TECHNICAL PAPERS ON
HEALTH AND BEHAVIOR MEASUREMENT

TECHNICAL PAPER 23

The Quality of Health Data Obtained in a New Survey of Elderly Americans

Charles F. Turner, Timothy K. Smith, Leslye K. Fitterman, Thomas Reilly, Kirk Pate, Michael B. Witt, A. Marshall McBean, Judith T. Lessler, and Barbara H. Forsyth

Reference Citation

Turner, C.F., T.K. Smith, L.K. Fitterman, T. Reilly, K. Pate, M.B. Witt, A.M. McBean, J.T. Lessler, and B.H. Forsyth. (1997) The quality of health data obtained in a new survey of elderly Americans: A validation study of the Medicare Beneficiary Health Status Registry (MBHSR). *Journal of Gerontology: Social Sciences* 52B(1):S49-S58.

Journal of Gerontology: SOCIAL SCIENCES
1997, Vol. 52B, No. 1, S49–S58

The Quality of Health Data Obtained in a New Survey of Elderly Americans: A Validation Study of the Proposed Medicare Beneficiary Health Status Registry (MBHSR)

Charles F. Turner,[1] Timothy K. Smith,[2] Leslye K. Fitterman,[3]
Thomas Reilly,[3] Kirk Pate,[2] Michael B. Witt,[2]
A. Marshall McBean,[4] Judith T. Lessler,[2] and Barbara H. Forsyth[1]

[1]Research Triangle Institute, Rockville, Maryland.
[2]Research Triangle Institute, Research Triangle Park, North Carolina.
[3]Health Care Financing Administration, Baltimore.
[4]Department of Health Management and Health Policy, School of Public Health, University of Minnesota.

The Medicare Beneficiary Health Status Registry (MBHSR) is a proposed new survey program that would collect health status indicators annually from large probability samples of Medicare beneficiaries. For reasons of economy, the MBHSR would use mail survey procedures with telephone follow-up of nonrespondents. Because of concerns about response rates and the validity and reliability of the data obtained by such methods, a large-scale (N = 1,922) field test was conducted. The field test assessed the validity of MBHSR survey reports of past medical treatment and conditions by comparing those reports with Medicare claims data. It assessed the (internal) reliability of MBHSR survey responses by comparing responses with logically related survey questions from the MBHSR. Analyses indicate that the MBHSR survey procedures using a combination of mail data collection with telephone follow-up of nonrespondents produced relatively high levels of sensitivity and specificity in identifying medical treatments and procedures previously recorded in Medicare claims data. In addition, the MBHSR Field Test obtained, in general, relatively high levels of internal consistency in survey reports.

THIS study assessed the quality of survey data collected during a field test of the Medicare Beneficiary Health Status Registry (MBHSR) — a proposed new federal survey program that would collect data annually on the health status of large probability samples of elderly Americans. We considered two aspects of data quality: *external validity*, which was assessed by comparing respondents' reports of medical conditions and procedures in the survey with Medicare records of their prior in-patient hospital stays for those same conditions and procedures; and *internal consistency*, which was assessed by calculating the proportion of cases in which logical (internal) inconsistencies occurred across two questions in the reporting of the same characteristic. The MBHSR questions covered a wide range of health-related characteristics, including medical conditions and procedures, use of alcohol and tobacco, disabilities, need for assistance in various daily activities, and use of assistive devices (grab bars, walkers, etc.).

The Survey Program

The proposed Medicare Beneficiary Health Status Registry was conceived as a system to enhance the utility of existing Health Care Financing Administration (HCFA) administrative data by adding survey data on health status collected from Medicare beneficiaries. Such a system could contribute to the formulation of national health policy by helping government agencies and health researchers understand the patterns of use and need for health services among elderly Americans. It could also be used to assess the effectiveness and quality of care provided under Medicare and the impact of changes in the provision of that care, as well as provide a basis for developing improved risk-adjustment procedures in setting reimbursement schedules for health care providers. As currently envisioned, the MBHSR would survey large samples of new Medicare enrollees (target sample size is 40,000 per year) with continuous recruitment of new cohorts over time. The registry would also include a database management system that could link survey information from each subject with Medicare administrative and claims data files for analytical purposes.

Over time, the annual accretion of the MBHSR database would offer a variety of unique analytical possibilities. For example, the data would allow researchers to continuously monitor the health status of the elderly population because the size of the sample would be large enough to detect changes specific to particular socioeconomic, racial, geographic, and other groupings. In addition, researchers could evaluate regional and other variations in the types of medical treatment obtained by Medicare enrollees with equivalent symptomatology, and assess the relative costs of medical care required over time for people who reported different classes of health conditions when they enrolled in Medicare.

Most of those analyses depend on the assumption that large, representative samples could be surveyed each year. Yet current budgetary constraints preclude HCFA from pursuing large-scale in-person or telephone data collection. For that reason, the MBHSR Field Test was developed to test a mailed self-administered questionnaire as the primary data collection mode. The field test also included telephone follow-up for people who did not respond to three waves of mail survey efforts. (For methodological purposes, personal interviews with a small, targeted sample of nonresponders to the telephone survey effort were part of the test as well.)

The MBHSR Field Test was designed to answer two major questions. First, can an adequate rate of response be obtained in a mail survey that makes sparing use of telephone follow-up? And second, will the self-reports of health conditions obtained from that kind of data collection be of adequate quality to permit confident use of the MBHSR data for research and policymaking?

Elsewhere we report in detail what were generally encouraging results in obtaining adequate levels of response for the elderly population at large and for key subpopulations, such as racial minorities and people with preexisting major medical conditions. Those results include the following findings:

- Response rates of 69 percent to 79 percent could be achieved with a strategy that combined multiple waves of mail contacts with telephone follow-up of nonresponders to the mail contacts.
- Survey questionnaires that were estimated to take as long as 30 minutes to complete achieved response rates of 77 percent to 79 percent; a questionnaire estimated to take 40 minutes to complete obtained a significantly lower rate of response (69 percent).
- A prenotification mailing (used in a random subset of the MBHSR Field Test sample) had only a minimal effect on response rates.
- Telephone follow-up reduced the potential biases that would have been introduced by the higher nonresponse rates of non-Whites to the mail data collection effort.

Details of these and other results from the MBHSR Field Test are presented elsewhere (Forsyth et al., 1996; Turner et al., 1994). This article reports on the quality of the MBHSR data. Specifically, we present evidence on the validity of the MBHSR measurements of self-reported medical procedures and chronic conditions as well as evidence on the internal consistency of responses given to logically related questions in the MBHSR.

Background

Several studies have shown that the validity of health survey measurements obtained from elderly respondents varies by the question topic, the characteristics of the respondent, and the mode of survey administration (see Aday, 1989; Herzog and Kulka, 1989). For example, among elderly males undergoing transurethral resections of the prostate, Doll and colleagues (1991) found varying levels of validity for responses about intimate, subjective, and factual topics. Both underreporting and overreporting of the utilization of physician services have been reported by Glandon, Counte, and Tancredi (1992) to be significantly related to the respondents' health status and their utilization of other health care services.

Although there have been some demonstrations of the effects of mode of survey administration on the quality of response among elderly adults (e.g., Landry et al., 1988), there is a dearth of such studies. A recent review of the literature (Smith and Biemer, 1991) found only 10 studies reporting on mail surveys of elderly populations. This lack of past research made it difficult for us to predict the levels of validity and reliability that should be expected if the proposed Medicare Beneficiary Health Status Registry were to rely heavily on mailed self-administered questionnaires to collect health data from elderly persons. For that and other reasons, we undertook a large-scale field test of the proposed design of the MBHSR.

METHODS

Sample design. — The MBHSR Field Test used a two-stage stratified sample design. The first stage was selected from national primary sampling units (PSUs) stratified by metropolitan statistical area (MSA) versus non-MSA. At the second stage, enrollees were sampled from the PSUs stratified by age group, race, gender, and history of surgical procedures and health conditions reported in the Medicare file of in-patient hospital stay records (see below). The intent of the sample design was to select a national sample that was representative of (a) persons who were new enrollees into Medicare and were between the ages of 65 and 66, (b) persons ages 76–80 who were enrolled in Medicare, (c) a subpopulation of persons within three primary sampling units purposively chosen for their proximity to the survey organization (Research Triangle Institute) in order to conduct a cost-efficient in-person reinterview component of the field test, and (d) an oversample of persons ages 76–80 who had at least one of 14 specified ICD-9-CM diagnoses or medical procedures recorded between 1984 and 1992 in HCFA's Medicare Provider Analysis and Review (MEDPAR) file of records on in-patient hospital stays. The application of this two-stage sample design resulted in the selection of 2,410 Medicare beneficiaries for inclusion in the MBHSR Field Test. By design, these beneficiaries overrepresented non-Whites in both age cohorts and overrepresented persons with prior health conditions in the older cohort. This oversampling was designed to permit an assessment of the proposed MBHSR procedures and data quality within populations of particular interest. (Note that persons enrolled in health maintenance organizations [HMOs] are not included in the MBHSR sample universe.)

Data collection procedures. — The MBHSR Field Test was conducted by Research Triangle Institute during the summer and early fall of 1993. The field test was designed as a randomized experiment to permit identification of an optimal design for the proposed MBHSR. Medicare beneficiaries (within each age cohort) were randomly assigned to groups to test two aspects of the data collection: questionnaire length and use of a prenotification mailing to alert sample members to the forthcoming survey. Three questionnaire lengths (short, medium, long) were used, and bene-

ficiaries were assigned to either receive or not receive a prenotification letter. (The short, medium, and long questionnaires were estimated to take 20, 30, and 40 minutes, respectively, to complete.)

MBHSR questionnaire. — The design of the questionnaire to be used in the proposed MBHSR survey began with the adaptation of relevant questions and scales from other instruments that have been successfully used in previous research on elderly populations. These included items from the Longitudinal Study of Aging questionnaire (Bureau of the Census, 1991b), the Long-Term Care Survey (Bureau of the Census, 1991a), the Medical Outcomes Study Health Status Questionnaire, and the Behavioral Risk Factor Survey (CDC, 1992). The final proposed design of the MBHSR questionnaire was informed by the results of an intensive program of cognitive testing of the items we adapted from these surveys. That testing resulted in numerous modifications to the preliminary MBHSR questionnaire. Among the many modifications were changes to alleviate respondents' problems with the survey, including (a) difficulties in following "skip" instructions, (b) failure to notice introductory statements preceding questions, (c) poor comprehension of negative modifiers, and (d) difficulties with unfamiliar technical terms.

Field procedures. — An experiment on the impact of survey prenotification was embedded in the MBHSR Field Test. At the start of the data collection period, a prenotification packet containing an introductory letter from the HCFA administrator, a letter from the RTI project director, a study brochure, and a return postcard to provide updated information on their mailing addresses was sent to one-half of the MBHSR Field Test sample. (The other half of the sample did not receive such prenotification. Analyses [Turner et al., 1994] indicate that prenotification had no significant impact on response to the MBHSR.) Three weeks later (May 20, 1993), the first of three mailings of questionnaire packets occurred. That packet included a letter from the RTI project director and a copy of the questionnaire.

Attempts to follow up nonrespondents were conducted by mail and telephone. Six weeks after the initial questionnaire mailing (July 6, 1993), a second questionnaire packet was mailed to the nonrespondents. A third packet was sent after another four weeks had passed (August 9, 1993). Both of the follow-up mailings included a revised letter from the RTI project director urging participation in the study. All nonresponse cases were assigned to the Telephone Survey Unit at RTI three weeks after the third questionnaire mailing (September 4, 1993). Since telephone numbers for the sample members were not available on the sample frame, these cases required tracing prior to the attempt to complete the telephone interview. This data collection strategy yielded an overall response rate of 79 percent when a medium-length questionnaire was used. (See Turner et al. [1994] for a summary of the response rates.) In the MSAs of Richmond, VA, Raleigh, NC, and Atlanta, GA, face-to-face interviews were attempted to elicit participation from nonrespondents to the mail and telephone data collection efforts. Because face-to-face follow-up is not expected to be included in the proposed MBHSR, the results of these interviews are not included in our analysis.

RESULTS

External Validation

The validation of self-reported medical conditions could be carried out only for the older cohort; new enrollees had no prior history of Medicare-covered hospital stays. For the older cohort, Medicare hospital stay records and, in particular, HCFA's (1993) Medicare Provider Analysis and Review (MEDPAR) files could be used as a validity criterion. The MEDPAR file contains records of Medicare-covered in-patient hospital stays. Each record represents an in-patient stay. The record may represent one claim or multiple claims depending on the length of a beneficiary's stay and the amount of in-patient services used during the stay. To the extent that respondent reports of procedures and conditions can be verified with information from the MEDPAR files, we have direct evidence of their measurement validity.

HCFA's MEDPAR records permitted us to identify patients who received the following procedures during a Medicare-covered hospital stay: cataract removal, gall bladder removal, coronary artery bypass surgery, pacemaker insertion, coronary angioplasty, breast surgery or breast removal, hysterectomy, hip replacement, and prostatectomy. Chronic conditions in the elderly population that are known risk factors for morbidity and that affect the quality of life and cost of health care were also selected for this validation study. The conditions included asthma, cancer, chronic obstructive pulmonary disease (COPD), diabetes, stroke, and heart attack. Indicators of these procedures and conditions were utilized in the sample design (e.g., to oversample persons with the selected conditions or procedures for the older cohort). Table 1 presents the text of the MBHSR survey questions that will be used in our analyses together with the ICD-9-CM codes against which they will be validated in HCFA's MEDPAR files.

Caveats. — Our external validity criterion provides something less than a true "gold standard." In particular, readers must bear in mind three major limitations of this validity criterion:

- The MEDPAR data available to us covered medical procedures performed and medical conditions treated during the time period 1984 through 1992 (based on date of discharge). It is thus possible for respondents to be accurately reporting medical events that occurred outside of this time period.
- The MEDPAR data, as with all datasets, are subject to a variety of errors ranging from clerical inaccuracies in transcription to inaccurate reporting by health care providers.
- The MEDPAR data reflect medical procedures performed and conditions treated in an in-patient hospital setting. It is thus possible for respondents to be accurately reporting events that occurred in an out-patient setting.

While the MEDPAR dataset, like all datasets, admits to measurement error, past assessments of this dataset have

Table 1. ICD-9-CM Codes and Conditions Used in Scanning the HCFA MEDPAR Database and Corresponding Survey Questions Administered in the Medicare Beneficiary Health Status Registry (MBHSR) Field Test, 1993

Conditions and Procedures	ICD Codes[a]	Wording of Survey Question
Coronary angioplasty	36.01, 36.02, 36.05	Have you ever had coronary angioplasty? That is where a small balloon is inflated in an artery leading to your heart in order to open or unplug the artery.
Pacemaker implant	37.80 to 37.83, 37.85 to 37.87[b]	Have you ever had a pacemaker? This is a small device placed under the skin near the shoulder or chest to help the heart beat normally.
Coronary bypass	36.1 to 36.3	Have you ever had a heart (coronary artery) bypass operation? This is where a blocked or plugged artery in your heart is replaced, usually with a section of vein taken from your leg or arm.
Hip replacement	81.51[c]	Have you ever had a total or partial hip replacement? Hip replacement is done for severe arthritis of the hip joint.
Cataract removal	13.1 to 13.8	Have you ever had a cataract removed from either of your eyes to improve your vision?
Mastectomy	85.4 to 85.48	How many times have you had either all or part of a breast removed because of breast cancer?
Gall bladder removal	51.22[d]	Have you ever had your gall bladder removed?
Diabetes	250 to 250.91	Has a health professional said that you should now be following a special diet to control diabetes, sometimes called sugar sickness?
Cancer	140 to 199.1 and 200 to 208.9	Have you ever been told by a doctor or nurse that you had cancer of any kind?
Asthma	493 to 493.91	Has a doctor or nurse told you that you currently have asthma?
Hysterectomy	68.3 to 68.8[e]	Have you ever had all or part of your womb removed? This is called a hysterectomy.
Heart attack	410 to 412[f]	Have you ever been told by a doctor or nurse that you had a heart attack?
Chronic obstructive pulmonary disease	490 to 492.8 and 496[g]	Have you ever been told by a doctor or nurse that you had emphysema, chronic bronchitis, or chronic obstructive lung disease?
Stroke	430 to 434.9[h]	Have you ever been told by a doctor or nurse that you had a stroke or brain hemorrhage?

Notes. The MBHSR Field Test embedded an experimental design within a probability sample of Medicare beneficiaries. The length of the questionnaire was varied across three treatment conditions that used short, medium, and long questionnaires. Thus, not all of the above survey questions were included in each survey. ICD-9-CM codes listed are those for 1992; ICD coding changes occur periodically, and thus there was some variation in the codes used in earlier years. We have not included reporting of hypertension (1992 ICD-9-CM codes 401 to 405.99) in our analyses because the survey questions did not ask about *past* conditions but rather: "Has a doctor or nurse told you that you should *now* be taking medicine for high blood pressure? [*emphasis in original*]."

[a]MEDPAR records carry five diagnosis fields and three procedure fields. The appearance of specified ICD-9-CM codes in *any* of the diagnosis fields (for medical conditions) or procedure fields (for procedures) was taken to indicate the presence of the condition or procedure.

[b]Codes for pacemaker do not include repair of pacemaker (ICD 37.89).

[c]Codes for hip replacement refer only to total hip replacement; partial hip replacement (ICD 81.52) and revision of hip replacement (ICD 81.53) are not included.

[d]Codes for gall bladder removal (cholecystectomy) refer to complete removal of gall bladder; partial excision (ICD 51.2) or laparoscopic cholecystectomy (ICD 51.23) are not included.

[e]Codes for hysterectomy do not include "other and unspecified hysterectomy" (ICD 68.9).

[f]Codes for heart attack refer to acute myocardial infarction only; congestive heart failure (ICD 428–428.9) and heart failure following cardiac surgery (ICD 429.4) are not included.

[g]Codes for COPD do not include associated lung diseases from exposure to environmental dusts (ICD 495).

[h]Codes for stroke do not include temporary conditions without residual effects (ICD 435) and acute, but ill-defined cerebrovascular disease (ICD 436).

produced some broadly encouraging results. Many researchers have found that dates of admission, major surgical procedures, and broad diagnosis categories on the claims are accurate over 80 percent of the time when validated against data from other sources. A study comparing cataract procedures reported to Medicare for reimbursement with medical records found that 99.7 percent of the cases reported to Medicare had, in fact, undergone cataract surgery (Javitt et al., 1993). Furthermore, Fisher et al. (1991, 1992) have found in several studies that 91 percent to 94 percent of the hip fracture cases could be identified from Part A Medicare claims. Fisher et al. also found, however, that medical procedures were more accurately reported than medical conditions in MEDPAR records — although the reporting of cancer and diabetes was an exception to this rule.

Strategy. — Our analyses use the Medicare administrative data as a "gold standard," recognizing, of course, that these data are not entirely error free. We began by constructing a two-by-two table for each condition and procedure. (Both the survey report and the indicator derived from Medicare claims were collapsed to binary variables indicating whether the condition or procedure had been reported. "Unsure" was offered as a response category for questions on coronary

bypass, coronary angioplasty, and implantation of a pacemaker. For this analysis, these answers were treated as negative responses.)

The resultant table cross-tabulated whether a particular medical procedure or treatment was reported by the respondent and whether or not the same treatment or procedure was coded in HCFA's (1993) MEDPAR files. We then calculated three statistics to characterize the results of this cross-tabulation for each medical condition or procedure. They are:

- Sensitivity: the proportion of survey respondents reporting the condition or procedure among all respondents who were coded as having had a previous Medicare claim for the condition or procedure.
- Specificity: the proportion of survey respondents who report not having had the condition or procedure among respondents who lack Medicare claims for the condition or procedure.
- Association: chi-square test (calculated across all respondents who answered the survey question) for independence between survey reports and Medicare claims for the condition or procedure.

In considering these results, "sensitivity" coefficients in our analyses have approximately the same meaning as in traditional analyses. They provide a quantitative answer to the question: Among those persons whom we know have had claims for a medical condition or procedure during 1984–1992, what proportion reported this procedure or condition in the survey? In contrast, the interpretation of our specificity coefficients differs due to the limitations of the data available to us. In particular, it should be noted that many survey respondents could be expected to report medical conditions or procedures they had prior to 1982 (or even prior to their enrollment in Medicare). Because of the frailty of this external criterion, one might, indeed, choose to refer to these results as "pseudo-specificity" coefficients.

Table 2. Sensitivity, Pseudo-specificity, and Test of Association Between Reporting of Selected Medical Conditions and Procedures in the Medicare Beneficiary Health Status Registry (MBHSR) Field Test and the Reporting of Claims for These Medical Conditions and Procedures in HCFA's MEDPAR Database of Medicare Claims[a]

	Sensitivity		Specificity[b]	
Measurement[b]	Coefficient	Base *N*	Coefficient	Base *N*
Procedures				
Coronary angioplasty	1.00	19	0.97	881
Pacemaker implant	1.00	17	0.98	888
Coronary bypass	1.00	31	0.96	873
Hip replacement	1.00	6	0.97	272
Cataract removal	1.00	17	0.72	871
Mastectomy	1.00	7	0.96	359
Gall bladder removal	0.89	44	0.87	855
Hysterectomy	0.75	16	0.63	550
Conditions				
Diabetes	0.86	96	0.92	804
Cancer	0.79	73	0.85	810
Asthma	0.79	19	0.96	564
Heart attack	0.72	117	0.90	777
Stroke	0.59	41	0.90	551
Chronic obstructive pulmonary disease	0.53	72	0.92	514

[a]The analysis is restricted to the cohort of Medicare enrollees aged 76 to 80 and uses unweighted frequency counts.

[b]Note that since the MEDPAR database included only conditions and procedures reported in claims for Medicare reimbursement, it is highly likely that many instances in which the respondent reported having had the condition or procedure in the survey but it did not appear in the MEDPAR database reflect conditions and procedures that predate Medicare eligibility. For this and other reasons discussed in the text, it might be more appropriate to view the specificity coefficients as "pseudo-specificity" coefficients. Note also that survey reports for all procedures and conditions have a significant association ($p < .01$) with MEDPAR indicators of the procedure or condition, based on $\chi^2_{(df=1)}$ test with correction for continuity.

Findings. — Table 2 presents the results of our external validation of the self-reports of medical conditions and procedures obtained in the MBHSR Field Test. It will be seen from Table 2 that there is a remarkably complete reporting for most of those medical procedures that we know to have been performed from Medicare records. The sensitivities are 100 percent for six of the eight medical procedures and 89 percent for one of the two remaining procedures (gall bladder removal). These results indicate that for instances in which Medicare indicated that a surgical procedure had been performed and paid for by the program, we found high rates of consistent reporting by respondents in the field test. Although the numbers of respondents having these procedures were small, it is reassuring to find that the survey instrument elicited complete reporting of coronary angioplasty, coronary bypass surgery, cataract surgery, hip replacement, mastectomy, and implantation of a pacemaker.

There is one exception to our findings of high levels of sensitivity for the reporting of medical procedures. This occurs for the reporting of hysterectomies. Four of 16 women who were coded as having had a hysterectomy in Medicare records did not report this in the survey. Although based on a very small sample size, the resultant sensitivity estimate of 75 percent suggests that these survey measurements may be of lesser quality than the measurements of other medical procedures.

This finding for the reporting of hysterectomies was a cause for concern. Since the procedure would have been performed within the prior 10 to 14 years, it seemed unlikely that normal memory lapses would be a likely explanation. To further explore this result, we first attempted to determine whether proxy reports might account for the diminished sensitivity of MBHSR reporting of hysterectomies. When proxy reports were excluded from the analysis, sensitivity rose from .75 (12 of 16) to .85 (11 of 13). We should note, however, that this result does not appear to be due to fallibility on the part of the proxy reporter. Rather, the two false-negative reports by proxies occurred in instances where the proxy reported that he or she was merely asking the questions and recording the respondent's verbal answers.

When cases with proxy reporting were excluded from the analysis, the sensitivity of MBHSR measurements of gall bladder removal also increased from .89 (39 of 44) to .97 (30 of 31). Since the other procedures all had sensitivities of 1.0,

exclusion of proxy reports did not affect our estimates of the sensitivity of these MBHSR measurements.

Reporting of most medical conditions yielded lower and more variable sensitivity estimates. These ranged between 72 percent and 86 percent for reporting of the following medical conditions: diabetes, asthma, cancer, and heart attack. For two medical conditions, stroke and chronic obstructive pulmonary disease (COPD), we found substantially lower sensitivity estimates. In particular,

- only 24 of 41 respondents treated for stroke (according to Medicare records) reported having had a stroke in response to the survey, for a sensitivity of 59 percent; and
- only 38 of 72 respondents treated for COPD (according to Medicare records) reported having COPD in response to the survey, for a sensitivity of 53 percent.

Elimination of proxy reports from our analysis of the reporting of conditions did not markedly affect estimates for any condition except stroke. For stroke, the estimated sensitivity declined from .59 (24 of 41) to .50 (14 of 28). This is an interesting result that suggests that the reporting of strokes when proxies were involved provided a more sensitive indicator (10 of 13 = .77) of this condition than when the respondents answered questions without the assistance of a proxy (14 of 28 = .50).

As Table 2 shows, our specificity estimates indicate that the survey reports obtained from the MBHSR instrument provide rather specific indicators of medical conditions and procedures. Specificity estimates averaged 88 percent for reporting of the eight medical procedures and 91 percent for reporting of the six medical conditions shown in Table 2. Indeed, there were only two measurements (cataract removal and hysterectomy) for which the estimates of specificity were less than 85 percent. The relatively high specificity estimates obtained in this analysis are particularly surprising since they were obtained from calculations that treat *accurate* reports of medical events that occurred prior to enrollment in Medicare as *errors*.

Finally, it should be noted from Table 2 that despite the relatively small number of respondents who reported having many of the conditions and procedures, there were statistically significant associations in every instance between the survey reports obtained in the MBHSR Field Test and the claims recorded in HCFA's (1993) MEDPAR files.

Internal Consistency Analysis

For our internal consistency analyses we calculated the percentage of disagreement for those items that asked for the same or similar information within the MBHSR questionnaire. Our internal consistency analysis began with the identification of 41 pairs of logically related health questions; these questions asked about the same or related events, conditions, or procedures for which certain patterns of response would be logically impossible. For example, a respondent indicating that he or she had never been diagnosed with diabetes should not record a response indicating that he or she was following a special diet recommended by a health professional in order to control diabetes. We first calculated the proportion of inconsistent responses to each pair of items using data from those respondents who had provided some response to both items in the pair. Respondents for whom responses to one item in the pair were not available, either due to inappropriately skipping an item or the respondent's inability or unwillingness to provide an answer, were excluded from the calculation. (Note that items selected for this analysis were not present in all questionnaire lengths.) Inconsistency rates for each age cohort were analyzed separately.

Findings. — Table 3 presents the results of our internal consistency analysis. This tabulation indicates that relatively high levels of consistent reporting occurred in instances where explicit checks were possible. Over all comparisons in both age cohorts, an average of only 2.3 percent of measurements was found to be inconsistent. MBHSR measurements were somewhat more inconsistent among the 76- to 80-year-old cohort than among the 65-year-olds; the mean (unweighted) inconsistency rates were 2.9 and 1.7 percent, respectively. While most differences in inconsistency rates between age cohorts were not statistically significant, in four of the five cases in which the differences exceeded chance expectations (at $p < .05$), the inconsistency rates were higher among the older cohort.

Inconsistency rates showed only modest variation across the MBHSR topic areas. Rates fluctuated in the two ranges: 0.6 to 2.6 percent for the cohort of 65-year-olds, and 2.2 to 3.6 percent for 76- to 80-year-olds. More instructive than fluctuations across topics are the fluctuations across individual measurements. Table 3 indicates that of the 41 MBHSR measurements, five yielded levels of inconsistent reporting greater than 10 percent in one or both age cohorts. These measurements were:

1. Physical health problems interfere with daily life (inconsistency rates of 9.5 [age 65] and 15.7 [ages 76–80]). (Questions asked about the "extent" that physical health interfered with social activities in the past 30 days (quite a bit, moderately, slightly, and not at all) [Q.D-1] versus "how often" physical health interfered with social activities in the past 30 days (all or most of time, some of time, a little of time, none of time) [Q.D-41].)
2. Emotional health problems interfere with daily life (inconsistency rates of 4.9 and 10.3). (Questions [D-2 and D-42] and response categories parallel those for physical health.)
3. Need for help with money management (inconsistency rates of 0.0 and 10.6). (Questions asked: [D-17] "Do you need help from another person to handle your money?" [response categories: 1: Yes, and 2: No], and [D-18] "How long have you needed help from another person to handle your money?" Response categories were: 1: I don't need any help to handle my money; 2: Less than 3 months; and 3: 3 months or more.)
4. Use of eyeglasses or contact lenses (inconsistency rates of 26.5 and 25.9). (Questions asked respondents: [D-38] "Which of the following do you currently use?" and [D-39] "Which other devices do you think would help you?" Each question contained a response category "eyeglasses or contact lenses.")
5. Use of dentures (inconsistency rates of 13.4 and 15.0). (Comparison uses same questions as for "eyeglasses"

Table 3. Rates of Internal Inconsistency for Measurements in the MBHSR Field Test, by Age Cohort and Question Topic

Measurements	Age 65			Ages 76–80			
	Percentage Inconsistent	*SE*	Base *N*	Percentage Inconsistent	*SE*	Base *N*	*p*-value
All Items							
Mean Inconsistency Rate (Unweighted)	1.7			2.9			
Health Conditions and Medical Procedures							
Adequacy of vision	2.5	0.59	460	0.7	0.29	414	<.01
Experience of pain in past 30 days	0.7	0.68	224	2.3	1.54	176	n.s.
Pain interferes with daily life	0.7	0.56	296	2.1	1.23	270	n.s.
Cataract surgery	0.2	0.21	745	0.6	0.37	634	n.s.
Diabetes	1.6	0.51	616	2.8	0.98	576	n.s.
Lung problems	0.0	0.00	294	0.5	0.40	272	n.s.
Physical health interferes with daily life	9.5	1.59	606	15.7	1.75	565	<.01
Emotional problems interfere with daily life	4.9	1.46	604	10.3	1.08	565	<.01
Chest pain when hurrying or when walking uphill	0.8	0.70	222	0.7	0.53	178	n.s.
Chest pain when walking at normal pace on level ground	0.3	0.30	221	0.0	0.00	173	n.s.
Mean Inconsistency Rate (Unweighted)	2.1			3.6			
Activities of Daily Life							
Amount of help needed taking medicines	5.0	2.17	217	6.1	1.28	171	n.s.
Need help with transportation	0.4	0.35	214	1.5	0.97	169	n.s.
Need help with shopping	0.0	0.00	216	10.6	0.88	172	<.001
Need help with money management	0.0	0.00	219	0.4	0.29	173	n.s.
Need help using telephone	0.0	0.00	216	1.2	1.17	172	n.s.
Need help with housework and minor home maintenance chores	0.8	0.72	222	0.0	0.00	170	n.s.
Need help dressing	0.0	0.00	219	1.1	0.90	169	n.s.
Need help bathing	0.0	0.00	220	0.0	0.00	175	n.s.
Need help getting out of bed	0.0	0.00	222	0.6	0.43	173	n.s.
Need help eating	0.0	0.00	219	0.8	0.87	168	n.s.
Mean Inconsistency Rate (Unweighted)	0.6			2.2			
Use of Assistive Devices							
Uses any assistive device	0.0	0.00	224	2.0	0.16	185	<.001
Would be helped by assistive devices	1.0	0.94	178	0.0	0.04	121	n.s.
Uses eyeglasses or contact lenses	26.5	4.21	175	25.9	6.54	120	n.s.
Uses a hearing aid	2.3	1.19	175	3.6	2.35	120	n.s.
Uses dentures	13.4	2.03	175	15.0	4.04	120	n.s.
Uses phone with enlarged numbers	0.0	0.00	175	0.5	0.39	120	n.s.
Uses amplifier for telephone	0.0	0.00	175	1.1	1.00	120	n.s.
Uses a wheelchair	0.0	0.00	175	0.6	0.55	120	n.s.
Uses a cane	0.0	0.03	175	1.8	1.10	120	n.s.
Uses crutches	2.3	1.36	175	0.0	0.00	120	<.10
Uses a walker	0.0	0.00	175	0.4	0.41	120	n.s.
Uses a back or leg brace	0.0	0.00	175	0.1	0.05	120	n.s.
Uses an + A4 artificial limb	0.9	0.91	175	0.0	0.00	120	n.s.
Uses grab bars in tub or shower	1.0	0.97	175	3.2	2.11	120	n.s.
Uses grab bars on toilet	0.0	0.00	175	0.1	0.04	120	n.s.
Uses adult diapers	0.0	0.00	175	1.7	1.11	120	n.s.
Uses a bedside commode	0.0	0.00	175	0.4	0.41	120	n.s.
Uses a raised toilet seat	0.0	0.03	175	2.6	1.84	120	n.s.
Mean Inconsistency Rate (Unweighted)	2.6			3.3			
Other Measurements							
Doctor's visit in past 12 months	1.1	0.34	624	1.8	0.81	591	n.s.
Uses tobacco	0.8	0.30	466	1.9	0.83	416	n.s.
Uses alcohol	2.7	0.70	348	3.4	1.39	254	n.s.
Mean Inconsistency Rate (Unweighted)	1.5			2.4			

Notes. n.s. = not significant. Percentages are weighted for projection to the sampled population. Standard errors take account of the complex sample design used in the MBHSR Field Test. *p*-values are for tests of differences in the mean percentage inconsistent across the two age strata; those values may be slightly conservative since the tests assume that the age strata are independent, whereas the design drew the older and younger cohorts from the same primary sampling units.

[see 4. above]; response category is listed as: "dentures [full or partial]."")

In considering the last two results (4 and 5), it is worth noting that these apparent inconsistencies arose from comparisons of responses to questions that asked (1) whether the respondent currently used glasses, contact lenses, or dentures, and (2) whether these same devices would be of (future) help to the respondent. The high inconsistency rates reflect a large number of respondents answering "yes" to both questions. Prior to field data collection, we conducted a series of cognitive interviews to identify problems that respondents had understanding the MBHSR questions and completing the questionnaires. During these cognitive interviews we found that respondents often did not understand the questionnaire's implied distinction between their *current use* and *potential benefit* of future use of such assistive devices. While we attempted to rectify this problem in the final draft of the MBHSR instrument, it appears that further revisions will be needed to these questions.

It should also be noted that the foregoing analyses focus on *substantive* response inconsistencies. They do not include procedural errors in which respondents failed to follow a "skip" instruction but nonetheless gave a *consistent answer* to a question that they should have skipped. While such procedural errors do not affect the quality of the resultant data, it is worth noting that the MBHSR Field Test results confirmed the findings of our presurvey program of cognitive testing of the MBHSR draft questionnaires. Substantial numbers of substantive errors were avoided by designing the MBHSR questionnaire to include response categories that would be appropriate for persons who failed to follow a prior "skip" instruction. For the 41 MBHSR measurements shown in Table 3, nonsubstantive "skip" errors occurred more than 10 percent of the time for 12 measurements in the 65-year-old cohort and for 14 measurements in the 76- to 80-year-old cohort. Indeed, in the 76- to 80-year-old cohort, nonsubstantive skip errors were observed more than 20 percent of the time for 10 of the MBHSR measurements.

Discussion

The proposed Medicare Beneficiary Health Status Registry would (for budgetary reasons) have to rely heavily on mail data collection for its execution. An MBHSR survey design that made heavy use of telephone or in-person interviewing would not be economically viable. Because of this constraint, there were serious concerns about the ability of the survey to obtain both adequate response rates and measurements of sufficient quality to support the proposed use of these data for research and policymaking.

The literature on mail surveys focusing on the elderly population is not overly abundant (Smith and Biemer, 1991), and there are good reasons to be concerned about the ability of a mail survey of the elderly population to generate high-quality measurements. Available data indicate, for example, that nearly 13 percent of Americans aged 65 and older report having some problems seeing even while wearing eyeglasses; another 1 percent report being blind in both eyes. Among males aged 75 to 84, the rate of vision impairment rises to 17 percent (NCHS, 1984).

Similar concerns exist regarding the possible effects of cognitive impairment on the quality of the data obtained in a mail survey. Regier and colleagues (1988) estimate that the prevalence of severe cognitive impairment among those aged 65 to 74 is approximately 4.2 percent for males and 1.9 percent for females. These rates increase to approximately 7 percent for both males and females aged 75 to 84. However, in face-to-face interviews of elderly persons who were living in private homes, Gibson and Aitkenhead (1983) found that poor mental status did not cause substantial trouble to the interviewers, and Carp (1989) reported similar experiences. (As pointed out by one reviewer, this apparent paradox may reflect the fact that caregivers may refuse a request for an interview with a cognitively impaired person or provide a proxy response.)

There is, of course, a large literature showing encouraging levels of internal reliability for measurements of health status obtained using self-administered questionnaires (e.g., Stewart and Ware, 1992). Expectations about the equivalence of results obtained using different survey modes — not to mention the combination of modes that the proposed MBHSR would include — were less certain when we began the current research, given our focus on the segment of the population aged 65 and over. Recently, McHorney, Kosinski, and Ware (1994) reported parallel research administering the SF-36 to a sample aged 18 and over. Respondents were randomly assigned to receive the SF-36 either by mail (with telephone follow-up of nonrespondents) or by telephone. McHorney, Kosinski, and Ware report finding that SF-36 health ratings were less favorable and reports of chronic conditions were more frequent among mail than among telephone respondents; however, there were only small differences in measurement reliability (i.e., internal consistency of SF-36 scales) and data completeness. Data collection costs, however, were 77 percent higher for the telephone survey, and this effort yielded a significantly and substantially lower response rate than mail data collection with telephone follow-up (69% versus 79%). The generalizability of such findings to the target population for the proposed MBHSR is, however, unknown.

Such concerns motivated our field test of the proposed MBHSR. The test was designed to assess both the feasibility and the quality of the data that would be obtained if the MBHSR were conducted as a mail survey with telephone follow-up. The results of this field test have been surprisingly positive with regard to both the feasibility of such a strategy and the quality of the resultant data (Turner et al., 1994). The evidence presented in this article indicates that despite adequate reasons to expect discordance between Medicare records and survey reports (including unreliability of both our survey measurements as well as the administrative records), the MBHSR instrument provided relatively high levels of sensitivity and specificity in the assessment of 12 medical events that could be tracked from Medicare records.

The two instances in which MBHSR measurements had relatively poor sensitivity (COPD and stroke) may be substantively unique. The fact that the sensitivity of stroke measurements was higher when proxies were involved (.77) than when respondents reported themselves (.50) prompts

the speculation that persons who have experienced strokes may be less reliable reporters of their medical histories. This speculation should be empirically testable when the proposed MBHSR begins surveying larger samples of the population, and we believe such research should be encouraged. The low sensitivity of MBHSR measurements of COPD, on the other hand, is rather perplexing. The MBHSR survey instrument asks about "emphysema, chronic bronchitis, or chronic obstructive lung disease," and it is possible that these disease terms were unfamiliar to respondents or that some of them (e.g., emphysema) may have challenged respondents' reading abilities. It is also possible that COPD may have been the *secondary* diagnosis on a hospital admission prompted by a different medical condition. If that were the case, COPD might have been of low salience to the respondent. Since our MEDPAR criterion included both primary and secondary diagnoses for each admission, this could translate into a lower sensitivity for the MBHSR measurements of COPD. (A parallel phenomenon may account for the lower sensitivity observed for reporting of medical conditions, in general, versus medical procedures.) At present, we have no sure way of discerning the true source of the low sensitivity for COPD measurements. We would suggest that future research should focus on (1) assessing respondents' understanding of the medical terminology used in the COPD question, and (2) distinguishing between primary and secondary diagnosis codes in future MEDPAR validity studies.

We would also note that, to the extent that medical procedures and diagnoses are well understood and highly salient to respondents, it is reasonable to expect high levels of measurement sensitivity. This would be consistent with our finding of 100 percent sensitivity for MBHSR measurements of medical events such as implantation of cardiac pacemakers, cataract removal, hip replacement, mastectomy, etc. In addition to providing relatively high levels of external validity, we would note that our analyses indicate that the MBHSR instruments provided (in most instances) relatively high levels of response consistency. Thus, we found that, where assessments were possible, respondents provided highly consistent answers to related questions in the MBHSR. While consistency levels were somewhat attenuated for the cohort aged 76 to 80 and there were five instances in which the MBHSR measurements do not perform as expected, the overall pattern of results engenders considerable confidence in the quality of the MBHSR data. Overall, the foregoing evidence of data quality coupled with the relatively high response rates obtained in the MBHSR Field Test suggest that mail survey procedures (with telephone follow-up of nonresponders) are a viable option for the collection of health data from elderly Americans.

ACKNOWLEDGMENTS

The MBHSR Field Test and the analyses reported in this article were supported under Contract 500-90-0053 between the Health Care Financing Administration and Research Triangle Institute. Dr. Charles Turner served as Principal Investigator, Dr. Thomas Reilly served as Project Officer, and Mr. Kirk Pate served as Project Director for the MBHSR Field Test from 1992 through completion of the project in 1995. Dr. Judith T. Lessler served as Principal Investigator and Project Director and Dr. Lynne Penberthy served as Project Officer for the MBHSR Field Test from its inception in 1990 through 1992. The present article draws on an unpublished final report (Turner et al., 1994) summarizing major findings from the MBHSR Field Test. With the exception of this note, we do not indicate the numerous places in which we have drawn on that report.

Address correspondence to Dr. Charles F. Turner, Research Triangle Institute, 6101 Executive Boulevard, Rockville, MD 20852. Internet: CFT@RTI.ORG

REFERENCES

Aday, L.A. 1989. *Designing and Conducting Health Surveys: A Comprehensive Guide*. San Francisco: Jossey-Bass.

Bureau of the Census. 1991a. *Long-term Care Survey: Community Questionnaire* (OMB No. 0607-0635: expiration June 30, 1991). Washington, DC: Bureau of the Census.

Bureau of the Census. 1991b. *Longitudinal Study of Aging Questionnaire* (OMB No. 0920-0219; expiration May 31, 1991). Washington, DC: Bureau of the Census.

Carp, F.M. 1989. "Maximizing Data Quality in Community Studies of Older People: A Methodological Issue." *American Journal of Epidemiology* 126:962–971.

CDC (Centers for Disease Control). 1992. *1992 Behavioral Risk Factor Questionnaire*. Atlanta, GA: CDC.

Doll, H., K. McPherson, J. Davies, A. Flood, J. Smith, G. Williams, M. Ginzler, M. Petticrew, and N. Black. 1991. "Reliability of Questionnaire Responses as Compared With Interview in the Elderly: Views on the Outcome of Transurethral Resection of the Prostate." *Social Science and Medicine* 33:1303–1308.

Fisher, E.S., J.A. Baron, D.J. Malenka, J.A. Barrett, W.D. Kniffin, F.S. Whaley, and T.A. Bubolz. 1991. "Hip Fracture Incidence and Mortality in New England." *Epidemiology* 2:116–122.

Fisher, E.S., F.S. Whaley, W.M. Krushat, D.J. Malenka, C. Fleming, J.A. Baron, and D.C. Hsia. 1992. "The Accuracy of Medicare's Hospital Claims Data: Progress Has Been Made But Problems Remain." *American Journal of Public Health* 82:243–248.

Forsyth, B.H., D.K. Pate, T.K. Smith, and L. Fitterman. 1996. "Reinterview Methods for Assessing and Improving the Quality of Data From a Medicare Population." In R. Warnecke (Ed.), *Health Survey Research Methods: Conference Proceedings*. DHHS Pub. No. (PHS) 96-1013. Hyattsville, MD: National Center for Health Statistics.

Gibson, D.M. and W. Aitkenhead. 1983. "The Elderly Respondent: Experiences From a Large-scale Survey of the Aged." *Research on Aging* 5:283–296.

Glandon, G.L., M.A. Counte, and D. Tancredi. 1992. "An Analysis of Physician Utilization by Elderly Persons: Systematic Differences Between Self-Report and Archival Information." *Journal of Gerontology: Social Sciences* 47:S245–S252.

HCFA (Health Care Financing Administration). 1993. *Medicare Provider Analysis and Review File (MEDPAR): Data Users Reference Guide*. Baltimore, MD: HCFA.

Herzog, A.R. and R.A. Kulka. 1989. "Telephone and Mail Surveys With Older Populations: A Methodological Overview." In M.P. Lawton and A.R. Herzog (Eds.), *Special Research Methods for Gerontology*. Amityville, NY: Baywood.

Javitt, J.C., A.M. McBean, S.S. Sastry, and F. DiPaolo. 1993. "Accuracy of Coding in Medicare Part B Claims: Cataract as a Case Study." *Archives of Ophthalmology* 111:605–607.

Landry, J.A., M.A. Smyer, J.G. Tubman, D.J. Lago, J. Roberts, and W. Simonson. 1988. "Validation of Two Methods of Data Collection of Self-Reported Medicine Use Among the Elderly." *The Gerontologist* 28:672–676.

McHorney, C.A., M. Kosinski, and J.E. Ware. 1994. "Comparisons of the Costs and Quality of Norms for the SF-36 Health Survey Collected by Mail Versus Telephone Interview: Results From a National Survey." *Medical Care* 32:551–567.

NCHS (National Center for Health Statistics). 1984. Unpublished data from the Supplement on Aging to the 1984 National Health Interview Survey. Hyattsville, MD: NCHS.

Regier, D.A., J.H. Boyd, J.D. Burke, D.S. Rae, J.K. Myers, N. Kramer, L.N. Robins, L.K. George, M. Karno, and B.Z. Locke. 1988. "One-Month Prevalence of Mental Disorders in the United States Based on

Five Epidemiologic Catchment Area Sites." *Archives of General Psychiatry* 45:977–986.

Smith, T.K. and P.P. Biemer. 1991. *Literature Review of Mail Surveys of the Elderly*. Report to the Health Care Financing Administration under Contract 500-90-0053, July 11, 1991. Research Triangle Park, NC: Research Triangle Institute.

Stewart, A.L. and J.E. Ware (Eds.). 1992. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach*. Durham, NC: Duke University Press.

Turner, C.F., B.H. Forsyth, S.C. Wheeless, M.B. Witt, D.K. Pate, T.K. Smith, and F.J. Mierzwa. 1994. *Field Test of Medicare Beneficiary Health Status Registry: Major Findings and Recommendations*. Final report to the Heath Care Financing Administration under Contract 500-90-0053, August 4, 1994. Research Triangle Park, NC: Research Triangle Institute.

Received October 31, 1995
Accepted August 20, 1996