“Real Time Sampling in Patient Surveys”

Ronaldo Iachan, ICF International

Deirdre F. Middleton, ICF International

Tonja Kyle, ICF International

1. Introduction

The Data Coordinating Center for HIV Supplemental Surveillance (DCC) at ICF International receives and manages data for two national HIV surveillance systems. These are the Medical Monitoring Project (MMP) which monitors those in care for HIV and the National HIV Behavioral Surveillance System (NHBS) which monitors populations at risk for HIV- high risk heterosexuals, men who have sex with men and injection drug users.

Before describing the pilot Real Time Sampling (RTS) study which was a component of MMP, it is useful to review both surveillance systems to have a broader perspective of how the pieces fit together and provide a picture of the HIV populations in care as well as populations at risk. All these populations are clearly hard to reach and the surveys involve sensitive topics.

National HIV Behavioral Surveillance System

NHBS was started in 2006 to collect data from individuals most at risk for HIV infection. With twenty funded project sites across the country, it targets three key populations, one per year: men who have sex with men (MSM), high risk heterosexuals (HET) and injection drug users (IDU). The sampling methodology for the HET and IDU cycles is respondent driven sampling (RDS) while the MSM cycle uses venue based sampling (VBS).

Respondent driven sampling starts with formative research where appropriate people, called seeds, are identified to be the first participants in the study and to start recruitment chains through their social networks. Each seed is screened for eligibility and if they are eligible, offered the opportunity to participate in an interview and take an HIV test. If they agree to participate, they are given coupons to distribute to potential participants in their social network, and each of those eligible participants is given the opportunity to distribute coupons and recruit additional eligible participants. If eligible and agree to participate, they have an interview and are offered an HIV test. Incentives are offered for the interview, HIV test, and recruiting an eligible participants. Recruitment continues until the desired sample size is reached.

Venue based sampling for the MSM cycle happens in three stages. After a formative research phase used to identify venues (such as bars, clubs or restaurants) where a high percentages of attendees are MSM, a frame of possible venues is creates where sampling can occur. In the first stage of sampling, venues are selected from this frame. In the 2nd stage of sampling, day-time periods when each selected venue is open are selected. These day-time periods were approximately 4 hour windows of time when the venue is available for researchers to recruit participants. In the third stage participants are selected at the venues during the day-time periods previously selected. If eligible and agree to participate, they have an interview and are offered an HIV test. We include a brief list of references for RDS and VBS methods as well as RTS methods.

Medical Monitoring Project

MMP is a surveillance system which focuses on behaviors and clinical outcomes among individuals infected with HIV and in care. The first full-scale data collection began in 2007 with 26 funded project areas, and there are now 23 funded project areas. In this paper we will first describe traditional MMP sampling and then the variations used for the pilot RTS study.

2. Traditional Sampling and RTS in MMP

In traditional MMP sampling, the sample is selected in three stages. In the first stage, project areas (states and cities) are selected with Probability Proportional to Size (PPS). The measure of size for project areas is based on 2002 AIDS cases. The first stage sample does not change from year to year in order to maintain the staffing and organization needed for yearly surveillance. In the second stage, a sample of facilities, "second-stage units" (SSUs) are selected. The facility sampling frame contains all facilities that provide HIV care to patients. It does not include facilities that only offer testing and referral services.

New facility samples are drawn every other year for each project area. These samples are selected with PPS using an estimated patient load (EPL) between Jan 1st and Apr 30th each calendar year, the population definition period (PDP), as the measure of size. Sampled facilities are asked to provide a patient list which contains all HIV patients seen in that time period. This patient list is used as the frame in the third sampling stage, when a sample of patients from the sample SSUs is drawn. Samples are 100-800 patients and drawn every summer for each project area. The patient sample is selected with approximately equal probabilities of selection.

Once patients are sampled, MMP staff attempt to recruit sampled individuals to participate in an interview and medical records abstraction (MRA). In many project areas, the staff can use surveillance authority for abstractions when they cannot locate or obtain consent from the patient.

A major challenge in MMP is maximizing response rates at both the facility and patient level. Some facilities find it burdensome to provide patient lists and assist MMP staff in recruiting patients. IRB issues may prevent MMP staff from contacting patients directly in some facilities. It is helpful to consider approaches that may reduce the burden on facilities in order to improve response rates. One of the main motivations for the RTS approach is the premise that this approach had the potential for reducing the burden on project area staff and on facility staff.

RTS Sampling in MMP

RTS method is a variation of methods for sampling in time and space that have also been used in a number of environmental and recreational user surveys. The methods are efficient for capturing a population of users or visitors as they enter or leave a facility. It involves a sample of site-period units defined in space and time.

For the RTS Pilot, first and second stage sampling remained unchanged but the method of sampling patients was modified. Rather than having a designated patient list at the beginning of the cycle, who would be sampled was based on who received care at sampled facilities during sampled time periods. Essentially, an additional level of selection- office period units- is added to the selection process. While project area and facility level sampling remained unchanged, within facilities the process is modified. Office-period units are selected using PPS sampling where size is calculated as the patient flow during that period (office hours of a particular day) in a particular office (office within the selected facility). The patient sampling frame is then the patients coming in for care during selected office period units.

A “site-period unit” describes the time and place where sampling of patients within that facility will occur for each sampling event. A facility may have different offices (sites) each with a separate check-in process, so each of those offices is considered a “site” within the facility. The “periods” consist of a block of time when that site is seeing patients. The patient frame is then those patients eligible for the study who are seen in the office-period unit.

Within facilities, office-period units are selected using PPS sampling where size is calculated as the patient flow during that period in a particular office within the facility. This means sampling events with higher expected numbers of patient visits have greater probabilities of selection. The patient sampling frame is then the patients coming in for care during selected office-period units.

This results in a nearly self-weighting sample where all eligible patients have approximately the same probability of selection if the MOS used in the first-stage PPS sampling is an accurate reflection of patient visits for each period. Adjustments are made based on the difference in actual patient visits and expected patient visits in the weighting, along with adjustments for multiplicity based on how often a sampled patient visits a facility.

Patients are sampled systematically as they enter the facility. The use of systematic sampling requires a careful count of eligible patients as they sequentially are admitted into the site. A designated enumerator counts eligible patients and flags every k^{th} patient who checks in during the sample period. An interviewer approaches each flagged patient and invites them to participate in the study.

3. RTS Fielding Process

Each month, a sample is drawn of site periods. To facilitate staffing, no more than one site period per day could be sampled. Modifications to the number of site periods sampled per month were adjusted throughout the cycle in order to account for response rates and how close estimated patient flows were to actual patient flows. This flexibility allowed for the site to efficiently reach their targeted number of interviews within the sampling period.

RTS fielding required two staff members- one to count and flag sampled patients (the enumerator) and one or more additional staff to schedule and conduct interviews. The staff records the time the first eligible patient entered the office and time first interview started, as well as several necessary counts in each event:

- Patients enumerated
- Patients sampled to participate
- Eligible patients interviewed

- Ineligible patients sampled
- Sampled patients who refused
- Interviews scheduled to occur

Patient response rates were improved to 77% with RTS compared to approximately 60% attained in this project area for the last cycle conducted with traditional methods.

4. Weighting RTS Data

This section describes the ongoing weighting process for RTS data. At the project area level, RTS data are weighted to account for the two sampling stages and for non-response. In addition, the procedures adjust for multiplicity. Finally, we combine the two components selected with RTS and the traditional methods.

Note first that the weighting at the level of each project area is distinct from the national level weighting which requires adjustment based on the project areas probability of selection. Since RTS was only conducted in one project area (PA), our discussion can be restricted to the weighting at the PA level.

The first-stage sampling weights reflect the sampling process for sample events. Conditional on the given sample facility, it is the reciprocal of the selection probabilities for sample events. These probabilities are the products of two selection probabilities:

- Sampling of sample days
- Sampling of site-periods for each sample day

As an illustration of the first-stage sampling weights, Table 1 presents the sampling weights for a given month (March). These combined first-stage weights reflect the probabilities of selection a day (WT1), and then a site-period within the day (WT2). The latter conditional PPS selection was conducted using the measure of size (MOS) shares of the site-period for the given sample day.

The second stage sampling weight reflects the sampling of patients within site periods. It is the reciprocal of the conditional probability of selection for patients selected with systematic random sampling in the site-period. Table 2 presents the count data used in computing the second-stage weights for site-periods selected for Weeks 1-15.

Once a base weight is calculated for each patient (including non-respondents), nonresponse adjustments are made using demographic data available for a large proportion of non-respondents.

We consider the following groups in order to calculate nonresponse adjustments:

A: did not respond, demographic information **is not** available

B: did not respond, demographic information **is** available

R: the set of respondents

First, we apply an initial adjustment factor, [W(A)+W(B)+W(R)]/[W(B)+W(R)] where W(A) is the weight sum over the set of patients who did not respond and for whom demographic information is not available. We then drop patients in subset A, and use weighting classes based on demographics with adjustment factor [W(B)+W(R)]/W(R).

Next, multiplicity adjustments are made. A patient may visit the RTS facility more than once during the population definition period (PDP), which means they have multiple chances of selection for the RTS sample. The patient is asked during their interview how frequently they are seen in that facility, and the response to this question is used to calculate the number of possible times they could have been sampled and make a multiplicity adjustment. Alternately, the patient could have visited another RTS facility or another MMP eligible facility in the frame during the PDP. Both of these factors can be determined from the survey data, and are used for multiplicity adjustments.

The final step in the weighting process is combining the RTS data weights, computed for the two RTS facilities, with the traditional study weights computed for all other, non-RTS sample (participating) facilities in the project area. The weights for two components are forced to sum the shares of the respective "sizes", i.e., the shares in the actual patient load (APL) for the two RTS facilitate and the non-RTS facilities. The actual patient load is computed from unduplicated patient lists obtained from participating facilities.

5. Conclusions

The pilot study demonstrated the feasibility and efficiency of selecting a real-time sample of patients in several stages with probability sampling methods from each sampled facility.

From a statistical perspective, the approach was flexible and adaptive allowing for fine tuning the sampling parameters in each monthly wave, which increased efficiency in reaching sampling goals. Sampling weights could be computed and adjusted for non-response and multiplicity.

RTS also showed the potential for covering segments of the target population that are typically not included in the MMP patient sample. Preliminary analyses of the potential reduction in non-response and non-coverage biases looked at some key variables. Specifically, we looked at the differences between the RTS sample estimates and the usual MMP sample estimates for these variables. Table 3 illustrates some significant differences that were found for key variables such as time since diagnosis. In this case, the RTS sample apparently included patients diagnosed a longer time ago, and that may have been sicker than the average sample patient. It included relatively fewer patients with more recent diagnoses (5-10 years ago, for example).

From a logistical point of view, the facility staff manifested a strong preference for the RTS procedures compared to the usual methods. This information was collected in surveys with facility personnel following the RTS data collection. The main reasons were that the RTS methods led to reduced burden and less intrusiveness in the daily operations of the facilities. Admittedly, implementation on a larger scale might not allow the degree of tailoring (or even "personalization") that was possible for the two study facilities.

RTS can represent a dynamic population as they actually use the health care facility. The traditional list sampling method for the HIV study leads to low response rates as selected patients can only get recruited much later and one patient at a time. The RTS approach can maximize coverage and response rates, minimize bias (represent the population) and simplify logistics of data collection. Participation rates may improve at the facility level as the methods reduce the burden facilities and project area staff incur in contacting patients. Participation rates may improve at the patient level as the methods lead to approaching patients in real time rather than much later following the visit to the facility. Bias is minimized because patients are sampled as they use the health care facility, ensuring that the sample is representative of those in care for HIV. The logistics of sampling are simplified because the times for recruitment and interviews are more easily scheduled and conducted. There is no need to spend countless hours attempting to contact or schedule with hard to reach patients from a predetermined list, and nonresponse related to patients moving or transferring to other facilities is eliminated.

Our future work will focus on combining weights with those for the traditional list-based approach in order to produce national MMP estimates which include both RTS and list based samples.

Table 1 First-stage sampling weight summary example

Facility	Sample Day	Site Period	Aggregate MOS	Site Period Probability of Selection	WT1	MOS Share	WT2
A	1-Mar	Tues. PM	118.8	0.67	1.49	58.68	2.02
E	2-Mar	Wed. AM	113.1	0.64	1.57	1.85	61.14
A	3-Mar	Thurs. PM	101.5	0.57	1.75	39.83	2.55
E	7-Mar	Mon. AM	120.6	0.68	1.47	12.76	9.45
A	8-Mar	Tues. PM	118.8	0.67	1.49	58.68	2.02
A	9-Mar	Wed. PM	113.1	0.64	1.57	67.90	1.67
A	14-Mar	Mon. PM	120.6	0.68	1.47	44.00	2.74
A	15-Mar	Tues. AM	118.8	0.67	1.49	45.92	2.59
A	16-Mar	Wed. AM	113.1	0.64	1.57	41.50	2.73
A	22-Mar	Tues. PM	118.8	0.67	1.49	58.68	2.02
A	23-Mar	Wed. PM	113.1	0.64	1.57	67.90	1.67
A	24-Mar	Thurs. PM	101.5	0.57	1.75	39.83	2.55
A	29-Mar	Tues. PM	118.8	0.67	1.49	58.68	2.02
A	30-Mar	Wed. AM	113.1	0.64	1.57	41.50	2.73

Table 2 Second-stage sampling weight summary example

MMP Real-Time Sampling Summaries (Weeks 1-15) for Facility A								
Week	**Expected Patient Flow**	**Actual Patient Flow**	**Enumerated Patient Count**	**Confirmed Patient Count**	**Selected**	**Refused**	**Interviewed**	**Ineligible**
1	128.1	65	47	46	7	0	6	0
2	90.9	42	20	21	3	0	3	0
3	113.8	46	22	24	6	1	5	0

4	67.9	13	9	9	3	1	2	0
5	130.1	58	52	51	14	1	13	0
6	160.3	81	56	56	10	3	6	0
7	141.5	87	55	58	8	2	5	1
8	158.7	97	53	54	6	0	6	0
9	86.0	50	35	35	7	2	5	0
10	114.0	76	51	51	10	0	7	2
11	100.0	57	33	33	5	1	3	1
12	166.0	84	52	52	14	4	5	3
13	60.0	38	25	25	3	0	3	0
15	58.7	16	8	8	2	0	2	0
1 to 15	*1576.0*	*810*	*518*	*523*	*98*	*15*	*71*	*7*

Table 3 Comparison of MMP traditional estimates and RTS-sample-based estimates for one example key variable

Count and Percent Years Post dx 0 to 4			
	Facility A	Facility E	Overall
MMP 2009	(8)22.86	(4)33.33	(12)25.53
MMP 2010	(9)23.68	(3)23.08	(12)23.53
MMP 2009,2010	(17)23.29	(7)28.00	(24)24.49
RTS 2011	(10)16.39	(5)22.73	(15)18.07

Count and Percent Years Post dx 5 to 9			
	Facility A	Facility E	Overall
MMP 2009	(5)14.29	(3)25.00	(8)17.02
MMP 2010	(5)13.16	(5)38.46	(10)19.61
MMP 2009,2010	(10)13.70	(8)32.00	(18)18.37
RTS 2011	(9)14.75	(3)13.64	(12)14.46

Count and Percent Years Post dx 10+			
	Facility A	Facility E	Overall
MMP 2009	(22)62.86	(5)41.67	(27)57.45
MMP 2010	(24)63.16	(5)38.46	(29)56.86
MMP 2009,2010	(46)63.01	(10)40.00	(56)57.14
RTS 2011	(42)68.85	(14)63.64	(56)67.47

References

Broadhead R, Heckathorn D, Weakliem D, Anthony D, Madray H, et al. (1998) “Harnessing peer networks as an instrument for AIDS prevention: results from a peer-driven intervention.” *Public Health Reports,* 113, 42–57.

Gile, K. (2011) “Improved Inference for Respondent Driven Sampling Data with Application to HIV Prevalence Estimation,” *Journal of the American Statistical Association,* 106 (493), 136-146.

Heckathorn D. (1997) “Respondent-driven sampling: a new approach to the study of hidden populations.” *Social Problems,* 44, 174-199.

Heckathorn D. (2002) “Respondent-driven sampling II: Deriving valid population estimates from chain referral samples of hidden populations.” *Social Problems,* 49, 11-34.

Iachan, R. (1989) “Issues in environmental survey design.” *Journal of Official Statistics,* 5, 323–335.

Iachan, R., M. L. Dennis. (1993) “A multiple frame approach to sampling the homeless and transient population.” *Journal of Official Statistics,* 9(4), 747–764.

Greene, J. M., C. L. Ringwalt, and R. Iachan. (1997) “Shelters for runaway and homeless youth: Capacity and occupancy.” *Child Welfare,* 76, 549–561.