2012

A Probability Sample for Monitoring the HIV-infected Population in Care in the U.S. and in Selected States

Martin R. Frankel
CUNY Bernard M Baruch College

AD McNaghten
Centers for Disease Control and Prevention

Martin F. Shapiro
The RAND Corporation

Patrick S. Sullivan
Emory University

Sandra H. Berry
The RAND Corporation

See next page for additional authors

A Probability Sample for Monitoring the HIV-infected Population in Care in the U.S. and in Selected States

Martin R. Frankel[1], A.D. McNaghten[*,2], Martin F. Shapiro[3,4], Patrick S. Sullivan[5], Sandra H. Berry[3], Christopher H. Johnson[2], Elaine W. Flagg[2], Sally Morton[6] and Samuel A. Bozzette[3,7]

[1]*Baruch College, The City University of New York, New York City, New York, USA*

[2]*Centers for Disease Control and Prevention, Atlanta, Georgia, USA*

[3]*The RAND Corporation, Santa Monica, California, USA*

[4]*University of California, Los Angeles, Los Angeles, California, USA*

[5]*Rollins School of Public Health, Emory University, Atlanta, Georgia, USA*

[6]*Research Triangle Institute, Research Triangle, North Carolina, USA*

[7]*The University of California, San Diego, San Diego, California, USA*

Abstract: Epidemiologic and clinical changes in the HIV epidemic over time have presented a challenge to public health surveillance to monitor behavioral and clinical factors that affect disease progression and HIV transmission. The Medical Monitoring Project (MMP) is a supplemental surveillance project designed to provide representative, population-based data on clinical status, care, outcomes, and behaviors of HIV-infected persons receiving care at the national level. We describe a three-stage probability sampling method that provides both nationally and state-level representative estimates.

In stage-I, 20 states, which included 6 separately funded cities/counties, were selected using probability proportional to size (PPS) sampling. PPS sampling was also used in stage-II to select facilities for participation in each of the 26 funded areas. In stage-III, patients were randomly selected from sampled facilities in a manner that maximized the possibility of having overall equal selection probabilities for every patient in the state or city/county. The sampling methods for MMP could be adapted to other research projects at national or sub-national levels to monitor populations of interest or evaluate outcomes and care for a range of specific diseases or conditions.

Keywords: HIV, sampling, representative, surveillance.

INTRODUCTION

The HIV epidemic is dynamic: during the years since the first cases were reported, it has been characterized by both epidemiological and clinical instability, as the disease spread rapidly, infected different populations, responded to treatment and was associated with a changing spectrum of opportunistic illnesses. Surveillance is an essential element in monitoring and planning responses to important diseases with such characteristics, and that is certainly the case for HIV. A challenge to public health is to determine how to conduct such surveillance to best answer questions about the status of the disease and its impact, given the changes that have occurred in disease spectrum, populations affected, and social correlates of the illness and its spread. Such surveillance also is highly relevant to the challenges in reforming the health care system. The comparative effectiveness of systems for delivering acute, chronic and preventive services is best understood in the context of the care that is provided in the full breadth of settings and systems, and to the full range of populations using such services.

AIDS has been reportable throughout the United States since the onset of the HIV epidemic, and both AIDS and HIV have been reportable in recent years [1]. The resulting core information is invaluable, but insufficient [2]. In response, the Centers for Disease Control and Prevention (CDC) has conducted a number of secondary surveillance studies. These were generally unlinked medical record and interview studies conducted in a relatively small number of urban sites. Though these studies provided much clinical and behavioral detail, they did not meet the growing need for integrated, representative, population-based data on clinical status, care, outcomes, and behaviors at the national level. The Medical Monitoring Project (MMP) is a CDC-sponsored supplemental surveillance project that is designed to meet that need by combining interviews with medical record abstractions from patients selected using nationally representative scientific probability sampling.

In this paper we describe a sampling strategy that provides both nationally representative estimates as well as estimates at the state level. These state level estimates cover virtually all large states as well as a sampling of smaller states. The strategy outlined in this paper (to provide both

*Address correspondence to this author at Emory University, Rollins School of Public Health, Department of Epidemiology, 1518 Clifton Road NE, Atlanta, GA 30322, USA; Tel: +1 404 727-8750; Fax: +1 404 712-8392; E-mail: a.d.mcnaghten@emory.edu

1874-6136/12

national and selected state estimates) provides a methodological model that may be applied to the study of other low incidence diseases in the United States as well as similar diseases in other countries. The sampling strategy described is somewhat different from more traditional probability sampling models for national estimates, where typically, primary sampling units consist of individual counties and metropolitan areas. However, the use of states as primary sampling units was consistent with the funding and data collection model. It also provided participating states with the direct benefit of having the ability to obtain valid state level estimates.

Probability Sampling

The Rationale

The goal of probability sampling is to generate unbiased estimates by developing a sample that accurately represents the whole target population rather than a subset selected according to some sort of rule. A probability sample is one in which each person in the target population has some probability of selection and in which the probability of selection is known for each person who is actually selected. That is, no person of interest is excluded and the number of people represented by each selected person is known.[1] Because of its ability to represent an entire population of interest, probability sampling is the method of choice for producing valid and unbiased statistical estimates for large populations and their subgroups.

Single Stage *vs* Multistage Probability Sampling

The simplest way of accomplishing a probability sample is a one-step process, or one stage sampling, by making random picks directly from a sampling frame containing a complete and current listing of all the persons in the population. However, this is impractical for most large populations, because it would result in a geographically dispersed sample which is difficult and expensive to implement. In addition, a complete and current list of population elements, such as HIV infected persons in care, may not be available. Fortunately, the construction of such a list is not required in order to develop a probability sample. Sampling can be accomplished by constructing a probability sample in stages, i.e., by making a series of random picks from each of a hierarchical series of sampling frames (e.g., geographic areas, health care providers, patients); these picks must be made in such a way that each selection is linked to a selection made in the previous sampling stage by virtue of having the frame for that pick be defined by the elements included in the previous selection.

The HIV Cost and Services Utilization Study (HCSUS) used a multistage probability sample of persons in care for HIV disease in the 48 contiguous United States in 1996, demonstrating that such an approach is feasible for studying HIV disease [3]. In the HCSUS, the hierarchy was as follows: 1) participants were selected from a complete list of patients being cared for by their provider, 2) providers were selected from a complete list of providers in their area, and 3) areas were selected from a complete list of all such areas in the United States. The MMP uses many of the features of the HCSUS sample design in developing a nationally representative probability sample of non-institutionalized persons receiving outpatient care for HIV infection in the 50 United States, the District of Columbia and Puerto Rico.

THE STRATEGY FOR DEVELOPMENT OF THE PROBABILITY SAMPLE IN MMP

For MMP, the population of interest is defined as all persons who are HIV-positive and have received any medical care from a known outpatient provider of HIV care during a specified period of time. We refer to this time period as the Population Definition Period (PDP).

The sample was selected in three stages. The overall approach is to randomly select from the sampling frames at each stage using probability proportional to size (PPS) sampling at all but the last stage. This method sets the selection probability for each unit according to the relative size of that unit: the larger the unit, the greater the probability of it being selected. This allows the calculation of selection probabilities in the last stage such that the overall probability of selection for each person in the sample is similar. This is desirable because each participant's data is similarly weighted, which is the most efficient circumstance.

The stage-I sampling frame consisted of 52 primary sampling units (PSUs): the 50 United States, the District of Columbia, and Puerto Rico. Twenty of these 52 PSUs were selected using geographic stratification and probabilities proportional to the estimated number of persons living with AIDS at the end of 2002. Eighty-one percent of all reported AIDS cases had been reported from those areas. Stage-II sampling frames were developed separately for the selected PSUs using a variety of appropriate methods. Six cities or counties reside within five of the 20 selected states but were considered separately in stage II (because their funding for HIV/AIDS surveillance activities, including MMP, is separate from that of the states in which they are situated). For these five states, the different funding areas may be thought of as separate sampling strata, each with its own second stage sampling frames. Thus, there are a total of 26 sampling strata, with corresponding frames. In each of these selected areas or frames, local MMP staff developed a comprehensive list of all outpatient facilities that manage HIV patients.

An outpatient facility was defined as any hospital-based or stand-alone clinic or health care facility, any group or private practice, or any grouping of these entities in which medical records or a medical record system is shared. Emergency departments and inpatient facilities were excluded from the facility sampling frames, and thus from MMP. Second-stage sampling was random with the probability of selection proportionate to size in 19 of the 20 PSUs and the 6 separately funded areas. In the other PSU (Delaware), a census of 21 facilities was necessary to obtain the minimum number of patients needed. Stage II resulted in the selection of 25 to 68 (mean = 41) facilities in each state or separately funded area.

The stage-III sampling frames were developed separately for each of the selected facilities. Frames consist of comprehensive listings of persons with known HIV infection who made at least one visit to one or more of the sampled facilities during the PDP. Selections were random and were

[1]The inverse of the probability of selection for a sampled person is the number of persons in the population represented by that sampled person.

performed in such a way as to maximize the possibility of having equivalent overall probabilities of selection for every patient in the project area (state or city/county), and with overall sampling rates calculated to obtain 10,000 selections overall and a minimum of 400 persons in each of the 20 PSUs.

This design ensures that each person in the population is given a non-zero probability of selection into the sample. It results in a valid probability sample because each element (member) of the target population is associated with a unit (facility) in the secondary sampling frame, and each of these is associated with a PSU.

FIRST STAGE OF SAMPLING

In most multistage probability samples, the first stage sampling units are defined so the resulting first stage sample comprises 20% or less of the defined units. In MMP, the percentage is much higher because there were a smaller number of PSUs. The decision to define a small number was driven largely by the MMP funding model. Under CDC Program Announcement 04155, entities eligible to receive funding for MMP were the 50 states, the District of Columbia, Puerto Rico, and the six cities/counties that receive separate funding for HIV/AIDS surveillance.

Identifying and Stratifying the Primary Sampling Units

The separately funded cities/counties are Chicago, IL; Houston, TX; Los Angeles County, CA; New York City, NY; Philadelphia, PA; and San Francisco, CA. The decision not to directly pick these cities was driven by a secondary goal of providing statewide data to participating state health departments. This was accomplished by folding the six cities/counties into the five states that contain them for the purposes of first stage sampling only (i.e., not for purposes of administering the project). This yielded 52 PSUs. The decision to sample 20 of the 52 PSUs was made based on funding availability and face validity rather than on any statistical argument.

In MMP, a systematic sampling with a random start was used to generate a random stratified proportional to size sample. In order to improve the reliability of the final sample, the PSUs were stratified into five groups - four geographic and one by size, then ordered by the PSU's measure of size (MOS). The MOS is an estimate of the number of persons in the population of interest that are contained in each unit of a sampling frame. For the first stage of MMP, the MOS were the CDC estimates of the number of adults and children living with AIDS at the end of 2002, current as of November 24, 2003. Although the target population for MMP is all persons diagnosed with HIV in care in the US, at the time the first stage was developed, there was no data system that collected information on HIV infected persons in care. Therefore, the estimated number of persons living with AIDS was used as a proxy measure of PSU size.

The systematic sampling procedure was as follows. We first created a list or pseudo-population of patients grouped by PSUs, arrayed by stratum and size. For example, if the largest PSU in the first stratum (PSU #1) had 1000 cases, then patients 1 to 1000 would be labeled "PSU #1;" and if PSU #2 had 2130 cases, then patients 1001 to 3130 would be labeled "PSU #2."

Selecting the Sample of States

Selection proceeds by picking a random start at the beginning of the list and then taking uniform steps forward through it. In MMP, the initial pick was to be a patient in the first 5% of the list with subsequent steps each being 5% forward through the list; 5% having been chosen because we wished to sample 20 PSUs. The PSU containing each patient picked in this procedure is included. Since the steps each contain 5% of the total cases, at least one step must fall among the patients that represent any PSU that has more than 5% of the total cases. These large PSUs are sampled with certainty. The process is iterative. The cases in certainty PSUs are removed, the step size recalculated based on the number of available and required PSUs, and the additional PSUs to be sampled with certainty are identified and removed. The cycle repeats until no further PSUs were sampled with certainty.

Describing the particular choices made in the MMP sample will elucidate this approach. The total number of cases, that is, the sum of the MOS, across all 52 PSUs was 384,070. Five percent of this total is 19,204. Thus any PSU with at least 19,204 cases was to be sampled with certainty. Four PSUs had at least this many cases: California, Florida, New York, and Texas (Table **1**).

We then excluded the four certainty PSUs, recalculated the total number of cases, and redefined the criteria for PSUs that would also be sampled with certainty. Sixteen PSUs remained to be sampled from 48 remaining PSUs. So, any PSU with at least 6.25%, or 12,473 of the remaining 199,569 cases, was to be sampled with certainty. Four PSUs had at least this many cases: Georgia, Illinois, New Jersey and Pennsylvania. Twelve PSUs remained to be sampled from the remaining 44 PSUs. So, any PSU containing at least 8.3%, or 11,860 of the remaining 142,321 cases was to be sampled with certainty. No PSUs contained this many cases, so no further PSUs were sampled with certainty.

The remaining 12 PSUs were to be sampled PPS in a stratified manner. This was accomplished by recreating the list according to stratum and size as described above, using the cases from the remaining 44 PSUs. Selecting randomly from such as list results in a sample that is: a) PPS because the proportion of entries that are grouped under a PSU reflects the proportion of cases contained in that PSU, and b) stratified because of the specific ordering.

We grouped the states into five strata: four based on region and one that grouped together states with few cases. The stratum of states with few cases was formed to minimize how many PSUs with few cases were included. Sampling cases from PSUs that contain very few cases is difficult and expensive to implement, so we defined a stratum of "small" PSUs in such a way that only one PSU would be sampled from that stratum. Since 12 PSUs were to be sampled, a stratum containing PSUs with close to but less than the "step" value of 8.3% (11,860) of all cases would likely contribute one PSU but could not contribute two or more to the sample. To define this stratum, the smallest PSU was chosen and PSUs were added from the list (taking the next smallest and so on) until the total number of included cases

was smaller than 11,860 but as large as possible. In this manner, 18 PSUs containing 11,118 were chosen for the small state stratum; adding the next smallest would have resulted in 12,915 cases in this stratum.

The remaining 26 PSUs (defined as "medium" size) were divided into four geographical strata based on the Census definitions of geographic regions: the Northeast region contained Census divisions New England and Middle Atlantic (states: CT, MA); Midwest contained Census divisions East North Central and West North Central (states: WI, MN, IN, MO, MI, OH); South contained Census divisions South Atlantic, East South Central and West South Central as well as Puerto Rico (states: AR, OK, KY, MS, AL, TN, SC, LA, NC, VA, DC, MD, PR); and West contained Census divisions Mountain and Pacific (states: OR, NV, CO, AZ, WA).

The geographic strata were ordered: Northeast, Midwest, South and West and then were followed by the small PSU stratum to form the pseudo population list. Within each stratum, the PSUs were ordered by size. In order to preclude possible sampling periodicity by size, within Northeast the PSUs were ordered smallest to largest; within Midwest largest to smallest; within South smallest to largest; within West largest to smallest; and within the small PSU stratum smallest to largest. The sampling frame resulting from the determination of the certainty picks, and from the stratification and ordering of the remaining PSUs, is shown in Table **1**.

PSUs eligible to be one of 12 picked by random PPS selection are listed after the certainty PSUs, beginning with Connecticut. As implied above, the initial pick had to be within the first 8.3% or 11,860 cases. A random number of 0.878 was chosen, thus identifying a random start of 0.878 x 11,860, or 10,413. Case number 10,413 is contained within Massachusetts, which therefore was the first randomly selected PSU. Stepping though the list at intervals of 11,860 resulted in the selection of one state in the Northeast, two each in the Midwest and the West, six in the South and, as anticipated, one from the grouping of states with few cases. Areas with "Yes" in the Sampled column in Table **1** comprise the final sample of states.

SECOND STAGE OF SAMPLING

The purpose of the second-stage sampling was to select facilities within the project areas. Although 20 PSUs were selected in the first stage of sampling, there were 26 project areas, as noted above, including the six cities/counties that are funded and administered separately in five of the 20 selected PSUs.

A facility sampling frame was developed individually in each of the 26 project areas. An eligible facility was defined as one known to provide HIV care, which was defined as having providers who prescribe antiretroviral therapy or order CD4 or HIV viral load tests. Providers who referred patients to other providers rather than managing their HIV medical care were not included in the facility sampling frame.

Constructing the Facility Sampling Frame

A variety of sources were used to identify facilities providing HIV care (including the number of individuals in care at those institutions). These included the HIV/AIDS Reporting System (HARS) databases, laboratory reporting databases and other local databases, including AIDS Drug Assistance Program and Medicaid databases. The HARS was the best source for identifying HIV care providers. It is the current version of the reporting system that all states have used for surveillance of new AIDS cases since 1985 and, since at least 2005, all new cases of HIV infection [4,5]. In 2005, when the first facility sampling frames were constructed, 18 of the 20 states conducting MMP had HIV viral load and/or CD4 reporting of any value or of all tests performed.

For the 2005 pilot data collection cycle, the following types of facilities were included on the facility sampling frame: outpatient and inpatient care facilities; Veterans Administration facilities; and state or local prisons and jails that met the definition of providing HIV care. The following types of facilities were not included: emergency departments (because they do not provide sufficient information on the standard of care), HIV counseling and testing sites and laboratories (which report HIV infection, but do not provide medical care for HIV infection), Federal prisons, military bases and institutions (project areas have no jurisdiction to obtain their medical records), and pediatric facilities (unless they provided HIV medical care to persons age 18 and older).

For the 2007 and subsequent data collection cycles, two other exclusions were made from the facility sampling frame to focus predominately on patients receiving outpatient care. It was decided not to include inpatient facilities or prisons or jails on the facility sampling frame. Inpatient facilities were excluded since they do not provide primary medical care for HIV infection. HIV-infected patients could have interfaced with inpatient care facilities for a variety of reasons. It would be prohibitively difficult to recruit providers who do not typically provide HIV care (but who may have prescribed antiretroviral medications or ordered CD4 counts or HIV viral load tests during the patient's inpatient stay). In addition, the likelihood of finding and recruiting patients who had only one encounter with the provider at an inpatient facility would be much lower than that for patients who have an ongoing relationship with a regular HIV care provider. In some instances, such as hospice or care in long term care facilities, primary medical care is provided; however, this care is different from outpatient care provided by other facilities on the facility sampling frame. Prisons and jails as providers of HIV care were excluded from the sampling frame because these facilities are not able to be accessed in all project areas due to Institutional Review Board (IRB) and other issues.

Once the lists of facilities from HARS and each of the supplemental sources were obtained, cleaned, and standardized, they were combined into a single facility sampling frame (FSF) for each project area, on which each facility only appears only once. Any outpatient facility that met the MMP facility definition and was a known provider of HIV medical care during the recent time periods used for each data source was eligible to be included on the FSF.Facilities that had not seen a patient with HIV during the time frame estimated patient loads (EPLs) were obtained (i.e., they had an EPL of 0) were included on the FSF, but patients were not sampled from these facilities.

Table 1. Sampling Frame of States, with Region, Measure of Size, Stratum, and Sample Indicator

Area of Residence	Measure of Size	Region*	PSU Stratum Size	Sampled
NEW YORK	63412	NE	Large (certainty)	Yes
FLORIDA	41015	S	Large (certainty)	Yes
TEXAS	27358	S	Large (certainty)	Yes
CALIFORNIA	52716	W	Large (certainty)	Yes
NEW JERSEY	15485	NE	Large (certainty)	Yes
PENNSYLVANIA	15362	NE	Large (certainty)	Yes
ILLINOIS	13718	MW	Large (certainty)	Yes
GEORGIA	12683	S	Large (certainty)	Yes
CONNECTICUT	6579	NE	Medium	No
MASSACHUSETTS	8025	NE	Medium	Yes
OHIO	5978	MW	Medium	No
MICHIGAN	5395	MW	Medium	Yes
MISSOURI	4838	MW	Medium	No
INDIANA	3429	MW	Medium	Yes
MINNESOTA	1818	MW	Medium	No
WISCONSIN	1797	MW	Medium	No
ARKANSAS	1837	S	Medium	No
OKLAHOMA	1908	S	Medium	No
KENTUCKY	2150	S	Medium	No
MISSISSIPPI	2602	S	Medium	Yes
ALABAMA	3660	S	Medium	No
TENNESSEE	5639	S	Medium	No
SOUTH CAROLINA	5863	S	Medium	Yes
LOUISIANA	6902	S	Medium	No
NORTH CAROLINA	7128	S	Medium	Yes
VIRGINIA	7443	S	Medium	Yes
DISTRICT OF COLUMBIA	8234	S	Medium	No
PUERTO RICO	10560	S	Medium	Yes
MARYLAND	11798	S	Medium	Yes
WASHINGTON	4889	W	Medium	Yes
ARIZONA	4316	W	Medium	No
COLORADO	3465	W	Medium	No
NEVADA	2502	W	Medium	No
OREGON	2448	W	Medium	Yes
NORTH DAKOTA	47	MW	Small	No
WYOMING	91	W	Small	No
SOUTH DAKOTA	99	MW	Small	No
MONTANA	181	W	Small	No
VERMONT	236	NE	Small	No
ALASKA	252	W	Small	No
IDAHO	262	W	Small	No

(Table 1) contd…..

Area of Residence	Measure of Size	Region*	PSU Stratum Size	Sampled
MAINE	492	NE	Small	No
NEW HAMPSHIRE	506	NE	Small	No
NEBRASKA	567	MW	Small	No
WEST VIRGINIA	599	S	Small	No
IOWA	686	MW	Small	No
RHODE ISLAND	1058	NE	Small	No
NEW MEXICO	1066	W	Small	No
UTAH	1085	W	Small	No
KANSAS	1113	MW	Small	No
HAWAII	1247	W	Small	No
DELAWARE	1531	S	Small	Yes

*NE = Northeast; S = South; MW = Midwest; W = West.

Facilities are sampled in the second stage in a manner analogous to the PSUs in the first stage. To accomplish PPS sampling at this stage, an EPL of adult HIV infected patients in each facility was also needed on the frame. The EPL is an estimate of the actual number of eligible patients that will be seen at a facility during the PDP for a given data collection cycle. In 2007, for each data source from which EPLs could be derived, a 4 month EPL was created using the most recent data from each data source as well as from facility contacts to accurately reflect the patient load for the January 1 through April 30, 2007 PDP. Data were obtained either from a data run or other record-based source or as a less precise estimate, typically provided by facility staff. A matrix, or table, of EPLs from each data source was constructed for all eligible facilities, and this matrix was used to create the FSF, which in turn was used to select facilities for the previous data collection cycles. During this step, the quality of the different EPLs obtained across the various data sources was evaluated in order to determine, for each facility, which EPL was the most accurate to use for facility sampling.

Facility Linkage

A sampled patient's overall selection probability is the product of the three stage-specific selection probabilities. It is desirable that this overall probability of selection be uniform. Such uniformity will result in greater statistical efficiency because there would be minimal variation in point estimates derived from the information that patients contributed. The result is that confidence limits for estimates derived from MMP data will thus be minimized. Facilities with very low EPLs, or small facilities, are problematic in this regard because achieving uniform selection probability may require the selection of more patients than they actually have. In MMP, this was handled by linking known small facilities to larger ones to create linked 'facilities' with combined EPLs that met or exceeded a minimum value.

Facilities designated as small were linked to one or more other facilities so that the small facility is selected for the sample only if the facilities to which it is linked also are selected. The desired minimum EPL across each project area ranges between 40 and 80, and depend in part on the distribution of EPLs across the entire FSF for that project area. Minimum values of 40 to 80 have been determined to be optimal for selecting the facility sample across project areas based on anticipated design effects and distributions of facility sizes.

In project areas of large geographic size, or with variations in facility attributes by region, this linkage was performed within pre-specified sub-regions to facilitate efficient use of project area resources during data collection, as well as to ensure facilities from every sub-region were selected.

Selecting the Facility Sample

Electronic files containing the FSF from each project area were sent to CDC using the CDC's Secure Data Network (SDN). All files sent to CDC are stripped of identifying information for each facility; facilities are identified only by unique numeric facility identification (ID) number, assigned by the project area. Facility ID numbers for all project areas are made unique by adding a 4-digit project area code in front of the assigned 4-digit facility ID number. The number of facilities sampled in each project area varied from 25 in Houston and Los Angeles to 68 in California (Table **2**). In Delaware a census of all 21 eligible facilities was used as the second stage sample. In other project areas, facilities were randomly sampled from among all facilities on the FSF. In the five states containing cities/counties that are separately funded for surveillance, a larger number of facilities was sampled in the second stage of sampling in order to provide more useful local data to the separately funded areas. Specifically, separate samples of facilities were selected within each of the six separately funded cities/counties as well as elsewhere in those states.

Of the 828 facilities sampled and eligible, 582 participated. Because this was a strict probability sample, no replacement facilities were sampled. Furthermore, in those project areas with lower facility response rates the sample size was not increased because of the strict probability sampling protocol. This might be considered in future years. Facility participation rates ranged from 65% to 100%, with a median of 91.4%.

Table 2. Facility and Patient Sample Sizes and Facility Response Rates

Area	# Facilities Sampled	# Facilities Sampled and Eligible	# Facilities Participating*	% Facilities Participating	Patient Sample Size
CALIFORNIA (rest of state)	68	56	28	69.6	500
LOS ANGELES, CA	25	23	21	91.3	400
SAN FRANCISCO, CA	30	29	24	89.7	400
DELAWARE	21	20	18	90.0	400
FLORIDA	60	49	31	85.7	754
GEORGIA	49	35	25	91.4	400
ILLINOIS (rest of state)	43	34	19	70.6	100
CHICAGO, IL	41	26	19	92.3	400
INDIANA	41	39	23	94.9	400
MASSACHUSETTS	39	38	35	94.7	400
MARYLAND	44	32	19	78.1	348
MICHIGAN	53	40	31	87.5	401
MISSISSIPPI	46	38	23	100.0	400
NORTH CAROLINA	43	36	20	91.7	400
NEW JERSEY	35	20	11	65.0	500
NEW YORK (rest of state)	44	42	33	95.2	200
NEW YORK CITY, NY	34	30	25	83.3	800
OREGON	60	37	23	91.9	400
PENNSYLVANIA (rest of state)	32	26	21	100.0	100
PHILADELPHIA, PA	28	22	20	100.0	400
PUERTO RICO	34	22	17	95.5	400
SOUTH CAROLINA	31	16	14	93.8	400
TEXAS (rest of state)	47	40	27	85.0	400
HOUSTON, TX	25	20	13	65.0	400
VIRGINIA	46	24	18	91.7	400
WASHINGTON	40	34	24	79.4	400
TOTAL	1059	828	582	87.2	10503
			Mean	87.4	
			Min	65.0	
			Max	100.0	
			Median	91.4	

*Includes only facilities with patients sampled.

THIRD STAGE OF SAMPLING

In the third and final stage of sampling, within each participating facility, eligible patients are sampled for inclusion in MMP. Participants are sampled from lists of patients seen at each facility during the PDP (i.e., January 1 through April 30 of the data collection year). The selection of the patient sample is done in a manner that results in an equal probability of selection method sample at the patient level. This means that patients are sampled from each facility with a third-stage sampling probability which, when multiplied by the second-stage selection probability, results in the same overall selection probability for every patient selected in the project area.

Each patient sample is only used for one data collection cycle. A new sample of patients is drawn from the participating facilities in each data collection cycle.

HIV-infected patients who received all of their care from emergency departments or inpatient facilities are excluded from MMP, given that these facilities are excluded from the FSF. For sampled patients, in addition to information

regarding their outpatient care, information on visits to emergency departments or inpatient facilities is also obtained during interviews, and/or may be documented in medical records.

Constructing the Patient Sampling Frame

At the end of the PDP, health department MMP staff request a list from each sampled facility of all HIV-infected adults who received medical care (defined as any visit to the facility for medical care or prescription of medications, including refill authorizations and immunizations) during the PDP. Patients are eligible for inclusion on the patient sampling frame if they were HIV infected, at least 18 years of age at the beginning of the PDP, and received care at a sampled and participating facility during the PDP.

Facilities construct patient lists using International Classification of Diseases (ICD-9 or ICD-10) codes for procedures, tests or prescriptions during the PDP, or in smaller facilities by reviewing appointment logs.

Patients are eligible for selection only at their first reported visit to the facility during the PDP to ensure that multiple visits to the same facility do not lead to multiple opportunities for selection. As facilities use different mechanisms to identify eligible patients, the lists are not unduplicated across facilities. To account for multiplicity - multiple patient visits to different facilities during the PDP - the interview includes questions about the number of different facilities visited during the PDP to allow for the adjustment of the multiplicity of probability of selection in the weighting process. Without this, persons visiting more facilities would have higher probabilities of selection which could lead to estimation bias.

For each facility, the actual count of patients seen during the entire PDP derived from a facility's patient list will differ from the selected best EPL used to construct the FSF. EPLs were obtained for a 4-month period, which should closely match the number of patients on the patient lists obtained for the 4 month PDP.

Selecting the Patient Sample

Patient sampling is conducted as soon as all patient lists have been received from the participating facilities. The file containing lists of HIV-infected patients seen during the PDP at all participating facilities is used to select the patient sample. The selected participant ID numbers are returned to the project area *via* the CDC's SDN after patient sampling has been completed; this set of participant IDs comprises the entire patient sample for the project area.

It was determined that 400 is the minimum sample size for a state to obtain total population estimates with an acceptable level of precision (assuming a moderate design effect [of between 2 and 4], or increase in variance of estimates due to using a multistage sampling design). This sample size was assigned to most of the states with the lowest AIDS prevalence. Sample sizes for states with moderate to high AIDS prevalence were determined based on the distribution of cases among the 20 sampled states and the 6 separately funded cities/counties in those states, in order to achieve a national sample size of approximately 10,000. States that have large numbers of prevalent AIDS cases were allocated larger sample sizes (California 1300; Florida 800; Illinois 500; New Jersey 500; New York 1000; Pennsylvania 500; and Texas 800). These project area sample sizes will allow national estimates at an acceptable level of precision (assuming a moderate design effect) for subpopulations as small as 5% of the total population of interest. Table **2** outlines the patient sample size selected for each MMP project area. Although patient sample sizes of 800 and 400 were selected for Florida and Maryland, respectively, errors in the estimation of patient loads during facility sampling frame construction resulted in reduced patient sample sizes in these areas. Patient selection was accomplished by systematic random sampling within facility. Changes in the probability of selection by oversampling in facilities with lower response rates were not attempted in order to maintain a constant probability of selection across all facilities in order to handle patient multiplicity.

It should be noted that in all stages of selection, the probability nature of the sample allows the computation of required probabilities of selection for selected patients:

Probability of Selection ($Patient_i$) = P1 x P2 x P3x M, where

P1 = the probability of selection for the state

P2 = the probability of selection for the facility associated with the i^{th} selected patient

P3 = the probability of selection for the i^{th} selected patient within the facility containing the patient.

M = the number of facilities the patient reports visiting during the population definition period.

PATIENT RECRUITMENT AND DATA COLLECTION

All patients selected for the sample should be recruited for enrollment in MMP. Persons selected during third-stage patient sampling may be offered enrollment through two general recruitment processes: MMP project area staff-contact enrollment or facility-referred enrollment. The recruitment strategy varies according to facility preference and state or local project area IRB requirements.

For MMP staff-contact enrollment, facilities provide project area MMP staff with contact information for patients selected for recruitment. After obtaining patient contact information, the MMP staff contact selected patients to describe the project and offer enrollment. Scripts are used by all project areas to ensure a standardized recruitment approach within project areas. Patients who are eligible for enrollment and agree to participate are scheduled for an interview at a location that is convenient for the patient and meets the need for patient privacy.

Patients recruited through facility-referred enrollment are initially contacted by staff of the facility from which they were sampled. This may be done by telephone, in person, through chart insert and/or letter mailed from the facility. If by telephone or in person, the facility staff describe the project briefly and ask permission to provide contact information to MMP staff so that enrollment can be completed, or the facility staff ask the patient to contact the MMP staff. If recruitment takes place *via* chart insert or

letter, the documents will describe the project briefly and will provide contact information to enable the participant to reach MMP staff.

Patients who agree to participate and consent to the interview are asked questions by a trained interviewer. The interview includes questions about patients' demographics, access to health care (including antiretroviral therapy), unmet needs for services, sexual behavior, drug and alcohol use, use of prevention services, and health and well-being. Following the interview, medical record abstraction is conducted on all sampled patients. Information obtained from medical charts by trained data abstractors includes patient demographics, insurance status, AIDS-defining and other illnesses, laboratory values, prescription of antiretroviral and other medications, and evidence of substance abuse. Many project area IRBs have determined that this abstraction can be done for all sampled patients as part of surveillance activities but in other project areas it can be done only if the patient agrees as part of the consent process.

DISCUSSION

The experience of MMP demonstrates that it is possible to develop a credible national probability sample approach for HIV-infected persons receiving medical care in the United States. Unique challenges, such as the need to identify all providers of HIV care in each project area and the estimated number of patients each provider serves, were met by state, city and county health departments working through their existing surveillance systems and relationships they have built with providers of HIV care over the years. However, due to the size of the task of constructing facility sampling frames, the voluntary nature of the project (which allows providers and patients to refuse participation at any stage), and IRB and individual facility constraints on patient recruitment, project areas had varying success in constructing facility and patient sampling frames and recruiting sampled providers and patients in the 2007 pilot year. As subsequent MMP data collection cycles have been implemented, project areas have adopted best practices for facility and patient sampling frame construction, as well as provider and patient recruitment, which have resulted in improved efficiency and response rates.

Since patients receiving HIV care are only included on the patient sampling frame if they attended sampled facilities during the 4-month PDP, it is possible that patients who attend HIV care less frequently may be underrepresented. However, an analysis of HIV patients' time to first annual HIV care visit found that for patients who had at least one HIV care visit in the previous year, 88% of patients had their first care visit within the first 4 months of the next year [6]. Therefore, the 4 month enrollment period should sufficiently reflect this patient population, even for those who do not frequently access care.

According to the Office of Management and Budget's (OMB) *Standards and Guidelines for Statistical Surveys*, non-response bias analyses should be planned for when the unit response is expected to be below 80% [7]. Based on these OMB guidelines, the goal of MMP was to obtain 80% overall response rates at both the national and state levels. The overall response rate at the national level is the product of the project area (stage I), facility (stage II), and patient (stage III) response rates. Achieving an overall response rate of at least 80% is ambitious, particularly in a pilot year, and may be difficult to achieve even once MMP is being conducted at peak efficiency. Therefore, MMP has also pursued a policy of collecting minimal data about each patient sampled (sex, age, race/ethnicity, mode of exposure to HIV, most recent CD4 count) using state or local HARS data to allow for an effective non-response analysis. However, in some cases project area IRBs have not allowed this without patient permission, and some facilities have been unwilling to provide patient information for sampled patients who chose not to participate.

Despite these challenges we expect that MMP will be an important step forward in providing nationally representative statistics about HIV patients in care in the U.S. We also note that there are several nationally representative surveys that have achieved response rates below 80% and that still produce estimates that are respected and used by the scientific and policy communities. For example, the coverage rate for the HCSUS was 68% [8], and the response rate for the Behavioral Risk Factor Surveillance System, a state-based system of random digit dialed telephone surveys on health risk behaviors, has declined over the years; the response rate in 2006 was approximately 51% [9]. Once higher response rates are achieved, we could compare the demographics of the sampled patients with the demographics of persons living with HIV from the project area's HIV/AIDS reporting system to ascertain the representativeness of their sample.

The MMP model, or similar models, could be adapted to enrolling and evaluating care and outcomes for a large range of chronic diseases for which comparative effectiveness data are desired. Sampling strategies can be adapted in response to priority research questions at the national level, or smaller geographic areas. Using such methods, it should be possible to gain an understanding of clinical effectiveness in the populations that are most relevant and most representative of the reference populations of greatest interest to health policy.

ACKNOWLEDGEMENT

Declared none.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

REFERENCES

[1] Nakashima AK, Fleming PL. HIV/AIDS surveillance in the United States, 1981-2001. J Acquir Immune Defic Syndr 2003; 32(Suppl 1): S68-S85.

[2] McNaghten AD, Wolfe MI, Onorato I, *et al*. Improving behavioral and clinical HIV/AIDS surveillance in the United States: the rationale for developing a population-based approach. PLoS One 2007; 2(6): e550.

[3] Shapiro MF, Berk ML, Berry SH, *et al*. National probability samples in studies of low-prevalence diseases. Part I: Perspectives and lessons from the HIV Cost and Services Utilization Study. Health Serv Res 1999; 34(5, Pt 1): 951-68.

[4] Centers for Disease Control and Prevention. Revision of the case definition of acquired immunodeficiency syndrome for national reporting--United States. MMWR Morb Mortal Wkly Rep 1985; 34: 373-5.

[5] Glynn MK, Lee LM, McKenna MT. The status of national HIV case surveillance, United States 2006. Public Health Rep 2007; 122 (Supp 1): 63-71.

[6] Sullivan PS, Juhasz M, McNaghten AD, Frankel MR, Bozzette SA, Shapiro MF. Time to first annual HIV care visit and associated factors for patients in care for HIV infection in 10 US cities. AIDS Care 2011; 23(6): 1-7.

[7] Office of Management and Budget (OMB). (2006) Standards and Guidelines for Statistical Surveys. (Office of Information and Regulatory Affairs, OMB, Washington, DC). Available online at http://www.whitehouse.gov/omb/assets/omb/inforeg/statpolicy/standards_stat_surveys.pdf [Accessed: March 15, 2010].

[8] Frankel MR, Shapiro MF, Duan N, *et al*. National probability samples in studies of low-prevalence diseases. Part II: Designing and implementing the HIV Cost and Services Utilization Study sample. Health Serv Res 1999; 34(5, Pt 1): 969-92.

[9] Fahimi M, Link M, Schwartz DA, Levy P, Mokdad A. Tracking chronic disease and risk behavior prevalence as survey participation declines: statistics from the Behavioral Risk Factor Surveillance System and other national surveys. Prev Chronic Dis 2008; 5(3): A80.

Received: April 15, 2011 Revised: August 22, 2011 Accepted: September 14, 2011