Medicaid Underreporting in the CPS: Results from a Record Check Study

Joanne Pascale[1], Marc Roemer[2], Dean Resnick[2]
[1]Statistical Research Division, US Census Bureau, 4600 Silver Hill Road, Washington, DC, 20233
[2]US Census Bureau

1. Introduction[1]

1.1 Medicaid Underreporting

For years, health researchers have been concerned about what has come to be called the Medicaid undercount – a persistent and consistent gap between administrative records and survey data regarding the number of people enrolled in Medicaid. Various studies have shown the gap to range between 10 and 30% (Card et al, 2001; Czajka and Lewis, 1999; Blumberg and Cynamon, 1999; Lewis et al, 1998; Klerman, Ringel and Roth, 2005). Because administrative records show higher counts of enrollees, many agencies even attempt to correct the shortfall by imputing coverage for persons who are deemed likely enrolled in Medicaid but whose coverage was not reported (Holahan et al, 1995; Brown et al, 1997; Lewis et al, 1998).

Survey underreporting is certainly a prime suspect in what is causing the gap, but there are a number of other possible contributors as well. First, administrative records often contain data for individuals outside the sample universe of the survey. For example, people who live in institutions or who have died before the interview may not be part of the survey's target population. Second, administrative records often contain duplicate records for individuals who changed residence or dropped out of the program and later re-enrolled. Third, the time frame of the record and survey data may not always match perfectly. For example, records may indicate an individual was covered at one point in time, but that coverage may have ended by the time the survey was administered. And finally, administrative records may include individuals receiving partial as well as comprehensive benefits, while the survey may cue respondents to report only comprehensive coverage.

While some state-level studies take some of these factors into account, until recently there was no national-level centralized database of Medicaid enrollees, thus it was not possible to account for certain differences between the record and survey data (such as duplicates across states). Furthermore, findings about the Medicaid undercount have been necessarily state-level, making generalizations to the nation as a whole very difficult. However, beginning with Fiscal Year 1999, the Balanced Budget Act (BBA) of 1997 requires states to submit all their eligibility and claims data to the Centers for Medicare and Medicaid Services (CMS) on a quarterly basis through the Medicaid Statistical Information System (MSIS). Owing in part to this new opportunity, the U.S. Census Bureau, CMS, the State Health Access Data Assistance Center, and the Assistant Secretary for Planning and Evaluation at the U.S. Department of Health and Human Services collaborated on a study of the CPS Medicaid underestimate in detail, bringing the administrative Medicaid data to bear directly on the CPS by individual-level comparison of records, while eliminating most of the universe and definitional differences between the two data sets (Davern et al, 2007). We believe this research project was the first of its kind. Its major finding after addressing many of the error sources mentioned above – that respondents' failure to report Medicaid coverage is the largest factor driving the underestimate – motivates and provides an analytic basis for this paper.

1.2 The Current Population Survey

While the undercount has been shown to affect several surveys, in this paper we focus on the Current Population Survey (CPS) for several reasons. First, major policy and funding decisions rely on these estimates (Blewett and Davern, forthcoming 2007). Second, the CPS produces the most widely-cited source of estimates on health insurance and the uninsured (Blewett et al, 2004). Third, perhaps for both of these reasons, several other federal and state surveys gauge their own estimates against the CPS and/or use questionnaires similar in structure to the CPS.

The CPS is an interviewer-administered household survey, fielded monthly, whose main purpose is to collect data about the labor force participation of the stateside, civilian, noninstitutional population. Once a year the Annual Social and Economic Supplement (ASEC) follows the basic monthly survey. In the ASEC, a single household respondent reports health insurance status for all household members via a series of eight questions, each on a different type or source of health insurance. Questions on three different sources of private coverage come first (employer-sponsored, directly-purchased and coverage from someone outside the household), followed by four questions on government-related plans (Medicare, Medicaid, SCHIP, and military plans), followed by a catch-all question about "any other plan." These core questions are asked at the household level (i.e.: "At any time during [calendar year] was anyone in this household covered by [plan type]?"). If "yes," a followup question determines which household members had the coverage. Note, however, that in a single-person household this structure reverts to a person-level design (i.e.: "At any time during [calendar year] were you covered by [plan type]?"). Regarding the time frame, the vast majority of interviews are conducted in March each year (with some interviews in February and April), and the questions ask about coverage "at any time" during the previous calendar year. An abbreviated list of questions is shown in Figure 1, while the complete set of questions (including followup questions about details of the coverage) is available at http://www.census.gov/apsd/techdoc/cps/cpsmar07.pdf.

Due in part to concerns over Medicaid underreporting, as well as data quality in general, cognitive testing of the CPS health insurance questions was conducted in 2004. In this test the questionnaire was administered to respondents and then a semi-structured set of retrospective probes was used to learn about the question-answer process from the respondent's perspective. Themes covered in probing included the time frame (i.e.: what months the respondent was thinking of when asked about coverage "during the last 12 months"), household members (which household members the respondent had in mind when answering questions phrased "Was anyone in this household covered by..."), and particular terms and phrases used in the questions (such as policyholder, Medicare, Medicaid and state-specific names for SCHIP). The testing was conducted mimicking actual CPS conditions as much as possible.[2]

In total 27 respondents with a range of demographic characteristics were interviewed (see Pascale 2007 for a full report on the methodology and results).

Results suggested three broad factors related to misreporting (Pascale, 2007). First, the overall CPS questionnaire structure was problematic. Respondents tended to miscategorize their coverage because they tried to "fit" it into questions that came early in the sequence, even though later questions were more appropriate. For example, when asked about employment-related plans a respondent reported that he was a policyholder, and then said his mother was also a policyholder, but it was later discovered that his mother's plan was Medicare. Second, some respondents did not attend to the 12-month reference period, but rather reported their current coverage status or their current "spell" of coverage. This type of reporting behavior could lead to under-reporting among those currently uninsured if they had had coverage at some point earlier in the 12-month period. Finally, the household-level approach seemed especially problematic for respondents in relatively large, complex, or non-traditional households. Respondents often forgot about certain household members entirely, or they had only a vague understanding of their coverage and hence had difficulty answering the CPS questions on particular type of coverage. Somewhat surprisingly, neither of these problems was necessarily associated with the "closeness" of the relationship between the respondent and the household member for whom he or she was reporting. Misreporting occurred for housemates and distant relatives, as well as parents, siblings, and live-in partners.

1.3 Shared Coverage Hypothesis

While somewhat surprising on first glance, in the context of health insurance eligibility rules and the changing complexity of household composition, the results are perhaps not so unusual. Health insurance eligibility generally revolves around some kind of "family unit". For example, for most private and military coverage the unit of eligibility is an adult, spouse and their children (up to age 18 or 21, depending on school enrollment). Medicaid eligibility is generally extended to certain adults and their dependent children, up to a certain age cut-off, depending on the state. And Medicare is an individual-based plan. In traditional households (i.e.: married couples with children), the coverage situation may be fairly simple – one spouse has employer-sponsored coverage which insures the other spouse and all the children – thus the respondent's reporting task for all household members would be fairly straightforward. However, it is not uncommon for spouses to carry their own insurance through their respective jobs, thus complicating the reporting task somewhat. And in non-traditional households there could be a particularly complicated mix of plan types covering the individual members, with no two individuals sharing the same plan type. For example, consider a household where the respondent has his own coverage through direct purchase, his live-in partner and her child have coverage through her job, his mother has Medicare, and his sister and her daughter have Medicaid. In this case the respondent would be asked to report on four different plans (direct purchase, employer-sponsored, Medicare, Medicaid), and he may not know enough about the source or name of the other household members' plans to report them accurately, at least in the terms the questionnaire uses.

Changes in household composition may exacerbate these types of reporting challenges, at least in multi-person households. In recent decades there has been a decline in "traditional" households (married couples with children); in 1960 the proportion of traditional households was 44% and by 1998 that proportion dropped to 25% (Casper and Bianchi, 2002). These traditional arrangements are being replaced by a combination of single-person households (25.5% in 2000, up from 17.1% in 1970) and "non-family households" (a householder sharing the unit with non-relatives) – 5.7% in 2000, up from 1.7% in 1970 (Fields and Casper, 2001).

These factors, and findings from the cognitive testing, lead us to a "shared coverage" hypothesis, which posits that respondents can more accurately report coverage for another household member if they both share the same type of coverage. With regard to Medicaid in particular, while the program does not have the same policyholder-dependent structure as private plans, parents who have Medicaid are often enrolled under the same account number as their children. Furthermore, among household members who are covered by Medicaid under different account numbers, there may be some shared knowledge with regard to program eligibility and enrollment procedures and use of services, hence an awareness among household members of each other's coverage status.

The current research sets out to learn more about the role of shared coverage in the accuracy of reporting, using administrative records as an indicator of true Medicaid status. The question we examine here is: if a CPS-sampled person was covered by Medicaid, is that coverage more likely to be reported if the respondent also had Medicaid than if the respondent did not have Medicaid? In taking up this question, we also consider other factors likely to affect reporting accuracy, including recency and duration of coverage. We proceed as follows: in the next section, we describe the two data sources used in the analysis, the steps taken to bring them together, and the methods used to perform the analysis. Section 3 describes the results of the analysis, and in Section 4 we discuss these results further in the context of relevant literature. Finally in Section 5 we discuss implications for future research.

2. Methods

2.1 The Linked Dataset and the Dependent Variable

To measure reporting accuracy we linked CPS survey data to Medicaid administrative records and deemed the records to be the "true" indicator of Medicaid enrollment. On the CPS side we used data reported for the year 2000, and we used a fairly liberal definition of what it meant to report Medicaid in the survey. Specifically, Medicaid enrollment was assigned for a given household member if the respondent reported that person as having been covered at any one of four survey items, all asking about government coverage (see bolded items 5, 6, 8 and 9 in Figure 1)[3]. Furthermore, we only considered cases where a direct response (yes) was recorded (i.e.: we did not use edited or imputed responses). On the records side, we used MSIS data for the year 2000 and we used a fairly conservative definition of being enrolled. Only persons enrolled with full benefits for at least one full day of the year were considered covered. This excludes from the analysis persons with only partial benefits and those with coverage only under the State Children's Health Insurance Program (SCHIP). The objective was to eliminate individuals from the records whose coverage may not have been reported in the CPS because the respondent deemed it not to be "true" Medicaid – either because benefits were limited or because there was some confusion between SCHIP and Medicaid. Combining the broad definition of reporting

Medicaid on the survey side and narrow definition of enrollment on the records side focuses our analysis on the "hard core" underreporters – those truly on Medicaid (versus some marginal or differently-named coverage) who failed to report it at every opportunity in the survey.

Linking each CPS-sampled person to an MSIS record required first assigning a unique identifier called a Protected Identity Key (PIK) to each person in each of the two datasets. The PIK has a one to one correspondence with Social Security Number, which it replaces to protect the privacy of this data element and the person's identity. If the SSN was available on the source dataset (i.e.: the CPS or the MSIS), it was verified through a process that compared date-of-birth on each record with the date-of-birth recorded for that SSN on the Social Security Administration's Number-Identification database ("Numident") and other administrative data files. CPS addresses and names were also compared; MSIS addresses and names were not available and instead sex-designation was compared. Matches over a certain threshold of quality were accepted as valid. For CPS records, when the SSN could not be confirmed, we used a probabilistic matching process with the Numident file to determine the correct value for it, when possible. Note, however, that if the respondent explicitly refused to provide an SSN in the CPS no search was done, consistent with US Census Bureau policy. Once the individuals on the CPS and MSIS had been assigned a PIK, merging the survey and administrative data sets was straightforward. The resulting merged dataset contained data from the survey side (such as reported Medicaid status, household composition, and demographic characteristics), as well as key data from the records side (such as enrollment status, duration of coverage, and Medicaid case ID information), and variables created for the model (discussed below) drew from both sources.

The March 2001 CPS contained 218,269 sampled persons, and we identified about 80% of them with a PIK, amounting to 173,967 individuals. Most unidentified persons were due to respondents' refusal to provide SSNs (about 13% of sampled persons). The MSIS file contained 45,737,631 persons, and the process identified about 90% of them with a PIK. Removing duplicate records (another 3%) brought the number of MSIS people eligible for matching to 39,911,501. When we merged the two datasets on common PIK the resulting analysis file contained 19,345 person records. This linked file represents people "known" to have been on Medicaid (according to the records) for whom we have an explicit "yes" response to the survey questions about Medicaid. Note that some individuals in the MSIS dataset may indeed be covered by Medicaid but because a PIK could not be verified, they could not be confirmed to be covered. The dependent variable, then, is whether those with a verified PIK and known to have been on Medicaid had that coverage reported for them in the CPS.

2.2 The Logistic Regression Model and Independent Variables

While this investigation began with a focus on shared coverage among household members, to control for and examine other factors related to Medicaid reporting accuracy, we built a well-fitting model that included those variables as well. Most of these variables related to the presumed saliency of Medicaid to the respondent in some way – that is, how important or aware the respondent may be of Medicaid, for him or herself and others in the household. In general we focused on five themes: household composition, recency of coverage, duration of coverage, receipt of Medicaid services, and demographics. While not directly related to saliency, demographics may control for other factors related to response accuracy and provide additional information about the relationship between the respondent and the person for whom he or she was reporting (called the "referent"). We explored various ways of operationalizing each of these themes as variables in the model, and tested several different versions to examine their individual effects and their contribution to the overall model.

To operationalize the concept of shared coverage we created a variable (called SHARED) to represent various relationships between the respondent and the referent. The first main distinction in this variable was whether the referent was one-and-the-same as the respondent (i.e.: a self-report) or whether the referent was reported on by someone else (a proxy report). For self-reporters we further distinguished those in single- versus multi-person households. Among the proxy reporters we created three categories: (1) both respondent and referent were covered by Medicaid and on the same Medicaid account; (2) both were covered by Medicaid but on different accounts; and (3) the referent was on Medicaid but the respondent was not. Table 1 shows the distribution of the sample across these categories.

The other variables included in the model were fairly straightforward. Regarding household composition, in addition to SHARED we included an indicator of whether another household member also had Medicaid within the year (OTHMEMB). For recency of coverage we included two measures: the most recent month enrolled in Medicaid (LAST_MNTH), and whether the referent was currently enrolled in Medicaid (i.e.: enrolled in the month in which the survey was conducted) – COV_NOW. For duration of coverage we constructed a variable which represents the proportion of days covered from January until the end of the last month enrolled (PCT_DAYS). The variable was defined this way in order to disentangle duration from recency of coverage. We also included an indicator of receipt of medical services within the year paid for by Medicaid (SERVICES). And finally, among demographics we included the sex of the respondent (R_SEX), and the age and race/ethnicity of the referent (REF_AGE and REF_RACE).

3. Results

3.1 Overview of Linked Dataset

First we provide descriptives and results of correct Medicaid reporting for the set of people in the linked dataset (Table 2), which indicate they are mostly female and often children, especially younger children. The rate of correct reporting of Medicaid status is 63.8% overall, with little difference between the sexes, but substantial differences by age. Reporting is best for those in the youngest age group (0-5) and those in the 45-64 year-old age range and worst for those in the 18-44 year-old range.

3.2 Logistic Regression Model Results

We used two tests to evaluate the quality of the model (Table 3). The Association of Predicted Probabilities and Observed Responses, which compares the order of each pair of persons in the dataset obtained by the model's predicted probabilities and by the zero-one flag existing in the dataset. This test shows correct ordering for 76.3% of pairs. The Hosmer and Lemeshow test ranks the persons in the dataset by the predicted probability, divides them into deciles, sums the probabilities in each decile to estimate the

expected number with correct and incorrect reporting, and performs a chi-square test against the actual number with correct and incorrect reporting. This test demonstrates the model is well-fitting.

Regarding the relative contributions of the independent variables, several interesting findings emerge from the Analysis of Maximum Likelihood Estimates (Table 4). All variables in the model show coefficients that are substantial and highly significant. To evaluate the relative importance of each variable in the model we ranked them according to the increase in the Chi-Square statistic of the Likelihood ratio resulting from each variable's inclusion in the model against an identical model with only this variable removed. This method allows us to rank the variables from the most to least impacting likelihood as shown in Figure 2. By this measure the strongest covariate of accurate Medicaid reporting was recency of coverage (LAST_MNTH); the more recently the referent had Medicaid, the more likely it was to be reported accurately. Duration of coverage (PCT_DAYS) was the second-strongest predictor, meaning that the longer the spell of coverage lasted, the more likely it was to be reported accurately. Receipt of medical services paid for by Medicaid (SERVICES) was also an important contributor to accurate reporting. After these three variables, various demographic and interaction terms (discussed more below) were significant predictors of accurate reporting and, while the shared coverage variable ranked last among the variables in the model, its effects were still substantial and highly discernible.

The estimated coefficients of SHARED demonstrate how coverage shared between the respondent and the referent (whether on the same or different cases) affects the accuracy of reporting. All these coefficients are greater than zero, indicating that a respondent with Medicaid coverage reports better than a respondent without coverage (the omitted group). Respondents in single-person households (SHARED=A) report better than respondents reporting for others (C, D, and the omitted category E), but somewhat surprisingly, reporting is substantially worse by respondents in multi-person households reporting for themselves (SHARED=B). These results are covered in more detail in the discussion section below.

4. Discussion

We initially set out to examine a specific hypothesis suggested from cognitive testing – whether shared coverage between the respondent and the referent was associated with reporting accuracy. Cognitive testing and other research had also suggested that recency of coverage could be a factor in reporting accuracy. We found strong empirical evidence for both these associations – recency of coverage being a more important factor than shared coverage. More specifically, cognitive testing found that when asked about coverage "at any time during the past 12 months" some respondents basically ignored the phrase on time period and instead thought of their current coverage status or spell of coverage; the record-check study found that respondents have a tendency to underreport coverage they have in the more distant past. Together these findings suggest that the underreporting problem may well stem from the questionnaire design – specifically a failure of the phrase "at any time during the past 12 months" to adequately motivate respondents to focus on the appropriate time period. Recall failure – even among those respondents who do focus on the appropriate time period – may also be a factor; indeed it could be a dominating factor. It's likely, though, that the observed underreporting is due to a combination of questionnaire design failure and recall error.

These findings on reference period are not without precedent. Other record-check studies on underreporting of safety-net benefits suggest that either recall error, questionnaire design or some combination is a contributing factor to underreporting, at least in surveys that employ a previous calendar year reference period. Resnick et al (2004) conducted a study linking administrative records of food stamps to survey data and found that "the lowest misreporting rate is for households receiving food stamps in the survey month: 21.2%." Among households last receiving food stamps more than four months prior to the survey interview the misreporting rate was 74.4%. A similar study on welfare (TANF) receipt found strong evidence that respondents "report program participation based on the situation at the time of the interview" not necessarily based on the 12-month time period specified in the questionnaire (Lynch, 2006). Both of these studies suggest that current status overrides attentiveness to the previous calendar year reference period for some respondents.

We turn now to some of the other variables that ranked in the model as important factors. Receipt of Medicaid services within the year had a very strong effect, and there is some support for this in the literature as well. Walden et al (1984) found an association between physician contact and knowledge of health plan benefits. With regard to demographics we can only offer speculations for the findings. Female respondents may report better than male respondents for young children since they may be more often the primary care-giver. There may also be a connection with doctor visits if there are more frequent visits for younger versus older children. Reporting may also be more accurate for younger children than older children because eligibility for Medicaid is more clear and consistent for younger children, while eligibility criteria vary as children age beyond 5 years old (in some states even 2 years old), perhaps confusing respondents reporting for the children. Regarding race and ethnicity, minorities may report less accurately due to language barriers or cultural differences.

Finally with regard to shared coverage, we found clear evidence that indeed when a respondent has Medicaid he or she is more likely to accurately report another household member's coverage. More specifically, we found that respondents reporting for themselves in single-person households are the most accurate reporters, and if the respondent does not have Medicaid but the referent does, reporting is the least accurate. Between the extremes the patterns of reporting accuracy depend on whether the referent had coverage in the survey month.

We were surprised by the finding that respondents covered by Medicaid are much more likely to report accurately in a single- versus multi-person household. This either stems from the fact that the reporting task is fundamentally different in the two settings – having to think only about one's own coverage, versus the coverage of all household members – or from the fundamentally different questionnaire stimulus – "...were you covered by Medicaid?" versus "...was anyone in the household covered by Medicaid?" Or some combination of factors could be at work. The literature offers some corroborating evidence for both possibilities. In tests of the decennial census roster questions there were occasional cases of respondents forgetting to include themselves on the roster of household members. Reasons appeared to be a mix of respondents getting distracted by rules (such as "list the owner/renter first"), thinking the question was just asking about others who they live with, or simply forgetting about themselves (Hunter, 2005; Hunter

and de la Puente, 2005; Childs et al, 2006). With regard to the household- versus person-level questions, prior research suggests that when household members' names are, by design, read by the interviewer, reporting of health insurance goes up (Blumberg et al, 2004; Hess et al,2001). However, there seems to be a trade-off given that the person-level design can increase the length of the survey and induce respondent fatigue. Pascale (2000) found that among large households (those with at least four members) reported Medicaid coverage rates were almost double in the household-version versus the person-level questionnaire design (11.5% vs. 6.2%). And Blumberg (2004 et al) found that longer administration times were related to higher rates of uninsurance and suggests that "respondent fatigue may contribute to higher uninsurance rates."

In sum these results strongly suggest the importance of saliency in accurate reporting. Recency and duration of coverage, receipt of Medicaid services, and shared coverage among household members all contribute toward a respondent's awareness of Medicaid in general, and toward each household member's Medicaid status. This greater awareness could well improve reporting. We can speculate that the effects of shared coverage are associated with events such as doctor visits, conversations about enrollment procedures or eligibility guidelines, and so on, that would inform the respondent of the referent's Medicaid status. While our study was limited to Medicaid, saliency factors may extend to other types of health insurance coverage as well. For example, a respondent with private coverage through a job (even as a dependent) may be knowledgeable about other household members who are also dependents because they share the same benefits package, list of eligible doctors, and so on. But such a respondent may not be aware of the status of household members who are not part of the same coverage unit because the other coverage types are not salient.

5. Future Research

With regard to underreporting of coverage in the more distant past, this research cannot disentangle the relative contributions of questionnaire design and recall error, and likewise a respondent's ability and motivation to recall events is intertwined with the questions that prompt that task. Therefore one obvious avenue to explore is modifying the questionnaire to better motivate respondents to focus on not just current status or spell but also the months further back in the reference period. One strategy would be to "anchor" respondents in their current coverage – since we do have some evidence that at least some respondents have this tendency anyway – and then design questions to "walk" the respondent back through time in some systematic way to get at past coverage.

Another questionnaire design feature that warrants further testing is the household-level phrase "was anyone in this household covered by…," which does not always prompt reporting for all household members, and also risks respondents failing to report themselves in multi-person households. One minor change would be to simply add "you" as follows: "Were you or was anyone in the household covered by Medicaid?" This, however, only has the potential to address underreporting for the respondent, not other household members. Research suggests that mentioning each person by name results in better reporting of insurance overall, but that this needs to be balanced with respondent burden. One strategy taking these factors into account could be to ask about the first person in the household by name, and any plans reported for that person would include follow-up questions to determine whether other household members are also covered on that same plan. Once the series is complete for the first person, it would repeat for the second person but in a way that harnesses any plans reported by the first person about the second person. For example, if the first person reported employer-based coverage and said that the second person was also on that plan, the series for the second person could simply ask: "Other than the employer-based coverage of [first person] does [second person] have any other type of health insurance?"

Finally, the finding that shared coverage between respondent and referent does indeed enhance reporting accuracy suggests that questionnaire design should exploit this. The CPS does, in that it asks about each plan type, one at a time, for all household members. But the utility of the approach may be compromised by the particular phrase "Was anyone in the household covered by..." The alternative approach discussed above – first identifying a plan covering a given household member and then asking if other household members are also covered – exploits the shared coverage reporting advantage and also prompts the respondent with each household member by name. Steps for further testing could explore this alternative approach, and examine whether it seems to hold promise for not just for Medicaid but for reporting of other types of health plans as well.

Notes

1. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.
2. Because the testing was conducted in the fall (versus March), the precise time frame employed by the CPS could not be replicated. Respondents were asked about coverage "at any time during the past 12 months." Regarding content, a subset of questions on work experience and government program participation were asked in an attempt to maintain some of the context of the CPS.
3. Items 5 and 6 are the primary questions on Medicaid and the State Children's Health Insurance Program (SCHIP) respectively. Item 8 is a catch-all item on "other" programs and Item 9 is a verification item. If the answer to Item 8 or 9 is "yes" then the respondent is asked to choose the appropriate coverage from a list of 15 plan types. If they choose Medicaid, SCHIP, or "other government," the household member is coded as having Medicaid (and since the "other" item includes state-specific government program names in the stem question, respondents who choose "other" from the plan type list are also coded as having Medicaid). For readers familiar with the CPS public use file, this is the logic used to create the variable called "MCAID."

Figure 1: Abbreviated Set of CPS Questions (bolded questions used to derive reported Medicaid)

1. These next questions are about health insurance coverage during the calendar year 2000. The questions apply to ALL persons of ALL ages. At any time in 2000, (were you/was anyone in this household) covered by a health plan provided through (their/your) current or former employer or union?
   - 1a. Who in this household were policyholders?
   - 1b. In addition to (you/name), who else in this household was covered by (name's/your) plan?

2. At anytime during 2000, (were you/was anyone in this household) covered by a plan that (you/they) PURCHASED DIRECTLY FROM AN INSURANCE COMPANY, that is, not related to current or past employment?
   2a. Who in this household were policyholders?
   2b. In addition to (you/name), who else in this household was covered by (name's/your) plan?
3. At any time in 2000, (were you/was anyone in this household) covered by the health plan of someone who does not live in this household?
   3a. Who was that?
4. At any time in 2000, (were you/was anyone in this household) covered by Medicare?
   READ IF NECESSARY: Medicare is the health insurance for persons 65 years old and over or persons with disabilities
   4a. Who was that?
5. **At any time in 2000, (were you/was anyone in this household) covered by Medicaid/(state name)?**
   READ IF NECESSARY: Medicaid/(state name) is the government assistance program that pays for health care.
   5a. Who was that?
6. **In (state), the (state SCHIP name) program (also) helps families get health insurance for CHILDREN. (Just to be sure,) Were any of the children in this household covered by that program?**
   READ IF NECESSARY: (state SCHIP name) is the name of (state)'s CHIP program. It is the same as the Children's Health Insurance Program, which helps pay for children's health care.
   6a. Who was that?
7. At any time in 2000, (were you/was anyone in this household) covered by TRICARE, CHAMPUS, CHAMPVA, VA, military health care, or Indian Health Service?
   NOTE: "CHAMPVA" IS THE CIVILIAN HEALTH AND MEDICAL PROGRAM OF THE DEPARTMENT OF VETERAN'S AFFAIRS.
   7a. Who was that?
   7b. What plan (were/was) (name/you) covered by?
8. **Other than the plans I have already talked about, during 2000, was anyone in this household covered by a health insurance plan (such as the [state name] plan or any other type of plan/of any other type)?**
   8a. Who has insurance?
   8b. What type of health insurance did (was/were) (name/you) covered by in 2000? Any other type of plan?
9. **I have recorded that (name/you) (was/were) not covered by a health plan at any time during 2000. Is that correct?**
   9a. Who should be marked as covered?
   9b. What type of health insurance (was/were) (name/you) covered by in 2000? Any other type of plan?

Figure 2: Ranking of Relative Importance of Independent Variables to Overall Model

1. LAST_MNTH:	Recency of coverage
2. PCT_DAYS:	Percentage of days enrolled from January until the last month enrolled
3. SERVICES:	Receipt of a medical service paid for by Medicaid within the calendar year
4. REF_RACE:	White non-hispanic
5. R_SEX:	Female respondent
6. OTHMEMB:	Another household member had full coverage
7. REF_AGE:	Age of referent
8. REF_AGE*R_SEX:	Age * sex of respondent
9. COV_NOW*SHARED	Covered in survey month * shared coverage between respondent and referent
10. COV_NOW	Covered in survey month
11. SHARED	Shared coverage between respondent and referent

Table 1: Shared Coverage Variable Categories and Percent Distribution

Self reports (respondent is the referent)	25.0%
A: Single-person household	6.0%
B: Multi-person household	19.0%
Proxy reports (respondent is reporting for the referent)	75.0%
C: Respondent and referent on Medicaid; same account number	21.1%
D: Respondent and referent on Medicaid; different account number	9.4%
E: Referent on Medicaid; respondent not on Medicaid	44.4%

Table 2: Persons in linked CPS-MSIS dataset by sex, age and Medicaid reporting accuracy

Sex and Age		Reporting in CPS	
		Incorrect	Correct
All		36.2	63.8
Female	58.7	37.2	62.8
Male	41.3	34.7	65.3
Age 00 - 05	19	29.9	70.1
Age 06 - 14	26.4	31.7	68.3
Age 15 – 17	6.6	36.3	63.7
Age 18 – 44	34.8	47.4	52.6
Age 45 – 64	6.3	29.8	70.2
Age 65+	6.9	40	60

Table 3: Logistic Regression Model Test Results

Association of Predicted Probabilities and Observed Responses

Percent Concordant	76.0	Somers' D	0.526
Percent Discordant	23.4	Gamma	0.529
Percent Tied	0.5	Tau-a	0.243
Pairs	86382894	c	0.763

Partition for the Hosmer and Lemeshow Test

		rptd_ma = 1		rptd_ma = 0	
Group	Total	Observed	Expected	Observed	Expected
1	1936	359	352.30	1577	1583.70
2	1936	713	708.24	1223	1227.76
3	1935	968	975.26	967	959.74
4	1950	1156	1175.55	794	774.45
5	2018	1324	1361.03	694	656.97
6	1930	1443	1418.50	487	511.50
7	1955	1507	1512.33	448	442.67
8	2047	1677	1663.28	370	383.72
9	1938	1665	1656.20	273	281.80
10	1700	1539	1527.83	161	172.17

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
7.6401	8	0.4694

Table 4: Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	SE	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.5336	0.1177	901.4917	<.0001
SHARED A	1	0.3936	0.1511	6.7883	0.0092
SHARED B	1	0.0197	0.0895	0.0483	0.8260
SHARED C	1	0.2011	0.0933	4.6492	0.0311
SHARED D	1	0.2493	0.1275	3.8238	0.0505
COV_NOW*SHARED A	1	0.9320	0.1701	30.0337	<.0001
COV_NOW*SHARED B	1	0.5326	0.0962	30.6251	<.0001
COV_NOW*SHARED C	1	0.4994	0.1075	21.5875	<.0001
COV_NOW*SHARED D	1	0.2576	0.1453	3.1441	0.0762
COV_NOW	1	0.4258	0.0654	42.4212	<.0001
SERVICES	1	0.6496	0.0413	246.9868	<.0001
LAST_MNTH	1	0.1388	0.00750	342.3694	<.0001
PCT_DAYS	1	1.0716	0.0699	234.6974	<.0001
R_SEX F	1	0.1987	0.0237	70.2325	<.0001
REF_AGE 00 - 05	1	0.3134	0.0443	50.1071	<.0001
REF_AGE 06 - 14	1	0.2032	0.0415	23.9140	<.0001
REF_AGE 15 - 17	1	0.0344	0.0667	0.2669	0.6054
REF_AGE 18 - 44	1	-0.3580	0.0404	78.5126	<.0001
REF_AGE 45 - 64	1	0.1071	0.0622	2.9604	0.0853
R_SEX*REF_AGE F 00 - 05	1	0.1504	0.0409	13.5491	0.0002
R_SEX*REF_AGE F 06 - 14	1	0.1514	0.0377	16.0881	<.0001
R_SEX*REF_AGE F 15 - 17	1	0.1531	0.0654	5.4822	0.0192
R_SEX*REF_AGE F 18 - 44	1	0.0503	0.0389	1.6704	0.1962
R_SEX*REF_AGE F 45 - 64	1	-0.1242	0.0592	4.3960	0.0360
OTHMEMB	1	0.3305	0.0443	55.6766	<.0001
REF_RACE (Not WNH)	1	-0.3721	0.0343	117.4543	<.0001

Effect	Point Estimate	95% Wald Confidence Limits	
SERVICES	1.915	1.766	2.076
LAST_MNTH	1.149	1.132	1.166
PCT_DAYS	2.920	2.546	3.349
OTHMEMB	1.392	1.276	1.518
REF_RACE (Not WNH)	0.689	0.644	0.737

Acknowledgments

We gratefully acknowledge funding support from the Robert Wood Johnson Foundation and from the Office of the Assistant Secretary for Planning and Evaluation for this project and ongoing research on how to best measure health insurance. We also thank Jeff Moore and Michael Davern for helpful comments on earlier drafts of this paper.

References

Blewett, Lynne A., Good, M.B., Call, K.T., Davern, M. (2004). "Monitoring the Uninsured: A State Policy Perspective," Journal of Health Politics, Policy and Law 29, no. 1: 107-145.

Blewett, Lynne A., Michael Davern, February 2007, Executive Summary: Evaluating Federal Funding Formulas: The State Children's Health Insurance Program. Journal of Health Policy Politics and Law, forthcoming 2007, Vol. 32 No. 3

Blumberg, Steve J., Osborn, L., Luke, J.V., Olson, L., Frankel, M.R. (2004). "Estimating the Prevalence of Uninsured Children: An Evaluation of Data from the National Survey of Children with Special Health Care Needs, 2001." National Center for Health Statistics. Vital Health Stat 2(136). Hyattsville, MD.

Blumberg, Stephen J., Marcie Cynamon (1999), "Misreporting Medicaid Enrollment: Results of Three Studies Linking Telephone Surveys to State Administrative Records." Proceedings of the Seventh Conference on Health Survey Research Methods, DHHS Publication No. (PHS) 01-1013

Brown, E. Richard, Hongjian Yu, Kenneth Fong, Roberta Wyn, William Cumberland, Reecka Levan (1997), "Adjusted Populated-Based Estimates of Medi-Cal Coverage: A Report to the California Department of Health Services," UCLA Center for Health Policy Research, August, 1997.

Caspar, Lynn, Suzanne Bianchi. 2002. Continuity and change in the American family. Thousand Oaks, CA: Sage Publications.

Card, David, Andrew K.G. Hildreth, Lara D. Shore-Sheppard (2001), "The Measurement of Medicaid Coverage in the SIPP: Evidence from California, 1990-1996." National Bureau of Economic Research, Working Paper 8514, October, 2001.

Childs, Jennifer Hunter, Eleanor Gerber, George Carter, Jennifer Beck (2006). "Cognitive Test of the 2006 NRFU: Round 2." Statistical Research Division Study Series Report, Survey Methodology #2006-05.

Czajka, John L., Kimball Lewis (1999), "Using National Survey Data to Analyze Children's Health Insurance Coverage: An Assessment of Issues," Report submitted by Mathematica Policy Research to the U.S. Department of Health and Human Services, May, 1999.

Davern, Michael, Lynne Blewett, Kathleen Call (2007), Presentation on preliminary results from "Research Project to Understand the Medicaid Undercount: Phase 2 Research Results: Examining Discrepancies between the National MSIS and CPS March Supplement (ASEC)," July 17, 2007, Washington, DC.

Fields, Jason, and L.M. Casper. 2001. "America's families and living arrangements." Current population reports P20-537, issued June 2001. U.S. Census Bureau, Washington DC.

Holahan, John, Colin Winterbottom, Shruti Rajan (1995), "A Shifting Picture of Health Insurance Coverage," Health Affairs, Volulme 14, No. 4, 1995

Hess, J., J. Moore, J. Pascale, J. Rothgeb, and C. Keeley (2001), "The Effects of Person-level vs. Household-level Questionnaire Design on Survey Estimates and Data Quality." Public Opinion Quarterly, Winter 2001, 65:574-584.

Hunter, Jennifer (2005). "Cognitive Test of the 2006 NRFU: Round 1." Statistical Research Division Study Series Report, Survey Methodology #2005-07.

Hunter, Jennifer and Manuel de la Puente (2005). "Report on the Cognitive Testing of the Space Saving Features, Roster Questions and Revised Instructions for the 2005 National Census Test." Statistical Research Division Study Series Report, Survey Methodology #2005-03.

Klerman, Jacob Alex, Jeanne S. Ringel and Beth Roth (2005). "Under-Reporting of Medicaid and Welfare in the Current Population Survey." RAND Labor and Population working paper series WR-169-3.

Lewis, Kimball, Marilyn Ellwood, John Czajka, (1998). "Counting the Uninsured: A Review of the Literature." Washington, D.C., The Urban Institute.

Lynch, Victoria (2006). "Causes of Error in Survey Reports About Who in the Household Gets Welfare." Unpublished Paper. College Park, MD: Joint Program in Survey Methodology.

Marquis, Kent H. and Jeffrey C. Moore (1990). "Measurement Errors in the Survey of Income and Program Participation (SIPP) Program Reports." Proceedings of the Bureau of the Census 1990 Annual Research Conference, Arlington, VA.

Pascale, J. 2007. Measuring Health Insurance in the U.S. Research Report Series (Survey Methodology #2007-11). US Census Bureau, Statistical Research Division, Washington DC.

Pascale, Joanne (2000). "Alternative Questionnaire Design Strategies for Measuring Medicaid Participation." Paper presented at the American Public Health Association, Boston.

Resnick, Dean, Love, S., Taeuber, C., Staveley, J.M. (2004), "Analysis of ACS Food Stamp Program Participation Underestimate." Paper presented at the 2004 Joint Statistical Meeting, Toronto.

Walden, Daniel C., Horgan, Constance M., Cafferata, Gail Lee (1984). "Consumer Knowledge of Health Insurance Coverage." Proceedings of the Fourth Conference on Health Survey Research Methods, Washington, DC.