Weighting Methods for the 2010 Data Collection Cycle of the Medical Monitoring Project

Lee Harding[1], Ronaldo Iachan[1], Christopher H. Johnson[2], Tonja Kyle[1], Jacek Skarbinski[2]
[1]ICF International, 11785 Beltsville Drive Suite 300 Calverton, MD 20705
[2]Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta GA 30333

Abstract
This paper describes the weighting procedures developed for the Medical Monitoring Project (MMP), a nationally representative surveillance system for HIV-infected adults receiving medical care in the United States. MMP employs a three-stage probability sample of states, medical facilities, and patients and collects data on sampled patients from three different data sources: face-to-face interviews, medical record abstractions (MRA), and an extract of data reported to the National HIV Surveillance System. In the 2009 data collection cycle, all three datasets were weighted independently for differential probabilities of selection and non-response and adjusted for multiple opportunities to enter the sampling frame(s). In 2010, we improved our weighting process by using a unified classification process for eligibility that uses information from all three datasets and parallel weight adjustments for the three sets of weights. We also used a single estimated population total to harmonize weight sums across the three sources. In addition to describing weighting methods, we describe design variables (strata and clusters) defined for variance estimation applicable to all MMP datasets.

Key Words: population-based surveys, health surveys, multistage sampling, non-response adjustment, weighting, HIV.

1. Introduction

The Medical Monitoring Project (MMP) is an ongoing supplemental surveillance project of HIV-infected adults at least 18 years of age receiving outpatient medical care within 23 project areas in 16 states and Puerto Rico. Behavioral data are collected using an in-person interview and selected clinical data are abstracted from medical records (Blair, 2011). In addition, data are collected via linkage with the National HIV Surveillance System and on characteristics of sampled facilities. The MMP data are weighted for probability of selection and non-response to be representative of adults who received outpatient medical care for HIV infection in the United States and Puerto Rico in January to April of the data collection cycle year. The rationale for establishing MMP and details on the sample design are described extensively in McNaghten (2007) and Frankel (2012).

Data from the 2009 data collection cycle were weighted, but improvements in the weighting process were made in the 2010 data collection cycle. The 2010 MMP weighting process benefited from the simultaneous availability of patient-level data for all data sources: an extract of data reported to the National HIV Surveillance System

known as the Minimum Data Set (MDS), Medical Record Abstraction (MRA), and Interview data. We were able to use an integrated approach to determining eligibility, defining the conceptual population, and computing the weights for components corresponding to each data source. In particular, we were able to use the union of all data sets in defining the eligible population and the total weight sum (the size of the conceptual population).

The patient-level weights within a project area consist of a base weight, three nonresponse adjustments, and one multiplicity adjustment. The national weights are the product of the final patient weights, computed for each project area, and the inverse of the project area probability of selection. The national weights also undergo a trimming adjustment. While no poststratification adjustments were used at the project area or national level, we adopted an alignment of the weight sums for all data sets. The weighting process also entailed devising a set of design variables, with strata and clusters created uniformly so as to be applicable across all data sets.

2. 2010 Weighting Process

2.1 Base Weights

The first step in the computation of survey weights within each project area is the computation of base weights that reflect the sampling design probabilities. Within each area (the primary sampling unit, or PSU), the design includes two sampling stages for the selection of sample facilities and sample patients. Accordingly, the patient base weight consists of two components.

The first component of the base weight is the facility base weight (W_1) calculated as the inverse of the probability of selection of the facility or grouped facilities (secondary sampling units, SSUs).

$$W_1 = \frac{1}{P_1}$$

where P_1 is the probability of selection of that SSU, given that the project area was selected. This is a conditional probability within the project area, given that the project area was selected. Thus, facilities that are part of the same SSU have the same facility weight.

The second component of the patient base weight is patient weight (W_2). This base weight is the inverse of the conditional probability of selection of the patient within a sample facility.

$$W_{2j} = \frac{1}{P_{2j}}$$

where P_2 is the probability of selection of patient *j*, given the facility or SSU was selected. The patient base weight is the product of these two weight components. We express this as

$$\widehat{W}_{2j} = W_{1j}W_{2j}$$

Introducing notation that will be convenient for expressing the weight at that stage as the product of weights for preceding stages,

$$\widehat{W}_{ij} = \widehat{W}_{(i-1)j} W_{ij}$$

where $\hat{W}_{1j} = W_1 j$. In other words, each $\hat{W}_{hj}$ is the product of the W_{ij} where $i \leq h$. Thus

$$\widehat{W}_{3j} = W_{1j} W_{2j} W_{3j}$$

and so forth.

Every sampled patient has a base weight, whether the patient is an eligible respondent, eligible nonrespondent, ineligible, or of unknown eligibility. Base weights are solely determined by the probability of selection, and are the same for every patient sampled within a facility.

The next sections describe a series of adjustments performed on the base weights. We first describe nonresponse adjustments, and then multiplicity adjustments.

2.2 Nonresponse Adjusted Weights

Once base weights are calculated for each patient, the next step is to adjust the weights for nonresponse. The primary goal of a nonresponse adjustment is to reduce bias. Nonresponse bias occurs when nonrespondents differ from respondents or when nonrespondents account for a large proportion of the population, resulting in differences in survey estimates.

While nonresponse adjustments reduce bias, they usually introduce added variation to the weights. A balance between bias reduction and the increase in variance must be considered when implementing a nonresponse adjustment. Nonresponse adjustments use the information available for the sampled cases. In general, the adjustment distributes the base weights of the nonrespondents to the responding sampled cases so that the sum of the adjusted weights equals the sum of the base weights.

Before developing nonresponse adjustments, it is helpful to clearly define what constitutes nonresponse at both the facility and patient levels. Facilities are classified as respondents (1), eligible nonrespondents (2), and ineligibles (3). The population definition period (PDP) is the reference period for the survey, which for the 2010 data collection cycle started January 1, 2010, and ended April 30, 2010. A respondent facility is one that submitted actual patient loads (APLs) for the PDP or reported that, although it was still in business, it had no patients during the PDP. If a facility submitted an APL and later would not provide access to some or all of the patients sampled, the facility was still deemed a respondent, and the patients were considered nonrespondents. As the APL would have been used to allocate patients to facilities, this approach preserves the initial sample as originally designed.

Several stages of nonresponse adjustments are applied to the patient level data during the MMP weighting process. Patients are classified into four categories: (1) eligible respondents, (2) eligible nonrespondents, (3) ineligible patients, and (4) patients for whom eligibility is unknown. A patient is deemed ineligible for weighting purposes if the patient was under 18 years of age, was not HIV positive, or did not receive care during the PDP.

For effective weight adjustments, weighting classes need to be sufficiently large and to be based on factors that are related to the key survey variables. Even more importantly, the weighting class variables need to be associated with response propensities. Bivariate and multivariate analyses for the response indicator suggested that facility size, facility type, and age were significant predictors of participation and therefore could be used to create the nonresponse adjustment categories. Weighting classes are redefined for each adjustment, and can vary by project area. There are three nonresponse adjustments for MMP: facility nonresponse adjustments, demographic nonresponse adjustments, and interview nonresponse adjustments.

2.2.1 Facility Nonresponse Adjustment

Weighting classes for nonresponse adjustments at the facility level were based on facility size. Facility size was obtained from the estimated patient loads (EPLs) projected for the 2010 PDP in the facility sampling frames. We defined two size classes in each project area, with approximately the same number of facilities in each. The cutoff between the two size strata was the median EPL within the area for eligible facilities, ranging from 15 to 388 patients.

The facility nonresponse adjustment can be described for a given weighting class in terms of each facility's EPL divided by its probability of selection. Where **a** is the sum of the EPL of each facility that responded divided by its probability of selection (i.e., that of the SSU to which the facility belongs) and **b** is the corresponding sum of the EPL for eligible sampled facilities that did not respond in the weighting class, the weight adjustment for all facilities in the weighting class is

$$W_3 = (\boldsymbol{a} + \boldsymbol{b})/\boldsymbol{a}$$

This adjustment applies to the patient base weight ($\hat{W}_{2j}$) described in section 2.1. According to our notation,

$$\hat{W}_{3j} = W_{1j}W_{2j}W_{3j}$$

2.2.2 First-stage Patient Nonresponse Adjustment

In the first-stage patient nonresponse adjustment, the 9,400 sampled patients are classified into five response categories: (1) eligible respondents, (2) eligible nonrespondents, (3) ineligible patients, and (4) patients for whom eligibility is unknown which here is further subdivided into (4.a.) nonrespondents with unknown eligibility and facility level data only and (4.b) nonrespondents with unknown eligibility with facility and MDS data. In the 2010 weighting process, we cross-referenced all the patient datasets to determine patient eligibility. If a patient was a nonrespondent to the interview and their disposition code indicated they had an unknown eligibility status, we checked for the presence of a MRA. If they had a MRA, we classified the patient as an eligible nonrespondent.

Nonresponse adjustments use information that is available for every sampled case. Only facility level data are available for all 9,400 sampled cases. The first-stage nonresponse adjustment classes used facility-level variables, from the Facility Attributes files. The weighting classes were defined within project areas, based on variables that were related to response propensity. The variables used to define weighting classes were not specific to the

project area, although the defined classes were. These included facility size (median size based on APL), university affiliation, and private practice status.

In 8 of the 23 project areas, the facility response rates differed by facility size. Two of the 23 project areas had differences in the facility response rate based on private practice status, and one area had differences based on university affiliation. No difference could be found in facility response rate based on the chosen facility variables in the remaining 12 project areas.

The first-stage nonresponse weight adjustment starts with the adjusted weights calculated for each patient *j*, $\hat{W}_{3j}$, described in Section 2.2.1. The adjustment is performed within a weighting class defined only by facility-specific variables. Within each weighting class, the adjustment is computed as the ratio of two sums of weights: the sum of the patient weights adjusted for facility nonresponse over all 9,400 patients sampled and the sum of these weights only for patients in response categories (1), (2), (3) and (4.b). Where *A* is the aggregation of respondents in response categories (1), (2), (3) and (4.b), *B* is the aggregation of nonrespondents in response category (4.a), and patient weights are summed over the relevant class,

$$W_{4j} = \frac{((\sum j \varepsilon A\, \widehat{W}_{3j}) + (\sum j \varepsilon B\, \widehat{W}_{3j}))}{(\sum j \varepsilon A\, \widehat{W}_{3j})}$$

This adjustment includes eligible and ineligible patients; ineligibles were used to estimate the percentage of patients with unknown eligibility who may actually be eligible.

The overall weight is then

$$\hat{W}_{4j} = W_{1j} W_{2j} W_{3j} W_{4j}$$

Now, every sampled patient with MDS, MRA or Interview data has a value for each of W_{1j} W_{2j} W_{3j} W_{4j}, and hence $\hat{W}_{4j}$.

The purpose of the first-stage patient nonresponse adjustment is to subset the sampled cases to records with demographic data. The patient's demographic data can come from one of three sources, the MRA, Interview or the MDS dataset. To make use of the best demographic data, especially when multiple sources were available, we prioritized the use of the data sources. MRA was used as the primary source for demographic information because the information was obtained from medical records. Interview data is obtained via self-report which can be unreliable if the participant is uncooperative. For each demographic variable used in the weighting, we therefore first relied on the MRA, then on the Interview data when these (but not the MRA) were available. Only when neither MRA nor Interview data were available, did we use the MDS data, information that is less recent and considered less reliable that the other sources.

After cross-referencing the three datasets, there were still records on each dataset with missing demographic variables. A nearest-neighbor hot deck imputation method was used to impute these missing values.

2.2.3 Second-stage Patient Nonresponse Adjustment

After the first-stage patient nonresponse adjustment, we have more information about the remaining sampled patients. During the second-stage patient nonresponse adjustment, we use both demographic and facility level data to create the weighting classes. In the second-stage adjustment, the weights of the patients with unknown eligibility are reduced to account for the possibility of their being ineligible. The second-stage adjustment factor is for patients with known eligibility. The sampled patients are classified into 3 categories: (1) eligible respondents, (2) eligible nonrespondents, and (3) ineligible patients. Patients previously classified in category (4.b) had previously been removed from the dataset.

The second-stage nonresponse weight adjustment starts with the adjusted weights calculated for each patient *j*, $\hat{W}_{4j}$, described in Section 2.2.2. The adjustment is performed within a weighting class defined by facility specific or demographic variables. Within each weighting class, the adjustment is computed as the ratio of two sums of weights: the sum of the patient weights adjusted for first-stage patient nonresponse for patients only in response categories (1) and (2), known eligible patients and the sum of these weights for patients in response categories (1), (2) and (3), known eligible and ineligible patients.

In other words, the adjustment can be expressed as follows:

$$W_{5j} = \frac{((\sum j \varepsilon A \, \widehat{W}_{4j})}{(\sum j \varepsilon B \, \widehat{W}_{4j})}$$

where *A* is the class of respondents in response categories (1) and (2) and *B* the class of nonrespondents in the response categories (1), (2) and (3). This adjustment includes eligible and ineligible patients; ineligibles were used to estimate the proportion of patients with unknown eligibility who may actually be eligible. The ineligible cases are then removed from the dataset.

2.2.4 Third-stage Patient Nonresponse Adjustment

The third-stage patient nonresponse adjustment was performed on four datasets. The respondent category is defined differently for each dataset. The respondent categories included in each of the four datasets were:

1. The union of MRA and Interview,
2. The MRA respondents only,
3. The Interview respondents only, and
4. The Overlap of MRA and Interview respondents.

The first dataset was used to create unified population estimates across all the MMP datasets.

The weighting classes for each of the third-stage patient nonresponse adjustments were the same across datasets. The weighting classes were based on a subset of categorical variables that were related to the patient response rates in each project area. The categorical variables considered were facility size, university affiliation, age 45+ years old or not, age 18-24 year old or not, private practice status, and

Hispanic ethnicity. Different age category variables were predictive in different project areas.

If A is the class of MMP respondents, B the class B of eligible nonrespondents with demographic data within an adjustment class, and h_j is an eligibility indicator, then the adjustment factor for MMP survey nonresponse among respondents with demographic data is:

$$W_{6j} = \frac{((\sum j \varepsilon A\ \widehat{W}_{5j}) + (\sum j \varepsilon B\ hj\ \widehat{W}_{5j}))}{(\sum j \varepsilon A\ \widehat{W}_{5j})}$$

.

2.3. Multiplicity Adjustment

A multiplicity adjustment is advisable when some patients may have a higher probability of selection because they received treatment at other eligible facilities during the PDP. As in every weight adjustment, the bias reduction potential needs to be balanced against any increased variability that may be induced by the adjustment. The adjustment applies to eligible facilities in the frame, whether or not they have been included in the sample.

Multiplicity adjustments typically involve a series of approximations to estimate the multiplicity factor; in the case of MMP, this adjustment is based on the number of visits, *m*, to eligible facilities during the PDP.

For weighting purposes the number of visits to eligible facilities during the PDP was primarily obtained using data from the MMP patient interview. Only a few patients indicated in their interview responses that they received care at two or more facilities during the PDP. During the weighting process for 2009 data, the multiplicity adjustment was applied independently for each of the patient level datasets. For the 2010 data, to obtain the best estimate of multiplicity, when a patient gave no indication during the interview of receiving care at a facility other than the one where he/she was sampled, we searched for additional Medical History Form records for this patient.Only 100 patients reported receiving care at two or more facilities during the PDP. It was determined that patients reporting receipt of care at three or more facilities during the PDP were outliers and *m* was capped at two.

We estimate the probability of selection by computing the probability of the complement, that is, the product of the probabilities of not including the patient in any of the *m* facilities:

$$\left(1 - p_j\right)^m$$

And the estimated probability of selecting him or her multiple times is then

$$1 - \left(1 - p_j\right)^m$$

For weighting purposes, we assume that each patient-facility visit would have the same weight, with the same probability of selection for each pair.

If we define $p_j = \frac{1}{\widehat{W}_{6j}}$ as an estimate of the adjusted probability of selection of patient j from any facility, where *m* is the number of facilities from which he or she could have been selected, then the weight can then be expressed as

$$\widehat{W}_{7j} = \frac{1}{1-(1-p_j)^m}$$

1. With a cap of $m = 2$,

$$W_{7j} = \frac{\widehat{W}_{6j}}{(2\widehat{W}_{6j} - 1)}$$

so that

$$\widehat{W}_{7j} = W_{7j}\widehat{W}_{6j} = \left(\frac{\widehat{W}_{6j}}{(2\widehat{W}_{6j} - 1)}\right)\widehat{W}_{6j}$$

Had there been any weights exceeding three times the median, the weights would have been capped at three times the median for the project area, and the weights would have been adjusted to add to $\hat{W}_{7j}$. As no project area weights were found to be larger than three times the median weight, no trimming was conducted at this level, and $\hat{W}_{7j}$ becomes the final patient weight.

This final weight is the starting point for the creation of national weights described in the following section.

3. National Weights

The computation of national weights started by accounting for the first stage of sampling (since States were PSUs). If the inverse of the probability of selection of the State from which patient j was sampled is W_{8j}, then the initial national weight for patient j is $\hat{W}_{7j}$ X W_{8j}.

To limit the variability in national weights, the initial national weights were trimmed within classes, k, defined by key demographics (age, race/ethnicity and gender) while preserving the weight sum in each class. The weights were capped at three times the median, u_k, of the $\hat{W}_{8j}$ for each class. The trimming also considered the median weight, u, over the entire sample, so the cap was the minimum of the two medians. In other words, the trimmed weight is min($\hat{W}_{8j}$, 3u,3uk).

Thus, the sum of the $\hat{W}_{8j}$ was obtained for every combination of race, age, and gender. Specifically, 12 weighting cells were defined in terms of race categories (African American, Hispanic, and other), age (over and under age 45), and gender.
The post-trimming adjustment factor for the national weights can be expressed as follows in terms of the indicators $f_{kj} = 1$ if patient j is in class k, and $f_{kj} = 0$ otherwise.

$$W_{9j} = \frac{min(\widehat{W}_{8j}, 3u, 3u_k)\left(\frac{\sum j\ \widehat{W}_{8j} f_{kj}}{\sum j\ min(\widehat{W}_{8j}, 3u, 3u_k) f_{kj}}\right)}{\widehat{W}_{8j}}$$

With the usual notation, the trimmed national weights were

$$\widehat{W}_{9j} = \widehat{W}_{8j} W_{9j}$$

4. Design Variables

In order to calculate variance estimates, we created design variables—i.e., strata and cluster variables—that account for the complex sampling design. We began with a matrix of responding SSUs in 2010. The matrix contains the SSUs, each SSU's probability of selection, and counts of responding patients within an SSU in each of the patient datasets, the MDS, MRA, Interview and Overlap datasets.

To create the project area design variables for 2010, we first sorted the matrix in descending order by the SSU probability of selection. Each SSU with a probability of selection of 1.0 (certainty facilities) was classified as a design stratum, and the patients within those facilities are classified as design clusters. A grouping of four noncertainty SSUs with similar probabilities of selection were classified as a stratum. Each SSU in the stratum was a cluster. By construction, there are at least two SSUs with one or more respondents in each stratum-survey combination (that is, for each of the four datasets).

In Figure 4.1 below, there are ten SSUs sorted in descending order by probabilities of selection. In 2009, we paired SSUs in each of the patient datasets to create the design strata and clusters. The 2009 groupings are indicated by the blue. This pairing created different design variables across all patient datasets. In 2010 we grouped four to five SSUs into one stratum across all datasets (MDS, MRA, Interview and Overlap) and each SSU is assigned a cluster number. In Figure 4.1, the first five SSUs were grouped into one stratum with five clusters and the next five SSUs were grouped into a second stratum with five clusters. In the first design stratum all five clusters will appear in the MDS and MRA datasets because each SSU has at least one respondent. Only the four clusters associated with SSUs 0001, 0002, 0003 and 0005 will appear in the Interview and Overlap datasets.

Figure 4.1: Process for creating design variables for the 2009 and 2010 weighting cycles

Facility ID	Facility Probability of Selection	Interview	MRA	Overlap	MDS
00010001	0.52	12	15	11	18
00010002	0.51	9	10	7	15
00010003	0.46	5	7	5	8
00010004	0.34	0	2	0	4
00010005	0.32	8	5	5	9
00010006	0.21	3	2	2	0
00010007	0.15	3	0	0	3
00010008	0.10	0	0	0	3
00010009	0.10	0	1	0	2
00010010	0.05	2	2	2	2

The change in the creation of design variables in 2010 allowed for the combination of datasets without the worry of strata with single clusters, which poses a problem for variance estimation that different software packages handle in different ways.

5. Conclusions

The availability of all patient-level datasets prior to the beginning of the 2010 MMP weighting process allowed for several innovations. The eligibility of each sampled patient was more accurately determined by cross-referencing the patient dispositions and the Interview and MRA datasets. We were able to use more information when creating the nonresponse adjustment cells. By using both the MRA dataset and Interview dataset we were able to apply the multiplicity adjustment more consistently across all the patient level datasets. Finally, we were able to create design variables for variance estimation that were consistent across the datasets.

References

Blair, J., McNaghten, A., Frazier, E., Skarbinski, J., Huang, P. & Heffelfinger, J. (2011). Clinical and Behavioral Characteristics of Adults Receiving Medical Care for HIV Infection --- Medical Monitoring Project, United States, 2007. *Morbidity and Mortality Weekly Report (MMWR),* **60,** 20.

McNaghten, A., Wolfe, M., Onorato, I., Nakashima, A., Valdiserri, R., Mokotoff, E., Romaguera, R., Kroliczak, A., Janssen, R. & Sullivan, P. (2007). Improving the representativeness of behavioral and clinical surveillance for persons with HIV in the United States: the rationale for developing a population-based approach. *PLoS One,* **2,** e550. Also available at: http://www.cdc.gov/hiv/pdf/research_mmp_McNaghten.pdf.

Frankel, M., McNaghten, A., Shapiro, M., Sullivan, P., Berry, S., Johnson, C., Flagg, E., Morton, S. & Bozzette, S. (2012). A Probability Sample for Monitoring the HIV-infected Population in Care in the U.S. and in Selected States. *Open AIDS Journal,* **Suppl 1** 67-76.