FIELD CODING COMPLEX DATA IN CAPI: AN INVESTIGATION OF THE USE OF DATABASE SEARCH PROCEDURES TO IDENTIFY AND CODE MEDICAL PROVIDERS IN A HOUSEHOLD SURVEY

Maria Elena Sanchez, Agency for Health Care Policy and Research, and Carmen J. Vincent, Westat
Carmen J. Vincent, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850

KEY WORDS: CAPI, unduplication, NMES

Introduction

This paper examines the feasibility of using computer-assisted search procedures in a CAPI interview administered by field interviewers in a national survey. The procedures were designed to help interviewers to identify medical providers mentioned during a household interview. The CAPI interview was part of the Screener round of the 1992 NMES Feasibility Study (NMES-FS). An indexed database of local health care providers (the Provider Database), in combination with search software, enabled interviewers to query the database for entries to match the providers mentioned by family respondents. These procedures were designed to improve upon the timeliness and reliability of procedures used in the 1987 NMES2 medical expenditure survey. This paper discusses the importance of provider identification for NMES surveys; describes the provider identification design modifications introduced in NMES-FS, including a description of the Provider Database, the search software, and the training interviewers received; and evaluates the success of the new provider identification techniques.

NMES2 and NMES-FS

The NMES2 survey provides national estimates of health care use and medical expenditures for persons living in households in the United States. The sample design consists of a national area probability sample of households, with oversampling of groups important to health care policy decisions, such as the elderly, Blacks and Hispanics, the poor and near poor, and persons over 65 with functional impairments. Dwelling units in NMES2 were screened for eligibility from October to early December of 1986 during the Screener round. The final sample selection of households took place at the home office in late December based on data collected in the Screener. Family respondents in NMES2 households thus selected were interviewed four additional times (Rounds 1-4) during 1987 and the early part of 1988. The NMES2 study observation period spanned twelve months (January 1 -- December 31, 1987).

In addition to interviewing households, NMES2 collected data from medical providers identified in the household interviews as having provided care to specific family members. The Medical Provider Survey (MPS) verified and supplemented medical and financial information obtained from household respondents.

The sample design for NMES-FS replicates the NMES2 design. The NMES-FS study design was patterned after NMES2, but modifications were introduced to evaluate alternative design options and new data collection methodologies applicable to future surveys similar to NMES2. Specifically, the study design for NMES-FS replicates half the NMES2 cycle, including a Screener round and two additional interviews with family respondents in sampled households (Rounds 1-2).

Design Changes to Improve Identification of Medical Providers

In NMES2, family respondents had difficulty supplying complete and accurate information (full name, address, telephone and specialty) to identify medical providers who had treated family members during 1987. The incomplete and inaccurate information about providers delayed the timely production of a correct and unduplicated list of medical providers for MPS use.

The MPS study design required complete provider names and addresses to use in mailing out the MPS advance packages and to make sure the interviewer would call at the right location for the MPS data collection. Accuracy of provider specialty information was necessary in order to exclude optometrists, chiropractors, podiatrists and similar practitioners from the list of providers eligible for MPS. These specialties, frequently confused with medical doctors by respondents, were out of scope in MPS. Finally, unduplication of provider mentions across household interviews was necessary to avoid contacting the same provider several times to collect information about different patients each time. To unduplicate, all sampled persons who received care from the same provider in NMES2 had to be identified as patients of a uniquely identified provider.

Two design changes were introduced in NMES-FS to test whether the kind of information required about providers for MPS could be elicited from family respondents earlier than in NMES2 and with greater accuracy and completeness:

1) Concurrent screening and final sample selection. Unlike NMES2, final sample selection in NMES-FS took place during the Screener round immediately upon completion of the screening interview that was administered using CAPI. Concurrent screening and final sample selection during Screener round was a prerequisite for the second important design modification.

2) Baseline Interview. The second NMES-FS design change that helped promote MPS goals was the Baseline Interview, also completed using CAPI. This interview, averaging 24 minutes in length, was administered in the sampled households immediately after the screening and sampling were completed, or as soon as possible thereafter. At the end of the Baseline Interview, the respondent was asked to enumerate the family's usual sources of medical care as well as the providers treating family members in the past 12 months. We reasoned that providers identified in this fashion had a high probability of delivering care to family members in the next two rounds of the NMES-FS.

The full names of the Baseline providers were elicited by provider type -- hospitals first, then clinics, then private doctor's offices. Then, using computer-assisted search procedures, the interviewer attempted to match each named provider to an entry in the database of medical providers for the PSU.

A definite advantage of doing the directory searches during the interview was the fact that respondents were able to assist the interviewer in making decisions whenever ambiguity was present. Respondents appear to have accepted the searching and matching activity without problems.

The Provider Database

The Provider Database was constructed from purchased files aggregated from telephone company yellow pages listings. The files were unduplicated by the vendor so that strict duplicates were excluded. Spelling and punctuation in the file entries had been standardized and cleaned prior to purchase. We purchased listings for specific medical facilities, specialties, and practice types for the counties comprising the selected Primary Sampling Units (PSUs.) The facilities, specialties, and practice types included in the Provider Database were physicians and surgeons (MDs and DOs); chiropractors; dentists; optometrists; podiatrists; clinics; health maintenance organizations (HMOs); hearing aid providers; home health care providers; hospitals; and nurses and nurses' registries.

The Provider Database on an interviewer's laptop included only the provider listings for the interviewer's assigned PSU, partly due to limitations of space on the CAPI laptop PCs. The 15 NMES-FS PSUs comprise a sample stratified to include areas of diverse population densities. Since the number of providers in an area is closely related to the population of the area, the size of the local Provider Database varied widely from PSU to PSU, with as few as 10 providers listed in the smallest PSU database and as many as 26,156 providers listed in the largest PSU. database.

A Provider Database entry included:

- The full name of the person or facility;
- The full address of the provider, including building or facility name;
- The provider's telephone number with area code;
- The provider's specialty (blank for hospitals and other facilities), and;
- A unique Provider ID;

The unique Provider ID, which was added to the provider records prior to indexing them, was used to unduplicate providers for the MPS, since mentions of the same provider in different Baseline interviews would bear the same unique provider ID when interviewers matched the provider's name to the unique database entry.

The search software allowed the interviewer to search the Provider Database by name (facility name or last name), by street name, by the full telephone number, or by a partial telephone number. After a search was specified, the software extracted the matching entries and displayed key elements of the matching entries in a scrollable "window". The interviewer could then choose to display the full entry of a likely match, to accept the entry as the correct match, to modify the entry with changes, or to reject the entry and continue searching. If none of the entries matched the provider identified by the respondent, the interviewer was prompted to add the full name and address of the provider. Interviewer-created entries for "added" and "modified" providers were inserted in the survey database rather than in the Provider Database, and were marked for unduplication coding at the home office.

Among the categories of known coverage shortfalls in the Provider Databases used in the field test are those facilities and providers who do not list themselves in telephone yellow pages. Included in this group are many state, local, and federal facilities, including health department clinics, military base hospitals, student health centers, employer clinics, and some (but not all)

VA hospitals. The Provider Database also did not include providers who were outside the geographic boundaries of the sampled PSUs.

Training Interviewers to Use Search Software

The majority of the 63 interviewers who worked on the NMES-FS Screener round of interviewing in Spring of 1992 were not familiar with CAPI or with computerized search procedures. Therefore, we needed to train most interviewers in the logic of indexed searching as well as to train them in the use of the NMES search software. The training consisted of a large-group lecture and demonstration of the search system; a look-and-do session in small groups, where interviewers followed the instructions of a demonstrator to conduct a search on their own PCs; self-paced individual exercises; and an evening tutorial session for those needing additional drill. Excluding the tutorial session, interviewers received approximately 4 hours of training on search logic and strategies, on using the search software, and on assessing the correctness of matches. In addition, many of the other CAPI exercises during the week-long training required searches of the Provider Database.

Before interviewer training, we pretested the Provider Databases as well as the search software extensively, so we were aware of some of the deficiencies and "styles" of the lists and their entries. We compiled job aid handouts containing searching "hints" and information about known list deficiencies tailored to each PSU. The hints section of the job aid included standard spellings and abbreviations used in the Provider Database. (For example, "Saint" in a facility name was always abbreviated ST without a period.) The deficiencies section included the names of any large facilities not listed in the database for a particular PSU.

System Specifications for the Search Software and CAPI Hardware

The search software for the Provider Database was written in C, using Vermont Views for the user interface. The Provider Database files were indexed using C-Trees. The search software and the CAPI software communicated by means of intermediate transaction files. A DOS batch file executed the search software when it was needed following the completion of the Baseline CAPI interview and the Closing/Tracing CAPI interview. Both the CAPI software and the search software were run on Compaq 286 LTE laptop computers with 40 megabyte hard drives.

Criteria for Assessing the Success of the Procedures

There are four dimensions that can be used to evaluate the success of the Baseline treatment involving requests for mentions of potential providers and computer-assisted search procedures. These are:

Criterion 1. The proportion of total Baseline mentions that were ultimately matched to the Provider Database (i.e., the proportion of Baseline mentions that were fully identified in the end). A low overall matching rate would decrease the utility of this exercise for future surveys.

Criterion 2. The proportion of provider mentions in Rounds 1 and 2 that involve a provider elicited and identified in the Baseline Interview. A high rate of use of Baseline providers in later rounds is required to meet the goal of minimizing delays through the prospective identification of providers for MPS.

Criterion 3. The ability of interviewers to search and match provider lists in the field, minimizing editing and review of Baseline provider mentions at the home office.

Criterion 4. The rate at which the right locations and providers are reached when Baseline-identified providers are contacted for MPS. This is partly a test of the quality of the purchased listings, but is also a test of the matches made by interviewers. A small NMES-FS MPS verification study is currently in the field to collect data for this evaluation, but results are not available for this paper.

Criterion 1. Matching the Household Providers in the Provider Database

One way to assess the usefulness of using a Provider Database to help identify providers mentioned in the Baseline Interview is to measure the completeness of the coverage of the Provider Database. The Provider Database used for the NMES-FS field test was constructed from a purchased list derived from yellow pages listings, as described above. The list was purchased from one source, and was not augmented by other listings or otherwise "improved".

We initially expected that the coverage of the Provider Database would be better for physicians practicing in private offices than for facilities such as hospitals or clinics. However, match rates for different types of providers are nearly the same. The overall match rate (the rate at which provider mentions are matched in the Provider Database) is 71.3% for all types of medical providers. The match rate for different types of providers is similar to the overall rate, with a rate of 71.3% for physicians in private offices and a rate just slightly higher for hospitals (72.3%) and slightly lower for clinics (68.9%.) The total number of provider mentions in the Baseline Interview was 2510 mentions by 975 families, or about 2.6 providers mentioned per family. More than 85% (2140 of 2510) of the provider mentions in the Baseline Interview were physicians practicing in private offices or group practices.

We also expected to find differences in the match rates for different geographical areas. We expected that areas with lower population density (i.e., the more rural areas) would have lower match rates than more densely populated areas, because people in rural areas would travel outside the PSU boundaries more often for health care. The match rate for the non-metropolitan PSUs is indeed lower than the overall match rate, with 55% of the non-metropolitan providers matched in the directory compared to 71% overall. Table 1 shows the match rates for Metropolitan and Non-metropolitan PSUs, based on Census classifications. For Metropolitan PSUs taken together, 76% of the provider mentions were matched in the Provider Database. However, when the Largest Metropolitan PSUs are examined separately from Other Metropolitan areas, the Largest Metropolitan areas have a lower match rate, 66%, and the Other Metropolitan areas have a higher match rate, more than 80%.

To learn more about the reasons providers were not matched, we examined the names and addresses of the 721 Baseline provider mentions not matched in the Provider Database, and assigned codes indicating why a match could not be found. The known deficiencies of the original source lists, which included shortfalls in types of facilities that often do not list themselves in yellow pages directories and the lack of coverage for providers not within the boundaries of the sampled PSUs, accounted for 42% of the unmatched provider mentions. In Table 2, the 721 provider mentions not matched in the Provider Database are classified according to the primary reason they were not matched, compared across PSUs with different population densities. Non-Metropolitan PSUs have about three times the proportion of unmatched provider mentions where the provider is outside the PSU, (61%) compared with the Other Metropolitan PSUs (22%) and the Largest Metropolitan PSUs (18%). In contrast, completeness of the database seems to be a greater problem in the Metropolitan PSUs; in the Largest Metropolitan PSUs, 74% of the unmatched providers were inside the PSU boundary, and in the Other Metropolitan PSUs, 67% were inside the PSU boundaries.

Criterion 2. Baseline Providers Who Were Referenced as Sources of Care in Later Rounds of the NMES-FS Household Survey

Collecting provider information in the Baseline Interview also should reduce the amount of unduplication coding necessary between rounds of the household survey. Eligibility for the MPS was conditioned on the provider being mentioned as a source of medical care during the household survey observation period, which includes a six month period covered by Rounds 1 and 2 of the NMES-FS survey, and a full twelve months in Rounds 1 through 4 of the NMES2 survey. To achieve a benefit from the early identification of providers in the Baseline Interview, it is necessary not only to collect and code the Baseline provider mentions. The identified Baseline providers would also need to be mentioned as providers of care in later rounds of the household survey.

About half (49%) of the providers referenced in Round 1 and 45% of the providers referenced in Round 2 were identified in the Baseline Interview. Of the 2437 providers referenced in either Round 1 or Round 2, 42% were first identified in the Baseline Interview, while 33% were identified in the Round 1 interview, and 24% were identified in the Round 2 interview. Note that the rate of referencing of Baseline provider mentions in the combined Rounds 1 and 2 is lower than the individual rounds, because 404 of the Baseline providers who were referenced in Round 1 were referenced again in Round 2.

Criterion 3. Interviewers' Success at the Searching and Matching Task

If the interviewer could not match a provider mention to a listing in the database, home office clerical staff used more intensive search methods to look for a match. The home office staff found matches for 32% (346) of the 1067 provider mentions for which interviewers could not find matches. The following analysis examines the interviewers' success at finding matches, restricting the analysis to those provider mentions that are "matchable" or "present in the database". This group includes all provider mentions the interviewers were able to match plus all provider mentions the home office staff matched later. Of the total number of matchable mentions (1789) the interviewers were able match 80.7% (1443). Interviewers had greater success at matching physicians in private practice in the database than they did at matching facilities such as clinics and hospitals. One reason for this may be that there is a greater discrepancy between the formal names of facilities and what the respondents call them. For example, "St. Joseph's Mercy Hospital" may be commonly known as "Mercy Hospital" among respondents. On the other hand, private-practice physicians, and even those in group practices, are referenced by their names both by respondents and in the yellow pages.

Across areas differing in population density, interviewers were most successful at matching the providers from non-metropolitan PSUs, where they matched 92% of the matchable providers. They were least successful in the largest metropolitan PSUs, where they matched only 70% of the matchable providers.

The search software allowed interviewers to search on provider name, name of street, and telephone number. Interviewers were trained to search first on provider name and then to search on the other fields if the name did not yield a match in the database. What was the match rate achieved by name searches? Was the match rate improved by searching on more than one modality? How frequently did interviewers use search modalities other than provider name? The discussion that follows is restricted to the 2,246 Baseline mentions that reference providers located inside the PSU boundaries. There is evidence that interviewers invested less effort in searching for providers that were clearly outside the PSU boundaries, including out of state.

Provider name searching was most prevalent, with 91% of the Baseline mentions searched by interviewers only on the provider's name. Searching on name only, interviewers were able to achieve a match rate of 61% for providers inside the PSU boundaries (1,357 Baseline mentions matched out of 2,246). The remaining 889 mentions were eligible to be searched on either street name or telephone number, but only 23% (198) of these provider mentions were searched further by interviewers.

Of the provider mentions searched only by name in the field and not matched by interviewers, 42% were matched by the home office, as shown in Table 3.

Conclusions

Our first conclusion relates to the potential for having interviewers use database search and query software to field code complicated information (such as, but not limited to, medical providers). Our conclusion, based on the results of the field test, is that interviewers can do this task. Interviewers were able to match a provider mention to a database entry about 80% of the time the provider was present in the database.

Our second conclusion pertains to the selection and construction of lists, and particularly provider lists, for a criterion database. The list we used for constructing the Provider Database was from a single source, and was unimproved by amendments or additions. It was also restricted to the geographical boundaries of the PSUs in the NMES-FS sample. Both of these constraints were the result of decisions to conserve the resources of the field study. With this unimproved list, we matched 71% of the providers mentions in the database. The match rate would be higher if efforts were made to improve the quality of the lists and if the lists were to include providers from areas surrounding the PSUs.

Our third conclusion relates to the early identification of providers for use in later rounds of NMES. The early identification of about 50% of the referenced providers helps to reduce the length of the Round 1 interview, which is the longest interview in NMES. However, for a variety of reasons respondents identified many more medical providers in the Baseline Interview than were referenced in later rounds as sources of care. To improve the efficiency of the early identification procedure, we need to find ways to avoid identifying these unused providers.

Early identification of providers in a Baseline Interview, particularly with the use of an on-line searchable Provider Database, has been beneficial to the NMES household survey. This effort has improved the timeliness, quality, and completeness of the household survey family provider rosters, and has contributed to the feasibility of starting the Medical Provider Survey of the NMES-FS shortly after the end of the Round 2 interviewing, rather than several months after the conclusion of the household survey interviewing, as occurred in NMES2. The detailed information collected about providers in the Baseline Interview may also impress upon survey respondents the importance of the detailed record-keeping the survey requires.

Table 1: Baseline Provider Mentions Matched and Not Matched in the Provider Database for PSUs with Different Population Densities

PSU population densities	Matched in the Provider Database	Not matched in the Provider Database	Total provider mentions
Metropolitan PSUs	1476 76.0%	465 24.0%	1941 100%
Largest metro PSUs	428 66.0%	211 33.0%	639 100%
Other metro PSUs	1048 80.5%	254 19.5%	1302 100%
Non-metropolitan PSUs	313 55.0%	256 45.0%	569 100%
All PSUs	1789 71.3%	721 28.7%	2510 100%

Table 2: Reasons Provider Mentions Were Not Matched in the Provider Database Across Different Population Density Areas

	Public facilities not in the Yellow Pages	Providers outside the PSU boundaries	Providers inside the PSU boundaries, not in the database	Total provider mentions
Largest metro PSUs	17 8.1%	38 18.0%	156 73.9%	211 100%
Other metropolitan PSUs	28 11%	56 22.1%	170 66.9%	254 100%
Non-metropolitan PSUs	8 3.1%	157 61.3%	91 35.6%	256 100%
Total mentions never matched	53 7.4%	251 34.8%	417 57.8%	721 100%

Table 3: Comparison of Outcomes by Search Modality Used by Interviewer for Baseline Mentions that Failed to Match Initially on Provider Name

	Action Taken By Interviewer After Failing To Match Baseline Mention On Provider Name		
Outcome of Match Attempt	Stopped Searching	Searched on Street or Telephone Number After Name	Never searched
Matched by Interviewer	N/A	86 43.4%	N/A
Matched by Home Office	285 41.9%	46 23.2%	2 20.0%
Never Matched	396 58.2%	66 33.3%	8 80.0%
Total Mentions Not Matched on Name	681 100.0%	198 100.0%	10 100.0%