TIME AND COSTS ASSOCIATED WITH INACCURATE ADMINISTRATIVE DATA

Karen A. CyBulski, Michael D. Sinclair, Frank J. Potter, Bidisha Ghosh, Barbara A. Kolln and Anne B. Ciemnecki
Mathematica Policy Research, P.O. Box 2393, Princeton NJ 08543-2393

KEYWORDS: Survey Response Rates, Time and Cost, Nonlocatability

I. Overview

Government agencies rely of surveys to collect perceptual information about public programs. Mathematica has conducted surveys for the Health Care Financing Administration[1] (HCFA – now the Centers for Medicare and Medicaid Services or CMS) to assess access to care and quality of care for Medicaid recipients who switched from traditional fee-for-service to managed health care. These were list frame surveys. There are two main sources of nonresponse for list frame surveys – respondents who cannot be located and those who refuse to participate. The focus of this paper is nonlocatability.

The response rate for a list frame sample is: completed interviews divided by eligible sample members. This definition implies that someone who cannot be located for an interview is included in the denominator. Therefore, if the sample frame contains incomplete or inaccurate contact information, the response rate will be low.

This paper will:

- Review results from surveys of Medicaid recipients, some of whom also collected Supplemental Security Income (SSI) or Temporary Assistance for Needy Families (TANF).
- Assess characteristics of contact data on administrative records as predictors for survey costs and response rates using both descriptive statistics and logistic standard multivariate regression techniques.
- Recommend ways to improve the quality of contact data.

[1]This research was funded by the U.S. Department of Health and Human Services (DHHS), Health Care Financing Administration (HCFA), with additional funding from the Assistant Secretary for Planning and Evaluation (ASPE) and the Substance Abuse and Mental Health Services Administration (SAMHSA). The contract numbers were 500-94-0047 and 500-95-0040.

II. Results from the Recent Surveys

The sample frames for these evaluations were state Medicaid files. The data set was comprised of eleven separate surveys in seven states, conducted between 1998 and 2001 (Ciemnecki et al, 2001; CyBulski and Ciemnecki, 2000) . A total of 12,131 interviews were conducted. The cooperation rate for these surveys was high. Nine out of ten people who were located completed the interview. Unfortunately, only 70 percent of the sample was located, even after considerable time and resources were spent. Thus, the overall response rate was only 61 percent (Table 1).

The low response rate was due, in part, to the quality of the contact data on the state files (Ciemnecki et al, 2000). Contact data on these files were often out of date, inaccurate, or incomplete. Their quality affected both survey quality and cost, resulting in response rates that were lower than desired, the potential for increased bias, longer field periods, and higher survey costs.

Some of the state's administrative records contained demographic data. These data were used to determine which demographic groups were under-represented in these surveys. Minorities, especially, Hispanics, Native Americans and African Americans were underrepresented, as were males, young and middle aged adults, those living in large households, and those with less than 12 years of education. Individuals who were under-represented in these surveys could be the most vulnerable and in need of high quality program services.

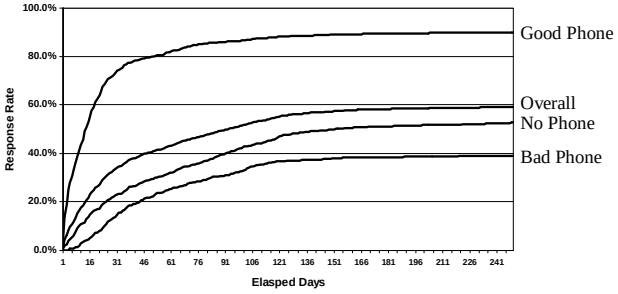
III. Factors Affecting Response Rates

In these sample frames, over half the sample had no phone number, about 20 percent had a bad phone number, and only a quarter had a good phone number (Figure 1). Among the cases with good telephone numbers, a 60 percent response rate was reached after two weeks of telephone interviewing. This response rate was the same as the final survey response rate. Within 5 weeks, the response rate had reached 80 percent. In contrast, by eight months, the end of the telephone interviewing period, neither the no phone number nor the bad phone number groups had reached a 60 percent response rate, while cases with a good phone number reached a 90 percent response rate.

TABLE 1
Completed Interviews, Cooperation Rates and Response Rates by State

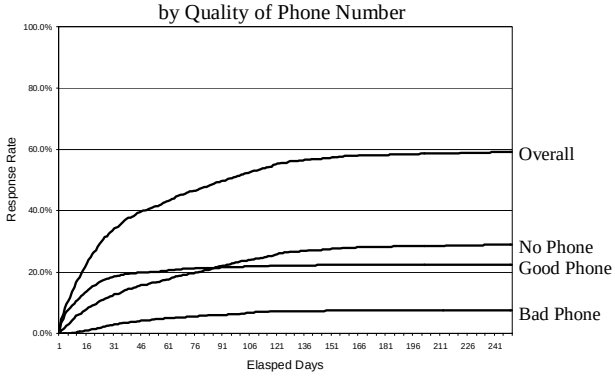
Survey	Completed Interviews	Nonlocatable Rate	Cooperation Rate	Response Rate
Minnesota	2,757	22%	90%	70%
Kentucky	1,779	23%	98%	75%
New York	2,248	37%	93%	57%
Tennessee	2,731	33%	92%	59%
Hawaii	835	36%	76%	45%
Rhode Island	747	27%	87%	58%
Oklahoma	1,034	31%	91%	56%
TOTAL	12,131	30%	91%	61%

Figure 1
Response Rate by Quality of Phone Number



Even though cases with a good phone number had a response rate of 90 percent, they only contributed 23 percent to the final response rate (Figure 2). This was because only 25 percent of the cases on the sample frame had a good phone number. The majority of the cases on the sample file did not have a phone number (55 percent). They contributed only 29 percent to the final response rate. Cases with incorrect telephone numbers (20 percent of the sample) contributed 8 percent to the final response rate.

Figure 2:
Contribution to Overall Response Rate
by Quality of Phone Number



Even after considerable effort, correct telephone numbers were located for only 70 percent of the sample. Significantly more cases were located that had a Social Security Number (70 percent) than cases that were missing a Social Security Number (62 percent) on the sample frame. Likewise, significantly more cases were located that had a standard address (55 percent) rather than a Post Office Box, rural route or general delivery (47 percent). Prior to the start of interviewing, all cases were sent to an outside vendor to conduct a computer match for a telephone number. As expected, a correct phone number was located for significantly more cases for which a number was found through the computer match (81 percent) than when a number was not found (61 percent).

IV. Modeling Time Spent on Different Activities

For six of the surveys we conducted a regression analysis to identify the amount of time spent on the various components of the data collection process. Specifically, we identified the amount of time spent on:

- Conducting the interview once participation was secured.
- Conducting refusal conversions once participation was refused.
- Conducting the first locating procedure to obtain a phone number when the case originally had a missing phone number.
- Conducting the first locating procedure to obtain a new phone number when the phone number on the sample file was incorrect.
- Conducting each subsequent locating effort (after accounting for either of the two prior situations above).

TABLE 2
Regression Model

	Hawaii[2]	Minnesota	Tennessee	New York City	Kentucky	Westchester County	Average	Average Hours
Number of Attempted Cases	1,438	1,742	1,854	2,352	2,413	1,708	1,918	
Model R-square value in percent	98.2	93.4	96.3	99.1	99.6	98.7	97.5	
Model Coefficients								
Completed Survey	24.3	31.6	17.6	27.2	27.6	24.7	25.5	0.42
Number of Refusal Outcomes	1.0	2.8	1.1	1.6	1.5	1.4	1.5	0.03
Frame Phone Number Incorrect	20.5	32.8	NA-none	36.3	NA-none	24.7	28.6	0.48
Time Spent on First Locating Effort Frame Didn't Provide Phone Number	17.8	35.1	NA[3]	24.7	NA[4]	23.4	25.3	0.42
Time Spent on First Locating Effort Addl. Number of Locating Attempts	17.3	27.1	13.5	25.8	45.7	25.6	25.8	0.43
Subtotals								
Number of Extra Locating Attempts	8.1	3.2	6.4	5.8	6.0	6.4	6.0	0.10
Total Time Added for No Phone	157.8	122.8	86.1	175.2	272.6	187.9	167.1	2.78
Total Time Added for Bad Phone	160.5	120.6	NA	186.8	NA	189.2	164.3	2.74

To conduct this analysis we obtained the computer assisted telephone interviewing (CATI) call history for each case. Our CATI system recorded the time spent on each call attempt and the outcome of the calling procedure. Hence, we were able to determine how much total time was spent on each case and whether certain activities (listed above) were necessary (and how many of each type occurred) to obtain final dispositions (e.g. completed eligible or ineligible interview, final refusal, or effort ended - if no contact had been made by the end of the field period).

We prepared a weighted standard regression model for each study to predict the total time spent on a case, in minutes, as a function of the case's history. To set up the model, we developed a summary record for each case that contained the total time spent on all attempts and the values for a series of event indicators, or event counts. For example suppose a case had three call attempts, with the first being a locating attempt at 30 minutes, as a result of not having an initial phone number. The second call resulted in a "ring, no answer" at 5 minutes, and the third, resulted in a successfully completed interview at 25 minutes. Therefore, the summary record showed a total of one hour as the time spent on the case and had a complete indicator with a value of one, and an initial phone missing indicator of one. Since no additional locating attempts beyond the first were required, the number of additional locating attempts was set to zero. In this approach, the beta coefficients from the model represented the amount of time spent various situations averaged over all cases in the study.

Because our locating efforts were conducted in a separate operation, the CATI system did not provide the amount of time spent on each locating attempt. On the other hand, the system did provide us with a record for each locating attempt, and as such, the total number of locating attempts spent on each case (each of these events initially had a time consumed value of 0). To prepare the data for the model in this situation, we assigned the average time spent on a locating event in each study to each of the locating attempt records. To compute the average for each study, we divided the total hours spent by locating staff by the total number of locating attempts on all cases.

In addition to the event indicators on the summary records, we also designed the models to control for differences in the demographic traits of the sampled persons that might be related to case time (e.g. age). For those studies for which we had demographic characteristics on the sampled cases, we induced demographic characteristics indicators in the model. We also used a normalized version of the survey weights (the inverse probablity of selection re-scaled to sum to the sample size) to conduct a weighted regression analysis. This helped to account

[2] Only data from the 2000 survey are included for Tennessee and Hawaii.

[3] For Tennessee, all of the cases did not have a starting phone number. In this case the number of additional locating attempts includes the initial locating effort.

[4] Like TN, in KY nearly all of the cases (all but 71 of 2,413) did not have a starting phone number. As a result, we eliminated this variable from the model so that the number of additional locating attempts includes the initial locating effort.

for the differential probabilities of selection in these studies and their relationship to characteristics that could influence the level of time spent to obtain final disposition.

The results from the six regression models are presented in Table 2. Table 2 also presents the coefficients associated with the case events. Overall the models provided an excellent fit with r-square values averaging 97.5 percent for all models. Table 2 also presents the average coefficient values across the six studies.

In summary, on average, we spent 0.42 hours conducting the interview. For cases that refused, we spent 0.03 hour converting the refusal. Based on the total hours consumed in all studies, and considering the time spent to conduct an interviews from these models, we estimate that interviewers spent an additional 0.81 hours dialing the phone and contacting respondents once a case was located. In contrast, if locating was needed, we spent nearly 3.0 hours on average finding sample members, or 2.4 times more time locating sample members than we did contacting and interviewing them.

V. Recommendations and Conclusions

Good telephone numbers on the Medicaid file led to increased response rates and decreased fielding time and cost. States can improve their contact data by collecting complete and accurate information and updating their electronic records frequently.

A simple, inexpensive way to update addresses is to include "address service requested" on mailings to clients. With this designation on outgoing mail, the post office returns new contact information to the agency. Another suggestion is to provide clients with post cards to mail to the state Medicaid office when contact information changes. Once the new contact information has been received, states could update their records. Also, the Medicaid offices could coordinate with other state agencies such as food stamp and welfare programs, that often update contact information more frequently.

Once collected, complete and accurate information needs to be entered in to the electronic files. States should include the complete address. This includes all lines of the address including apartment numbers and zip codes (which were entirely missing from one state's data). Also, the spacing of the address is very important – 11 19th St. should not be confused with 1 119th St.

In addition, phone numbers need be complete, including the area code. Changes to area codes need to be updated. One of the states had no area codes on the file. Since good telephone numbers help ensure a high response rate, states should ask for the phone numbers of client who report an unlisted number. If a client does not have a telephone, states can ask for the phone number of a friend or relative who would be willing to pass on a message.

In conclusion, surveys are important sources of information about public programs, and accurate contact information contributes to higher quality data at lower costs. The inability locate respondents affects the quality of the data and increases the survey costs. It took 2.4 times as long to find a respondent than to interview one whose whereabouts was known.

Nonlocatability also affects the response rate. It takes much less time to reach respectable response rates when phone numbers are accurate. Within 5 weeks we were able to reach an 80 percent response rate for cases with a good number; whereas even after 8 months cases without a phone number or ones with a bad number did not event reach 50 percent.

Some of the changes that would improve the quality of the information are low cost and easy to implement. Although these recommendations would certainly help survey researchers, they have wide-reaching benefits for policy makers, states, and ultimately beneficiaries of the programs, because programs can best serve people if they can keep in touch with them.

References

Ciemnecki, Anne B., Karen A. CyBulski, John W. Hall, and Barbara A. Kolln. "Evaluation of Five Section 1115 Medicaid Reform Demonstrations: Survey Methodology." April 2001. Princeton, NJ: Mathematica Policy Research, Inc.

Ciemnecki, Anne, Karen CyBulski, and Judith Wooldridge. "Consequences of Inadequate Medicaid Administrative Data on Survey Response Rates and Medicaid Health Policy Issues." Paper presented at the 128th American Public Health Association Annual Meeting, Boston, MA, November 12-16, 2000.

CyBulski, Karen A., and Anne B. Ciemnecki. "Interviewing Populations with Disabilities by Telephone: Survey Design and Operations." American Statistical Association Survey Research Methods Section Proceedings, 2000, pp. 1016-1021.