

UCLA

UCLA Previously Published Works

Title

Methodological Challenges Associated with Patient Responses to Follow-up Longitudinal Surveys Regarding Quality of Care

Permalink

<https://escholarship.org/uc/item/82d196w4>

Journal

Health Services Research, 38(6p1)

ISSN

0017-9124

Authors

Kahn, Katherine L
Liu, Honghu
Adams, John L
et al.

Publication Date

2003-12-01

DOI

10.1111/j.1475-6773.2003.00194.x

Peer reviewed

Methodological Challenges Associated with Patient Responses to Follow-up Longitudinal Surveys Regarding Quality of Care

Katherine L. Kahn, Honghu Liu, John L. Adams, Wen-Pin Chen, Diana M. Tisnado, David M. Carlisle, Ron D. Hays, Carol M. Mangione, and Cheryl L. Damberg

Objective. To illustrate, using empirical data, methodological challenges associated with patient responses to longitudinal surveys regarding the quality of process of care and health status, including overall response rate, differential response rate, and stability of responses with time.

Data Sources/Study Setting. Primary patient self-report data were collected from 30,308 patients in 1996 and 13,438 patients in 1998 as part of a two-year longitudinal study of quality of care and health status of patients receiving care delivered by 63 physician organizations (physician groups) across three West Coast states.

Study Design. We analyzed longitudinal, observational data collected by Pacific Business Group on Health (PBGH) from patients aged 18–70 using a four-page survey in 1996 and a similar survey in 1998 to assess health status, satisfaction, use of services, and self-reported process of care. A subset of patients with self-reported chronic disease in the 1996 study received an enriched survey in 1998 to more fully detail processes of care for patients with chronic disease.

Data Collection/Extraction Methods. We measured response rate overall and separately for patients with chronic disease. Logistic regression was used to assess the impact of 1996 predictors on response to the follow-up 1998 survey. We compared process of care scores without and with nonresponse weights. Additionally, we measured stability of patient responses over time using percent agreement and kappa statistics, and examined rates of gender inconsistencies reported across the 1996 and 1998 surveys.

Principal Findings. In 1998, response rates were 54 percent overall and 63 percent for patients with chronic disease. Patient demographics, health status, use of services, and satisfaction with care in 1996 were all significant predictors of response in 1998, highlighting the importance of analytic strategies (i.e., application of nonresponse weights) to minimize bias in estimates of care and outcomes associated with longitudinal quality of care and health outcome analyses. Process of care scores weighted for nonresponse differed from unweighted scores ($p < .001$). Stability of responses across time was moderate, but varied by survey item from fair to excellent.

Conclusions. Longitudinal analyses involving the collection of data from the same patients at two points in time provide opportunities for analysis of relationships between

process and outcomes of care that cannot occur with cross-sectional data. We present empirical results documenting the scope of the problems and discuss options for responding to these challenges. With increasing emphasis in the United States on quality reporting and use of financial incentives for quality in the health care market, it is important to identify and address methodological challenges that potentially threaten the validity of quality-of-care assessments.

Key Words. Methodological challenges, longitudinal studies, quality of care

In recent years, longitudinal data analysis has been heralded as a means to allow us to learn about the relationships between what is done for a patient (process of care) and what happens to them (outcomes of care) (Donabedian 1988, 1968; Blumenthal 1996; Miller and Luft 1997; Fremont et al. 2001; Cleary 1999; Brook, McGlynn, and Cleary 1996; Hays et al. 1999). The collection and analysis of data from the same patient at two or more points in time can provide information about the processes and outcomes of care that cannot be ascertained with cross-sectional data analyses, even when they are collected at more than one point in time. However, reliance on data from multiple time points within patient generates challenges regarding differential nonresponse, as well as the stability of patient reports with time. Although these methodological issues have been extensively studied, particularly with respect to panel economic data (Kasprzyk et al. 1989), few applications have documented how quality scores might vary as a function of nonresponse or stability (Allen and Rogers 1997; Cleary et al. 1998; Zaslavsky, Zaboriski, and Cleary 2000).

The conduct of large follow-up longitudinal patient studies in community settings provides the opportunity to examine patient- and organization-level data in terms of response rate, response bias, and stability of responses across time. Regardless of whether longitudinal quality of care evaluations occur at the level of the individual provider, physician group, or health plan level, the distribution of patient response to follow-up surveys are central to the

Address correspondence to Katherine L. Kahn, M.D., Professor of Medicine, Associate Division Chief for Research, UCLA School of Medicine, Division of General Internal Medicine and Health Services Research, 911 Broxton Plaza, Los Angeles, CA 90095-1736. Dr. Kahn is also a Senior Scientist for RAND Health in Santa Monica, CA. Honghu Liu, Ph.D., is an Associate Adjunct Professor, UCLA School of Medicine, Division of General Internal Medicine and Health Services Research, Los Angeles. John L. Adams, Ph.D., is a Statistician at RAND Health, Santa Monica. Wen-Pin Chen, M.S., is a Programmer/Analyst, UCLA School of Medicine, Division of General Internal Medicine and Health Services Research, Santa Monica.

inferences that can be made. We provide empirical data about and discuss policy implications of longitudinal data we observed in a large study of care and outcomes for patients clustered by medical organization across a two-year time window.

METHODS

Overview of Study

The Pacific Business Group on Health (PBGH), a large nonprofit business coalition of 47 private and public sector purchasers of health care represents approximately three million employees, dependents, and retirees, and \$4 billion in annual health care expenditures (Pacific Business Group on Health 1989) with most of its members' employees and dependents enrolled in health maintenance organizations (HMOs), point of service (POS), or preferred provider organization (PPO) plans (Damberg and Bloomfield 1997; Pacific Business Group on Health Online 2002). Across the last decade, PBGH has systematically collected and publicly reported quality data with the goal of driving improvements in the health system. In 1996, PBGH initiated a two-year longitudinal evaluation of the quality of care delivered by West Coast physician groups. The Physician Value Check Survey Project surveyed 61,998 patients who received care from 63 associated managed care physician groups (medical groups and independent practice associations [IPAs] in California, Oregon, and Washington). The survey asked patients to report on their satisfaction with care, receipt of preventive care services, the management of hypertension and/or hypercholesterolemia, and health-related quality of life. The PBGH tracked this cohort by again surveying patients in 1998 with the goal of assessing the longitudinal changes in satisfaction with care, process of care, and outcome of care as measured by changes in health-related quality of life. The objective was to hold medical groups accountable for the long-term functioning and well-being of their patient population.

Sample

All HMO commercially insured and Medicare-risk contract senior plan members between the ages of 18 and 70 years enrolled in any of the 63 physician groups, who had at least one physician encounter during calendar year 1995 defined the 1996 study population. The eligible HMO participants were supplemented with a sample of 4,000 eligible participants aged 18–70 years with at least one visit in 1995 and receipt of care from nonmanaged care

PPO providers. The size of the 1996 population (i.e., the sampling frame) was 1,170,242 patients.

From each physician group, 1,000 patients were randomly selected, with oversampling of individuals between ages 50 and 70 years old to increase the likelihood of detecting two-year changes in health-related quality of life. Of the 1,000 enrollees associated with each medical organization, 700 were drawn from the 50 to 70 year-old group and 300 from the 18 to 49 year-old group. The final sample size reflected the need to ensure enough respondents at Time 2 (i.e., 1998) to detect differences between physician groups in the outcomes of interest, assuming a 50 percent response rate at Time 1 and approximately 20 percent attrition in enrollment over time (Damberg and Bloomfield 1997).

1996 and 1998 Patient Surveys

All patients who responded to the 1996 survey were invited to participate in either an updated 1998 version of the four-page 1996 survey (called the 1998 Core survey) or an enriched data collection with a nine-page survey that supplemented Core Survey questions with additional severity and process-of-care measures to understand process of care associated with chronic conditions (known as the Chronic Condition Survey). Patients who specified with the 1996 survey that they had at least one of four chronic conditions (diabetes, ischemic heart disease, asthma or COPD, or low back pain) were randomized to receive either survey. Patients who did not indicate they had one of the specified chronic conditions on the 1996 survey were sent the updated version of the four-page 1996 survey (i.e., the 1998 Core).

Field Work

In 1996, patients selected for the study sample were sent a pre-alert postcard, the survey, a standard cover letter printed on the letterhead of the patient's physician group and signed by that group's medical director using first-class mail, in a standard envelope. Two weeks after sending the first survey, a reminder postcard and then a replacement survey was sent to all patients who had not responded; six weeks after sending the initial survey, phone follow-up was initiated to encourage nonrespondents to complete the survey. As an incentive to complete the survey, respondents were eligible to receive one of fifty \$100 cash prizes.

In 1998, a similar protocol was followed for the standard four-page survey sample, with a third copy of the survey mailed to all nonrespondents. For the chronic condition survey, a similar protocol was used except

the phone follow-up was initiated earlier, at two weeks after the mailing of the survey with more frequent follow-up calls than had occurred in 1996.

Variables Used to Predict Patient Response in 1998

All variables available from the 1996 survey that made clinical sense as possible predictors of response to a 1998 survey were included in the model predicting the likelihood of responding to the 1998 survey: patient demographics (age, gender, race, ethnicity, highest level of education, household income); insurance (use of a copayment); cigarette use (never, former, or current smoker); comorbidity (i.e., the presence or absence of a history of hypertension, myocardial infarction, heart disease such as angina or heart failure, chronic lung disease, diabetes, cancer, migraine headache, chronic allergies, seasonal allergies, arthritis, sciatica, ulcers, hemorrhoids, dermatitis, kidney problem, limitations in the use of limbs, blindness or trouble seeing, blurred vision, deafness, epilepsy, thyroid problems), and health-related quality of life (general health now as compared with one year ago; SF-12 physical and mental health summary scores) (Ware, Kosinski, and Keller 1996). Additionally, the 1996 survey queried patients about use of health services (i.e., number of visits to the doctor, hospitalizations, and use of out-of-plan services during the prior 12 months; use of urgent care services during the prior 6 months); continuity with providers (i.e., number of years enrolled in their current health plan, number of years receiving care from their current medical doctor[s]); experiences with the health care system (i.e., delays in medical care while waiting for approval), receipt of counseling by the provider for any of the following (exercise, nutrition, smoking, injury prevention, motor vehicle safety, alcohol and substance abuse, sexually transmitted diseases, contraception), and patient ratings and satisfaction with health care.

We examined the associations between response in 1998 (yes or no) and specifications of 1996 variables for the longitudinal cohort who responded in 1996 and were invited to respond in 1998. (Descriptions of categories of ordered and categorical data, as well as descriptions of continuous variables are presented later in Table 2). Comorbidity was entered into the regression as a variable representing a count of up to 21 comorbid conditions.

Patient ratings of care from 1996 were entered using three separate scales, each scored with a possible range from 0 to 100 with 100 representing maximum satisfaction for patient self-report of the doctor's quality of care,

satisfaction with health plan, and satisfaction with access to and promptness of care.

Because patients were sampled from medical organizations we anticipated responses could cluster within medical organizations; accordingly we used administrative data to specify each patient as belonging to one of 63 physician organizations, either a medical group, an IPA, or a PPO. We also tested response rates in 1998 as a function of medical organization type (with two dummy variables) and as a function of the identity of the medical organization with a dummy for each of the medical organizations.

We tested three models for response to the Core survey and three models for response to the Chronic Condition Survey: the parsimonious model, which included demographics, health plan use, and comorbidity was tested because other researchers interested in lessons from this analysis may only have access to a smaller number of variables; the intermediate model supplemented the parsimonious model with 1996 SF-12 physical and mental health summary scores; the enriched model used all available 1996 variables to best understand how underlying patient characteristics and baseline patient report of care experiences and health-related quality of life might influence response to a two-year follow-up longitudinal survey. This model included demographics, insurance, comorbidity, SF-12 summary score, satisfaction, patient ratings of their doctor, health plan, and other factors, enrollment in a medical group (omitted variable), IPA, or PPO.

Statistical Methods and Modeling

Missing data were filled with hot deck imputation methods based on randomly selected nonmissing values from patients who reported similar age, gender, race, and ethnicity in 1996 for variables with >5 percent rates of missing (Little and Rubin 1987).

We compared the marginal means or proportions of 1996 patient-level variables between respondents and nonrespondents. We then used multiple logistic-regression analysis (Hosmer and Lemeshow 1989) to assess the independent effects of patient characteristics, their organization type (medical group versus IPA versus PPO), and dummies for each of their 49 organizations on response rates. We adjusted for the underlying cluster effect of patients within medical organization on standard error estimates of model parameters using the technique of Huber/White correction (White 1980) and tested the significance of model parameter estimates. We present univariate statistics and adjusted odds ratios obtained from the logistic regression models. A *p*-value of

less than .05 (two-tailed testing) was considered to indicate statistical significance.

To calculate adjusted differences of mean predicted probabilities of response across categories of patient characteristics (such as male versus female), we used the final regression model fitting each variable with the mean for both categorical and continuous variables (Neter, Wasserman, and Kutner 1985).

To address patient nonresponse in 1998, we constructed nonresponse weights with the weight = $1/(\text{predicted probability of response in 1998})$ from the logistic prediction model. We tested whether the models predicting five separate measures of 1998 process varied as a function of 1996 patient characteristics when nonresponse weights both were not and were included in the regression model.

Observed Stability in Patient Reports with Time

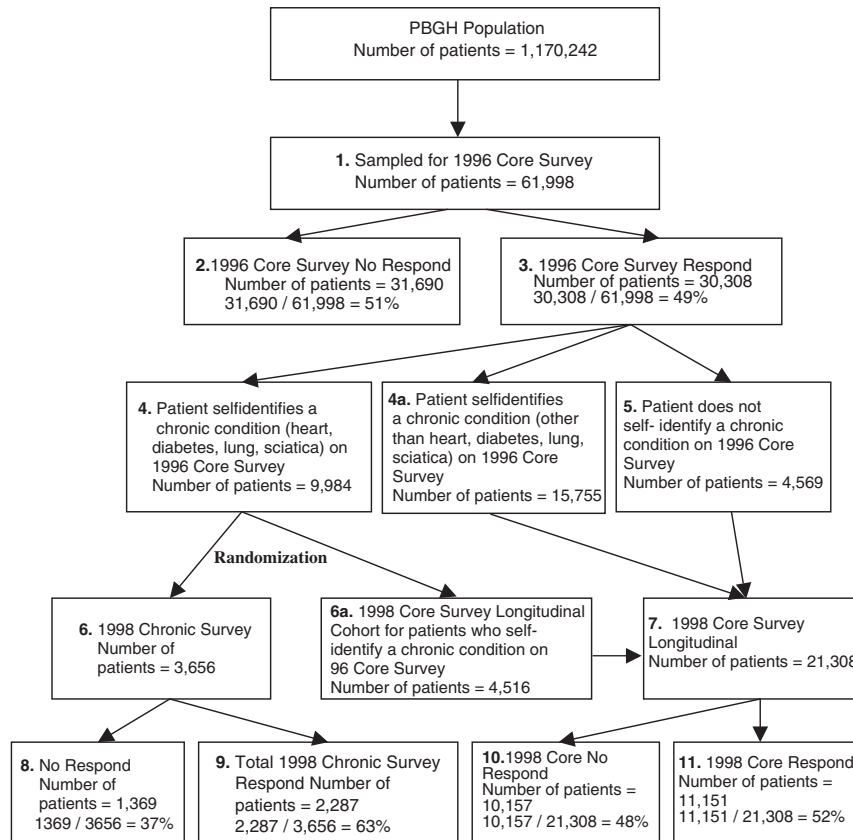
We used survey data from the longitudinal cohort of patients who provided data in both 1996 and 1998 to assess patient report of demographics, presence ever of a comorbidity (heart disease like angina, cancer, high blood pressure, diabetes, migraine headache, high blood pressure, and high cholesterol), counseling of the patient by the provider regarding eating less fat and/or using less salt due to hypertension, and use of medications for hypertension and/or high cholesterol in both 1996 and 1998. We studied marginal frequencies for the same cohort across two time periods separated by 24 months. We present the rates of agreement for the presence or absence of the condition assessed, and unweighted kappa score for demographic and clinical data (Fleiss 1981). Using the nomenclature of Landis and Koch, we use kappa terms to describe the relative strength of agreement between baseline and follow-up responses: kappa 0.81–1.00 (almost perfect); 0.61–0.80 (substantial); 0.41–0.60 (moderate); 0.21–0.40 (fair); 0.00–0.20 (slight); <0.00 (poor) (Landis and Koch 1977). We also observed the frequency with which patients report a different gender across the two time windows.

RESULTS

Patients and Medical Organizations Participating as Survey Respondents

From the PBGH population of 1,170,242 patients, 61,998 patients from California, Washington, and Oregon were mailed the 1996 survey (Table 1). Of the 63 physician groups who agreed to participate in the 1996 survey, 11

Table 1: Cohort Tree



groups elected not to participate in the 1998 survey including three that were no longer in business in 1998, one that had an extremely low response rate in 1996, and seven that declined to participate in 1998 without explanation (including three with very low SF-12 and performance scores in 1996). The 1996 patient response rate was 49 percent with 30,308 patients from 63 groups providing completed patient surveys in 1996.

The 9,984 1996 survey respondents who self-identified diabetes, ischemic heart disease, asthma or COPD, or low back pain were randomized in 1998 to receive either the enriched nine-page chronic survey (raw response rate 63 percent in 1998, standard deviation [SD] 11.40, group-level response

rate range 42–100 percent) or the standard four-page core survey (response rate 55 percent in 1998). Patients who did not report one of the specified chronic conditions in 1996 were sent the standard four-page core survey, which showed a response rate of 52 percent (SD 6.53, group level response rate range 39–66 percent). These latter patients were merged with the chronic patients who were randomized to the core survey for an overall core 1998 survey response of 52 percent for 11,151 patients from 49 medical organizations (SD 6.13, group-level response rate range 39–66 percent).

1998 Follow-up Survey Respondents Differ from 1996 Survey Respondents

Table 2 shows rates of survey response in 1998 stratified by 1996 survey data. For example, using a bivariate comparison (*Raw Rate of Response* column in Table 2), we see the survey response in 1998 was 53 percent for patients with the mean 1996 sample age of 53 years, 35 percent for patients with an age one standard deviation below the 1996 mean age (i.e., ≤ 41 years) and 66 percent for patients with age one standard deviation above the 1996 mean age (i.e., ≥ 66 years).

The univariate statistics for predicted probability of survey response at follow-up are mean (.52, SD .13), median (.54), range (.73 from .10 to .83). Adjusted for listed 1996 patient characteristics and participation in type of medical organization (medical group, IPA, or PPO) as well as for clustering of patients within medical organization (Adjusted Rate of Response column in Table 2), we see substantial variations in 1998 response rates across multiple domains of 1996 patient characteristics.

In 1998, the adjusted response rate for women and men was 54 percent and 50 percent, respectively. White patients responded in 1998 with an adjusted rate of 54 percent compared with lower rates for African Americans (44 percent), Hispanics (51 percent), Asians or Pacific Islanders (51 percent), and other multiracial or not specified (46 percent) ($p < .01$). Adjusted rates of survey response in 1998 increased with more education, and with more income. Some of the patients the health care system is most interested in learning about were less likely to respond in 1998 (Table 2). For example, 1998 adjusted response was 52 percent for former smokers, 54 percent for never smokers, and 48 percent for current smokers ($p < .05$). Patients who reported general health much worse now as compared with one year ago responded in 1998 with an adjusted rate of 44 percent compared with 52 percent for those with more stable health status with time ($p < .001$). Response rate varied as a

Table 2: 1998 Survey Raw Rate and Predicted Probability of Response as a Function of 1996 Variables

1996 Characteristics	Response Categories	N ¹ = 21,308	Unadjusted		Predicted Probability of Survey Response at Time 2	
			Raw Rate of Response (%) ²	P-value ³	Probability of Response (%)	P-value ⁴
Age in 96 ⁵	>=1SD less than Mean (i.e., <=41 years)	3,819	35	< .0001	42	< .0001
	(41–66 years)	13,593	53		52	
	>=1SD more than Mean (i.e., >=66 years)	3,896	66		62	
Gender	Male	7,819	53	0.44	50	< .0001
	Female	13,457	52		54	
Race/ Ethnicity	White	16,154	55	<.0001	54	< .01
	African American	711	45		44	
	Hispanic	1,808	43		51	
	Asian or Pacific Islander	1,510	48		51	
	Other	437	43		46	
Education (highest year completed)	1–12<High school	8,084	51	<.0001	50	< .01
	13–16 some college	5,819	51		52	
	17 > college graduate	7,441	55		55	
Household income	< \$30,000	4,286	52	NS	52	< .05
	>= \$30,000-<= \$60,000	10,274	53		53	
	> \$60,000	6,748	52		52	
Smoking habits	Never	10,663	52	<.0001	54	< .05
	Former	7,834	55		52	
	Current	2,811	48		48	
Comorbidity count	Mean Comorbidity Count - 1 SD (i.e., 0)	3,902	45	< .0001	49	< .0001
	Mean Comorbidity Count is 2 (SD = 2)	13,351	53		52	
	Mean Comorbidity Count + 1 SD (i.e., 4)	4,055	58		57	
Patient awareness of clinical problem	MD ever said you had high blood pressure problem—yes	7,080	57	<.0001	53	NS
	MD ever said you had high blood pressure problem—no	14,228	50		52	
	MD ever said you had high cholesterol problem—yes	7,006	56	<.0001	52	NS
	MD ever said you had high cholesterol problem—no	14,302	50		52	
Insurance	1 No copayment with visits	2,603	56	<.0001	53	NS
	2–5 Any copayment with visits	18,705	52		52	
Continuity with providers and services	No. of years in current health plan < 2	4,601	48	<.0001	49	< .05
	No. of years in current health plan =2	4,158	49		50	
	No. of years in current health plan 3–<6	6,567	53		54	
	No. of years in current health plan >=6 years	5,982	57		55	
Health status	General health now as compared with 1 year ago: Much better, no change, somewhat worse	20,913	53	<.0001	52	.001
	General health much worse now as compared with 1 year ago	395	39		44	
Current clinical status (i.e., self-report of outcomes)	Current high BP—yes	1,135	55	.0988	50	NS
	Current high BP—no	20,173	52		52	
	Current high cholesterol—yes	3,369	57	<.0001	54	NS
	Current high cholesterol—no	17,939	51		52	
Experience with health care system	Long delay experienced in medical care while waiting for approvals	1,762	46	<.0001	50	.02
	Short or no experience of delay while waiting for approvals	19,546	53		53	

function of 1996 SF-12 scores. We observed a similar two-percentage-point decrement in response rate in association with a one standard deviation decrease in 1996 SF-12 physical and mental scores.

Table 2: *Continued*

Receiving medical care from current doctor(s) ⁴	<=2 years	7,567	47	<.0001	51	<.0001
	3-<10 years	9,045	53		52	
	>=10 years	4,696	59		55	
Counseling services provided	Any specified counseling	15,330	53	<.0001	53	NS
	No specified counseling	5,978	49		51	
96 SF-12 - Physical ⁴	Mean SF-12-Physical - 1 SD (i.e., 39)	3,699	53	.004	49	<.0001
96 SF-12 - Physical ⁴	Mean SF-12-Physical is 48.7 (SD = 10)	16,506	53		53	
96 SF-12 - Physical ⁴	Mean SF-12-Physical + 1 SD (i.e., 58)	1,103	47		54	
96 SF-12 - Mental ⁴	Mean SF-12-Mental - 1 SD (i.e., 43)	3,365	46	<.0001	48	<.0001
96 SF-12 - Mental ⁴	Mean SF-12-Mental is 51.6 (SD = 9)	16,456	53		53	
96 SF-12 - Mental ⁴	Mean SF-12-Mental + 1 SD (i.e., 61)	1,487	56		55	
Rating MD ⁴	Mean MD Rating - 1 SD (i.e., 53)	3,305	44	<.0001	48	<.0001
Rating MD ⁴	Mean MD Rating is 73.1 (SD = 20)	14,031	53		52	
Rating MD ⁴	Mean MD Rating + 1 SD (i.e., 92.9)	3,972	57		55	
Rating Health Plan ⁴	Mean Health Plan Rating - 1 SD (i.e., 41)	2,695	45	<.0001	50	
Rating Health Plan ⁴	Mean Health Plan Rating is 55 (SD = 14)	15,636	53		52	
Rating Health Plan ⁴	Mean Health Plan Rating + 1 SD (i.e., 69)	2,977	56		55	.001
Rating Others ⁴	Mean Other Rating - 1 SD (i.e., 37)	3,103	43	<.0001	52	
Rating Others ⁴	Mean Other Rating is 58.4 (SD = 22)	14,688	53		52	NS
Rating Others ⁴	Mean Other Rating + 1 SD (i.e., 80)	3,517	56		52	
No. Visits to doctor past 12 months	Visits to doctor during past 12 months: 0-1	2,809	50	.002	51	
	Visits to doctor during past 12 months: >= 2	18,499	53		53	.028
Hospitalization during last 12 months	No hospitalizations during past 12 months	18,884	53	<.0001	53	<.0001
	Any hospitalizations during past 12 months	2,424	47		49	
No. visits outside health plan during last 12 months	Any visits outside health plan during past 12 months	3,719	50	.002	52	NS
	No visits outside health plan during past 12 months	17,589	53		52	
Used urgent care services during past 12 months	No urgent care visits during past 12 months	16,579	53	<.0001	53	
	Any urgent visits during past 12 months	4,729	49		51	.04
Organization type	Medical Group	12913	53	.035	53	
	IPA	6,793	51		51	
	PPO	1,602	52		49	.01

¹ N is provided as the number of patients who provided data in 1996 regardless of response in 1998.

² Raw rate of response is unadjusted.

³ The *p*-values associated with the raw rates of response are reported after using the technique of Huber to adjust for clustering of patients within physician groups.

⁴ The adjusted *p*-values are reported after using the technique of Huber to adjust for clustering of patients within physician groups.

⁵ This variable is entered into the model as a continuous variable. However, for exposition purposes, we present categorical data in the Response Category columns.

Response rates in 1998 were significantly lower for patients who were dissatisfied in 1996. For example, each one standard deviation point increase in 1996 satisfaction ratings of the physician was associated with 1998 adjusted response rate increases of two percentage points ($p < .0001$).

Considering multiple patient characteristics using the logistic regression model, we can compare adjusted 1998 response rates of two patients with identical demographic data. The 1998 adjusted response rate is 52 percent for a patient with the mean 1996 score on each of their SF-12 physical, SF-12-

mental, ratings of doctors, plans, and other satisfaction measures (using 0–100 point scales). This contrasts with a 1998 adjusted response rate of 64 percent for an idealized patient with identical demographics but with a higher score in 1996 (by one standard deviation for each of these measures).

Using a likelihood ratio chi square test, we tested the significance of the model without and with organizational variables and found a statistically significant difference when the organization dummies are included in the model ($p < .0001$). Table 2 shows response varied by medical organization with 1998 response rates lower in IPA and PPO than in medical groups ($p < .05$). The model was robust with stable patient-level coefficients and tests of significance when dummy variables were included for each of the medical organizations or when the organizations were identified as medical group, IPA, or PPO. We observed little variation in the C-statistic for the regression models when medical organization type (C-statistic = .65) or dummies for the medical organizations (.66) are included in the model as a supplement to patient characteristics (.65).

Using Weights to Adjust for Nonresponse Bias

We compared the weighted and unweighted regression models estimating quality of 1998 process scores based upon data from the 1998 survey data as a function of 46 patient characteristics reported in 1996. Specifically, we ran these comparisons for five separate types of process scores in 1998: an aggregate measure of 16 explicit process criteria pertinent to cognitive data gathering by providers; an aggregate measure of five explicit process measures pertinent to technical explicit process; an implicit patient-report rating of providers; rates of colorectal cancer screening for eligible patients; and a process measure pertinent to advising patients about the use of new medications. Comparison of the weighted and unweighted 1998 predicted process scores adjusted for 1996 patient characteristics show significant differences with $p < .0001$. Comparisons of the weighted and unweighted regression models showed differences of more than 10 percent of the value of the coefficients for more than half of the 46 coefficients across each of the models (not shown). For example, for technical explicit process we noted a greater than 10 percent change in the regression coefficients for 29 of the 46 (63 percent) variables included in the model predicting response to the follow-up survey including: gender, ethnicity, education, income, smoking habit, specified comorbidities, SF-12 physical, and use of services such as number of doctor visits.

This difference in coefficients indicates that the models for predicting process are different without and with weights. We also studied the distribution of standard deviations using the unweighted regression as the base model and the weighted regression as the comparison model. We found 13 coefficients from the weighted model were within one SD above the coefficients from the unweighted model, and an additional 14 coefficients from the weighted model were within one SD below the coefficients from the unweighted model. Two are more than one SD above and one is more than one SD below the coefficients from the unweighted model. Finally, we checked whether the ranking of the medical organizations with respect to process scores changed as a function of the application of the nonresponse weights. After comparing organization rank without and with weights for 42 organizations, we found between 5 and 9 organizations (depending upon the process measure being evaluated) varied their ranking. Half of the organizations improved their rank and half lowered their rank.

Stability of Patient Responses

Demographic data were very stable with time with a kappa score of .92 for gender, .82 for education, and .81 for ethnicity (Table 3). When asking a patient about the presence *ever* of a comorbidity we expect the patient report to either be stable, or with the passage of time to change the response from no to yes if the patient developed a new condition. In fact, we did observe this with more patients reporting the presence *ever* of each comorbid variable in 1998 than in 1996. Although one might expect a similar pattern with patient report of *ever* using specified medications and/or counseling, we observed a decrease with time in patient report of medication use *ever* and in patient report of provider counseling *ever* (4). Kappa across time were moderate for reports of comorbidity, and use of medication and were less good (.46 and .27) for reports of counseling.

Despite the high kappa score for gender, the 1998 survey data show that 4 percent of survey respondents specified a different gender in 1998 than in 1996. This includes 4 percent of patients from Core survey cohort and 3 percent of patients from the Chronic Condition survey cohort. Though the kappa is high at .92, in aggregate 4 of 100 patients providing longitudinal data probably represent an across-gender change in survey respondent from the baseline to the follow-up survey (assuming these gender switches only very rarely involve intended surgical gender switches) (Michel, Mormont, and Legros 2001).

Table 3: Stability of Patient Responses from 1996 to 1998 (N = 11,151 Patients for Four-Page Core Survey)

<i>Variable</i>	<i>Frequency</i>		<i>Concordant Responses</i>		<i>Discordant Responses</i>		<i>Frequency of Data Missing</i>		<i>Simple Kappa</i>
	1996	1998	Yes '96 Yes '98	No '96 No '98	Yes '96 No '98	No '96 Yes '98	1996	1998	
DEMOGRAPHICS									
Gender	64	63	61	35	2	2	1	0.1	.92
White	79	79	79	17	2	3	0	0	.81
African/Black	3	3	3	97	0	0			
Hispanic	7	7	6	92	1	1			
Asian	7	7	7	93	0	0			
Other	5	5	1	92	3	4			
Education <high school	7	7	5	92	1	2	0	5	.828
Education is some college	57	54	50	40	5	4			
Education is >= 4 year college graduate	37	39	35	58	3	4			
EVER COMORBIDITY									
Ever smoke	51	50	47	47	3	3	3	2	.887
Ever heart disease like angina	2	9	2	91	0	7	1	1	.267
Ever cancer	5	9	4	90	1	5	1	1	.534
Ever high blood pressure	30	37	26	59	4	11	1	1	.676
Ever diabetes	4	8	3	91	0	5	1	1	.551
Ever migraine	8	11	5	86	3	6	1	1	.512
Ever high blood pressure	37	38	31	57	5	6	1	2	.755
Ever high cholesterol	37	39	30	54	7	9	5	3	.656
EVER MEDICATIONS									
Ever use meds for hypertension	82	53	81	10	4	4	65	40	.670
Ever use meds for high cholesterol	36	28	34	50	4	11	66	38	.688
EVER COUNSELING									
Doctor ever tell you to eat less fat	81	44	57	13	24	6	65	3	.275
Doctor ever tell you to eat less salt due to hypertension	65	38	52	24	16	9	67	39	.469

DISCUSSION

During the last decade, there has been growing support for quality-based competition in the health care market to stimulate improvements in the quality of care (Brook and Appel 1973; Angell and Kassirer 1996; Dudley and Luft 2001; Kahn et al. 2002). Longitudinal analyses involving the collection of data from the same patients at two points in time has been heralded as one means to improve the reliability and validity of the measurement of relationships

between processes and outcomes of care. Empirical analyses using longitudinal data can provide opportunities for us to understand the scope and magnitude of threats to validity using longitudinal data for assessing the quality of care. We used a large study of patient report of demographics, process of care, and health status across 49 medical organizations in three states and two time periods to evaluate attrition in the number of patients and participating organizations, possible biases in follow-up 1998 survey response rates, effect of weights on Time 2 follow-up surveys, and the stability of patient responses with time.

Survey responses of 50 to 60 percent have for a long time been considered suboptimal (Groves and Couper 1998). The reality is that response rates, across all types of patient report surveys, have been declining over the past 10 years and are likely to continue to do so as health survey researchers compete with telemarketers and households with people under increasing time pressures (Kasprzyk et al. 1989). As others have demonstrated (Fowler et al. 2002; Hox and de Leeuw 2002; Dijkstra and Smit 2002; Singer 2002), we show that survey-fielding methods can influence response rates. We found a response rate of 63 percent for a nine-page survey, eight percentage points higher than for a four-page survey including only a subset of the items fielded on comparable patients with a chronic disease. This likely resulted from the intensified survey efforts associated with the fielding of the longer follow-up survey. Although intense fielding efforts can be implemented, achievement of higher response rates is very difficult and costly and many studies are not going to have resources to achieve higher response rates, particularly with follow-up surveys.

The high prevalence of suboptimal response rates across many settings in which health status and quality are assessed highlight the importance of considering the characteristics of patients who do not respond to follow-up surveys. We need to know more about how the exclusion of their data at follow-up influences assessments of care and outcomes and what can be done about it. Our analyses demonstrate substantial variations across physician groups in longitudinal survey response consistent with previously reported response bias by demographics as has been previously appreciated (Zaslavsky, Zaborski, and Cleary 2000; Groves and Couper 1998; Etter and Perenger 1997; Allen and Rogers 1997). However, our study also demonstrates the importance in predicting response to the follow-up survey of patient comorbidity, functional status, and satisfaction as reported in the baseline survey. Typically, quality of care reports (e.g., quality report cards) do not specify details of response rate at each round of data collection, characteristics

of nonresponders, efforts made to gain follow-up survey participation for traditional nonresponders, or analytic strategies (e.g., application of non-response weights). We recommend consideration and reporting of these specifications be regarded as standard for longitudinal data analyses as they apply to quality of care and health status reports to better understand how Time 2 data represent patients survey at Time 1 (Solberg et al. 2002; Zaslavsky, Zaborski, and Cleary 2002).

Key predictors of satisfaction, good process of care and outcomes at Time 2 are known to be satisfaction, good process of care, and health status at Time 1. Accordingly, quality reports could yield misleading estimates for patients or groups differentially participating in the follow-up survey if this were not accounted for by nonresponse weights (Zaslavsky, Zaborski, and Cleary 2002; Lasek et al. 1997). We used three analytic evaluations of the effect of the application of nonresponse weights on process scores. Each suggests that process scores adjusted with nonresponse weights differ from nonweighted scores. Even with three different efforts to quantify the effect of weighting on process scores, it is challenging to know how much difference the application of nonresponse weights make. The application of nonresponse weights seems appropriate since they can be helpful in redistributing the relative importance of each observation and this can result in the analytic finding being more representative of the original population. Conceptually, this seems useful as we try the best we can with available data, to represent the quality of process scores for patients involved in the baseline patient self-report survey.

Many factors influence the stability of patient reports over time. The mere passage of time allows increased opportunities for demographics to change (e.g., with the addition of educational opportunities, the affiliation with a different ethnic group or changes in annual income), for the occurrence of new comorbid conditions, for the use of new medications, or the receipt of new interventions (such as counseling by providers). For events that happened once, the passage of time also presents the problem of recall bias where patients over- or underestimate the occurrence of an event. To quantify these elements, we tested the stability of patient reports with time and found moderate or better agreement for reports of demographics, comorbidities, and use of medications, but less good agreement for interventions that may have occurred remotely in time. This serves as a reminder to researchers to minimize key analyses that rely upon patients' recall of remote events, particularly if they may have been isolated, individual interventions rather than continuous events.

We were surprised to find 497 (4 percent of 13,438) patients with longitudinal data who clearly indicated a different gender in their 1998 survey as compared with 1996. With longitudinal data analysis it is often assumed, after targeting a household and a patient within that household, that the patient who responded to the survey at Time 1 will also be the patient who responds to the longitudinal survey at Time 2. We have documented with our report of gender switch that this is clearly not always the case. Based upon careful scrutiny of all available data, we have been able to isolate examples where it appears that one individual completed the survey at baseline and another household member of the opposite gender completed the survey at Time 2. It is not known how often the identity of a survey respondent switches such that a different respondent completes the Time 1 and Time 2 surveys. Although this is expected when surveys invite proxy respondents, in this and other surveys the intent is to survey (as we explained in the survey instructions) *only* the originally sampled individual.

We have identified identity switch after noting a self-report discrepancy in gender across longitudinal surveys. Other variables may need to be examined to more fully understand the magnitude of identity switch in longitudinal surveys even after verification that the survey was completed within the targeted household. The problem could arise if the respondent did not realize the follow-up survey was intended for the household member who completed the baseline survey. Alternatively, a household member (e.g., spouse) could attempt to complete the 1998 survey on behalf of their (ill) spouse whom they know is the target for the survey. During the completion of the survey, they may erroneously switch to complete the survey on behalf of themselves.

We consider apparent gender switches as a special case of stability in the identity of the respondent. We do not know how often other forms of patient switch occur in longitudinal surveys. Because of this switch problem, we recommend that several items be placed in both the baseline and follow-up survey to aid in the verification that the intended (same) patient is completing all components of the longitudinal survey. Often researchers try to avoid repeating collecting data about fixed patient characteristics (e.g., gender) across longitudinal surveys to minimize the respondent burden. However, we have identified a reason to enrich, rather than to minimize, the number of fixed patient characteristics that should be collected across surveys from the same patient. Although respondents from the same household may have value as two unique respondents in separate cross-sectional analyses of each year, in most instances they should be excluded from longitudinal analyses unless

proxy responses are considered valid. Although biases in survey research have been known for years, the science concerning the methodological challenges associated with longitudinal data survey analysis pertinent to quality of care and health status studies is still unfolding. We have presented empirical data to document the scope of the problems and raise discussions regarding options for responding to these challenges.

We have made some recommendations. Most importantly we call for more public disclosure of the magnitude of these challenges across reports of quality of care and health status so that we can gauge the validity of quality of care measurements. With increasing emphasis in the United States on competition in quality in the health care market, it is important that we identify and respond to methodological challenges that might threaten the validity of quality of care analyses.

REFERENCES

- Allen, H. M., and W. H. Rogers. 1997. "The Consumer Health Plan Value Survey: Round Two." *Health Affairs* 16 (4): 156–66.
- Angell, M., and J. P. Kassirer. 1996. "Quality and the Medical Marketplace: Following Elephants." Editorial. *New England Journal of Medicine* 335 (12): 883.
- Blumenthal, D. 1996. "Part 1: Quality of Care: What Is It?" *New England Journal of Medicine* 335 (12): 891–4.
- Brook, R. H., and F. A. Appel. 1973. "Quality-of-Care Assessment: Choosing a Method for Peer Review." *New England Journal of Medicine* 288 (25): 1323–29.
- Brook, R. H., E. A. McGlynn, and P. D. Cleary. 1996. "Quality of Health Care. Part 2: Measuring Quality of Care." *New England Journal of Medicine* 335 (13): 966–70.
- Cleary, P. D. 1999. "The Increasing Importance of Patient Surveys. Now That Sound Methods Exist, Patient Surveys Can Facilitate Improvement." *British Medical Journal* 319 (7212): 720–1.
- Cleary, P. D., J. Lubalin, R. D. Hays, P. F. Short, S. Edgman-Levitan, and S. Sheridan. 1998. "Debating Survey Approaches." 17 (1): 265–8.
- Damberg, C., and L. Bloomfield. 1997. "Physician Value Check Survey Final Report." San Francisco: Pacific Business Group on Health.
- Dijkstra, W., and J. H. Smit. 2002. "Persuading Reluctant Recipients in Telephone Surveys." In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. Little, pp. 121–34. New York: Wiley.
- Donabedian, A. 1968. "Promoting Quality through Evaluating the Process of Patient Care." *Medical Care* 6 (3): 181–202.
- . 1988. "The Quality of Care: How Can It Be Assessed?" *Archives of Pathology and Laboratory Medicine* 121 (11): 1145–50.
- Dudley, R. A., and H. S. Luft. 2001. "Managed Care in Transition." *New England Journal of Medicine* 344 (14): 1087–93.

- Etter, J., and T. Perenger. 1997. "Analysis of Nonresponse Bias in a Mailed Survey." *Journal of Clinical Epidemiology* 50 (10): 1123–8.
- Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*, 2d ed. New York: Wiley.
- Fowler, F. J. Jr., P. M. Gallagher, V. L. Stringfellow, A. M. Zaslavsky, J. W. Thompson, and P. D. Cleary. 2002. "Using Telephone Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members." *Medical Care Research Review* 40 (3): 190–200.
- Fremont, A. M., P. D. Cleary, J. L. Hargraves, R. M. Rowe, N. B. Jacobson, and J. Z. Ayanian. 2001. "Patient-Centered Processes of Care and Long-Term Outcomes of Myocardial Infarction." *Journal of General Internal Medicine* 16 (12): 800–8.
- Groves, R. M., and M. P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.
- Hays, R. D., J. A. Shaul, V. S. Williams, J. S. Lubalin, L. D. Harris-Kojetin, S. F. Sweeney, and P. D. Cleary. 1999. "Psychometric Properties of the CAHPS 1.0 Survey Measures. Consumer Assessment of Health Plan Study." *Medical Care* 37 (3, supplement): MS22–31.
- Hosmer, D. W. Jr., and S. Lemeshow. 1989. *Applied Logistic Regression*. New York: Wiley.
- Hox, J., and E. de Leeuw. 2002. "The Influence of Interviewer's Attitude and Behavior on Household Survey Nonresponse: An International Comparison." In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. Little, pp. 103–20. New York: Wiley.
- Kahn, K. L., J. L. Malin, J. Adams, and P. A. Ganz. 2002. "Developing a Reliable, Valid, and Feasible Plan for Quality-of-Care Measurement for Cancer: How Should We Measure?" *Medical Care* 40 (6, supplement 3): 73–85.
- Kasprzyk, D., G. J. Duncan, G. Kalton, and M. P. Singh. 1989. *Panel Surveys*. New York: Wiley.
- Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–74.
- Lasek, R. J., W. Barkley, D. L. Harper, and G. E. Rosenthal. 1997. "An Evaluation of the Impact of Nonresponse Bias on Patient Satisfaction Surveys." *Medical Care Research Review* 35 (6): 646–52.
- Little, R. J. A., and D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.
- Michel, A., C. Mormont, and J. J. Legros. 2001. "A Psycho-Endocrinological Overview of Transsexualism." *European Journal of Endocrinology* 145 (4): 365–76.
- Miller, R. H., and H. S. Luft. 1997. "Does Managed Care Lead to a Better or Worse Quality of Care?" *Health Affairs* 16 (5): 7–25.
- Neter, J., W. Wasserman, and M. H. Kutner. 1985. *Applied Linear Statistical Models*, 2d ed. Homewood, IL: Irwin.
- Pacific Business Group on Health. 1989. "Pacific Business Group on Health Online" [accessed on September 6, 2002]. Available at <http://www.pbgh.org>.
- Singer, E. 2002. "The Use of Incentives to Reduce Nonresponse in Household Surveys." In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. Little, pp. 163–78. New York: Wiley.

- Solberg, L. I., M. B. Plane, R. L. Brown, G. Underbakke, and P. E. McBride. 2002. "Nonresponse Bias: Does It Affect Measurement of Clinician Behavior?" *Medical Care Research Review* 40 (4): 347-52.
- Ware, J. Jr., M. Kosinski, and S. D. Keller. 1996. "A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity." *Medical Care Research Review* 34 (3): 220-33.
- White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817-38.
- Zaslavsky, A. M., L. Zaborski, and P. D. Cleary. 2000. "Does the Effect of Respondent Characteristics on Consumer Assessments Vary Across Health Plans?" *Medical Care Research Review* 57 (3): 379-94.
- . 2002. "Factors Affecting Response Rates to the Consumer Assessment of Health Plans Study Survey." *Medical Care Research Review* 40 (6): 485-99.