Real Time Sampling in Patient Surveys

Ronaldo Iachan,[1] Tonja Kyle,[1] Deirdre Farrell[1]

[1]ICF Macro, 11785 Beltsville Drive, Suite 300 Calverton, MD 20705

Abstract

The article describes a real-time sampling (RTS) methodology used for patient surveys in a clinical setting. This approach for sampling from a target population who utilize a facility of interest can maximize coverage and response rates, minimize bias and simplify the logistics of data collection. This is a multistage sampling method using site-period units as the first stage of sampling, and systematic random sampling of patients as the second stage. A "site-period unit" describes the time and place where sampling of patients within that facility will occur for each sampling event. Site period units are selected with probabilities proportional to size (PPS) based on the estimated patient flow.

This methodology produces a nearly self-weighting patient sample as all eligible patients have approximately the same probability of selection. Weighting can be completed with 3 steps. To start, first stage sampling weights are calculated as the reciprocal of the selection probabilities for sample events. Second stage sampling uses the reciprocal of the conditional probability of selection for patients selected with systematic random sampling in the site-period. The final step of weighting the sample involves adjustments for non-response and multiplicity.

This paper describes both the data collection and weighting for this methodology.

Key Words: real-time sampling, multistage sampling, weighting, survey, facility sampling

1. Introduction

This paper presents the real time sampling (RTS) methods used in a pilot study within the Medical Monitoring Project (MMP). The ICF Macro team developed the methods in collaboration with the CDC and in consultation with the Philadelphia MMP personnel.

Selection for participation in MMP is made on three levels- project areas, facilities and patients- with the goal of all eligible patients having an equal chance of selection and participation in the study. There are multiple ways this objective can be achieved. The methodology for selecting project area sites and facilities is described elsewhere and is not changed in the RTS design. For this paper we will describe how the patient frame is developed using RTS and how the data is weighted.

Using the original methodology for MMP, the patient frame within a selected facility is all patients utilizing the facility for HIV care during a specified time frame who meet eligibility requirements. Patients are selected at random from this patient list at the beginning of the data collection cycle for participation and efforts are made to recruit them into the study over the course of several months.

For RTS, the methods for selecting projects area sites and facilities remain unchanged, but the method for selecting patients is modified. Essentially, an additional level of selection- office period units- is added to the selection process. Within facilities, 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. This paper describes this methodology and how to weight the data.

2. Sampling

We consider two main sampling stages in real time sampling (RTS):

1) A sample of office-period units, or sampling events, where offices are nested within facilities and periods are nested within days, and

2) A sample of patients selected from the first-stage sampling units (site-period units)

The sampling design is premised on the selection of events with probabilities proportional to size (PPS), an approach that uses measures of size (MOS) for each event to select with greater probabilities those sampling events with higher expected numbers of patient visits. The PPS design has many advantages from both logistical and statistical perspectives.

Statistically, it generates a patient sample that is nearly self-weighting, i.e., all eligible patients have approximately the same probability of selection, and therefore, approximately equal sampling weights. Self-weighting samples provide maximum precision for a given sample size by minimizing the effects of unequal weighting.

While a number of alternative approaches are possible for the sampling at each stage, we developed a design described that seemed to be most practical and acceptable for the facilities involved and allowed for a single two-person team for data collection. To ensure that data collection can be conducted by one single team, the design allows only one sample site-period to be selected for a given sample day. Therefore, the design considers the selection of sample days at the first stage, followed by the selection of one single site-period unit per sample day.

An additional design feature that enhances flexibility is the selection of the sample in waves. This approach allows adjustment of the sampling parameters for each wave (month) using the data from the previous waves. Eligibility, response rates and how well the estimated patient flow numbers approximate the reality on the ground for the month are some of the factors considered. The calibration of sampling parameters is especially useful in the final month of fielding the RTS sample.

2.2 Sampling events (site-periods)

2.2.1 Measures of size

The measures of size necessary for the PPS sampling are typically available from the patient flow data matrix provided by each participating facility for each of its offices. The basic matrix, reproduced in Table 2.1 below, can be simplified to distinguish only two periods within each day, a morning (am) period and an afternoon (pm) period. The flow data can be further improved with the availability of unduplicated data for the previous year for each facility. These data would presumably provide the most accurate measure of size (MOS) for the RTS selection of sampling events with PPS.

Table 2.1 Patient Flow Data form for each office used to calculate measure of size.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
7a-8a							
8a-9a							
9a-10a							
10a-11a							
11a-12p							
12p-1p							
1p-2p							
2p-3p							
3p-4p							
4p-5p							
5p-6p							
6p-7p							

2.2.2 Sample Sizes

For the first study, the first-stage sampling (site-periods) was designed to select 4 independent monthly samples for the 4-month period. The design of the first study included the selection of 16 sample days for each of the 4 months in the period for a total of 64 sample days. We projected the selection of three eligible patients per site-period, and per sample day.

2.2.3 Event sample selection

The sample of site-periods was selected from a first-stage sample of days selected with probabilities proportional to size (PPS).[1] For the sample of days, both the list of eligible days for the month and the associated size measures were derived from the patient flow table completed for the office by the facility.

Specifically, the measures of size (MOS) were the marginal totals for the days of the week. In other words, the MOS for the ith day, S(i) say, was the total aggregated for the day over all periods and all offices (across facilities). The selection probability for the ith day may be computed as

$$P_s = \frac{3S(i)}{S(t)} \qquad (2.1)$$

Here, S(t) is the total size measure over all days listed for the month, and n=3 is the sample size assigned to each sample event (site-period pair).

The list of eligible days for the month was then arrayed with their size measures, and any certainty days identified prior to the PPS selection of sample days. Certainty days are those days whose large measure of size lead to their selection with certainty into the sample. For example, it was possible the flow data would show a very large volume for Mondays, and that as a result, every Monday in the month could end up as a certainty day.

Once the sample days are determined, we can switch to the selection of eligible site-period pairs with probabilities proportional to size (PPS).[2] The size measures are again derived from the patient flow

[1] We also considered a simple random sample of days for those situations where the number of eligible days in the month is only slightly larger than the number of days to be selected for the month (n=16 in principle).
[2] The sample days can be distributed to facilities first as a heads-up for the data collection schedule.

matrix. For each pair, the selection probabilities are conditional on the given sample day, so that overall the probabilities are proportional to size (flow).

Specifically, the conditional probability that office-period unit (j,k) be selected within sample day "i" is computed in terms of the number of patients expected for this unit in the given day, S(i, j, k),

$$P(j,k; i) = S(i,j,k)/ S(i) \qquad (2.2)$$

Therefore, by multiplication of the two equations above, (2.1) and (2.2), the overall probability for the unit (i,j,k) is proportional to the size measure S(i,j,k).

It is possible that the flow measures not correspond exactly to the estimated patient load provided by the facility at the time of facility and sample size selection. Thus, facility A may have estimated a certain load during the entire period, but the estimated loads by days and offices may not add up to that load. If that is the case, we adjust the site-period loads to correspond to the load for the total facility over the four months.

A second possible adjustment is that the randomizing element could result in a larger or smaller sample size than anticipated for one facility or another. With this approach, we draw the office-periods for the first two months and examine the results prior to the beginning of the patient selection, and correct any over-sampling or under-sampling which may result from the randomization process.

We considered an alternative that would select offices sequentially, and adjust the probability of selection depending on the number of patients already selected for each facility. This method would guarantee the desired sample size in each facility (not office) at the expense of greater complexity in the sampling procedures and in the calculation of the weights.

2.2 Sampling patients

Patients were selected with systematic random sampling, a method that produces an equal probability sample, and a patient sample that is approximately self-weighting. The proviso is that we select an approximately constant number of patients in each sample office-period unit; we targeted a fixed number of three eligible patient selections per sampling event (n=3).

To implement the systematic sampling procedure, we need to estimate the sampling interval, k= N/n, where n=3 eligible selections and N is the total number of patients with appointments in the period. While this total, N, is initially estimated from the patient flow data, it is updated and adjusted using the appointment book for the office at the start of the sampling event period.

The details of the updated computation of the sampling interval are provided in Table 2.2. This table reproduces the directions given to data collectors to compute the sampling interval on site using the number of scheduled appointments for the period. The computation also uses the No-show rate (NS Rate), a composite average rate of no-shows and cancellations calculated for the facility.

Table 2.2 Formula and materials used by data collectors to compute sampling interval on site for a given sampling event.

$$Formula = \frac{(Scheduled\ Appointments) - (Scheduled\ Appointments) \times (NoShow\ Rate)}{3}$$

SAMPLING INTERVAL: _____ (Random Number)=_____

Complete the following rows using the information provided by ICF Macro

Expected patient flow for this sampling event	
Original sample target for this sampling event?	3
Original sampling interval	

Complete the following two rows prior to the start of the sampling event

Updated patient flow provided prior to the start of the sampling event	
Cancellation/ No Show Rate	
Revised sampling interval	

Complete each of the following after the sampling event

Total number of patients enumerated during the sampling event	
Total number of patients sampled to participate during the sampling event	
Total number of eligible patients interviewed during the sampling event	
Total number of sampled patients found to be ineligible during the sampling event	
Total number of sampled patients who refused MMP during the sampling event	
Time first eligible patient entered the office during the sampling event	
Time first interview started during the sampling event	
Total number of interviews that took place on-site at the office during or immediately following the sampling event	
Total number of interviews scheduled to occur at a time and place different from the sampling event	

To select a systematic sample, patients need to be enumerated as they enter the office and check in. In both facilities, we developed a mechanism to identify likely eligible patients without any screening but with some help from facility staff. Every k^{th} eligible patient was recruited, so that interview appointments could be made for the time following the medical appointment. If the patient contacted is not eligible, the next patient was contacted until an eligible patient is identified.[3]

Data collectors also obtained the total number of appointments for eligible patients for each sample event. These totals are useful for validation and for weighting. Table 2.3 provides a summary of the various patient counts obtained for each sample event. These totals, which are shown up to Week 15 for illustration, are defined as follows:

- Expected Patient Flow: Measures of size used in the PPS sampling of events
- Actual Patient Flow: Numbers obtained immediately prior to enumeration and data collection for each sample event

[3] We discussed the feasibility of adding a screener and consent forms to the forms filled out by every patient as they check into the facility. While this approach could allow the inclusion of facilities that have a mixed population not comprised exclusively of HIV patients, these types of facilities would present serious issues, and are not recommended for RTS

- Enumerated Patient Count: Total number of (likely) eligible patients enumerated by data collectors during the period for the event
- Confirmed Patient Count: Total number of eligible patients recorded in appointment book obtained following the sample period (event), used for validation and weighting

Table 2.3 RTS Summary Patient Counts

Facility	Week	Expected Patient Flow	Actual Patient Flow	Enumerated Patient Count	Confirmed Patient Count
Fac A	1	128.1	65	47	46
Fac A	2	90.9	42	20	21
Fac A	3	113.8	46	22	24
Fac A	4	67.9	13	9	9
Fac A	5	130.1	58	52	51
Fac A	6	160.3	81	56	56
Fac A	7	141.5	87	55	58
Fac A	8	158.7	97	53	54
Fac A	9	86.0	50	35	35
Fac A	10	114.0	76	51	51
Fac A	11	100.0	57	33	33
Fac A	12	166.0	84	52	52
Fac A	13	60.0	38	25	25
Fac A	15	58.7	16	8	8
Fac A	*1 to 15*	*1576.0*	*810*	*518*	*523*
Fac B	3	13.2	20	14	14
Fac B	4	26.0	35	25	25
Fac B	5	10.5	15	10	10
Fac B	6	10.5	14	10	10
Fac B	10	13.0	18	12	12
Fac B	11	4.0	14	9	9
Fac B	13	10.5	16	12	12
Fac B	14	6.9	12	6	6
Fac B	*1 to 15*	*94.6*	*144*	*98*	*98*
Fac A and Fac B	**1 to 15**	**1670.6**	**954**	**616**	**621**

Table 2.4 presents the number of selections and completed interviews for all sample events selected up to Week 15. The table also shows the numbers of patients who were scheduled to complete an interview outside the sampling period, and those interviews actually completed from this subset of interviews scheduled "outside" the period.

Table 2.4 RTS Summary Interview Counts

Facility	Week	Selected	Refused	Interviewed	Ineligible	Outside	Interviewed Outside
Fac A	1	7	0	6	0	1	
Fac A	2	3	0	3	0	0	
Fac A	3	6	1	5	0	0	
Fac A	4	3	1	2	0	0	
Fac A	5	14	1	12	0	1	1
Fac A	6	10	3	6	0	1	
Fac A	7	8	2	4	1	1	1
Fac A	8	6	0	5	0	1	1
Fac A	9	7	2	4	0	1	1
Fac A	10	10	0	3	2	5	4
Fac A	11	5	1	3	1	0	
Fac A	12	14	4	3	3	4	2
Fac A	13	3	0	3	0	0	
Fac A	15	2	0	2	0	0	
Fac A	*1 to 15*	*98*	*15*	*61*	*7*	*15*	
Fac B	3	4	2	2	0	0	
Fac B	4	7	0	6	0	1	
Fac B	5	3	2	1	0	0	
Fac B	6	3	0	3	0	0	
Fac B	10	4	0	3	1	0	
Fac B	11	3	0	2	0	1	1
Fac B	13	4	2	2	0	0	
Fac B	14	3	0	2	0	1	
Fac B	*1 to 15*	*31*	*6*	*21*	*1*	*3*	
Fac A and Fac B	**1 to 15**	**129**	**21**	**82**	**8**	**18**	**11**

The probability of selection of a patient also depends on the number of times a patient could be selected using this procedure, i.e., the multiplicity for the given sample patient. Our approach included screening out patients who had visited the facility previously during the period. By making every patient eligible for one and only one visit, we would virtually guarantee the same probability of selection for every patient. It is difficult, however, to ascertain two types of multiplicity:

a) Visits to the same facility, or another facility, taking place in the period after the RTS interview;

b) Visits to other facilities in the frame taking place any time in the period

To estimate the number of visits to the same facility in the period by a same patient, we will obtain and use the unduplicated patient list at the end of data collection (summer 2011). These multiplicities will be used in multiplicity weight adjustments for the sample patients.

3. Weighting

The weighting process includes the creation of base weights, adjustments of the weights for non-response and for multiplicity, and trimming. The first step in the weighting process is to compute base weights, a step described in Section 3.1. Section 3.2 describes the series of weight adjustments that can be performed for the weights.

Base weights, also known as design weights or sampling weights, will be obtained for respondents and non-respondents in the RTS sample. Ineligible patients were identified and eliminated during the sampling and data collection process, so that RTS base weights can be assigned to eligible respondents and eligible non-respondents (refusals).

3.1 Base Weights

Sampling weights, or base weights, computed for eligible sample RTS patients have two basic components associated with the two main sampling stages described earlier, the sampling of events and the sampling of patients.

3.1.1 First-stage sampling weights

First-stage sampling weights are computed as the reciprocal of the selection probabilities for sample events. As described in the sampling sections (see Section 2.1, for example), these probabilities are themselves the products of two selection probabilities corresponding to the sampling of sample days, and the sampling of site-periods (events) for each sample day.

Tables 3.1 and 3.2 illustrate the computation of these weights for the March sample. Table 3.1 details the weights computed for the selection of sample days. Table 3.2 provides the weights computed for the selection of sample site-periods, or events, within the given sample days. It is worth recalling that for the first selection, the measure of size (MOS) is an aggregate for the given day of the week. For the second selection, meanwhile, the MOS is the share of the site-period for the given sample day.

Table 3.1 Example of first-stage sampling weights component #1: sampling of days (March sample).

Date	MOS	Probability	Sampling Weight
3-Mar-2011	101.5	0.573	1.746
7-Mar-2011	120.6	0.681	1.469
8-Mar-2011	118.8	0.670	1.491
9-Mar-2011	113.1	0.638	1.567
10-Mar-2011	101.5	0.573	1.746
14-Mar-2011	120.6	0.681	1.469
17-Mar-2011	101.5	0.573	1.746
18-Mar-2011	82.8	0.467	2.140
21-Mar-2011	120.6	0.681	1.469
22-Mar-2011	118.8	0.670	1.491
23-Mar-2011	113.1	0.638	1.567
25-Mar-2011	82.8	0.467	2.140
28-Mar-2011	120.6	0.681	1.469
29-Mar-2011	118.8	0.670	1.491

Table 3.2 Example of first-stage sampling weights component #2: sampling of site-periods within sample days (March sample).

Date	Facility	Day Period	MOS	Expected Hits	Sampling Weight
3-Mar-2011	A	Thursday AM	41.6	0.410	2.442
7-Mar-2011	E	Monday AM	12.8	0.106	9.455
8-Mar-2011	A	Tuesday PM	58.7	0.494	2.025
9-Mar-2011	A	Wednesday PM	67.9	0.600	1.666
10-Mar-2011	A	Thursday AM	41.6	0.410	2.442
14-Mar-2011	A	Monday AM	60.2	0.499	2.003
17-Mar-2011	A	Thursday PM	39.8	0.392	2.548
18-Mar-2011	E	Friday AM	3.6	0.043	23.000
21-Mar-2011	A	Monday PM	44.0	0.365	2.741
22-Mar-2011	A	Tuesday AM	45.9	0.387	2.587
23-Mar-2011	A	Wednesday AM	41.5	0.367	2.726
25-Mar-2011	A	Friday PM	32.2	0.389	2.573
28-Mar-2011	A	Monday AM	60.2	0.499	2.003
29-Mar-2011	E	Tuesday AM	10.5	0.088	11.319

3.1.2 Second-stage sampling weights

The second-stage sampling involved the selection of eligible patients with equal probabilities from within the set of eligible patients showing up for an appointment at the given sample event (site-period). The second-stage sampling weight is computed as the reciprocal of this conditional probability of selection for patients selected with systematic random sampling in the site-period (event).

Specifically, the weight is the ratio of the total number of eligible patients enumerated for the event and the number of eligible selections. Recall from Section 1.2 that the best measure for the numerator of this weight is the confirmed patient count obtained from the office at the end of the sample period data collection.

3.2 Weight Adjustments

We apply non-response adjustments for the RTS sample that are separate from any other sample component. For one study, these weights need to be combined with the non-RTS sample. In addition, weights are adjusted for multiplicity that may arise due to the increased chances of selection for patients that visit the RTS facility more than once, as well as other RTS facilities and non-RTS facilities more generally.

Using the adjusted weight, $\hat{W}_j$, for patient j, we define the estimated probability of selection for patient j as follows:

$$P_j = \frac{1}{\hat{w}_j} \qquad (3.1)$$

Then, the probability of the patient being selected one or more times can be estimated as follows:

$$P_{jj} = 1 - (1 - P_j)^m \qquad (3.2)$$

Here, "m" is the number of facilities from which the patient could have been selected. Thus the adjusted weight is the inverse of this probability estimated in (3.2). The multiplicity associated with visits to the same facility during the period will be ascertained from an unduplicated patient file obtained from each RTS facility after the end of the data collection period.

4. Conclusion

The pilot study demonstrated the feasibility and efficiency of selecting a probability sample of patients in several stages from each sample facility. The approach was flexible and adaptive allowing for fine tuning the sampling parameters in each monthly wave. It should be noted that a cooperative relationship between the facility staff and data collection team was essential to the success of this project. Thus, we recommend meetings with the facility staff that would be involved, and dry runs of the RTS procedures in the study facilities before commencing data collection. The most evident downside is the additional burden this methodology imposes on technical personnel to monitor the continuous data collection and the sampling and weighting.

Acknowledgements

We would like to thank our colleagues at the Centers for Disease Control and Prevention, the Philadelphia MMP data collection team and the staff from the medical facilities whose assistance was essential to the success of this study.