The Costs and Benefits of Improving Response Rates of the CAHPS® Medicare Fee For Service Survey

Larry N. Campbell
RTI International, Research Triangle Park, NC 27709

Linda L. Dimitropoulos
RTI International, Chicago, Il 60601

Lisa Carpenter
RTI International, Research Triangle Park, NC 27709

G. Gordon Brown
RTI International, Research Triangle Park, NC 27709

Key words: Telephone non-response follow-up, CAHPS Fee for Service, survival analysis

Introduction

This research explores the costs and benefits of increasing the level of effort (defined as number of call attempts) during the telephone follow-up phase of an annual mail survey of beneficiaries enrolled in the Medicare Fee-For-Service plan (MFFS) (a.k.a. Original Medicare). The CAHPS® (Consumer Assessment of Health Plans Study) MFFS Survey is one of three CAHPS surveys of Medicare beneficiaries conducted annually by the Centers for Medicare & Medicaid Services (CMS). The CAHPS surveys fulfill a requirement of the Balanced Budget Act of 1997 to provide information to Medicare beneficiaries on the quality of health services provided through the Original Medicare Plan and to compare this information to similar information collected from beneficiaries enrolled in Medicare managed-care health plans.

The CAHPS-MFFS serves as an important source of information for Medicare beneficiaries to use when making health care decisions. It is critical that the survey reach an appropriate level of response to conduct the necessary subgroup analyses. The project team has conducted a number of studies to identify methods to increase response rates by using resources efficiently.

For the 2003 CAHPS-MFFS Survey, we selected a stratified random sample of 178,950 beneficiaries from more than 30.3 million Medicare FFS beneficiaries residing in the U.S., Puerto Rico and the Virgin Islands in September, 2003 who had been enrolled in Medicare for at least six months and who did not have a representative payee. The survey is conducted primarily as a mail survey with an inbound computer-assisted telephone interview (CATI) option. Beneficiaries who did not respond to either of the first two mailings were placed into a telephone follow-up if a telephone number could be obtained for them. Beneficiaries for whom a telephone number could not be obtained were followed up by mail.

The non-response follow-up to a mail survey, by telephone or special delivery mail, is the final step in the standard mail survey methodology and has been shown to reliably boost response among non-respondents to earlier attempts (Dillman, Christenson, Carpenter and Brooks, 1974; Dillman, 1978; 2000; Heberlein and Baumgartner, 1978). In Dillman’s very early work, the follow-up was conducted by certified mail which reportedly increased response rates by an average of 13%. More importantly, certified mail seemed to get response from people who were older, less educated and reported lower incomes thereby reducing non-response error on those variables. Moore and Dillman (1980) investigated alternatives to certified mail by testing special delivery and telephone follow-up calls and found that, like certified mail, both produced a significant increase in response over first class mail. Dillman attributes the affect to the “out-of-the-ordinary” mailing procedures.

An important issue with a telephone follow-up of non-respondents is determining the maximum number of follow-up calls to make.
The level of effort of the telephone follow-up is closely tied to issues of data quality and survey efficiency. This study explored the following research questions using the 2003 CAHPS-MFFS data:

- Are we penetrating an under-represented population segment with additional calls?
- To what degree do we experience diminishing returns with the number of call attempts? What are the cost implications?
- Are the CAHPS measures affected?

**Method**

**Data issues**

Our telephone subcontractor supplied a file containing a detailed call history of each beneficiary participating in the telephone follow-up. Each telephone attempt was assigned a status code. The type of code determines if another attempt would be made. Complete and final codes such as refusals, beneficiaries found to be institutionalized, and beneficiary not available for duration of survey, result in no more contact. Whereas, a pending code, such as no answer, busy, or not available at the moment, would lead to additional attempts.

To determine the number of attempts for each sample member we used the counter from the subcontractor’s computer-aided interviewing software which follows industry standards. This is the counter that they used to determine when the maximum number of attempts has been reached. Although the counter is always a whole number, it may take two or more of a certain type of call, like a busy signal, to increment the counter.

At the beginning of the telephone portion of our survey we set a maximum of 10 attempts. However, midway through the telephone follow-up we decided to extend data collection for 2 weeks and increased the maximum to 12 attempts. This decision was made because we wanted to improve our lagging response rate and it was within our budget to do so.

The subcontractor provided a call history for all 77,245 cases sent to telephone follow-up, however, we removed certain individuals from our analysis for a total of 67,413 cases. We excluded those who completed a mail survey during telephone follow-up and those who we later found to be deceased. Sample members from Puerto Rico or the Virgin Islands were removed from analysis because it has been our experience that these individuals respond differently than other sample members and we wanted to avoid this confounding effect.

Demographic variables used in the modeling, race, age, gender, dual eligibility, disabled, come from the CMS Enrollment Database (EDB), and are available for all sample members. The urban county indicator is based on the Beale code of the individual’s home county. Question variables are taken directly from responses to the 2003 CAHPS-MFFS Survey.

**Why Survival Analysis.**

We used survival analysis to address whether we were penetrating an underrepresented population segment and to what degree we experience diminishing returns with the number of call attempts (is this accurate?). Survival analysis is appropriate for ‘time to event’ data when there are censored observations in the data. The event in this instance is completing the survey. Final codes, those that result in no further contact but are not a complete, were considered censored. Due to the nature of the response, survival analysis is preferred over other regression methods. Analysis methods such as linear regression and logistic regression don’t adequately handle censored observations. Failure to account for the censored individuals results in biased regression coefficients, poor hypothesis tests and misleading inference.

We used the discrete proportional hazards model (ref) to fit the data, modeling the level of effort to achieve a completed survey. The level of effort was defined as the number of attempts ranging from 1 to 13. A 13 was given if the maximum number of attempts was made without a final code. All observations remaining in the population after the 12th attempt were considered censored at the 13th attempt. This decision was made to satisfy modeling assumptions. The covariates included in the model were race, age category, dual eligibility, disability indicator, and urban county indicator.

**Mathematical description of model**

The discrete proportional hazards model will fit discrete data for a finite number of time periods, $m$. In our study let $i=1, 2, \ldots, 67,413$ represent
individuals and \( j = 1, 2, \ldots, 13 \) represent the calling period. We assume that the \( i \)-th individual censored in the \( j \)-th interval is censored at the beginning of the \( j \)-th interval. The discrete model assumes the following proportional hazard function, under which the probability that the \( i \)-th individual with \( p \)-dimensional covariate vector \( x \) has survived (not a complete or censor) the \((j-1)\)-th interval and has an event (complete) in the \( j \)-th interval is

\[
P(j|x) = 1 - \left(1 - \lambda_j \right) \exp(\beta' x)
\]

where \( \beta' \) is the vector of regression coefficients and \( \lambda_j \) is the baseline hazard for the \( j \)-th interval.

Let \( t_i \) represent the time an individual leaves the population either by a event (complete) or censor. The probability that a given individual \( i \) is in the population at time 1, does not have an event at times 1 to \( t_{i-1} \), and does have an event at time \( t_i \) is

\[
P(t_{2i} = k \mid x_j) = \left[1 - \left(1 - \lambda_i \right) \exp(\beta' x_j) \right] \prod_{j=1}^{t_i-1} \left[1 - \lambda_j \right] \exp(\beta' x_j)
\]

SUDAAN’s SURVIVAL procedure (Research Triangle Institute, 2004) was used to fit all models. This procedure is capable of fitting the discrete proportional hazards model and adjusting for sample design issues inherent in the CAHPS survey such as stratification and unequal weighting.

In several situations we wished to report the predictive margins (direct standardization) for various categorical variables. The methodology for predictive margins is given in Korn and Graubard, 1999 and a synopsis is given here. The concept is to calculate an expected response for a given level of a categorical variable controlling for all other variables in the model. [Is more explanation required?] In this situation, an ‘expected response’ is the vector of probabilities of under going an event (complete) at the each of the twelve sampling times. These are, in effect, the \( \lambda_j \) for a given group and can be used to construct the cumulative response probabilities given in Section ***.[Need graph of response probs] SUDAAN was not capable of directly calculating the predicted margins for the discrete proportional hazards model. A SAS macro was used that manipulated SUDAAN output to accomplish this task. The general concept was to fit the full model and calculate response probabilities for each level of race. The response probabilities were then adjusted by race using the population distributions of the covariates used in the full model. The results produced race specific response probabilities.

One question we wanted to address was the proportionality assumption inherent in the survival model that was used. For example, we wanted to know if there was a race by sampling period interaction. If such an interaction were significant, this would imply that different elements of the population were being ‘completed’ at during different sampling periods. Models were fit separately for each level of each categorical variable of interest. The response probabilities for each attempt, 1-12, were output. Since attempt 13 reflects those who had no response after 12 attempts, the response probability for attempt 13 was set to zero. From our previous overall model we obtained the hazard ratios for each covariate. We then calculated every individual’s response probability at each attempt using these \( \lambda_j \) and hazard ratios. The mean response propensities for each attempt were then compared graphically. [It would be nice to show a graph illustrating this.]

Attempts versus covariates – test for interactions. Need to compare the slopes of the lines. Constant differences can be detected using the model; differences in slope will need to be done outside of the model. This has been done visually. It can be done mathematically but will probably be somewhat ad hoc.

Larry: Do any tests need to be performed that have not been?
Tests of the interactions; specifically the Age*attempts interaction–what does it mean. Can it be reasonably interpreted? It is a quadratic effect and rises at both ends of the model. It makes for an odd effect. Most interactions were not significant.

Show results of full models? [Perhaps in results section]. Perhaps showing the results from at least one of the models is a reasonable approach to take. Show only the parameters of interest?
Reduction of bias- compare the first 6 periods vs the last 6 periods. *This test has not been conducted as far as I know. Should not be hard to do.*

Test drop off – lambda(1-6) vs lambda(7-12)

Model trend - Linear trend

Discuss how we examined the effect on CAHPS measures

**Results Larry/Lisa/Gordon**

Overall percent response- extrapolate out to higher values. Show “diminishing returns” as a function of total percentage of people who complete. *This hasn’t been done that I know of. Will need to create the test to do this.*

Predicted overall cumulative vs Predicted conditional response. Explain the differences: Overall is percent of entire population that is a complete. Conditional is percent of available population for the xth call that is a complete.

Odd flip up at the end of sampling, 11th and 12th calling period.

Predicted Cumulative Graphs: some, if not all, showed be shown, very informative.

Effect on CAHPS measures: Most are unaffected. Age is problematic. How to interpret quadratic trend? Artifact of the data? Show some results?

**Discussion**

It is possible that additional call attempts, particularly when the number of attempts approaches the maximum, will yield responses from a group of beneficiaries whose demographics are different from those who responded to fewer attempts. Since it is desirable to have the respondents fully represent the population sampled, it makes sense to consider additional call attempts to reach an under-represented population segment.

With respect to the demographic variables we examined—age, race, gender, dual eligibility, and disability, we did not find that additional call attempts resulted in penetrating a population segment with different demographic characteristics. For example, Figure 3 describes cumulative response rates for five racial categories. If there were a difference in response rates based on numbers of call attempts among the racial groupings, we would expect to observe plot lines that deviate markedly from parallel. While the plot line in Figure 3 for the Asian race category does reflect a significant difference from the other racial classifications in response rate, it is reasonably parallel to the other four categories which clearly, in this plot, are parallel.

We observed the same general result for age, gender, dual eligibility, and disability, further supporting the assertion that response distributions resulting from a maximum number of call attempts do not differ from those distributions that result from fewer attempts. Thus, there is insufficient justification for the use of increased call attempts as a strategy for reaching an under-represented segment of this population.

The economic principle of diminishing returns is relevant to the interpretation of our findings. We can expect that, at some point, additional call attempts will produce a negligible improvement in response rate, and will therefore cease to be a cost effective strategy.

Using the CAHPS-MFFS survival modeling of the predicted conditional response rate for each attempt, we have produced Figures 1 and 2. With a maximum conditional response rate of 5.0% (on attempt 2) and a minimum rate of 3.0% (on attempt 10), this survey does not show any significantly diminishing response rate for an increase in attempts up to the maximum of 12. The downsloping curve indicates a slight decline in response rate with additional attempts. However, on the basis of response rate alone, cost aside, there is no point on this curve beyond which one would not choose to continue with further attempts, since each further attempt is producing a significant response.

We believe the slight tailing up of the curve for attempts 10-12 is an artifact of our increasing the maximum number of calls from 10 to 12 midway through the telephone follow-up. We also used refusal conversion specialists for many of the calls for attempts 10-12.

Figure 3, for the cumulative overall cumulative overall response rate, also demonstrates that there are not diminishing returns for at least the first 12 calls. With diminishing returns, we would observe a general flattening of the curve with increasing number of attempts. In this case, what we observe is a
curve that continues to slope upwards, even for a maximal number of call attempts.

A second critical decision factor in determining the number of attempts to make is how much it costs. We have provided a cost analysis in Figure 5, that allows us to evaluate the benefit of an additional telephone attempt using the marginal costs of a completed survey.

Using the total number of call attempts, 500,000, assigning a relative cost of 1 unit for the average call attempt, and 13.7 units for a completed telephone survey, we produced the marginal costs per completed survey for each number of attempts. For instance, the marginal cost per completed survey of 34 units for attempt number 2, is the incremental cost, above all attempt number one costs, to obtain each survey completed with 2 attempts.

As with the response rate, if there were significantly diminishing returns for additional attempts with respect to cost, we would observe that the marginal cost per completed survey would increase substantially with additional attempts. While we do observe some increase in marginal cost per complete, it is only a moderate increase. The message here is that if there is sufficient budget to support additional attempts, there will not be a significant degradation in the cost efficiency of obtaining completed surveys.

As with the response rate, if there were significantly diminishing returns for additional attempts with respect to cost, we would observe that the marginal cost per completed survey would increase substantially with additional attempts. While we do observe some increase in marginal cost per complete, it is only a moderate increase. The message here is that if there is sufficient budget to support additional attempts, there will not be a significant degradation in the cost efficiency of obtaining completed surveys.

Our final question is whether or not we obtain different answers to the core CAHPS questions from those who answer with many attempts as compared to those who answer with fewer. Without such a difference in the CAHPS measures, we could remove the concern of response bias and our desire for precision would remain as the driving statistical factor for how many attempts to make. Our modeling of the CAHPS measures, taking into account the same demographic variables of age, race, gender, dual eligibility, and disabled, showed that the number of attempts had no significant effect on the CAHPS outcomes. Whether someone responded on the first telephone attempt, or on the twelfth, the CAHPS measures were not different, as adjusted by various demographics.

Conclusion

When making a choice of how many attempts to make in the telephone follow-up of the CAHPS-MFFS annual survey, whether bias is your issue (under-penetrated population segments), or your concern is diminishing returns and cost, it makes sense to set the maximum number of attempts at 12. Without either a significantly diminishing response rate or increasing marginal costs for a completed survey, the threshold for the maximum number of attempts, where the costs exceed the benefits, is beyond 12. As long as the budget is sufficient, the maximum number of attempts on this survey should be set no lower than 12.

Emphasize CAHPS measure are not affected.

Overall percent is diminishing, conditional may not be.

Few, if any, interactions.

References

Dimitropoulos, L.L., Campbell, L.N. & Iannacchione, V.G. (2003). The Effect of Method of Delivery and Response Propensity on Response to a Non-response Follow-up