages and 0.9 units for continuous variables. The absolute values of the relative difference are less than 10%, except for those with very low prevalence (Underweight, CVD, and MI) where the estimates are not stable.

**Output 2.2‑1**

**Baseline Characteristics of HCHS/SOL Target Population using Data from Visit 1 (Baseline) and Visit 3 “Exams Only” Participants**

|  |  | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **HCHS/SOL Target Population Estimates based on Visit 1 Sample (N=16415 for Visit 1 Data)** | | **HCHS/SOL Target Population Estimates based on Visit 3 Sample (N=9090 for Visit 3 Data)** | |  | |
| **Baseline Characteristics** | **N** | **Mean or % (95% CI)** | **N** | **Mean or % (95% CI)** | **Difference** | **Relative Difference** |
| **Age (years)** | 16415 | 41.06 (40.6, 41.5) | 9090 | 41.13 (40.5, 41.7) | 0.07 | 0.00 |
| **Sex at birth(%)  Male** | 6580 | 47.87 (46.8, 48.9) | 3166 | 47.87 (46.3, 49.4) | 0.00 | 0.00 |
| **Female** | 9835 | 52.13 (51.1, 53.2) | 5924 | 52.13 (50.6, 53.7) | 0.00 | 0.00 |
| **Education (%)  Less than high school** | 6207 | 32.35 (31.0, 33.7) | 3319 | 30.36 (28.6, 32.1) | -1.99 | -0.06 |
| **High school graduate** | 4180 | 28.20 (27.1, 29.3) | 2261 | 27.51 (26.1, 28.9) | -0.69 | -0.02 |
| **Greater than high school** | 5937 | 39.46 (37.9, 41.1) | 3478 | 42.14 (40.1, 44.1) | 2.68 | 0.07 |
| **Hispanic/Latino background(%)  Cuban** | 2348 | 20.02 (16.7, 23.3) | 1320 | 19.81 (16.4, 23.2) | -0.21 | -0.01 |
| **Dominican** | 1473 | 9.94 (8.6, 11.3) | 836 | 9.96 (8.4, 11.5) | 0.02 | 0.00 |
| **Mexican** | 6472 | 37.37 (34.2, 40.6) | 3690 | 37.13 (33.9, 40.4) | -0.25 | -0.01 |
| **Puerto Rican** | 2728 | 16.15 (14.6, 17.7) | 1337 | 15.98 (14.3, 17.7) | -0.17 | -0.01 |
| **Central American** | 1732 | 7.40 (6.3, 8.5) | 984 | 7.63 (6.3, 8.9) | 0.22 | 0.03 |
| **South American** | 1072 | 4.98 (4.4, 5.6) | 656 | 4.97 (4.3, 5.7) | -0.02 | -0.00 |
| **Other** | 503 | 4.13 (3.6, 4.7) | 245 | 4.54 (3.7, 5.4) | 0.40 | 0.10 |
| **Annual family income(%)  <$20,000** | 7207 | 41.85 (40.1, 43.6) | 3932 | 40.45 (38.2, 42.7) | -1.40 | -0.03 |
| **$20,000-$50,000** | 6119 | 36.88 (35.6, 38.2) | 3553 | 37.64 (35.8, 39.5) | 0.76 | 0.02 |
| **>$50,000** | 1601 | 11.70 (10.2, 13.2) | 898 | 12.80 (10.8, 14.8) | 1.09 | 0.09 |
| **Not reported** | 1488 | 9.57 (8.8, 10.3) | 707 | 9.11 (8.1, 10.1) | -0.46 | -0.05 |
| **Marital status(%)  Single** | 4522 | 34.64 (33.3, 36.0) | 2189 | 33.96 (32.2, 35.7) | -0.67 | -0.02 |
| **Married or living with partner** | 8436 | 48.82 (47.3, 50.4) | 5003 | 50.22 (48.2, 52.3) | 1.39 | 0.03 |
| **Separated divorced, or widowed** | 3369 | 16.54 (15.6, 17.5) | 1869 | 15.82 (14.5, 17.1) | -0.72 | -0.04 |
| **Health insurance(%)** | 8172 | 50.54 (48.7, 52.4) | 4552 | 52.64 (50.4, 54.9) | 2.10 | 0.04 |
| **US residence >= 10 Years(%)** | 12490 | 72.34 (70.5, 74.2) | 6966 | 72.82 (70.6, 75.0) | 0.48 | 0.01 |
| **Language preference(%)  Spanish** | 13119 | 74.86 (73.0, 76.7) | 7545 | 75.09 (72.9, 77.3) | 0.23 | 0.00 |
| **English** | 3296 | 25.14 (23.3, 27.0) | 1545 | 24.91 (22.7, 27.1) | -0.23 | -0.01 |
| **Systolic BP (mmHg)** | 16401 | 119.92 (119.4, 120.4) | 9085 | 119.24 (118.7, 119.8) | -0.68 | -0.01 |
| **Diastolic BP (mmHg)** | 16394 | 72.19 (71.9, 72.5) | 9080 | 71.95 (71.5, 72.3) | -0.24 | -0.00 |
| **Hypertension (%)** | 4937 | 24.19 (23.0, 25.4) | 2730 | 23.80 (22.4, 25.2) | -0.39 | -0.02 |
| **Treated for hypertension(%)b** | 3464 | 68.94 (66.8, 71.0) | 1962 | 70.10 (67.6, 72.6) | 1.17 | 0.02 |
| **Total cholesterol(mg/dL)** | 16248 | 194.32 (193.2, 195.4) | 9022 | 194.52 (193.0, 196.1) | 0.20 | 0.00 |
| **LDL-cholesterol(mg/dL)** | 15918 | 119.74 (118.8, 120.7) | 8866 | 120.29 (119.0, 121.6) | 0.54 | 0.00 |
| **HDL-cholesterol(mg/dL)** | 16246 | 48.48 (48.2, 48.8) | 9022 | 48.70 (48.3, 49.1) | 0.22 | 0.00 |
| **eGFR** | 16131 | 106.92 (106.3, 107.5) | 8960 | 107.78 (107.1, 108.5) | 0.86 | 0.01 |
| **Treated for hypercholesterolemia(%)c** | 1629 | 24.36 (22.6, 26.1) | 1119 | 24.08 (22.1, 26.1) | -0.28 | -0.01 |
| **BMI kg/m2** | 16344 | 29.36 (29.2, 29.5) | 9064 | 29.27 (29.1, 29.5) | -0.09 | -0.00 |
| **Obesity Status (%)  Underweight (BMI<18.5 kg/m2)** | 130 | 1.16 (0.9, 1.4) | 47 | 0.99 (0.6, 1.4) | -0.17 | -0.15 |
| **Normal (BMI 18.5-25 kg/m2)** | 3191 | 22.07 (21.1, 23.1) | 1622 | 21.58 (20.2, 22.9) | -0.49 | -0.02 |
| **Overweight (BMI 25-30 kg/m2)** | 6116 | 37.19 (36.0, 38.4) | 3539 | 38.58 (37.1, 40.1) | 1.39 | 0.04 |
| **Obese (BM>=30 kg/m2)** | 6907 | 39.58 (38.3, 40.9) | 3856 | 38.85 (37.2, 40.5) | -0.73 | -0.02 |
| **Fasting glucose(mg/dL)** | 16220 | 102.20 (101.4, 103.0) | 9010 | 102.00 (100.9, 103.1) | -0.21 | -0.00 |
| **Diabetes - definition #2 (%)d** | 3218 | 14.88 (14.1, 15.7) | 1738 | 14.88 (13.8, 16.0) | -0.00 | -0.00 |
| **Diabetes - definition #4 (%)e** | 3227 | 14.85 (14.0, 15.7) | 1744 | 14.83 (13.8, 15.9) | -0.02 | -0.00 |
| **Treated for diabetes(%)f** | 1836 | 53.77 (51.3, 56.2) | 956 | 51.77 (48.2, 55.3) | -2.00 | -0.04 |
| **Waist circumference (cm)** | 16349 | 97.37 (96.9, 97.8) | 9064 | 97.16 (96.6, 97.7) | -0.21 | -0.00 |
| **Current Smoker (%)** | 3166 | 21.37 (20.3, 22.5) | 1545 | 20.51 (19.1, 21.9) | -0.86 | -0.04 |
| **Asthma (%)** | 2637 | 17.37 (16.4, 18.4) | 1420 | 17.55 (16.2, 18.9) | 0.18 | 0.01 |
| **COPD (%)** | 488 | 2.78 (2.4, 3.1) | 252 | 2.65 (2.2, 3.1) | -0.13 | -0.05 |
| **CVD (%)** | 858 | 4.72 (4.2, 5.2) | 420 | 4.14 (3.5, 4.7) | -0.58 | -0.12 |
| **MI (%)** | 384 | 2.34 (2.0, 2.7) | 187 | 1.90 (1.5, 2.3) | -0.44 | -0.19 |
| **Hearing Loss (%)** | 2799 | 15.06 (14.2, 15.9) | 1491 | 14.13 (13.1, 15.2) | -0.93 | -0.06 |

Abbreviations: BMI: body mass index; BP: blood pressure; LDL: low density lipoprotein; HDL: high density lipoprotein; COPD: chronic obstructive pulmonary disease; CVD: cardiovascular disease; MI: myocardial infarction.

a All values (except N) weighted for study design and non-response.

b Denominator is restricted to participants with hypertension at baseline (Unweighted Visit 1: N=4937, Visit 3: N=2730).

c Denominator is restricted to participants with hypercholesterolemia at baseline (Unweighted Visit 1: N=5332, Visit 3: N=3775).

d ADA guideline plus scanned/transcribed medication use.

e ADA guideline plus self-reported medication use.

f Denominator is restricted to participants with diabetes (ADA guideline plus self-reported diabetes) at baseline (Unweighted Visit 1: N=3384, Visit 3: N=1833).

Source: HC331511 (18SEP24 using INV2 data)

**Output 2.2‑2**

**Baseline Characteristics of HCHS/SOL Target Population using Data from Visit 1 (Baseline) and Visit 3 “All” Participants**

|  |  | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **HCHS/SOL Target Population Estimates based on Visit 1 Sample (N=16415 for Visit 1 Data)** | | **HCHS/SOL Target Population Estimates based on Visit 3 Sample (N=9864 for Visit 3 Data)** | |  | |
| **Baseline Characteristics** | **N** | **Mean or % (95% CI)** | **N** | **Mean or % (95% CI)** | **Difference** | **Relative Difference** |
| **Age (years)** | 16415 | 41.06 (40.6, 41.5) | 9864 | 41.13 (40.6, 41.7) | 0.07 | 0.00 |
| **Sex at birth(%)  Male** | 6580 | 47.87 (46.8, 48.9) | 3471 | 47.87 (46.5, 49.2) | -0.00 | -0.00 |
| **Female** | 9835 | 52.13 (51.1, 53.2) | 6393 | 52.13 (50.8, 53.5) | -0.00 | -0.00 |
| **Education (%)  Less than high school** | 6207 | 32.35 (31.0, 33.7) | 3617 | 30.62 (28.9, 32.3) | -1.73 | -0.05 |
| **High school graduate** | 4180 | 28.20 (27.1, 29.3) | 2465 | 27.56 (26.2, 28.9) | -0.64 | -0.02 |
| **Greater than high school** | 5937 | 39.46 (37.9, 41.1) | 3745 | 41.82 (40.0, 43.7) | 2.36 | 0.06 |
| **Hispanic/Latino background(%)  Cuban** | 2348 | 20.02 (16.7, 23.3) | 1392 | 19.82 (16.5, 23.1) | -0.20 | -0.01 |
| **Dominican** | 1473 | 9.94 (8.6, 11.3) | 922 | 9.95 (8.4, 11.5) | 0.01 | 0.00 |
| **Mexican** | 6472 | 37.37 (34.2, 40.6) | 4033 | 37.23 (34.0, 40.5) | -0.14 | -0.00 |
| **Puerto Rican** | 2728 | 16.15 (14.6, 17.7) | 1477 | 16.11 (14.4, 17.8) | -0.04 | -0.00 |
| **Central American** | 1732 | 7.40 (6.3, 8.5) | 1048 | 7.63 (6.3, 8.9) | 0.23 | 0.03 |
| **South American** | 1072 | 4.98 (4.4, 5.6) | 699 | 5.00 (4.3, 5.7) | 0.02 | 0.00 |
| **Other** | 503 | 4.13 (3.6, 4.7) | 264 | 4.25 (3.5, 5.0) | 0.12 | 0.03 |
| **Annual family income(%)  <$20,000** | 7207 | 41.85 (40.1, 43.6) | 4294 | 41.46 (39.4, 43.5) | -0.39 | -0.01 |
| **$20,000-$50,000** | 6119 | 36.88 (35.6, 38.2) | 3819 | 37.22 (35.5, 39.0) | 0.34 | 0.01 |
| **>$50,000** | 1601 | 11.70 (10.2, 13.2) | 976 | 12.34 (10.6, 14.0) | 0.64 | 0.05 |
| **Not reported** | 1488 | 9.57 (8.8, 10.3) | 775 | 8.97 (8.0, 9.9) | -0.59 | -0.06 |
| **Marital status(%)  Single** | 4522 | 34.64 (33.3, 36.0) | 2424 | 33.91 (32.2, 35.6) | -0.72 | -0.02 |
| **Married or living with partner** | 8436 | 48.82 (47.3, 50.4) | 5397 | 50.34 (48.4, 52.3) | 1.52 | 0.03 |
| **Separated divorced, or widowed** | 3369 | 16.54 (15.6, 17.5) | 2008 | 15.75 (14.5, 17.0) | -0.80 | -0.05 |
| **Health insurance(%)** | 8172 | 50.54 (48.7, 52.4) | 4936 | 51.86 (49.8, 53.9) | 1.32 | 0.03 |
| **US residence >= 10 Years(%)** | 12490 | 72.34 (70.5, 74.2) | 7548 | 72.32 (70.2, 74.4) | -0.01 | -0.00 |
| **Language preference(%)  Spanish** | 13119 | 74.86 (73.0, 76.7) | 8142 | 75.72 (73.7, 77.7) | 0.86 | 0.01 |
| **English** | 3296 | 25.14 (23.3, 27.0) | 1722 | 24.28 (22.3, 26.3) | -0.86 | -0.03 |
| **Systolic BP (mmHg)** | 16401 | 119.92 (119.4, 120.4) | 9858 | 119.33 (118.8, 119.9) | -0.59 | -0.00 |
| **Diastolic BP (mmHg)** | 16394 | 72.19 (71.9, 72.5) | 9853 | 72.00 (71.6, 72.4) | -0.19 | -0.00 |
| **Hypertension (%)** | 4937 | 24.19 (23.0, 25.4) | 2951 | 23.72 (22.3, 25.1) | -0.47 | -0.02 |
| **Treated for hypertension(%)b** | 3464 | 68.94 (66.8, 71.0) | 2122 | 70.33 (68.0, 72.7) | 1.39 | 0.02 |
| **Total cholesterol(mg/dL)** | 16248 | 194.32 (193.2, 195.4) | 9787 | 194.87 (193.5, 196.3) | 0.55 | 0.00 |
| **LDL-cholesterol(mg/dL)** | 15918 | 119.74 (118.8, 120.7) | 9614 | 120.59 (119.4, 121.8) | 0.84 | 0.01 |
| **HDL-cholesterol(mg/dL)** | 16246 | 48.48 (48.2, 48.8) | 9787 | 48.59 (48.2, 49.0) | 0.10 | 0.00 |
| **eGFR** | 16131 | 106.92 (106.3, 107.5) | 9717 | 107.56 (106.8, 108.3) | 0.64 | 0.01 |
| **Treated for hypercholesterolemia(%)c** | 1629 | 24.36 (22.6, 26.1) | 1186 | 23.65 (21.7, 25.6) | -0.71 | -0.03 |
| **BMI kg/m2** | 16344 | 29.36 (29.2, 29.5) | 9835 | 29.30 (29.1, 29.5) | -0.06 | -0.00 |
| **Obesity Status (%)  Underweight (BMI<18.5 kg/m2)** | 130 | 1.16 (0.9, 1.4) | 62 | 1.20 (0.8, 1.6) | 0.04 | 0.03 |
| **Normal (BMI 18.5-25 kg/m2)** | 3191 | 22.07 (21.1, 23.1) | 1776 | 21.58 (20.3, 22.9) | -0.49 | -0.02 |
| **Overweight (BMI 25-30 kg/m2)** | 6116 | 37.19 (36.0, 38.4) | 3795 | 37.77 (36.3, 39.2) | 0.58 | 0.02 |
| **Obese (BM>=30 kg/m2)** | 6907 | 39.58 (38.3, 40.9) | 4202 | 39.45 (37.8, 41.1) | -0.13 | -0.00 |
| **Fasting glucose(mg/dL)** | 16220 | 102.20 (101.4, 103.0) | 9776 | 102.10 (101.1, 103.1) | -0.10 | -0.00 |
| **Diabetes - definition #2 (%)d** | 3218 | 14.88 (14.1, 15.7) | 1897 | 15.26 (14.2, 16.3) | 0.38 | 0.03 |
| **Diabetes - definition #4 (%)e** | 3227 | 14.85 (14.0, 15.7) | 1904 | 15.19 (14.2, 16.2) | 0.34 | 0.02 |
| **Treated for diabetes(%)f** | 1836 | 53.77 (51.3, 56.2) | 1040 | 51.70 (48.3, 55.1) | -2.07 | -0.04 |
| **Waist circumference (cm)** | 16349 | 97.37 (96.9, 97.8) | 9837 | 97.23 (96.7, 97.7) | -0.13 | -0.00 |
| **Current Smoker (%)** | 3166 | 21.37 (20.3, 22.5) | 1685 | 21.00 (19.6, 22.4) | -0.37 | -0.02 |
| **Asthma (%)** | 2637 | 17.37 (16.4, 18.4) | 1525 | 17.07 (15.8, 18.3) | -0.29 | -0.02 |
| **COPD (%)** | 488 | 2.78 (2.4, 3.1) | 273 | 2.76 (2.3, 3.3) | -0.01 | -0.00 |
| **CVD (%)** | 858 | 4.72 (4.2, 5.2) | 454 | 4.05 (3.5, 4.6) | -0.67 | -0.14 |
| **MI (%)** | 384 | 2.34 (2.0, 2.7) | 201 | 1.86 (1.5, 2.3) | -0.48 | -0.20 |
| **Hearing Loss (%)** | 2799 | 15.06 (14.2, 15.9) | 1609 | 14.10 (13.1, 15.1) | -0.97 | -0.06 |

Abbreviations: BMI: body mass index; BP: blood pressure; LDL: low density lipoprotein; HDL: high density lipoprotein; COPD: chronic obstructive pulmonary disease; CVD: cardiovascular disease; MI: myocardial infarction.

a All values (except N) weighted for study design and non-response.

b Denominator is restricted to participants with hypertension at baseline (Unweighted Visit 1: N=4937, Visit 3: N=2951).

c Denominator is restricted to participants with hypercholesterolemia at baseline (Unweighted Visit 1: N=5332, Visit 3: N=4026).

d ADA guideline plus scanned/transcribed medication use.

e ADA guideline plus self-reported medication use.

f Denominator is restricted to participants with diabetes (ADA guideline plus self-reported diabetes) at baseline (Unweighted Visit 1: N=3384, Visit 3: N=1997).

SOURCE: HC331511 (18SEP24 using INV2 data)

1. Longitudinal Analysis: Introduction

In this chapter, we introduce the groundwork for conducting **longitudinal analysis** with repeated measures for HCHS/SOL data involving more than two clinic visits. We begin by discussing missing visits and related methodologies in longitudinal analysis. Next, we explore data management strategies for longitudinal analysis. Finally, we provide guidance on creating an analytic dataset, including sample code for dataset generation. All examples will utilize data from the first three HCHS/SOL clinic visits.

For how to conduct **longitudinal analysis** for HCHS/SOL datainvolving only two clinic visits, for example, Visit 1 and Visit 3 data only or Visit 2 and Visit 3 data only, focusing on modelling the difference, rate of change, incident event odds ratio, or incidence rate, please refer to *HCHS/SOL Analysis Methods - Visit 2* and use Visit 3 sampling weights.

## Methods for Addressing Missing Visits

Participants missing follow-up visits is a common phenomenon in any longitudinal study. Data for participants who missed follow-up visit(s) will be missing. It can lead to biased estimates and reduced precision if missing visits are not accounted for properly. The missingness mechanism behind missing visits can be grouped into three categories: **Missing Completely at Random (MCAR)**, **Missing at Random (MAR)**, **Missing Not at Random (MNAR)**.

MCAR occurs when the probability of a participant missing a visit is independent of both observed and unobserved data. In other words, a participant missing a visit is a result of completely random events that are unrelated to any participant characteristics or outcomes of interest, regardless of whether they are observed or unobserved. MCAR can be partially verified if no significant differences are found when comparing the characteristics of participants with complete visits to those with missing visits. However, this verification is limited to observed variables and cannot rule out relationships with unobserved data. When MCAR holds, a complete case analysis which drops the missing records and uses only the data from participants who completed all visits, is expected to provide valid inference of the true population parameters. This approach is the default in most statistical software. However, MCAR is a strong assumption that rarely holds in practice. Moreover, using only the complete cases leads to a loss of efficiency (larger standard errors) with the extent of efficiency loss depending on the proportion of missing data.

MAR occurs when the probability of a participant missing a visit depends on observed data, but not on unobserved data. In other words, a participant missing a visit is a result of factors that are related to observed participant characteristics or outcomes of interest, but not to unobserved characteristics or outcomes that would have been collected at a missing visit. When MAR holds, statistical methods that properly account for the observed data associated with missingness can provide valid inference of the true population parameters. MAR is a less stringent assumption than MCAR and is often more plausible in longitudinal studies.

MNAR, also known as informative or non-ignorable missingness, occurs when the probability of a participant missing a visit depends on unobserved data, even after accounting for the observed data. In other words, a participant missing a visit is a result of factors that are related to unobserved participant characteristics or outcomes of interest, including those that would have been collected at a missing visit. When MNAR holds, standard statistical methods, even those that account for observed data, can provide biased inference of the true population parameters. Handling MNAR often requires more complex approaches that jointly model the outcome and missingness process, such as selection models or pattern mixture models. What approach to use depends on the scientific question of interest. MNAR is the most challenging missing data mechanism to address, and its presence cannot be definitively determined from the observed data alone. Therefore, sensitivity analyses are recommended to assess the robustness of findings under different MNAR scenarios.

In HCHS/SOL, the baseline cohort (N=16,415) has been followed over time. About 71% of the original cohort participated in Visit 2 (N=11,623). About 60% of the original cohort participated in Visit 3 (N=9,864), out of which 9,090 participated in the in-person exam and 774 had phone interview only. For participants who did not participate in Visit 2 or/and Visit 3 or dropped out of the study, they are considered as having missing visits. An overview of missing visits with respect to the baseline cohort is presented in two ways (**Output 3.1‑1**): (1) for Visit 3 in-person participation only, and (2) for ALL Visit 3 participation, including those with phone interview only. We assume the missing-visit mechanism is MAR and describe appropriate methods to address this type of missingness in each respective chapter.

**Output 3.1‑1: Missing Visits Overview**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Visit 1** | **Visit 2** | **Visit 3**  **Exam Only** | **N** | **%** | | ✓ |  |  | 4134 | 25.2 | | ✓ | ✓ |  | 3191 | 19.4 | | ✓ |  | ✓ | 658 | 4.0 | | ✓ | ✓ | ✓ | 8432 | 51.4 | | **Sum** | | | **16415** | **100** | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Visit 1** | **Visit 2** | **Visit 3**  **All** | **N** | **%** | | ✓ |  |  | 3905 | 23.8 | | ✓ | ✓ |  | 2646 | 16.1 | | ✓ |  | ✓ | 887 | 5.4 | | ✓ | ✓ | ✓ | 8977 | 54.7 | | **Sum** | | | **16415** | **100** | |

### Multiple Imputation

**Multiple Imputation (MI)** is a widely used strategy to handle missingness in both outcome and covariates, particularly under the MAR assumption. MI involves creating multiple plausible imputed datasets, analyzing each dataset separately, and then combining the results using specific rules, e.g., Rubin's rules (Rubin, 2018). This approach accounts for the uncertainty in the imputed values, leading to valid statistical inferences. For a detailed introduction, please refer to *Flexible Imputation of Missing Data* by Stef van Buuren (van Buuren, 2018).

Within the MI framework, various methods can be used to create the imputed datasets. One popular and flexible method is **Fully Conditional Specification (FCS)**, also known as **Multiple Imputation by Chained Equations (MICE)**. FCS operates through a sequence of univariate imputation models, assuming the existence of a joint distribution for all variables. This approach makes FCS suitable for datasets with arbitrary missing patterns. The method works by imputing missing values on a variable-by-variable basis, using iterative cycles to refine imputations. This process preserves relationships between variables in the imputed data and captures complex interdependencies. FCS can accommodate various types of variables (continuous, binary, categorical) within the same imputation model. Additionally, the method allows for the inclusion of auxiliary variables in the imputation model, potentially improving the quality of imputations.

For the FCS/MICE imputation process, based on the type of variable being imputed, the following regression methods can be used:

* Continuous: Linear regression
* Binary: Logistic regression
* Categorical (ordinal): Ordered logistic regression (proportional odds)
* Categorical (nominal): Multinomial (polytomous) logistic regression

A key point in MI is to appropriately specify the variables related to the missing mechanism in the imputation model under the MAR assumption. Our simulation results showed: when the imputation model is under-specified, the resulting estimates can be biased, and the inference can be invalid; when the imputation model is correctly specified or over-specified, the resulting estimates are approximately unbiased, and the inference is valid.

Because the true missingness model is unknown in practice, the CC recommends including all covariates from the main analytic model, along with any additional variables that may be related to the probability of missingness (e.g., design variables), in the MI procedure.

### Inverse Probability Weighting

**Inverse Probability Weighting (IPW)** is an alternative approach for handling missing visits that, like MI, provides unbiased estimation under the assumption that visit participation is MAR after conditioning on observed data. Whereas MI replaces missing data with plausible values, IPW adjusts the contribution of observed data by up-weighting participants who are similar to those who did not participate in a given visit.

In the HCHS/SOL, the released sampling weights for **Visit 1, Visit 2, and Visit 3** already accounted for nonresponse due to missing the specific visit. For Visit 1, please refer to *HCHS/SOL Analysis Methods at Baseline*. For Visit 2, please refer to *HCHS/SOL Analysis Methods – Visit 2*. For Visit 3, as described in **Section 2.1**, the Visit 3 participation probabilities were estimated using a CART model based on baseline and follow-up characteristics, and their inverses formed the non-response–adjusted weights. These released weights therefore serve as **design-based nonresponse-adjusted weights**, accounting simultaneously for unequal selection probabilities at baseline and for the respective differential visit participation.

When the true missingness mechanism depends solely on categorical or discrete variables that are well represented in the CART-defined strata, using released sampling weights can provide valid finite-population inference for longitudinal analysis. However, CC simulation studies showed that when the true missingness mechanism depends on continuous variables, the CART-based weights alone may not fully capture variation in visit participation, because the tree algorithm partitions the sample into broad cells and can only approximate continuous effects coarsely.

In such cases, an IPW adjustment based on a **Generalized Linear Model (GLM)** can be used as an alternative approach to account for visit-specific nonresponse. In this approach, visit participation is modeled using a GLM with a logistic link, where predictors, including continuous ones, are selected to correctly specify the visit participation mechanism. Separate GLM models are fitted for Visit 2 and Visit 3 to estimate visit-specific probabilities of participation, and the inverses of these estimated predicted probabilities are multiplied by the released Visit 1 sampling weights, which already account for nonresponse at Visit 1, to form the Visit 2 and Visit 3 GLM-based IPW-adjusted weights. This **GLM-based IPW approach** provides an alternative to the CART-based nonresponse adjustment used in the Visit 2 and Visit 3 released sampling weights.

Because the GLM-based IPW approach depends on the covariates specified in the model, complete data on those covariates are required for all participants. It is therefore recommended to impute any missing covariate values using MI before fitting the GLMs to estimate visit-specific response probabilities. The CC recommends that predictors included in the GLMs and subject to imputation have minimal missingness, ideally less than 5%. This same practice was applied in the *HCHS/SOL Baseline Physical Activity Data Overview, Methods & Guidelines*, where missing covariates were multiply imputed prior to constructing the Actical IPW-adjusted weights to account for missing accelerometer data.

Because the true missingness model is unknown in practice, the CC recommends including in the MI model all variables that will be used in the GLMs predicting visit participation, along with any additional variables that may be related to the probability of missingness (e.g., design variables). For the GLM model itself, include covariates that may be associated with visit participation but represent information available prior to the occurrence of missingness.

### MI vs. IPW

Both MI and IPW are valid approaches for handling missing visits under the MAR assumptions, and each has distinct advantages depending on the research objectives.

**The primary focus of MI is on modeling the joint relationships among study variables and fully utilizing all available information. When the imputation model includes sampling weights and design variables, MI produces design-consistent estimates while preserving the relationships among outcomes and covariates across visits. However, MI can be computationally intensive, particularly when applied to large longitudinal datasets or outcomes with substantial missingness. Although simulation studies have shown that MI can still yield reliable and unbiased estimates even when outcome missingness is high, investigators may reasonably be concerned about the plausibility of the imputed values, especially for clinical variables that can have complex or bounded distributions, even if overall parameter estimates remain unbiased.**

**The GLM-based IPW approach relies only on the observed data and requires imputation of covariates (if needed) but not outcomes, thereby avoiding concerns about the plausibility of imputed outcome values. This approach directly models visit** participation **and applies inverse probability weights on top of the released sampling weights to adjust for differential participation, maintaining design-consistent inference. While GLM-based IPW is typically less efficient than MI, because it does not borrow information across participants or time points, it is straightforward to implement, transparent, and more robust when imputation of outcomes is problematic or difficult to justify clinically.**

## Visit 1 Sample vs. Visit 3 Sample

In longitudinal analyses involving multiple clinic visits, two analytic samples may be defined: (1) the **Visit 1 Sample**, which includes all participants enrolled at baseline (N=16,415) and incorporates all available data from any visit; and (2) the **Visit 3 Sample**, which is limited to participants who completed the latest visit (e.g., Visit 3 Exam Only N=9,090, i.e., excluding those with phone interview only) and incorporates all available data from any visit for these participants. Under the HCHS/SOL inferential framework, both analytic samples can be used to estimate parameters for the same finite target population. When missing data are properly addressed using methods appropriate to each analytic sample, analyses based on either sample are expected to yield statistically equivalent estimates of the same finite-population parameters.

In **Output 3.2‑1**, we summarize the extent of missingness for variables in the analytic dataset based on the Visit 1 Sample. In **Output 3.2‑2**, we summarize the extent of missingness for variables in the analytic dataset based on the Visit 3 Sample.

Output 3.2‑1: Visit 1 Sample (N=16,415), Extent of Missingness

|  |  |  |
| --- | --- | --- |
| Variable | # Observed | # Missing |
| Baseline/Visit 1 | | |
| AGEGROUP\_C6 | 16414 | 1 |
| BKGRD1\_C7NOMISS | 16415 | 0 |
| CENTERNUM | 16415 | 0 |
| SEX | 16414 | 1 |
| WEIGHT\_FINAL\_NORM\_OVERALL | 16415 | 0 |
| SBP5\_V1 | 16400 | 15 |
| BMI\_V1 | 16343 | 72 |
| HYPERTENSION2 | 16412 | 3 |
| US\_BORN | 16341 | 74 |
| EMPLOYED | 16108 | 307 |
| EDUCATION\_C3 | 16323 | 92 |
| Visit 2 | | |
| WEIGHT\_NORM\_OVERALL\_V2 | 11623 | 4792 |
| YRS\_BTWN\_V1V2 | 11623 | 4792 |
| SBP5\_V2 | 11591 | 4824 |
| BMI\_V2 | 11245 | 5170 |
| HYPERTENSION2\_V2 | 11620 | 4795 |
| Visit 3 | | |
| WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 | 9090 | 7325 |
| YRS\_BTWN\_V1V3 | 9864 | 6551 |
| SBP5\_V3 | 9046 | 7369 |
| BMI\_V3 | 8758 | 7657 |
| HYPERTENSION2\_V3 | 9087 | 7328 |

Output 3.2‑2: Visit 3 Sample (N=9,090), Extent of Missingness

|  |  |  |
| --- | --- | --- |
| Variable | # Observed | # Missing |
| Baseline/Visit 1 | | |
| AGEGROUP\_C6 | 9089 | 1 |
| BKGRD1\_C7NOMISS | 9090 | 0 |
| CENTERNUM | 9090 | 0 |
| SEX | 9089 | 1 |
| WEIGHT\_FINAL\_NORM\_OVERALL | 9090 | 0 |
| SBP5\_V1 | 9084 | 6 |
| BMI\_V1 | 9063 | 27 |
| HYPERTENSION2 | 9089 | 1 |
| US\_BORN | 9070 | 20 |
| EMPLOYED | 8978 | 112 |
| EDUCATION\_C3 | 9057 | 33 |
| Visit 2 | | |
| WEIGHT\_NORM\_OVERALL\_V2 | 8432 | 658 |
| YRS\_BTWN\_V1V2 | 8432 | 658 |
| SBP5\_V2 | 8416 | 674 |
| BMI\_V2 | 8246 | 844 |
| HYPERTENSION2\_V2 | 8432 | 658 |
| Visit 3 | | |
| WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 | 9090 | 0 |
| YRS\_BTWN\_V1V3 | 9090 | 0 |
| SBP5\_V3 | 9046 | 44 |
| BMI\_V3 | 8758 | 332 |
| HYPERTENSION2\_V3 | 9087 | 3 |

## Recommendations for Marginal Approaches

Assuming the missing-visit mechanism is MAR, the CC recommends the following marginal (population-averaged) approaches for the analysis of longitudinal HCHS/SOL data. These approaches are organized by outcome type, visit sample, inferential framework, software procedure, and method for addressing missing visits. These recommendations are informed by extensive simulation studies carried out by the CC, the full results of which will be presented in a separate report.

#### Continuous Outcomes

If using the Visit 1 Sample (N = 16,415):

* Complex-Survey GEE with MI
* Model-Based GEE with MI
* Complex-Survey GEE with GLM-Based IPW
* Model-Based GEE with GLM-Based IPW

If using the Visit 3 Sample, excluding those with phone interview only (N = 9,090):

* Complex-Survey GEE with GLM-Based IPW
* Model-Based GEE with GLM-Based IPW

#### Binary Outcomes

If using the Visit 1 Sample (N = 16,415):

* Complex-Survey GEE with GLM-Based IPW
* Model-Based GEE with GLM-Based IPW

If using the Visit 3 Sample, excluding those with phone interview only (N = 9,090):

* Complex-Survey GEE with GLM-Based IPW
* Model-Based GEE with GLM-Based IPW

The next two subsections summarize the general procedures underlying these recommendations, and **Section 4** provides corresponding sample code for implementation in the illustrative analytic example.

### General Procedure for GEE with MI

**Step 1 (Impute):** Generate *m* imputed datasets from the wide-format analytic dataset using FCS/MICE; Impute each variable (outcome and covariates) with missing values that appears in the main model of interest.

**Step 2 (Transform then Fit):** Transform each imputed dataset from wide to long-format; apply GEE to each transformed dataset for the longitudinal outcome.

**Step 3 (Combine):** Combine the results from the *m* separate analyses using Rubin's rules to obtain final estimates and standard errors, accounting for variability both within and between the imputed datasets.

The CC recommends using m = 10 imputations. The imputation model should include all variables from the main analytic model, along with any additional variables that may be related to the probability of missing a clinic visit, even if they are not part of the main model. Design variables should also be included, as they capture key aspects of the sampling design and may be associated with visit participation. Performing imputation in the wide format helps preserve relationships among variables across visits, providing a more comprehensive representation of the longitudinal structure and maintaining the temporal dependencies and correlations between measurements at different time points.

### General Procedure for GEE with GLM-based IPW

**Step 1 (Impute):** Generate *m* imputed datasets from the wide-format analytic dataset using FCS/MICE; Impute all variables to be used in the IPW model if they have any missingness.

**Step 2 (Estimate IPW):** Fit a GLM within each imputed dataset to estimate the probability of participating in each visit that participants may miss (e.g., Visit 2 or Visit 3). After fitting the GLM in each imputed dataset, average the **linear predictors (logits)** across all m imputations for each participant to obtain a single pooled mean linear predictor. Transform this pooled linear predictor to the probability scale using the logistic function:

This predicted probability represents each participant’s estimated likelihood of participating in a specific visit. The inverse of this probability is then multiplied by the released visit-specific sampling weight to create the GLM-based IPW-adjusted weight:

The resulting IPW-adjusted weights are then merged back into the wide-format analytic dataset for subsequent steps.

**Step 3 (Transform then Fit):** Transform the wide-format dataset to long format and apply GEE to the updated dataset for the longitudinal outcome using the GLM-based IPW-adjusted weights.

The CC recommends using m = 10 imputations and applying appropriate imputation methods for continuous, binary, and categorical variables. The purpose of the imputation in this approach differs from that described in **Section 3.1.1**. Here, imputation is performed only for covariates used as predictors in the IPW model, rather than for all variables in the main analytic model (including outcomes). This targeted imputation ensures complete data for estimating visit-participation probabilities and allows all participants to contribute to the estimation of participation probabilities. The IPW model should include baseline covariates that are associated with both visit participation and study outcomes, while excluding sampling weights as predictors. In addition, variables measured after the occurrence of missingness should be omitted to avoid conditioning on factors that may themselves be influenced by non-participation.

## Analytic Dataset

The following code generates the analytic dataset "sol\_wide.sas7bdat", a wide-format SAS dataset with all participants from the baseline cohort (N=16,415). This dataset will be used for the illustrative analytic examples. This dataset is created by importing variables needed for the examples from relevant investigator files (e.g., blood pressure measurements from “sbp” files) from each visit, renaming some with visit-specific suffixes (e.g., \_V1, \_V2, \_V3) to accommodate the wide format. Because the analytic examples are based on measures from the in-person exam, WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3, the Visit 3 sampling weights for Exam Only participants, is imported.

The modified 7-level reclassification of Hispanic/Latino background, BKGRD1\_C7\_NOMISS, is created to incorporate missing data into the "Mixed/Others" category. The binary indicator for Visit 2 participants, PARTICIPANT\_V2, is created. The binary indicator for Visit 3 participants with in-person exam component, PARTICIPANT\_EXAMONLY\_V3, is merged from the Visit 3 dataset, with non–Visit 3 participants coded 0 (missing recoded to 0).

A list of variables in the analytic dataset is presented in **Output 3.4‑1.**

%let v1\_inv = inv5; /\* version of V1 INV file \*/

%let v2\_inv = inv3; /\* version of V2 INV file \*/

%let v3\_inv = inv2; /\* version of V3 INV file \*/

/\* Visit 1 \*/

data analys\_v1 (rename = (BMI = BMI\_V1 SBPA5 = SBP5\_V1));

merge part\_derv\_&v1\_inv. sbpa\_&v1\_inv.;

by ID;

keep PSU\_ID HH\_ID ID STRAT WEIGHT\_FINAL\_NORM\_OVERALL

CENTERNUM SEX BKGRD1\_C7 AGEGROUP\_C6

US\_BORN EMPLOYED EDUCATION\_C3 BMI SBPA5 HYPERTENSION2;

**run**;

/\* Visit 2 \*/

data analys\_v2 (rename = (SBP5 = SBP5\_V2));

merge part\_derv\_v2\_&v2\_inv. sbp\_v2\_&v2\_inv.;

by ID;

keep ID WEIGHT\_NORM\_OVERALL\_V2

BMI\_V2 YRS\_BTWN\_V1V2 SBP5 HYPERTENSION2\_V2;

**run**;

/\* Visit 3 \*/

data analys\_v3 (rename = (SBP5 = SBP5\_V3));

merge part\_derv\_v3\_&v3\_inv. sbp\_v3\_&v3\_inv.;

by ID;

keep ID WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 PARTICIPANT\_EXAMONLY\_V3

BMI\_V3 YRS\_BTWN\_V1V3 SBP5 HYPERTENSION2\_V3;

**run**;

/\* Analytic Dataset (wide-fomrat) \*/

data sol\_wide;

merge analys\_v1 analys\_v2 analys\_v3;

by ID;

/\* recoded background \*/

BKGRD1\_C7NOMISS = BKGRD1\_C7;

if BKGRD1\_C7NOMISS < **.Z** then BKGRD1\_C7NOMISS = **6**;

drop BKGRD1\_C7;

/\* V2 participant indicator \*/

if WEIGHT\_NORM\_OVERALL\_V2 < **.Z** then PARTICIPANT\_V2 = **0**;

else PARTICIPANT\_V2 = **1**;

/\* Set missing to 0 for non-V3 participants \*/

if missing(PARTICIPANT\_EXAMONLY\_V3) then PARTICIPANT\_EXAMONLY\_V3 = **0**;

**run**;

Output 3.4‑1: Variables in the Analytic Dataset

|  |  |
| --- | --- |
| Variable | Description |
| Design | |
| PSU\_ID | Primary Sampling Unit (Block Group) ID |
| STRAT | Stratification Variable ID |
| HH\_ID | Secondary Sampling Unit (Household) ID |
| ID | Participant ID |
| Baseline/Visit 1 | |
| AGEGROUP\_C6 | Age Groups, Visit 1:  1=Ages 18-24, 2=Ages 25-34, 3=Ages 35-44, 4=Ages 45-54, 5=Ages 55-64, 6=Ages 65+ |
| BKGRD1\_C7NOMISS | Hispanic/Latino Background, Visit 1:  0=Dominican, 1=Central American, 2=Cuban, 3=Mexican, 4=Puerto-Rican, 5=South American, 6=More than one heritage/Other, DK/Refused, Missing |
| CENTERNUM | Participant's Field Center, Visit 1:  1=Bronx, 2=Chicago, 3=Miami, 4=San Diego |
| SEX | Sex, Visit 1:  0=Female, 1=Male |
| WEIGHT\_FINAL\_NORM\_OVERALL | Overall Sampling Weights, Visit 1 |
| SBP5\_V1 | Average Systolic (mm Hg), Visit 1 |
| BMI\_V1 | BMI (kg/m2), Visit 1 |
| HYPERTENSION2 | Hypertension using NHANES definition, Visit 1:  0=No, 1=Yes |
| US\_BORN | Born in mainland US, Visit 1:  0=Not born in 50 US States/DC, 1=Born in 50 US States/DC Only |
| EMPLOYED | Employment Status, Visit 1:  1=Retired and not currently employed, 2=Not retired and not currently employed, 3=Employed part-time (<=35 hours/week), 4=Employed full-time (>35 hours/week) |
| EDUCATION\_C3 | Education Status, Visit 1:  1=Less Than High School, 2=High School or Equivalent, 3=Greater than High School or Equivalent |
| Visit 2 | |
| PARTICIPANT\_V2 | Visit 2 participants Indicator |
| WEIGHT\_NORM\_OVERALL\_V2 | Overall Sampling Weights, Visit 2 |
| YRS\_BTWN\_V1V2 | Elapsed time between visits 1 and 2 (years) |
| SBP5\_V2 | Average Systolic (mm Hg), Visit 2 |
| BMI\_V2 | BMI (kg/m2), Visit 2 |
| HYPERTENSION2\_V2 | Hypertension using NHANES definition, Visit 2:  0=No, 1=Yes |
| Visit 3 | |
| PARTICIPANT\_EXAMONLY\_V3 | Visit 3 participants with in-person exam Indicator |
| WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 | Overall Sampling Weights, excluding those with phone interview only, Visit 3 |
| YRS\_BTWN\_V1V3 | Elapsed time between visits 1 and 3 (years) |
| SBP5\_V3 | Average Systolic (mm Hg), Visit 3 |
| BMI\_V3 | BMI (kg/m2), Visit 3 |
| HYPERTENSION2\_V3 | Hypertension using NHANES definition, Visit 3:  0=No, 1=Yes |

## Data Management: wide-format and long-format

For longitudinal data, there are two ways to format the data for analysis, wide-format and long-format. In the **wide-format data**, each participant has one record with separate variables for repeated measures at each follow-up visit. For example, BMI measurements at Visits 1, 2, 3 would be represented as three distinct variables: BMI\_V1, BMI\_V2, and BMI\_V3. In contrast, in the **long-format data** there is only one variable with the measurement (BMI) and a variable to identify the clinic visit (VISIT), and there are multiple records per participant, one for each visit. For example, a participant would have one record for BMI at Visit 1, another record for BMI at Visit 2, and a third record for BMI at Visit 3.

1. Longitudinal Analysis of Continuous Outcomes

In this chapter, we outline the recommended methods for conducting longitudinal analysis of HCHS/SOL data with repeated measures involving more than two clinic visits, focusing on modeling a continuous outcome over time. To illustrate the proposed methods, systolic blood pressure is used as an example, with sample code provided in SAS/SUDAAN, Stata, and R.

## Illustrative Analytic Example

### Model Specification and Covariates

As an analytic example for illustration, we define the main model of interest as a longitudinal analysis examining the effect of time-varying BMI on systolic blood pressure over time across the three clinic visits (long-format: SBP5; wide-format: SBP5\_V1, SBP5\_V2, SBP5\_V3) in the HCHS/SOL target population. The primary predictor of interest is BMI over time across the three clinic visits (long-format: BMI; wide-format: BMI\_V1, BMI\_V2, BMI\_V3), while adjusting for the following covariates:

* Baseline demographic factors: 6-level age group (AGEGROUP\_C6), 7-level re-classification of Hispanic/Latino background (BKGRD1\_C7NOMISS), field center (CENTERNUM), sex (SEX), US-born status (US\_BORN), 4-level employment status (EMPLOYED), and 3-level education level (EDUCATION\_C3)
* Time-related factor: years elapsed from Visit 1 (long-format: TIME; wide-format: YRS\_BTWN\_V1V2, YRS\_BTWN\_V1V3)

Formulaically, the main model of interest is:

,

where is the link function appropriate for the distribution of (e.g., use identity link for the continuous outcome) for participant at visit (for covariates only at baseline, is omitted).

The coefficient for the TIME variable would be interpreted as: On average, among individuals with the same values of the covariates included in the model, the expected systolic blood pressure increased by mmHg for each year that elapsed since Visit 1, or equivalently, for each year of aging.

### Implementation of MI

Following the procedure in **Section 3.3.1**, **Step 1**: Use FCS/MICE to generate *10* imputed datasets from the wide-format analytic dataset "sol\_wide"; if using the Visit 3 Sample, first subset to participants who completed Visit 3 Exam-Only component, i.e., PARTICIPANT\_EXAMONLY\_V3 = 1. Impute each variable (outcome and covariates) with missing values using the following FCSregressions:

* Linear regression: SBP5\_V1, BMI\_V1, YRS\_BTWN\_V1V2, SBP5\_V2, BMI\_V2, YRS\_BTWN\_V1V3, BMI\_V3, SBP5\_V3
* Binary logistic regression: US\_BORN, SEX
* Ordered logistic regression (proportional odds): EMPLOYED, AGEGROUP\_C6
* Multinomial (polytomous) logistic regression: EDUCATION\_C3

Covariates without missing values but included in the main model (e.g., BKGRD1\_C7NOMISS) are also specified in the MI process to preserve their associations with other variables. In addition, the imputation model includes the following design variables: center (CENTERNUM), the Visit 1 overall sampling weights (WEIGHT\_FINAL\_NORM\_OVERALL), Visit 2 overall sampling weights (WEIGHT\_NORM\_OVERALL\_V2), and Visit 3 overall sampling weights for clinic or home exams only (WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3).

### Implementation of GLM-based IPW

Following the procedure in **Section 3.3.2**, **Step 1**: Use FCS/MICE to generate *10* imputed datasets from the wide-format analytic dataset "sol\_wide"; if using the Visit 3 Sample, first subset to participants who completed Visit 3 Exam-Only component, i.e., PARTICIPANT\_EXAMONLY\_V3 = 1. Impute all variables with missing values that will be used as predictors in the IPW models. In the analytic example, only baseline covariates that are not time-varying are included so that the same set of IPW-adjusted weights can be applied to both the Visit 1 and Visit 3 samples for simplicity. The following FCS regression models are used:

* Binary logistic regression: US\_BORN, SEX
* Ordered logistic regression (proportional odds): EMPLOYED, AGEGROUP\_C6
* Multinomial (polytomous) logistic regression: EDUCATION\_C3

Covariates without missing values but included in the main model (e.g., BKGRD1\_C7NOMISS) are also specified in the MI process to preserve their associations with other variables. In addition, the imputation model includes the following design variables: center (CENTERNUM), and the Visit 1 overall sampling weights (WEIGHT\_FINAL\_NORM\_OVERALL).

**Step 2**: Within each imputed dataset, fit a logistic regression model to estimate the probability of participating in Visit 2 (PARTICIPANT\_V2) among all participants. Consistent with the imputation model, the same set of baseline covariates that are not time-varying are included as predictors. For each participant, average the fitted linear predictors across imputations and transform this pooled value to the probability scale to obtain predicted probability of Visit 2 participation (RR\_V2). Analogously, within each imputed dataset, fit a logistic regression model to estimate the probability of participating in the Visit 3 Exam-Only component (PARTICIPANT\_EXAMONLY\_V3), using the same set of baseline covariates as in the Visit 2 model, plus Visit 2 participation status (PARTICIPANT\_V2) to capture the sequential visit process. Average the fitted linear predictors across imputations and transform them to the probability scale to obtain the predicted probability of Visit 3 participation (RR\_V3). As noted in **Output 3.2‑1** and **Output 3.2‑2**, all variables imputed in this analytic example have less than 5% missingness.

Next, combine the pooled Visit 2 and Visit 3 response probabilities with the released Visit 1 sampling weights to construct the Visit 2 and Visit 3 GLM-based IPW-adjusted weights.

For Visit 2, assign the IPW-adjusted weight to participants who participated in the visit (PARTICIPANT\_V2 = 1) as

where WEIGHT\_FINAL\_NORM\_OVERALL is the released Visit 1 sampling weight, and RR\_V2 is the estimated probability of participation at Visit 2 obtained from the GLM-based model.

For Visit 3 (Exam-Only), assign the IPW-adjusted weight to participants who participated in the in-person exam (PARTICIPANT\_EXAMONLY\_V3 = 1) as

where RR\_V3 is the estimated probability of participation at Visit 3 (Exam-Only) obtained from the GLM-based model.

Participants who did not participate in a given visit do not have the corresponding visit-specific IPW-adjusted weight (i.e., missing).

### Choice and Specification of MI and IPW Models

In the analytic example, a single FCS/MICE imputation specification is used to jointly impute all variables included in the MI procedure. Similarly, in the GLM-based IPW procedure, a single FCS/MICE imputation specification is used to impute predictors of visit participation, and the same set of baseline covariates is included in the logistic regressions for estimating participation at both Visit 2 and Visit 3. The Visit 3 model additionally includes Visit 2 participation to account for the sequential nature of visit participation. This unified approach is adopted for simplicity and to maintain consistency across the example, allowing the illustration to focus on the overall analytic workflow rather than on model-specific variations.

In practice, investigators may tailor the MI and IPW model specifications based on the analytic objectives, data structure, and observed patterns of missingness. For instance, the set of covariates included in the IPW models may differ across visits to more accurately capture visit-specific factors influencing participation. The overarching goal is to specify imputation and weighting models that adequately reflect the mechanisms driving missingness and visit non- participation.

## SAS/SUDAAN

In this section, we provide sample code for the illustrative analytic example implementing both GEE with MI and GEE with GLM-based IPW in SAS and SAS-callable SUDAAN, for both the Visit 1 Sample and the Visit 3 Sample. The MI and GLM-based IPW procedures are identical across implementations, using the same FCS/MICE imputation specification for the MI approach, and the same combination of FCS/MICE imputation specification and IPW model specification for the GLM-based IPW approach. Subsequently, SUDAAN procedures are used for the design-based complex-survey analyses that account for the complex HCHS/SOL sampling design, while the corresponding SAS procedures illustrate the alternative model-based (non-survey) implementation.

In the following subsections, we first describe the common MI and GLM-based IPW procedures implemented in SAS, followed by the recommended analytic methods organized by sample type, inferential framework, modeling approach, and strategy for addressing missing visits.

### MI

This section describes the shared MI procedure common to all relevant implementations presented in **Section 4.2**. The same FCS/MICE imputation specification applied to the wide-format data, followed by the same transformation from wide to long format, is used across both the SAS and SAS-callable SUDAAN implementations, as well as for both the Visit 1 and Visit 3 samples.

**proc** **mi** data=sol\_wide nimpute=10 seed=**2024** out=sol\_mi\_wide;

class AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX

US\_BORN EMPLOYED EDUCATION\_C3;

var AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX

US\_BORN EMPLOYED EDUCATION\_C3

WEIGHT\_FINAL\_NORM\_OVERALL SBP5\_V1 BMI\_V1

WEIGHT\_NORM\_OVERALL\_V2 YRS\_BTWN\_V1V2 SBP5\_V2 BMI\_V2

WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 YRS\_BTWN\_V1V3 SBP5\_V3 BMI\_V3; fcs reg(SBP5\_V1 BMI\_V1

WEIGHT\_NORM\_OVERALL\_V2 YRS\_BTWN\_V1V2 SBP5\_V2 BMI\_V2

WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 YRS\_BTWN\_V1V3 SBP5\_V3 BMI\_V3);

fcs logistic(US\_BORN SEX EMPLOYED AGEGROUP\_C6 /\* link=logit\*/);

fcs logistic(EDUCATION\_C3 / link=glogit);

**run**;

The procedure **proc** **mi** performs MI. The nimpute option specifies the number of imputations. The seed option sets a random seed for reproducibility (i.e., obtain the same results every time the code is run). The out option outputs sol\_mi\_wide, a single dataset that contains all the imputed data stacked, containing an imputation number identifier *\_*IMPUTATION\_= 1, 2,…10automatically generated by SAS.

The class statement specifies the categorical variables. The var statement specifies all variables to be used in the imputation model. The fcs statement specifies the following FCS regressions: reg, linear regression for continuous variables; logistic (with the default logit link), binary logistic regression for binary variables and ordered logistic regression for ordinal variables; logistic specifying link=glogit, multinomial logistic regression for nominal variables.

/\* Reshape imputed wide data to long (VISIT-level rows) \*/

**data** sol\_mi\_long;

set sol\_mi\_wide;

/\* Visit 1 \*/

VISIT = **1**;

SBP5 = SBP5\_V1;

BMI = BMI\_V1;

TIME = **0**;

PARTICIPANT\_BY\_VISIT = **1**; /\* everyone observed at V1 \*/

output;

/\* Visit 2 \*/

VISIT = **2**;

SBP5 = SBP5\_V2;

BMI = BMI\_V2;

TIME = YRS\_BTWN\_V1V2;

PARTICIPANT\_BY\_VISIT = PARTICIPANT\_V2;

output;

/\* Visit 3 (Exam-only definition) \*/

VISIT = **3**;

SBP5 = SBP5\_V3;

BMI = BMI\_V3;

TIME = YRS\_BTWN\_V1V3;

PARTICIPANT\_BY\_VISIT = PARTICIPANT\_EXAMONLY\_V3;

output;

**run**;

For Step 2, the data step transforms the wide-format imputed dataset sol\_mi\_wide (16415\*10 observations because of 10 imputed files) into long format sol\_mi\_long (16415\*10\*3 observations because of 10 imputed files and 3 visits) by assigning the visit-specific variables to their generic long-format versions and creates an indicator variable VISIT to indicate to which visit an observation belongs. For Visit 1, TIME is set to 0, and the participation indication PARTICIPANT\_BY\_VISIT is set to 1.

### GLM-based IPW

This section describes the IPW procedure common to all relevant implementations presented in **Section 4.2**. This approach applies MI to covariates used as predictors in the IPW models, followed by logistic regressions to estimate visit participation probabilities and calculate inverse-probability weights. The same FCS/MICE imputation specification and visit-specific modeling steps are implemented consistently across both the SAS and SAS-callable SUDAAN versions and for both the Visit 1 and Visit 3 samples.

/\* = MI on baseline covariates only = \*/

**proc** **mi** data=sol\_wide nimpute=**10** seed=**2024** out=sol\_mi\_for\_ipw; /\* keep in WORK \*/

class AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX US\_BORN EMPLOYED EDUCATION\_C3;

var AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX

US\_BORN EMPLOYED EDUCATION\_C3 WEIGHT\_FINAL\_NORM\_OVERALL;

fcs logistic(US\_BORN SEX EMPLOYED AGEGROUP\_C6/\* link=logit \*/);

fcs logistic(EDUCATION\_C3 / link=glogit);

**run**;

/\* = Visit 2: logistic on PARTICIPANT\_V2 with baseline covariates = \*/

**proc** **logistic** data=sol\_mi\_for\_ipw descending noprint;

by \_Imputation\_;

class AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX

US\_BORN EMPLOYED EDUCATION\_C3 PARTICIPANT\_V2;

model PARTICIPANT\_V2 =

AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX

US\_BORN EMPLOYED EDUCATION\_C3 WEIGHT\_FINAL\_NORM\_OVERALL;

output out=pred\_v2\_imp(keep=\_Imputation\_ ID xb\_v2) xbeta=xb\_v2;

**run**;

**proc** **means** data=pred\_v2\_imp nway noprint;

class ID;

var xb\_v2;

output out=pred\_v2\_bar(drop=\_type\_ \_freq\_) mean=xb\_v2\_bar;

**run**;

/\* = Visit 3: logistic on PARTICIPANT\_EXAMONLY\_V3 with baseline + V2 = \*/

**proc** **logistic** data=sol\_mi\_for\_ipw descending noprint;

by \_Imputation\_;

class AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX

US\_BORN EMPLOYED EDUCATION\_C3 PARTICIPANT\_V2 PARTICIPANT\_EXAMONLY\_V3;

model PARTICIPANT\_EXAMONLY\_V3 =

AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX

US\_BORN EMPLOYED EDUCATION\_C3 WEIGHT\_FINAL\_NORM\_OVERALL

PARTICIPANT\_V2;

output out=pred\_v3\_imp(keep=\_Imputation\_ ID xb\_v3) xbeta=xb\_v3;

**run**;

**proc** **means** data=pred\_v3\_imp nway noprint;

class ID;

var xb\_v3;

output out=pred\_v3\_bar(drop=\_type\_ \_freq\_) mean=xb\_v3\_bar;

**run**;

/\* = Merge RRs back to sol\_wide and compute IPWs = \*/

**proc** **sort** data=sol\_wide; by ID; **run**;

**proc** **sort** data=pred\_v2\_bar; by ID; **run**;

**proc** **sort** data=pred\_v3\_bar; by ID; **run**;

**data** sol\_ipw\_wide;

merge sol\_wide(in=a)

pred\_v2\_bar(rename=(xb\_v2\_bar=lp\_v2))

pred\_v3\_bar(rename=(xb\_v3\_bar=lp\_v3));

by ID;

if a;

/\* Predicted response rates \*/

if not missing(lp\_v2) then RR\_V2 = **1**/(**1**+exp(-lp\_v2));

if not missing(lp\_v3) then RR\_V3 = **1**/(**1**+exp(-lp\_v3));

/\* IPW-adjusted weights (only for observed attendees) \*/

if PARTICIPANT\_V2=**1** then

WEIGHT\_IPW\_V2 = WEIGHT\_FINAL\_NORM\_OVERALL / RR\_V2;

else WEIGHT\_IPW\_V2 = **.**;

if PARTICIPANT\_EXAMONLY\_V3=**1** then

WEIGHT\_EXAMONLY\_IPW\_V3 = WEIGHT\_FINAL\_NORM\_OVERALL / RR\_V3;

else WEIGHT\_EXAMONLY\_IPW\_V3 = **.**;

**run**;

/\* = reshape + visit-specific weight= \*/

/\* Starting point: WORK.SOL\_IPW\_WIDE from Step 2 merge \*/

**data** sol\_ipw\_long;

set sol\_ipw\_wide;

length VISIT **8** TIME **8** SBP5 **8** BMI **8** WEIGHT\_IPW\_BY\_VISIT **8** PARTICIPANT\_BY\_VISIT **8**;

/\* Visit 1 row \*/

VISIT = **1**;

SBP5 = SBP5\_V1;

BMI = BMI\_V1;

TIME = **0**;

WEIGHT\_IPW\_BY\_VISIT = WEIGHT\_FINAL\_NORM\_OVERALL; /\* baseline weight \*/

PARTICIPANT\_BY\_VISIT = **1**; /\* everyone observed at V1 \*/

output;

/\* Visit 2 row \*/

VISIT = **2**;

SBP5 = SBP5\_V2;

BMI = BMI\_V2;

TIME = YRS\_BTWN\_V1V2;

WEIGHT\_IPW\_BY\_VISIT = WEIGHT\_IPW\_V2; /\* IPW-adjusted V2 weight \*/

PARTICIPANT\_BY\_VISIT = PARTICIPANT\_V2;

output;

/\* Visit 3 row (Exam-only definition) \*/

VISIT = **3**;

SBP5 = SBP5\_V3;

BMI = BMI\_V3;

TIME = YRS\_BTWN\_V1V3;

WEIGHT\_IPW\_BY\_VISIT = WEIGHT\_EXAMONLY\_IPW\_V3; /\* IPW-adjusted V3 (exam-only) \*/

PARTICIPANT\_BY\_VISIT = PARTICIPANT\_EXAMONLY\_V3;

output;

**run**;

### Application to Visit 1 Sample

As described in **Section 3.2**, the **Visit 1 Sample** includes all participants enrolled at baseline (N=16,415) and incorporates all available data from any visit, regardless of whether a participant participated in later clinic visits.

#### Complex-Survey GEE with MI

Design based estimate are performed in SUDAAN, but it does not have a native BY statement. Therefore, getting combined estimates from multiple imputed datasets involves looping over each dataset and using the MIANALYZE procedure to generate the final pooled values. This is the reason for the macro GEE\_MI called below to proceed with the analysis of the continuous response.

%include "GEE\_MI.sas";

\* Convert cluster ID to a numerical variable;

**data** data;

set sol\_ipw\_long;

hh\_id\_num=input(substr(hh\_id, **2**),**8.**);

**run**;

\* Call the GEE\_MI macro ;

%***GEE\_MI***(data=data,

strata=strat,

psu=hh\_id\_num,

wt=weight\_final\_norm\_overall,

response=sbp5,

covars=bmi agegroup\_c6 bkgrd1\_c7nomiss centernum sex us\_born employed education\_c3 time,

class= agegroup\_c6 bkgrd1\_c7nomiss centernum sex us\_born employed education\_c3,

class\_ref= agegroup\_c6=**6** bkgrd1\_c7nomiss=**3** centernum=**4** sex=**0** us\_born=**0** employed=**1** education\_c3=**1**);

The **GEE\_MI** macro produces estimates for both continuous and binary responses and requires an input dataset (data) containing the \_imputation\_ index as column, along with design variables specifying stratification (strata), primary sampling units (psu), and sampling weights (wt). The user defines the continuous dependent variable (response), independent covariates (covars), and categorical predictors (class) with their respective reference categories (class\_ref). The optional argument nimpute (default = 10) indicates the number of imputations to perform. For each imputed dataset, the macro fits the model, extracts estimates, and produces pooled coefficients and standard errors that account for both within- and between-imputation variability. Note that prior to executing the GEE\_MI macro, the PSU information needs to be transformed into a numerical variable to prevent issues with the SUDAAN procedure.

Within the **GEE\_MI** macro, and for the continuous response, the PROC REGRESS procedure is employed to fit a generalized estimating equations (GEE) model that accounts for the complex survey design. The option filetype=SAS enables SUDAAN to read and write standard SAS datasets. The correlation structure is specified as r=independent. Robust variance estimation is conducted using semethod=zeger, applying the Zeger “sandwich” estimator to ensure consistent standard errors even if the working correlation model is incorrect. The NOTSORTED option eliminates the need for prior sorting by design variables. Finally, the statement

output beta sebeta / filename=<filename> filetype=sas replace;

saves the estimated regression coefficients and standard errors into a SAS dataset for subsequent multiple-imputation pooling and analysis and replace it if already exists. The macro requires to be saved in SAS format and it will contain the code presented below.

**%macro** GEE\_MI(data, strata, psu, wt, response, covars, class, class\_ref, nimpute=**10**);

\* Print the consolidated dataset;

proc sql noprint;

select count(distinct &response.) into: n\_cats trimmed

from &data. where &response is not missing;

quit;

\* Print the consolidated dataset;

%do j=**1** %to &nimpute.;

data db; set &data.(where=(\_imputation\_ = &j.)); run;

%if &n\_cats = **1** %then %do;

%put ERROR: Review the response variable before continuing.; %abort cancel;

%end;

%else %if &n\_cats. = **2** %then proc rlogist;

%else %if &n\_cats. > **2** %then proc regress;

data=db filetype=sas r=independent semethod=zeger notsorted;

nest &strata. &psu.;

weight &wt.;

class &class.;

model &response.=&covars.;

reflevel &class\_ref.;

output beta sebeta / filename=est\_mi\_&j. filetype=sas replace;

run;

data betas\_mi\_&j.(rename=(beta=Estimate sebeta=StdErr));

set est\_mi\_&j.;

\_imputation\_=&j.; Parm = cats('V\_',modelrhs);

run;

%end;

\* Merge the betas and se across imputation;

data outparms; set betas\_mi\_:; run;

proc sql noprint; select max(modelrhs) into:maxrhs from outparms; quit;

%let vlist=;

%do i=**1** %to &maxrhs.;

%let vlist=&vlist. V\_&i.;

%end;

data outparms; set outparms(drop=modelrhs procnum modelno); run;

proc sort data=outparms; by \_imputation\_; run;

\* Combine estimates using Rubin’s rules;

proc mianalyze parms=outparms;

modeleffects &vlist.;

ods output ParameterEstimates=betas\_mi(keep=Parm Estimate StdErr tValue Probt);

run;

proc datasets library=work nolist; delete outparms; quit;

\* Create table with variable name, labels and merge;

%***var\_levels***(data=&data., covars=&covars., class=&class.);

proc sort data = betas\_mi; by Parm; run;

proc sort data = data\_out; by Parm; run;

data betas\_out(drop=Parm); merge data\_out betas\_mi; by Parm; run;

proc sort data = betas\_out; by Variable ClassVal; run;

\* Print the consolidated dataset;

title 'Pooled Beta Estimates using SUDAAN';

proc print data = betas\_out nobs; run;

**%mend** GEE\_MI;

The output XXXX contains the results from the macro **GEE\_MI**. Observe that the estimate for BMI is 0.0197 with a standard error of 0.025. This positive coefficient indicates that, on average, among individuals with the same covariates included in the model and the same amount of time since Visit 1, every 1-unit increase in BMI within a person is associated with a 0.197-unit increase in SBP5. This effect is statistically significant (p < .0001).

Output XXXX: SUDAAN, Parameter Estimates from GEE with MI

Variable ClassVal Estimate StdErr tValue Probt

INTERCEPT . 128.113806 1.162781 110.18 <.0001

AGEGROUP\_C6 1 -27.237640 0.891231 -30.56 <.0001

AGEGROUP\_C6 2 -24.900034 0.878139 -28.36 <.0001

AGEGROUP\_C6 3 -19.653613 0.904229 -21.74 <.0001

AGEGROUP\_C6 4 -13.298408 0.906271 -14.67 <.0001

AGEGROUP\_C6 5 -7.093930 0.834566 -8.50 <.0001

AGEGROUP\_C6 6 0 0 . .

BKGRD1\_C7NOMISS 0 2.309152 0.779266 2.96 0.0035

BKGRD1\_C7NOMISS 1 1.559880 0.633601 2.46 0.0144

BKGRD1\_C7NOMISS 2 0.407996 0.742664 0.55 0.5834

BKGRD1\_C7NOMISS 3 0 0 . .

BKGRD1\_C7NOMISS 4 0.869791 0.638490 1.36 0.1741

BKGRD1\_C7NOMISS 5 -1.880510 0.701913 -2.68 0.0083

BKGRD1\_C7NOMISS 6 1.918489 1.060808 1.81 0.0793

BMI . 0.196995 0.025862 7.62 <.0001

CENTERNUM 1 1.149023 0.657448 1.75 0.0820

CENTERNUM 2 -1.810191 0.478265 -3.78 0.0002

CENTERNUM 3 4.646581 0.754652 6.16 <.0001

CENTERNUM 4 0 0 . .

EDUCATION\_C3 1 0 0 . .

EDUCATION\_C3 2 -1.080721 0.401703 -2.69 0.0078

EDUCATION\_C3 3 -2.422101 0.392110 -6.18 <.0001

EMPLOYED 1 0 0 . .

EMPLOYED 2 -0.490358 0.920357 -0.53 0.5963

EMPLOYED 3 -0.684703 0.961289 -0.71 0.4794

EMPLOYED 4 0.229406 0.912882 0.25 0.8025

SEX 0 0 0 . .

SEX 1 5.546536 0.318894 17.39 <.0001

TIME . 0.397621 0.023166 17.16 <.0001

US\_BORN 0 0 0 . .

US\_BORN 1 -0.026674 0.491953 -0.05 0.9568

#### Model-Based GEE with MI

\* Fit GEE simultaneously in 10 imputed datasets ;

**proc** **genmod** data=sol\_mi\_widetolong;

by \_IMPUTATION\_;

class HH\_ID AGEGROUP\_C6(ref = '6') BKGRD1\_C7NOMISS(ref = '3')

CENTERNUM(ref = '4') SEX(ref = '0') US\_BORN(ref = '0') EMPLOYED(ref = '1') EDUCATION\_C3 (ref = '1');

weight WEIGHT\_FINAL\_NORM\_OVERALL;

model BMI = AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX US\_BORN

EMPLOYED EDUCATION\_C3 SBP5 TIME/ dist=normal;

repeated subject = HH\_ID /corr=ind;

ods output GEEEmpPEst=betas\_mi;

**run**;

The **proc** **genmod** procedure fits GEE to sol\_mi\_widetolong. The analysis is performed separately for each imputation through specifying in the by statement the imputation number identifier *\_*IMPUTATION\_*.* Reference levels can be specified in the class statement, e.g., AGEGROUP\_C6(ref = '6') sets level 6 as the reference. The weight statement specifies Visit 1 overall sampling weights (WEIGHT\_FINAL\_NORM\_OVERALL) for weighted GEE. The model statement specifies BMI as the outcome and includes all covariates of interest, assuming a normal distribution through dist=normal. The repeated statement defines the clustering variable subject=HH\_ID for household clusters. corr=ind specifies an independent working correlation structure. The ods output outputs the parameter estimates in the output object GEEEmpPEst to the dataset betas\_mi.

/\* Step 3 \*/

**proc** **mianalyze** parms(classvar=level)=betas\_mi;

class AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX US\_BORN

EMPLOYED EDUCATION\_C3;

modeleffects INTERCEPT AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM

SEX US\_BORN EMPLOYED EDUCATION\_C3 SBP5 TIME;

**run**;

For Step 3, the **proc** **mianalyze** procedure combines the MI results in betas\_mi using Rubin's rules. parms with the classvar=level option is needed to correctly identify the classification levels of variables specified in the class statement. The modeleffects statement lists all the effects in the model, including the intercept and all covariates specified in **proc** **genmod** from Step 2.

After removing redundant rows and columns (for the classification levels) from the output, parameter estimates with household clusters are displayed in **Output 4.2‑2**. Based on the results, the estimate for systolic blood pressure (SBP5) is 0.0293 with a standard error of 0.0045. This positive coefficient suggests that, on average, among individuals with the same values of the covariates included in the model (such as baseline age, sex, background, center, US-born status, employment, and education), and the same amount of time that have elapsed since Visit 1 or equivalently the same amount of aging (represented by 'TIME'), every 10 units (mm Hg) increase in systolic blood pressure within a person is associated with a 0.293 units (kg/m2) increase in BMI. This effect is statistically significant (p < .0001).

Output 4.2‑2: SAS, Parameter Estimates from GEE (household clusters) with MI



#### Complex-Survey GEE with GLM-Based IPW

#### Model-Based GEE with GLM-based IPW

### Application to Visit 3 Sample

As described in **Section 3.2**, the **Visit 3 Sample** is limited to participants who completed the latest visit (e.g., Visit 3 Exam Only; N = 9,090) and includes all available data from any visit for these participants.

#### Complex Survey GEE with GLM-based IPW

#### Model-Based GEE with GLM-based IPW

## Stata

### Application to Visit 1 Sample

#### Model-Based GEE with MI

**Note**: in Stata example, we provide results using subject clusters (ID) instead of household clusters (HH\_ID) as the MI procedure from Stata has the limitation that the specified weights need to be constant within the panel variable, which is not the case if using household clusters.

import sas using "sol\_wide.sas7bdat",clear

set seed 2024

rename WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 WEIGHT\_EXAMONLY\_V3

mi set flong

\*\* Extent of Missingness \*\*

mi misstable summarize

In Stata, the analysis dataset first needs to be loaded into working memory. This can be done using the use command for Stata datasets (with a ".dta" file extension) or the import command if the dataset is in a different format. import sas command loads the SAS dataset "sol\_wide.sas7bdat". The clear option ensures that any existing data in memory is cleared before importing the new dataset. The set seed command sets a specific random seed for reproducibility of any subsequent random processes. The rename command is used to shorten the name of the variable WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 to WEIGHT\_EXAMONLY\_V3, as a name of an imputation variable is not allowed to contain more than 29 characters in Stata.

The mi set flong command sets up the data for MI in the "flong" (full long) style, which is one of Stata's formats for storing multiply imputed data. The mi misstable summarize command examines the extent of missingness in the dataset and **Output 4.3‑3** presents the results.

Output 4.3‑3: Stata, Extent of Missingness

A screenshot of a table

Description automatically generated

\*\* Step 1 \*\*

mi register imputed SBP5\_V1 BMI\_V1 US\_BORN EMPLOYED EDUCATION\_C3 WEIGHT\_NORM\_OVERALL\_V2 YRS\_BTWN\_V1V2 SBP5\_V2 BMI\_V2 WEIGHT\_EXAMONLY\_V3 YRS\_BTWN\_V1V3 SBP5\_V3 BMI\_V3

mi register passive AGEGROUP\_C6 BKGRD1\_C7NOMISS CENTERNUM SEX WEIGHT\_FINAL\_NORM\_OVERALL

mi impute chained (regress) SBP5\_V1 BMI\_V1 WEIGHT\_NORM\_OVERALL\_V2 YRS\_BTWN\_V1V2 SBP5\_V2 BMI\_V2 WEIGHT\_EXAMONLY\_V3 YRS\_BTWN\_V1V3 SBP5\_V3 BMI\_V3 (logit) US\_BORN (ologit) EMPLOYED (mlogit) EDUCATION\_C3 = i.AGEGROUP\_C6 i.BKGRD1\_C7NOMISS i.CENTERNUM i.SEX WEIGHT\_FINAL\_NORM\_OVERALL, add(10)

For Step 1, The mi register imputed command specifies all variables to be imputed. The mi register passive command identifies variables that are not imputed but are used in the imputation model. The mi impute chained command performs multivariate imputation using FCS methods: linear regression regress for continuous variables; logistic regression logit for binary variables (US\_BORN); ordered logistic regression ologit for ordinal variables (EMPLOYED); multinomial logistic regression mlogit for nominal variables (EDUCATION\_C3). The add(10) option specifies that 10 imputed datasets will be created. The non-imputed variables to be included in the imputation model are specified at the end after the = sign, with the *i.* prefix indicating the classification/categorical variables.

\*\* Step 2 \*\*

\* Renaming BMI variables for easy reshape

rename BMI\_V1 BMI1

rename BMI\_V2 BMI2

rename BMI\_V3 BMI3

\* Renaming SBP5 variables for easy reshape

rename SBP5\_V1 SBP51

rename SBP5\_V2 SBP52

rename SBP5\_V3 SBP53

\* Renaming years between visits for easy reshape

rename YRS\_BTWN\_V1V2 TIME2

rename YRS\_BTWN\_V1V3 TIME3

\* Creating a new variable TIME1 and setting it to 0 for all

generate TIME1 = 0

\* Reshape data from wide to long;

mi reshape long BMI SBP5 TIME, i(ID) j(VISIT)

For Step 2, after MI, visit-specific variables are renamed to facilitate reshaping the data from wide to long format by modifying their suffixes (from \_VX to X), so they can be recognized by Stata as to which visit they are referring to. For instance, BMI\_V1, BMI\_V2, and BMI\_V3 are renamed to BMI1, BMI2, and BMI3. The time since Visit 1 variable for Visit 1 (TIME1) is created and set to 0. The mi reshape long command transforms the data from wide to long format. The *i(ID)* option specifies that ID is the variable that uniquely identifies subjects across visits, and the *j(VISIT)* option creates an indicator variable VISIT to indicate to which visit an observation belongs. In Stata, fitting GEE and combining the MI results using Rubin's rules are done with a single command, explained in Step 3.

\*\* Step 3 \*\*

encode ID, gen(ID\_NUM)

mi xtset ID\_NUM

mi estimate: xtgee BMI ib6.AGEGROUP\_C6 ib3.BKGRD1\_C7NOMISS ib4.CENTERNUM ib0.SEX ib0.US\_BORN ib1.EMPLOYED ib1.EDUCATION\_C3 SBP5 TIME [pw=WEIGHT\_FINAL\_NORM\_OVERALL], family(gaussian) corr(independent)

For Step 3, the encode command encodes the ID variable into the numeric format as a new variable ID\_NUM. This is necessary so that the mi xtset command declares the data to be longitudinal (panel) data, with ID\_NUM specified as the panel variable.

The mi estimate: xtgee command fits GEE and automatically combines the results across imputed datasets using Rubin's rules. Categorical variables are indicated by the prefix *ib*. and numeric values can be appended to indicate the reference level, e.g., ib6.AGEGROUP\_C6 sets level 6 as the reference. The [pw=WEIGHT\_FINAL\_NORM\_OVERALL] option applies Visit 1 overall sampling weights as probability weights for weighted GEE. The family(gaussian) option specifies a Gaussian (normal) distribution for the dependent variable, and corr(independent) specifies an independent working correlation structure for the GEE model.

Parameter estimates with subject clusters are displayed in **Output 4.3‑4**. The point estimates, standard errors, and confidence intervals with household clusters from other software and those with subject clusters from Stata are similar with only slight differences in this example, and no impact on statistical significance.

Output 4.3‑4: Stata, Parameter Estimates from GEE (subject clusters) with MI

A screenshot of a computer

AI-generated content may be incorrect.

## R

### Application to Visit 1 Sample

#### Model-Based GEE with MI

## Set up ##

library(haven)

library(dplyr)

library(tidyr)

library(skimr)

library(mice)

library(glmtoolbox)

library(mitml)

sol <- read\_sas("sol\_wide.sas7bdat")

# Reference levels

sol$SEX <- relevel(factor(sol$SEX), ref='0')

sol$CENTERNUM <- relevel(factor(sol$CENTERNUM), ref='4')

sol$AGEGROUP\_C6 <- relevel(factor(sol$AGEGROUP\_C6), ref='6')

sol$BKGRD1\_C7NOMISS <- relevel(factor(sol$BKGRD1\_C7NOMISS), ref='3')

sol$US\_BORN <- relevel(factor(sol$US\_BORN), ref='0')

sol$EMPLOYED <- relevel(factor(sol$EMPLOYED), ref='1')

sol$EDUCATION\_C3 <- relevel(factor(sol$EDUCATION\_C3), ref='1')

## Examine the extent of missingness ##

skim(sol)

In R, necessary libraries need to be loaded first. These include: 'haven' for reading data formats from other software; 'dplyr' and 'tidyr' for data manipulation 'skimr' for data summaries; 'mice' for MI using FCS; 'glmtoolbox' for GEE; 'mitml' for additional MI tools. The *read\_sas* function reads the SAS dataset "sol\_wide.sas7bdat" into R. The *relevel* function converts categorical variables to factors with specified reference levels, e.g., relevel(factor(sol$AGEGROUP\_C6), ref='6') sets level 6 as the reference. This ensures that subsequent analyses use the correct reference categories for these variables. Finally, the *skim* function examines the extent of missingness in the dataset. **Output 4.2‑5** presents part of the results.

Output 4.2‑5: R, Extent of Missingness

A screenshot of a computer

AI-generated content may be incorrect.

## Step 1 ##

# Set up MI using MICE

predMatrix <- quickpred(sol, include = c("AGEGROUP\_C6", "BKGRD1\_C7NOMISS", "CENTERNUM", "SEX", "WEIGHT\_FINAL\_NORM\_OVERALL", "SBP5\_V1", "BMI\_V1", "US\_BORN", "EMPLOYED", "EDUCATION\_C3", "WEIGHT\_NORM\_OVERALL\_V2", "YRS\_BTWN\_V1V2", "SBP5\_V2", "BMI\_V2", "WEIGHT\_NORM\_OVERALL\_EXAMONLY\_V3 ", "YRS\_BTWN\_V1V3", "SBP5\_V3", "BMI\_V3"))

methods <- make.method(sol)

# choose imputation methods, default for continuous variables is PMM

for (i in seq\_along(methods)) {

if (methods[i] == "pmm") {

methods[i] <- "norm"

}

}

# Modify the method for binary and categorical variables specifically

methods[c("US\_BORN")] <- "logreg"

methods[c("EMPLOYED")] <- "polr"

methods[c("EDUCATION\_C3")] <- "polyreg"

# Perform MI

imputed\_data\_wide <- mice(sol, method = methods, predictorMatrix = predMatrix, m = 10, seed = 2024)

For Step 1, the *quickpred* function creates a prediction matrix, with *include* option specifying variables to be included in the imputation model. The *make.method* function sets up the default FCS methods. For continuous variables, the method is changed from the default predictive mean matching *pmm* to linear regression *norm*. In terms of other variable types, specify: logistic regression *logreg* for binary variables (US\_BORN); ordered logistic regression *polr* for ordinal variables (EMPLOYED); multinomial logistic regression *polyreg* for nominal variables (EDUCATION\_C3). The *mice* function performs MI, with the following options: imputation methods *method*; predictor matrix *predictorMatrix*; number of imputations *m*; random seed for reproducibility *seed*. The process results in a list object, stored as 'imputed\_data\_wide', that contains all the imputed data with the imputation identifier 'imp'.

## Step 2 ##

# Combine all imputed datasets into one data frame

imputed\_data\_combined <- complete(imputed\_data\_wide, "long")

# Transform the combined data from wide to long format

imputed\_data\_long\_combined <- imputed\_data\_combined %>%

pivot\_longer(

cols = starts\_with(c("BMI\_", "SBP5\_")),

names\_to = c(".value", "VISIT"),

names\_pattern = "(.\*)\_(V\\d)"

) %>%

mutate(

VISIT = as.numeric(gsub("V", "", VISIT)),

TIME = case\_when(

VISIT == 1 ~ 0,

VISIT == 2 ~ YRS\_BTWN\_V1V2,

VISIT == 3 ~ YRS\_BTWN\_V1V3

)

)

# Split the combined long data back into individual imputed datasets

imputed\_data\_long\_list <- split(imputed\_data\_long\_combined, imputed\_data\_long\_combined$.imp)

# Initialize lists to store GEE results

model\_list <- list()

# Fit GEE to each transformed imputed dataset

for (i in 1:10) {

imputed\_data\_long\_i <- imputed\_data\_long\_list[[i]]

# Fit GEE

model\_list[[i]] <- glmgee(

BMI ~ AGEGROUP\_C6 + BKGRD1\_C7NOMISS + CENTERNUM + SEX + US\_BORN + EMPLOYED + EDUCATION\_C3 + SBP5 + TIME,

data = imputed\_data\_long\_i,

id = HH\_ID,

corstr = "independence",

weight = WEIGHT\_FINAL\_NORM\_OVERALL,

family = gaussian(link = "identity")

)

}

For Step 2, the *complete* function combines all items in the list object from Step 1 into a single data frame. The *pivot\_longer* function transforms the combined data from wide to long format. This transformation creates separate rows for each visit, with variables like BMI and SBP5 now having a single column each, and a new VISIT column indicating the visit number. The time since Visit 1 (TIME) for Visit 1 is set to 0. The *split* function splits the long-format data back into a list object based on the imputation identifier 'imp'. Within a *for* loop, the *glmgee* function applies GEE to each of the transformed imputed datasets in the list object . The option *id = HH\_ID* specifies household (HH\_ID) clusters. The *corstr = independence* option sets the working correlation structure to independence. The *weight = WEIGHT\_FINAL\_NORM\_OVERALL* option applies the Visit 1 overall sampling weights in weighted GEE. The option *family = gaussian(link = identity)* specifies that the model assumes a Gaussian (normal) distribution for the outcome. The results are stored in a list object 'model\_list'.

## Step 3 ##

pooled\_results <- mitml::testEstimates(model\_list, fun = summary)

# Create a data frame of coefficients

coefficients\_df <- data.frame(

name = rownames(model\_list[[1]]$coefficients),

round(pooled\_results$estimates,4)

)

coefficients\_df

For Step 3, the *testEstimates* function from the *mitml* package pools the results with Rubin's rules. To include variable names in the output, which are not provided from the *testEstimates* function, a data frame 'coefficients\_df' is created. This data frame combines the variable names extracted from the 'model\_list' object (from the coefficients in GEE fitting) with the rounded pooled estimates (4 decimal places), providing a more interpretable summary of the results.

Parameter estimates (formatted to include variable names) accounting for household clusters are displayed in **Output 4.4‑6**;

Output 4.4‑6: R, Parameter Estimates from GEE (household clusters) with MI

A screenshot of a computer

Description automatically generated

1. Longitudinal Analysis of Binary Outcomes

References

Lavange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J., Criqui, M. H., & Elder, J. P. (2010). Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, *20*(8), 642-649. <https://doi.org/10.1016/j.annepidem.2010.05.006>

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22. <https://doi.org/10.1093/biomet/73.1.13>

Rubin, D. B. (2018). Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition* (pp. 29-62). Chapman and Hall/CRC.

Sterba, S. K. (2009). Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration. *Multivariate Behavioral Research*, *44*(6), 711-740. <https://doi.org/10.1080/00273170903333574>

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press, Taylor & Francis Group. <https://books.google.com/books?id=bLmItgEACAAJ>