Measuring Clinical Performance: Comparison and Validity of Telephone Survey and Administrative Data

Betsy L. Thompson, Patrick O'Connor, Raymond Boyle, Michael Hindmarsh, Nagi Salem, Katrina Wynkoop Simmons, Edward Wagner, John Oswald, and Suzanne M. Smith

Objective. To compare and validate self-reported telephone survey and administrative data for two Health Plan Employer Data and Information Set (HEDIS) performance measures: mammography and diabetic retinal exams.

Data Sources/Study Setting. A telephone survey was administered to approximately 700 women and 600 persons with diabetes randomly chosen from each of two health maintenance organizations (HMOs).

Study Design. Agreement of survey and administrative data was assessed by using kappa coefficients. Validity measures were assessed by comparing survey and administrative data results to a standard: when the two sources agreed, that was accepted as the standard; when they differed, confirmatory information was sought from medical records to establish the standard. When confirmatory information was not available ranges of estimates consistent with the data were constructed by first assuming that all persons for whom no information was available had received the service and alternately that they had not received the service.

Principal Findings. The kappas for mammography were .65 at both HMOs; for retinal exam they were .38 and .40. Sensitivity for both data sources was consistently high. However, specificity was lower for survey (range .44 to .66) than administrative data (.99 to 1.00). The positive predictive value was high for mammography using either data source but differed for retinal exam (survey .69 to .78; administrative data .99 to 1.00).

Conclusions. Administrative and survey data performed consistently in both HMOs. Although administrative data appeared to have greater specificity than survey data the validity and utility of different data sources for performance measurement have only begun to be explored.

Key Words. Data quality, performance measures, preventive services, quality of care

Interest in measuring the quality of health care services has grown rapidly over the past decade (Angell and Kassirer 1996; Epstein 1995). The Health Plan Employer Data and Information Set (HEDIS) developed by the National Committee for Quality Assurance (NCQA 1997) has become a determinant of the services managed care organizations offer and what programs they target for improvement (Angell and Kassirer 1996). In 1995 the Jackson Hole Group launched the Foundation for Accountability, another national performance measurement effort (Skolnick 1997a). The Joint Commission on Accreditation of Healthcare Organizations (Skolnick 1997b) is also developing a performance measurement set. Performance measures are being used for marketing purposes, quality-improvement efforts, accreditation purposes, and consumer comparisons between providers and plans (Blumenthal 1996; Chernew and Scanlon 1998; Spoeri and Ullman 1997).

Despite this interest performance measurement is a nascent science with several unresolved issues including (1) a lack of data for measuring certain aspects of health care quality, (2) uncertainties about how the various data collection sources being used compare with one another, (3) discrepancies in the quality of data being used, and (4) uncertainties about how to adjust measures to make valid comparisons between various providers and plans (Eddy 1998; Epstein 1995; Iezzoni 1997; Localio, Hamory, Sharp, et al. 1995; Spoeri and Ullman 1997). We undertook this study to help address the first three of these concerns. Specifically we wanted to estimate (1) how well self-reported data from a telephone survey, which allows for collection of numerous measures not available from administrative data, perform for collection of selected HEDIS measures that are currently estimated from administrative data; (2) how survey and administrative data compare with

---

This work was supported through a task order contract with The HMO Group funded by the Centers for Disease Control and Prevention.

Address correspondence to Betsy L. Thompson, M.D., M.S.P.H., 4 Shadow Wood Place, Alamosa, CO 81011 (fax 719/589-0690; e-mail betsyt@vanion.com). Patrick O'Connor, M.D., M.P.H. and Raymond Boyle, Ph.D. are from HealthPartners, Group Health Foundation, Minneapolis. Michael Hindmarsh, M.A. and Edward Wagner, M.D., M.P.H. are from the Group Health Cooperative of Puget Sound, Center for Health Studies, Seattle. Nagi Salem, Ph.D. and John Oswald, M.P.H. are from the Minnesota Department of Health, Minneapolis. Katrina Wynkoop Simmons, Ph.D. is from the Washington State Department of Health, Olympia. Suzanne M. Smith, M.D., M.P.H. and Dr. Thompson are from the Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Adult and Community Health, Health Care and Aging Studies Branch. This article, submitted to *Health Services Research* on May 13, 1999, was revised and accepted for publication on May 16, 2000.

one another; and (3) the quality of both survey and administrative data against a data standard. Finally, we wanted to explore interests in health care quality that state and federal public health agencies share with managed care organizations. We selected two performance measures to examine in this report—mammography screening and retinal exams for persons with diabetes—because they are priority concerns for the agencies involved and are routinely collected through both the Behavioral Risk Factor Surveillance System (BRFSS) and HEDIS.

METHODS

This study was conducted as a collaborative effort among the Minnesota Department of Health, HealthPartners Research Foundation, the Washington Department of Health, Group Health Cooperative of Puget Sound, and the Centers for Disease Control and Prevention (CDC). The study subjects were enrollees of either HealthPartners, a 750,000-member network-model health maintenance organization (HMO) in the Twin Cities area of Minnesota or Group Health Cooperative of Puget Sound, a 500,000-member staff-model HMO in the greater Puget Sound area. Institutional review board approval for this project was obtained from both HMOs and CDC.

Methods used to define the sample populations and determine whether mammography or retinal exam was performed are consistent with HEDIS 2.5 (NCQA 1995; see Table 1). We estimated that we would need approximately 600 completed interviews for each sample to have 95 percent confidence in our estimates of sensitivity and specificity within .02 to .04 points. Because we expected that we would be unable to contact 10 to 15 percent of the sample and would get a 10 to 20 percent refusal rate we drew random samples of approximately 1,100 eligible plan members. In Minnesota the survey firm made multiple attempts within a randomly selected subsample of approximately 900 members (based on the assumptions above) and never used the larger sample of 1,100. In Washington all telephone numbers in the larger sample were used rather than limiting calls to a subsample initially. Fewer attempts were made to each number in Washington, resulting in a lower response rate. The final sample sizes, response rates, and cooperation rates are shown in Table 2. We used administrative data to compare differences in age, sex (for diabetes sample only), and receipt of retinal exam or mammography according to HEDIS methodology between survey respondents and nonrespondents.

Table 1: Data Sources and Information Used for Defining Sample Populations and Determining Whether Mammography and Retinal Exams Were Performed

	Definition	*Data Source*
Mammography sample	Women aged 52 to 64 years continuously enrolled for at least two years	Enrollment databases
Mammography performed (administrative data)	CPT-4 code 76090, 76091, or 76092; *or* revenue code 401 or 403; *or* ICD-9 procedure codes 87.36 or 87.37; *or* revenue code 320 or 400 *and* ICD-9 diagnosis code 174.xx, 198.81, 217, 233.0, 611.72, 793.8, V10.3, or V76.1 during preceding 24 months	Claims and encounter data
Mammography performed (telephone survey data)	1. A mammogram is an x-ray of each breast to look for breast cancer. Have you ever had a mammogram? 2. (If yes to 1) How long has it been since you had your last mammogram?	Survey
Diabetes sample	Enrollees 31 to 64 years of age continuously enrolled for at least one year, with at least two ICD codes 250.xx or treated with a diabetes-specific medicine* in the preceding year	Discharge, claims, encounter, and pharmacy data
Retinal exam performed (administrative data)	CPT-4 codes 92002, 92004, 92012, 92014, 92018, 92019, 92225, 92226, 92235, 92250 in the previous 12 months	Claims and encounter data
Retinal exam performed (telephone survey data)	When was the last time you had an eye exam in which the pupils were dilated? This would have made you temporarily sensitive to bright light.	Survey

*Insulin, sulfonylureas, or biguanides.

Table 2: Sample Sizes and Response Rates for Mammography and Diabetes Samples

	Minnesota	*Washington*
Mammography		
Total eligible people	21,351	20,492
Random sample selected	925	1,132
Deceased/ineligible	2	17
Refusals	80	146
Unable to contact	98	300
Completed interviews	745	669
Cooperation rate*	90.3%	82.1%
Response rate†	80.7%	60.0%
Diabetes		
Total eligible people	6,275	5,436
Random sample selected	810	1,134
Deceased/ineligible	6	0
Refusals	77	150
Unable to contact	122	324
Completed interviews	605	660
Cooperation rate*	88.7%	81.5%
Response rate†	75.2%	58.2%

*Completed interviews/(completed + refusals).

†Completed interviews/(completed + refusals + noncontacts).

Telephone survey items and methodology were based on the BRFSS, a state-based survey coordinated by CDC and conducted continuously in all 50 states, the District of Columbia, and several U.S. territories (Powell-Griner, Anderson, and Murphy 1997). We also asked respondents where services were provided so that services provided outside of the enrollees' current HMO could be verified; these questions are not asked on the regular BRFSS. Survey research firms that regularly conduct the BRFSS carried out the telephone survey by using list samples provided by the HMOs. A letter was sent to potential participants one week in advance of the first attempted telephone call to inform HMO enrollees of the project and give them the opportunity to refuse participation before the call was placed. At least ten attempts at different times of day were made to reach enrollees; directory assistance was used to locate new or correct numbers. Active conversion of initial refusals was not attempted.

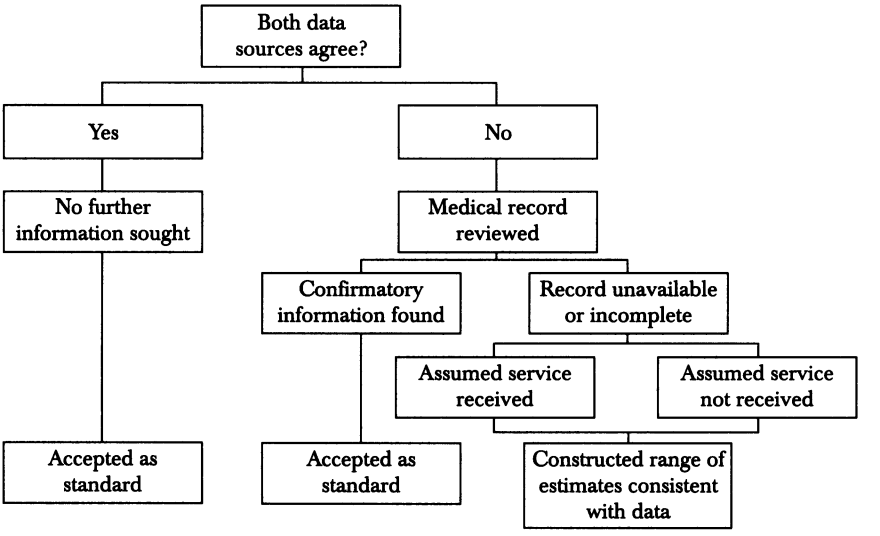
Administrative data methods were based on HEDIS 2.5 specifications that were in place at the time the study was initiated (NCQA 1995). We measured mammography screening within the preceding two years among

women 52 to 64 years of age who had been continuously enrolled for at least two years and retinal exams within the preceding 12 months among persons 31 to 64 years of age with diabetes who had been continuously enrolled for at least one year. Because there may be a one- to three-month lag between the time mammography or an eye exam is conducted and the time the documentation for the service appears in the databases used for this study administrative data were not analyzed until three months after the telephone survey was completed.

Agreement between survey and administrative data was assessed by using overall agreement and Cohen's kappa statistic, which adjusts for agreement caused by chance (Fleiss 1988). Because there is no gold standard for this project we constructed a standard to validate the two data sources in the following way (see Figure 1). If both telephone survey and administrative data agreed that the service was either received or not received, that was accepted as accurate. When the two sources differed further information from the medical record or telephone contact with clinics (if outside the HMO) was sought. When further information was not available for discrepant cases a range of validity estimates consistent with the data was calculated by first assuming that all those with discrepant findings and no confirmatory information had received the service, then that they had not received the service. Sensitivity, specificity, positive predictive values, and negative predictive values were calculated for both data sources against this constructed standard.

We examined two possible reasons for disagreement between the two data sources: the effect of "telescoping" on survey responses and the use of different administrative databases, currently allowed by HEDIS, for identifying persons with diabetes. Telescoping occurs when a person recalls that a particular event happened more recently than it actually did (Sawyer, Earp, Fletcher, et al. 1989; Sudman and Bradburn 1973). In addition to the comparison of administrative and survey reports of mammography during the previous two years we compared survey reports of a mammogram during the previous two years with administrative records for the preceding three years to measure the extent of telescoping. Similarly, for retinal exam—where the HEDIS measure covers the previous year—we also compared survey reports of an eye exam during the previous year to administrative data for the previous two years. To examine the effect of using different administrative databases for identification of persons with diabetes we looked at retinal exam rates among those identified with ICD-9 codes (Health Care Financing Administration 1994), those identified using prescription data, and those identified using either database.

Figure 1: Flow Diagram of Standard Construction Based on Agreement Between Administrative and Survey Data and Availability of Confirmatory Information from Medical Record



RESULTS

Respondents and nonrespondents did not vary by age but women in the Minnesota diabetes sample were more likely to respond to the survey than men (79.5 percent vs. 71.2 percent; $p = .007$). Survey respondents were also more likely to have received mammography than nonrespondents (82.0 percent vs. 64.8 percent in Minnesota; 73.5 percent vs. 64.4 percent in Washington; $p < .05$) based on available administrative data. Retinal exams rates did not vary between respondents and nonrespondents.

Among enrollees at the two HMOs 83 percent and 90 percent of survey respondents reported receiving a mammogram compared with 74 percent and 82 percent from administrative data; 73 percent and 79 percent of survey respondents reported receipt of a retinal exam compared with 49 percent and 53 percent from administrative data. Approximately 20 percent of enrollees in Minnesota and fewer than 5 percent of enrollees in Washington reported receiving a mammogram or retinal exam from a provider outside of their health plan. Survey and administrative data agreed on whether the service was

provided for 91.5 percent and 88.2 percent of mammography respondents in Minnesota and Washington, respectively, and for 70 percent of respondents with diabetes in both sites. Kappa scores were also similar in both sites but were higher for mammography (.65 in both sites) than for retinal exams (.40 and .38).

Confirmatory information was available for 159 of the 515 enrollees for whom survey and administrative reports were discordant. When administrative data indicated that the service was received, confirmatory information usually upheld the administrative data (for nine of ten women in the mammography sample and 23 of 28 patients with diabetes). However, when administrative data indicated that the service was not received, approximately half of the self-reports of receiving the service were confirmed (15 of 38 in the mammography and 45 of 83 in the diabetes samples).

Both data sources yielded consistently high sensitivities across measures and sites (see Table 3). Specificity was significantly and consistently lower for survey methodology compared with HEDIS methodology. For retinal exam positive predictive values were substantially higher for administrative than for survey data. Negative predictive values were consistently high across sites and measures.

In regard to telescoping 15 of 62 (24.2 percent) women in Minnesota and 35 of the 56 (62.5 percent) women in Washington who reported having a mammogram in the previous two years but were classified as not having

Table 3: Validity of Survey and Administrative Data Sources Compared with a Constructed Standard

	Sensitivity	*Specificity*	*Positive Predictive Value*	*Negative Predictive Value*
Retinal exam				
Minnesota survey data*	.95–.96	.52–.59	.69–.78	0.89–0.92
Administrative data*	.83–.92	.98–1.00	.98–1.00	0.79–0.92
Washington survey data	.97	.44	.69	0.93
Administrative data	.94	.99	.99	0.93
Mammography				
Minnesota survey data*	1.00	.57–.66	.92–.94	0.99–1.00
Administrative data*	.96–.99	.99–1.00	1.00	0.82–0.96
Washington survey data*	.98	.63–.65	.89–.90	0.92
Administrative data*	.97–.98	1.00	1.00	0.92–0.95

*Range of values consistent with the data are presented by assuming that those with discrepant findings from the two data sources and no confirmatory information had received the service, then that they had not received the service.

a mammogram by the standard had received a mammogram within the previous three years according to administrative records. In Washington 65 of the 164 (39.6 percent) respondents with diabetes who reported receiving a retinal exam within the previous year but were classified as not having the exam by the standard had a retinal exam in the previous two years according to administrative records.

Using different databases for identifying persons with diabetes in Minnesota had little effect on agreement or kappa scores but did affect retinal exam rates. Restricting our analysis to persons identified by ICD-9 codes only, we found that survey and HEDIS methodologies agreed for 70.1 percent of respondents with a kappa of .37. When restricted to persons identified from pharmacy data only the two sources agreed for 70.3 percent with a kappa of .41. Estimates of retinal exam rates varied by as much as a 5-percent absolute or 10-percent relative difference depending on the database used and were highest if based solely on ICD-9 codes. The use of different databases for identifying persons with diabetes similarly affected retinal exam rates estimated from survey and administrative data sources (see Table 4).

DISCUSSION

There is interest in expanding performance measurement for specific chronic diseases as well as for certain subpopulations of patients (Angell and Kassirer 1996; Galvin 1998; Lansky 1998). Developing such measures has proven problematic for several reasons, most notably the lack of available, high-quality, and comparable data. Efforts have been hampered because of the inability to completely identify the subpopulation at risk (e.g., persons with asthma), by difficulties ascertaining the particular event accurately (e.g., low birth weight), or because data are not available from medical records or administrative data (e.g., health education or health risk behavior data).

Table 4: Effect of Using Different Databases for Identifying Diabetics on Estimates of Retinal Exam Rates by Survey and Administrative Data Sources

Data source(s) used	N	*Survey (%)*	*HEDIS (%)*
ICD-9 and prescription	605	73.2	49.1
Prescription only	532	72.8	50.1
ICD-9 only	402	77.9	54.6

Measurement sets such as HEDIS have focused on measures of health care quality that can be obtained readily from administrative data, primarily preventive services that are well recorded, such as mammography and Pap smear, and limited aspects of certain chronic diseases, such as retinal exams for persons with diabetes. Virtually no data demonstrate that such measures adequately describe the overall quality of care delivered by a health care organization (Brook, McGlynn, and Cleary 1996; Eddy 1998).

Our data document that both survey and administrative data can provide valid estimates for receipt of mammography but that administrative data appear to outperform survey data for estimating receipt of retinal exams among persons with diabetes (see Table 3). Consistent with previous studies we found some evidence that self-report leads to overreporting of receipt of these services because of telescoping (McGovern et al. 1998; Sawyer, Earp, Fletcher, et al. 1989; Sudman and Bradburn 1973). However, it is also possible that the administrative method and the standard we constructed underestimate services received either because an appropriate procedure code was not assigned even though the service was provided or because the service was received outside of the enrollees' current health plan and could not be verified. While we are reluctant to describe our constructed standard as "gold," we do believe that it has clear advantages over simply accepting administrative data or medical record data alone as truth. Nevertheless, our standard is limited by the same factors that limit the usefulness of medical record and administrative data (Iezzoni 1997). It is important to determine both the magnitude and causes of discrepancies between administrative and survey reports. For example, we found that 24 to 62 percent of the discrepancies were because of telescoping rather than reporting that an event that had never occurred had occurred. This has important implications for interventions that might be implemented.

It has proven difficult to accurately define subpopulations of interest for health care performance measures. HEDIS 2.5 allowed plans some flexibility in their choice of data sources for identifying enrollees with diabetes. We found that the choice of definitions affected the rates of retinal exams by as much as 10 percent. This degree of difference needs to be considered when comparing data from plans that use different data sources to identify persons with diabetes. The similar performance of both methodologies across sites and services suggests that current efforts to compare quality of care across plans are reasonable when standard definitions and methods are used.

Our study may be limited in its generalizability to other HMOs. Both study HMOs have relatively sophisticated clinical databases and decades of

experience in managed care. It seems likely that plans with less-sophisticated data systems would be even more likely to underestimate the receipt of preventive services when using administrative data alone. We also found evidence that survey respondents differed from nonrespondents although it is difficult to say from these data how those differences affected our comparisons.

When making decisions about data sources for performance measurement the inherent strengths and weaknesses of survey and administrative data sources should be considered. A notable strength of surveys is their ability to provide data that are not available from other sources. As with most HMOs the participating plans do not have administrative data about socioeconomic status or race. In addition measures of satisfaction, access, health risk behaviors, functional and health status, as well as the receipt of certain preventive services (e.g., foot exams or counseling services) can be collected through surveys and are rarely available from administrative data. Another advantage of surveys is that they can be conducted independent of the plan being assessed. Theoretically this could result in increased standardization among plans and less "gaming" of measures. Finally, information about services received outside of the health plan can be collected through surveys.

On the other hand, it seems likely that administrative data give more valid estimates of certain measures, particularly if it is a procedure, utilization, or other type of measure that is nearly universally and accurately captured in administrative data. Administrative data can be subjected to external audits to improve standardization (Spoeri and Ullman 1997). It is difficult to make any definitive recommendations about costs of various data sources. Although the costs to initiate a new survey are high there is growing consensus that consumer surveys are a necessary tool in quality measurement (e.g., Blumenthal 1996; Lansky 1998), and the cost to add an additional question is nominal. Administrative data are inexpensive to collect only if the necessary electronic databases are available, accurate, and accessible.

If we want to develop performance measures that adequately and accurately measure the quality of health care delivered—be it for a specific subset of patients such as persons with diabetes or the entire population served—more work is needed to develop and test measures. How valid are they? Do they measure what we think they measure? What is the most efficient and effective way to use different data sources to measure quality? It also seems likely that performance measurement efforts will be enhanced by involving participants from various sectors and with varying perspectives. Although several groups

are represented on the decision-making committees of NCQA (1997) and the Foundation for Accountability (1996) performance measurement sets have been criticized for being too narrowly focused on employers' needs (Galvin 1998; Lansky 1998; Thier and Gelijns 1998). This study supplies a piece of the performance measurement puzzle. Continued methodologic evaluation is critical if we are going to continue to move performance measurement forward with the hope that we will actually be able to measure quality.

REFERENCES

Angell, M., and J. P. Kassirer. 1996. "Quality and the Medical Marketplace—Following Elephants." *New England Journal of Medicine* 335 (12): 883–85.

Blumenthal, D. 1996. "Part 1: Quality of Care—What Is it?" *New England Journal of Medicine* 335 (12): 891–93.

Brook, R. H., E. A. McGlynn, and P. D. Cleary. 1996. "Part 2: Measuring Quality of Care." *New England Journal of Medicine* 335 (13): 966–70.

Chernew, M., and D. P. Scanlon. 1998. "Health Plan Report Cards and Insurance Choice." *Inquiry* 35 (1): 9–22.

Eddy, D. M. 1998. "Performance Measurement: Problems and Solutions." *Health Affairs* 17 (4): 7–25.

Epstein, A. 1995. "Performance Reports on Quality—Prototypes, Problems, and Prospects." *New England Journal of Medicine* 333 (1): 57–61.

Foundation for Accountability. 1996. "In Practice—Health Risks." *Accountability* 1 (1): 1–20.

Fleiss, J. L.1988. *Statistical Methods for Rates and Proportions, 2nd edition.* New York: John Wiley & Sons.

Galvin, R. S. 1998. "Are Performance Measures Relevant?" *Health Affairs* 17 (4): 29–31.

Health Care Financing Administration. 1994. *International Classification of Diseases, 4th edition, 9th revision, clinical modification.* Washington, DC: U.S. Department of Health and Human Services, Public Health Service.

Iezzoni, L. I. 1997. "Assessing Quality Using Administrative Data." *Annals of Internal Medicine* 127 (8): 666–74.

Lansky, D. 1998. "Measuring what Matters to the Public." *Health Affairs* 17 (4): 40–41.

Localio, A. R., B. H. Hamory, T. J. Sharp, S. L. Weaver, T. R. TenHave, and J. R. Landis. 1995. "Comparing Hospital Mortality in Adult Patients with Pneumonia: A Case Study of Statistical Methods in a Managed Care Program." *Annals of Internal Medicine* 122 (2): 125–32.

McGovern, P. G., N. Lurie, K. L. Margolis, and J. S. Slater. 1998. "Accuracy of Self-Report of Mammography and Pap Smear in a Low-Income Urban Population." *American Journal of Preventive Medicine* 14 (3): 201–08.

National Committee on Quality Assurance. 1995. *HEDIS 2.5.* Washington, DC: NCQA.

———. 1997. *HEDIS 3.0.* Washington, DC: NCQA.

Powell-Griner, E., J. E. Anderson, and W. Murphy. 1997. "State- and Sex-Specific Prevalence of Selected Characteristics—Behavioral Risk Factor Surveillance System, 1994 and 1995." *CDC Surveillance Summaries, Morbidity and Mortality Weekly Report* 46 (SS-3): 1–31.

Sawyer, J. A., J. A. Earp, R. H. Fletcher, F. F. Daye, and T. M. Wynn. 1989. "Accuracy of Women's Self-Report of Their Last Pap Smear." *American Journal of Public Health* 79 (8): 1036–37.

Skolnick, A. A. 1997a. "A FACCT-Filled Agenda for Public Information." *Journal of the American Medical Association* 278 (19): 1558.

———. 1997b. "Joint Commission Begins Tracking Outcome Data." *Journal of the American Medical Association* 278 (19): 1562.

Spoeri, R. K., and R. Ullman. 1997. "Measuring and Reporting Managed Care Performance: Lessons Learned and New Initiatives." *Annals of Internal Medicine* 127 (8): 726–32.

Sudman, S., and N. M. Bradburn. 1973. "Effects of Time and Memory Factors on Response in Surveys." *Journal of the American Statistical Association* 68 (344): 805–15.

Thier, S. O., and A. C. Gelijns. 1998. "Improving Health: The Reason Performance Measurement Matters." *Health Affairs* 17 (4): 26–28.