How Valid Are Medical Records and Patient Questionnaires for Physician Profiling and Health Services Research?: A Comparison with Direct Observation of Patient Visits

Kurt C. Stange; Stephen J. Zyzanski; Tracy Fedirko Smith; Robert Kelly; Doreen M. Langa; Susan A. Flocke; Carlos R. Jaén

Medical Care, Vol. 36, No. 6. (Jun., 1998), pp. 851-867.

Stable URL:
http://links.jstor.org/sici?sici=0025-7079%28199806%2936%3A6%3C851%3AHVAMRA%3E2.0.CO%3B2-1

Medical Care is currently published by Lippincott Williams & Wilkins.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/lww.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Wed Apr 25 20:56:59 2007

MEDICAL CARE
Volume 36, Number 6, pp 851–867
©1998 Lippincott–Raven Publishers

How Valid Are Medical Records and Patient Questionnaires for Physician Profiling and Health Services Research?

A Comparison With Direct Observation of Patient Visits

KURT C. STANGE, MD, PHD,*†‡§ STEPHEN J. ZYZANSKI, PHD,*§ TRACY FEDIRKO SMITH, PHD†
ROBERT KELLY, MD, MS,*§‖ DOREEN M. LANGA, BA,*†
SUSAN A. FLOCKE, PHD,*‡§ AND CARLOS R. JAÉN, MD, PHD¶

OBJECTIVES. This study was designed to determine the optimal nonobservational method of measuring the delivery of outpatient medical services.

METHODS. As part of a multimethod study of the content of primary care practice, research nurses directly observed consecutive patient visits to 138 practicing family physicians. Data on services delivered were collected using a direct observation checklist, medical record review, and patient exit questionnaires. For each medical service, the sensitivity, specificity, and Kappa statistic were calculated for medical record review and patient exit questionnaires compared with direct observation. Interrater reliability among eight research nurses was calculated using the Kappa statistic for a separate sample of videotaped visits and medical records.

RESULTS. Visits by 4,454 patients were observed. Exit questionnaires were returned by 74% of patients. Research nurse interrater reliabilities were generally high. The specificity of both the medical record and the patient exit questionnaire was high for most services. The sensitivity of the medical record was low for measuring health habit counseling and moderate for physical examination, laboratory testing, and immunization. The patient exit questionnaire showed moderate to high sensitivity for health habit counseling and immunization and variable sensitivity for physical examination and laboratory services.

CONCLUSIONS. The validity of the medical record and patient questionnaire for measuring delivery of different health services varied with the service. This report can be used to choose the optimal nonobservational method of measuring the delivery of specific ambulatory medical services for research and physician profiling and to interpret existing health services research studies using these common measures.

Key words: physician profiling; medical record review; survey research; primary care; health services research; research methodology. (Med Care 1998;36:851–867)

Valid data on the delivery of outpatient medical care, particularly primary care, are critically important to practitioners, researchers, and policy makers to guide practice, research measurement, and health policy decisions.[1,2] In addition, valid data on outpatient medical and preventive service

*From the Department of Family Medicine, Case Western Reserve University, Cleveland, Ohio.

†From the Department of Sociology, Case Western Reserve University, Cleveland, Ohio.

‡From the Department of Epidemiology & Biostatistics, Case Western Reserve University, Cleveland, Ohio.

§From the Ireland Cancer Center at Case Western Reserve University and University Hospitals of Cleveland, Cleveland, Ohio.

‖From the Department of Family Medicine, MetroHealth Medical Center, Cleveland, Ohio.

¶From the Center for Urban Research in Primary Care, and Departments of Family Medicine and Social & Preventive Medicine, State University of New York at Buffalo.

Supported by a grant from the National Cancer Institute (1RO1 CA 60862) and by a Robert Wood Johnson Generalist Physician Faculty Scholar Award to Dr. Stange.

Address correspondence to: Kurt C. Stange, PhD, Department of Family Medicine, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106.

delivery increasingly are sought by third-party payers and health care consumers to assess the performance of health care plans and individual physicians.[3–6]

Despite the need for accurate data, the commonly used measures of medical record review and patient survey have not been widely validated, in part because of the lack of a gold standard. Direct observation has been proposed as such a standard, but it has not been used previously because of its cost and potential intrusiveness.[7,8] Previous attempts to measure the rate of primary care service delivery typically have used claims data, physician self-report, medical record review, and patient surveys. Each of these measurement methods have different limitations and potential sources of error.

Claims data, although potentially limited in detail and accuracy for some data elements, increasingly are used to characterize delivery of both inpatient and outpatient medical care.[9–16] Claims data have a number of practical advantages and have been shown to provide useful profiles of many aspects of care.[11,14,16–18]

> To the extent that such files contain information on the use and content of health services, they may be useful in medical effectiveness research. However, they frequently lack critical clinical and patient-level information. Another limiting factor is the absence of validation of many data elements.[19]

Many studies have used physician self-report of usual practices to estimate the rate of delivery of various services, particularly preventive services. This literature has shown that physicians tend to overreport their delivery of preventive services when compared with medical record review or with patient survey.[20–23]

Medical record review has been used to measure many aspects of outpatient care, particularly quality and process outcomes, and has the advantage of easy accessibility.[24] Comparisons of medical record review with tape-recorded patient visits and videotaped visits have shown highly variable rates of concordance, depending on the particular service being assessed.[25–27] In general, the medical record tends to underreport delivery of services compared to review of recorded visits.

Patient surveys of receipt of various medical services have been compared with medical record review radiology records and physician reports.[28–37] These studies have shown moderately wide variability in the degree of concordance of patient report of service delivery, depending on the particular service.

A few studies have attempted to simultaneously compare multiple methods of measuring the rate of service delivery.[23,31,38] In a study comparing the concordance between physician interviews, patient reports, medical records, and videotaped encounters for detection of medication regimens prescribed for patients with chronic obstructive pulmonary disease, all four methods were in agreement only 36% of the time.[27] A study of multiple methods for measuring cancer screening rates of family physicians found chart audits and patient surveys to be highly correlated, but physician self-report tended to overreport delivery of cancer screening services.[23] Disagreement between patients and physicians on what occurs during an encounter has been found to vary widely with different services.[30]

The limited data on the accuracy of many measures of the delivery of outpatient medical services leaves researchers, administrators, health care purchasers, and readers of the medical literature in a quandary. Researchers and administrators need to know the most valid method for measuring the delivery of different services. Users of the medical literature need to know the accuracy of the measures being used to interpret and apply the findings of studies of physician practices. Multimethod research that incorporates different measures of service delivery in real world settings has been proposed as a method for understanding and ultimately improving the practice of medicine in primary care settings.[20,39–42]

The current study was undertaken to examine the interrater reliability and validity of the commonly used and relatively inexpensive medical record review and patient questionnaire methods compared with a gold standard of direct observation of the outpatient visit. The focus of this study was on comparing different measures of delivery of patient services, particularly preventive services, during individual patient visits to primary care physicians.

Methods

Sites and Subjects

In the summer of 1994, family physician members of the Ohio Academy of Family Physicians who practice within a 50-mile radius of Cleveland and Youngstown were invited to participate in a

study of the content of family practice and to become members of a practice-based network designed to serve as a laboratory for research on primary care. Physicians not practicing in family practice settings and full-time academic physicians were excluded, with the exception of 30 members of the faculty of the Northeast Ohio Universities Colleges of Medicine (NEOUCOM), who practice in community sites that function as training practices for family practice residents.

Each participating physician was visited by one of four teams of two research nurses while providing outpatient care during 2 days between October, 1994 and August, 1995. Each physician's observation days were separated by an average of 4 months to maximize variation in seasonal reasons for patient visits. The study coordinator scheduled representative patient care days and asked the office representative to schedule patients in their customary fashion. Consecutive patients seen on observation days were informed about the study in the waiting room before meeting with their physician and were enrolled if they gave consented.

Data Collection Procedures

Multiple strategies were used to minimize the possibility of a Hawthorne effect, that is, the possibility that the presence of a nurse-observer would alter the phenomena being studied. Physicians were told to follow their usual scheduling and patient care procedures. To avoid biasing their behavior, physicians were informed that the study would use multiple methods to examine the content of the ambulatory patient visit, but no specific hypotheses or study goals were shared with the physicians, office staff, or patients. In addition, the observation of consecutive patients made it impossible for physicians to spend more time or to provide more services than their usual routine without severely compromising their ability to stay on schedule. The research nurses instructed the physicians to ignore them during the observed visit, so that they could be "like a fly on the wall." Patients likewise were told to ignore the nurses' presence. The research nurses observed the visits from the least obtrusive corner of the room, from a position that avoided eye contact with either the physician or the patient. Because the presence of a nurse is a normal occurrence during many outpatient visits to physicians, the majority of patients and physicians reported that the presence of the nurse-observer did not change their behavior during the visits observed for the study.

Before the beginning of data collection, the eight research nurses were trained extensively in the use of all research instruments. This 7-week training included their initial involvement in discussions of the theoretical basis, measurement intent, and final refinement of the measures. Training involved practice with data collection, initially using videotaped medical visits and medical records from family practices not involved in the larger study. Later in the training, the research nurses practiced the entire data collection protocol at family practice office sites that were not participating in the larger study.

During the course of the data collection, the research nurses met for one half day every other week to discuss any problems with the data collection procedures at the study sites and to simultaneously but independently code videotaped patient visits and medical records from sites not participating in the larger study. Data from these 16 videotaped patient visits and copies of 19 medical records of different patients were used to assess interrater reliability.

The research nurses collected data on the content and context of the office visit, using the following measures:

1. Direct observation of the patient visit using a modified[42] version of the Davis Observation code. The Davis Observation code categorizes time use during every 15-second interval of each patient visit into 20 different behavioral categories;[27]
2. A direct observation checklist of services delivered during the patient visit;
3. A patient exit questionnaire;
4. Medical record review;
5. A practice environment checklist;
6. Billing data on CPT[44] and ICD9-CM diagnoses;
7. A physician questionnaire; and
8. Ethnographic field notes.[27,43,45]

Measures were linked by specific confidential identification numbers for the patient, physician, and date.

Each physician was visited by a team of two research nurses during two patient care observation days and 2 days during which medical records of observed patients were abstracted. During the

two patient care observation days, one research nurse accompanied the physician during all visits by consenting patients. This nurse recorded her direct observation of the content of the visit using the Davis Observation Code and direct observation checklist. The other research nurse obtained consent from patients in the waiting room and gave participating patients a questionnaire at the end of their visit. Patients were asked to complete the questionnaire in the waiting room and to give it to the research nurse to be placed in a confidential envelope. If they were unable to stay, patients were instructed to complete the questionnaire as soon as possible after the visit and to mail it to the study research office in a confidential prepaid envelope. Parents or guardians of children younger than 13 years old were asked to complete the questionnaire for their children. Patients 13 to 17 years old were given the option of completing the questionnaire themselves or with help from a parent or guardian. All patients were offered help in clarifying questionnaire items by the research nurse or by calling the study office on a toll-free number. Patients were sent a reminder postcard within 1 week of their visit. Nonrespondents were sent a second questionnaire within 1 month of their visit.

Medical record review data were obtained by the research nurses on a day subsequent to each observation day. Seventy-nine percent of medical record reviews were performed by a research nurse who had not observed the actual visit. For medical records that were reviewed by the same nurse who had observed the patient visit, nurses were trained to review the medical record independently from any recollection they may have had of the observed visit. Because of the many observed visits and medical record reviews that intervened between the observation and medical record review days at most offices, the research nurses reported that they had little recollection of the specific visits for which they were reviewing medical records, and they were able to review and code the medical records independently.

The practice environment checklist about multiple aspects of the practice organization was completed by the research nurse teams based on direct observation and interview of key office informants, such as the office manager, during both the patient care observation and medical record review days. Billing data on the observed visits were obtained from the responsible office personnel after the observation day. Ethnographic field notes were based on brief "field jottings" and were dictated by the research nurses immediately after each visit to the practice.[46] The research nurse teams often shared car rides home and dictated the notes together on hand-held dictaphones, often reflecting on each other's observations. Two thousand pages of text thus were dictated to critique the study methods and to provide additional insights into the office culture and factors that were not measured adequately by the quantitative instruments.

After the first round of data collection, in which each physician was visited once, the research instruments were expanded based on the early ethnographic findings and on input from the entire team. Physician questionnaires were distributed only after each physician had completed the second observation day to avoid biasing their behavior during the study.

Measures

For this report, comparisons involve the direct observation checklist, patient exit questionnaire, and medical record review instrument. Each of these instruments contained similar measures of whether or not particular services were delivered during the observed visit, in addition to other data items. Items designed to assess the delivery of services during outpatient visits formed the basis for comparisons of the different vantage points on measurement: direct observation, medical record review, and patient report of the delivery of specific services. These items measure different domains of services, including physician history taking, physical examination, health habit counseling, diagnostic or screening testing, immunization, referral, and reason for visit.

For the direct observation checklist, the research nurse observing the office visit checked a box for each service that was observed to have been performed or ordered during each physician–patient encounter. In addition, for some services the research nurse indicated whether or not the service had been performed in response to a patient's symptoms or chronic medical condition.

Similarly, for the medical record review, the research nurses indicated whether or not particular services were noted on the chart note for the observed visit. Medical record data also were collected on delivery of services during the past year and other specific time intervals for certain serv-

ices. Additional data were collected on a number of factors, including demographics, number of chronic illnesses and medications, number of years as a patient of the practice, number of visits in the past year, and presence of specific illnesses.

The patient exit questionnaire asked a wide variety of questions, including whether or not a list of services was provided during the observed office visit. Demographic questions ascertained the patients' age, sex, race, educational level, and marital status. Health status was measured with a modified[42] five-item version (alpha = 0.81) of the MOS 6-item Health Survey.[43] These items used a 5-point Likert response format to ask about global health status and health limitations in everyday physical activities, emotional problems, limitations in work because of physical or emotional problems, and bodily pain during the 4 weeks before the visit.

Reason for visit was measured with the typology from the National Ambulatory Care Survey and was ascertained by direct observation, medical record review, and patient exit questionnaire.[47,48] For the purposes of this report, reason for visit was collapsed into the broad categories of acute illness, chronic illness, or well-care visit. Current Procedure Technology (CPT) codes were assigned by the research nurses to each visit based on direct observation and medical record review using established guidelines.[44] During the medical record review, the nurses also rated the components of the visit that led to the assignment of a CPT code by American Medical Association guidelines: extent of history, complexity of medical decision-making, extent of examination, and nature of presenting problem.

Analyses

The representativeness of the physician sample was calculated by comparing the demographics of participating physicians with those of members of the American Academy of Family Physicians (AAFP).[49]

Several methods were used to assess the representativeness of the patient sample. First, characteristics of participating patients and visits were compared with similar data obtained from the National Ambulatory Medical Care Survey.[50] In addition, the research nurses recorded observable characteristics of patients who declined to participate, including any reason that patients gave for declining. Finally, a subsample of 12 of the participating physicians reviewed the medical records of their patients who declined participation. For each patient, the physician recorded the patients' demographics and number of years as a patient of the practice. The physicians also noted their belief about why the patient declined to participate based on the physician's knowledge of the patient and the characteristics of the patient's visit on the observation day. Among patients who agreed to have their outpatient visits observed, the characteristics of patients who returned questionnaires were compared with nonreturners using the observation and medical record data. T tests were used for comparisons involving continuous variables, the Wilcoxon Rank Sum test was calculated for highly skewed ordinal variables, and χ^2 tests were calculated for ordinal variables.

Initial descriptive analyses used data from multiple sources. The physician sample was described based on physician self-report on the physician questionnaire. Data to describe the patient sample were obtained from the patient exit questionnaire. Additional data on the patient's type of insurance were obtained from billing data collected by each office site for each patient visit. Visit characteristics were described from direct observation data, including the research nurses' assessment of the reason for visit and the length of the visit, as timed during the collection of the Davis Observation Code data on time that the physician spent in direct patient contact. Descriptive data from the practice environment checklist were used to characterize the office settings.

Analyses were carried out using the multiple rater kappa statistic to establish the research nurse interrater reliabilities.[51] These analyses used data from the research nurses' review of the 16 videotaped patient visits and 19 different copied medical records during 2-week intervals throughout the study. The multiple rater kappa coefficient was calculated for delivery of individual services using direct observation and chart audit form as a conservative measure of the generalized agreement among the team of eight research nurses rating the "presence" or "absence" of delivery of services.[52–54] The multiple rater kappa calculates weighted averages of the pairwise proportions of observed agreement and pairwise proportions of agreement expected by chance.[55] Each pairwise estimate is weighted by the number of subjects rated by a particular judge pair. Kappa coefficients between 0.81 to 1.00 are considered almost perfect agreement, those between 0.61 to 0.80 are

considered very high agreement, those between 0.41 to 0.60 are considered moderate agreement, and coefficients between 0.21 to 0.40 are considered only fair agreement.[56] Only services that were present on at least two videotaped patient encounters or medical records are presented.

Analyses of the concordance of direct observation with the medical record review and with the patient exit questionnaire were used to calculate sensitivity, specificity, and kappa coefficients.[57,58] These analyses compared the reference standard of direct observation with the medical record review and the patient exit questionnaire. Analyses were restricted to eligible patient groups. For example, data on provision of Pap smears are presented only for female patients who were at least 13 years old. To improve the stability of the estimates of sensitivity, specificity, and kappa, data are presented only for services with at least 30 observations. A sample size of 30 was selected as a lower bound for using the large sample normal approximation to estimate confidence intervals for proportions. With a sample size of 30, a two-sided 95% confidence interval for a single proportion would extend approximately 0.15% from the observed proportion for expected proportions between 0.05 and 0.95.[56]

Results

Based on power calculations to test the main hypotheses of the overall study, a sample size of 120 physicians had been targeted. Of the 531 physicians invited, 138 volunteered to participate. Table 1 describes characteristics of the physicians, patients, practices, and outpatient visits observed. Physicians were demographically similar to active practicing members of the American Academy of Family Physicians (AAFP) in age (AAFP mean = 45 years), percentage in rural locations (AAFP = 25%), and number of patients seen per week (AAFP mean = 103).[49] The study sample represented recent demographic trends in family physicians in that participating physicians were more likely to be female (AAFP = 21%) and residency trained (AAFP = 73%). Patient characteristics were similar to characteristics of patients coming to see family physicians participating in the 1992 National Ambulatory Care Survey (NAMCS) in age (NAMCS = 38 years) and the percentage of females (NAMCS = 60%).[50] Patients in our study were slightly more

TABLE 1. Characteristics of the Sample

	% or Mean ± SD
Physician characteristics (n = 128)	
Age	43 (7.6)
Sex (% female)	28
Marital status (% married)	88
Years in current practice	11 (7.8)
Type of practice (% solo)	22
Year graduated from medical school	1979 (7.6)
Family practice residency completed (% yes)	89
Number of patients seen per week	104 (63)
Patient characteristics (n = 3,287)	
Age	41 (24.2)
Sex (% female)	61
Marital status (% married)	54
Self-reported ethnicity (% white)	87
Self-reported education level (%)	
Some high school	30
High school graduate	28
Some college or associate degree	23
College graduate	19
5-item self-reported health status (1 = poor, 5 = excellent)	3.8 (0.8)
Years as a member of the observed practice	
1st visit (%)	8
< 1 (%)	13
1–3 (%)	27
4–6 (%)	20
7–10 (%)	12
10+ (%)	21
Insurance type (%)	
Managed care	36
Indemnity (fee for service)	20
Medicaid	7
Medicare	23
Other	7
None	7
Visit characteristics (n = 4454)	
Major reason for visit (%)	
Acute illness	58
Chronic illness	23
Well care	12
Other	7
Length of visit (min)	10 (5.8)
Practice characteristics (n = 84)	
Type of practice (%)	
Single specialty group	53
Solo	30
Multiple specialty group	8
Residency training practice	6
Health maintenance organization	2

likely to have a managed care type of insurance (NAMCS = 21%). Our directly observed length of visit was shorter (10 minutes versus 15 minutes) than reported by physicians in the NAMCS. This most likely represents physicians in the NAMCS reporting patient visit-related time not spent in direct patient contact.

Of 4,994 patients seen by their family physicians during the two observation days with each physician, 4,454, or 89%, agreed to have their visits observed. Patients who declined to participate were estimated by the research nurses to be slightly older (mean estimated age of refusers = 45 years versus 41 years for participants, $P = 0.01$). A similar percentage of refusers were female (60% versus 62%, $P = 0.42$) and white (87% for both refusers and participants, $P = 0.79$). Among the 37% of patients who gave a reason for not participating, the most common reasons were privacy (13% of total of all refusers), a "personal" reason for the visit (8%), gynecologic examination (5%), feeling too sick (2%), and being a new patient and not yet comfortable with the doctor (2%). In addition, 11 patients (2% of nonparticipants) were not enrolled because they were minors who did not have a parent or guardian present to give consent, and four patients (1% of nonparticipants) were not enrolled because language barriers inhibited informed consent.

Twelve participating physicians provided additional information on their patients who declined to participate. This subsample of 54 patients was older than participating patients ($P < 0.001$), but similar in sex, race, and number of years as a patient. The physician attribution of the patients' reason for nonparticipation revealed patient concerns about privacy as the most common reason (39%), followed by anxiety (11%), embarrassment (7%), gynecologic reason for visit (7%), and shyness (6%).

Medical records were available for review for 4,432 (99.5%) of the 4,454 observed visits. Patient exit questionnaires were returned by 3,283 patients, for a 74% response rate. Patients who returned questionnaires were more likely than nonreturners to be older, female, white, married, and to have a greater number of chronic illnesses and a longer relationship with the practice (all $P < 0.01$). Smokers and patients being seen for an acute illness were less likely to return exit questionnaires ($P < 0.01$).

Table 2 shows the interrater reliability for the eight research nurses' measurement of various services during direct observation of training videotapes of patient visits and for medical record review. The kappa values were generally in the high range, but were lower for services that were not overt to an observer or for which considerable amounts of research nurse judgment was involved.[56] For example, interrater reliability was high for all physical examination items and for most laboratory testing and counseling items. Interrater reliability was moderate, however, for direct observation of physician ordering of a chemistry panel, which might not have been explicitly stated by the physician during the visit, and for the research nurses' judgment of whether family or social history was obtained. Interrater reliabilities were lower but still acceptable for the nurses' rating of the extent of history taking, extent of physical examination, and complexity of decision making by the physician. The concordance of nurses' assessment as to whether a service was performed because of symptoms or illness was very high, with the exception of attributions about the reason for taking a family or social history, which required more judgment.

The concordance between medical record review and direct observation is shown in Table 3. The concordance between patient exit questionnaire and direct observation is depicted in Table 4. The degree of concordance varied based on the specific medical service. For most services, the specificity was high for both the medical record and the patient exit questionnaire, in part because most services rarely are performed during an individual visit.

The sensitivity of the medical record was low for ascertaining whether health habit counseling was performed, with the highest sensitivities being for performance of alcohol and tobacco histories and discussion of estrogen replacement therapy or contraception. The sensitivity of the medical record was high for most physical examination items and moderate to high for most laboratory testing and immunization items. The sensitivity of the medical record for documentation of any referrals was only 58% and was even lower for referrals to nonphysicians. The nurses' judgment of the reason for visit from the medical record was moderately highly concordant with the reason for visit as judged by direct observation for the major categories of acute illness, chronic illness, and well care.

The patient exit questionnaire (Table 4) showed moderate to high sensitivity for most health habit counseling items. The sensitivity of patient report

TABLE 2. Interrater Reliability (Among Eight Research Nurses)

	Direct Observation (128 observations on 16 videotaped patient visits)	Number of Videotapes on Which Service Was Observed	Chart Audit (152 reviews of 19 medical records)	Number of Charts on Which Service Was Recorded
Health habit counseling				
Exercise advice	0.70	5	1.00	2
Smoking cessation	0.77	7	—	0
Other diet counseling	0.65	3	0.53	2
Physical examination				
Lung	1.00	9	1.00	6
Abdomen	1.00	6	0.97	6
Height/length	1.00	4	1.00	3
Weight	1.00	4	0.91	14
Blood pressure	1.00	5	1.00	11
Head/neck	0.85	12	0.91	9
Heart	0.85	8	0.92	4
Extremities	0.81	7	0.88	6
Skin examination	0.76	4	0.92	4
Back	0.75	3	1.00	1
Neurological examination	0.74	5	1.00	9
Rectum	—	0	0.92	2
Bimanual pelvic examination	—	0	0.71	2
Lab testing				
Urinalysis	1.00	2	—	0
Blood glucose	0.85	1	1.00	3
Cholesterol	0.79	2	1.00	1
EKG	—	0	1.00	2
Hematocrit/hemoglobin	—	0	1.00	2
History taking				
Tobacco history	0.83	5	—	0
Family history	0.41	4	1.00	1
Social history	0.39	13	0.65	8
Patient demographic information				
Sex	1.00	15	0.93	17
Marital status	NA		0.96	12
Race				
Hispanic	1.00	1		
Black	0.89	5		
White	0.82	9	NA	

(*Continues*)

of having been advised about exercise, birth control, dental health, passive tobacco exposure, and accident prevention was less than 50%. The sensitivity of the patient questionnaire was high for some physical examination and screening items that would be overt and memorable for a patient, such as breast and pelvic examinations and Pap smears. An exception was the low rate of report-

TABLE 2. (Continued)

	Direct Observation (128 observations on 16 videotaped patient visits)	Number of Videotapes on Which Service Was Observed	Chart Audit (152 reviews of 19 medical records)	Number of Charts on Which Service Was Recorded
General information: chart				
Previous myocardial infarction or stroke			1.00	17
Diabetes			1.00	18
Years with practice			1.00	4
Smoker			0.85	17
Number of nurses visits in last year			0.83	18
Drugs prescribed			0.82	18
Centralized database used on chart			0.78	18
Old records present on chart			0.68	18
Diagnosis of depression			0.67	18
Thickness of chart (cm)			0.64	18
Legibility			0.62	18
Number of physician visits in last year			0.62	6
Prevention flow sheet on chart			0.50	18
Other flow sheet on chart			0.57	18
General information: direct observation				
Other family present during visit	0.93	10		
Problem of other family members discussed	0.65	2		
Referral information				
Nonphysician out of office	0.78	15	0.41	18
Nonphysician in-office	0.72	15	0.85	18
Another physician	0.57	16	0.80	18
Any referral	0.76	15	0.90	18
Reason for visit	0.81	15	0.85	18
Category of service	0.88	15	0.97	18
Extent of history*	NA		0.34	9
Complexity of medical decision*	NA		0.39	17
Extent of examination*	NA		0.43	18
Nature of presenting problem*	NA		0.42	18
CPT codes				
New patient	1.00	2	0.25	6
Established patient	0.79	14	0.36	17

(Continues)

ing of testicular examination by male patients, for which similar findings were reported by Brown and Adams.[59] The sensitivity of patient report for less overt or memorable examination items, such as back, skin, or head and neck examination was lower. Some tests, such as electrocardiograms and urinalysis, were reported with moderate sensitivity, whereas patient report was quite insensitive to physician ordering of a hematocrit or hemoglobin tests. Immunizations were reported with a variable sensitivity, depending on the particular immunization. Patient report of referral to another physician was moderately sensitive, but the sensitivity of report of referral to nonphysicians was low. The concordance was only moderate between the patients' reported reason for visit and the rea-

TABLE 2. (Continued)

Judgment on Whether or Not a Service Was Performed Because of Symptoms or Illness (service)	Direct Observation (128 observations on 16 taped patient visits) (kappa)	Number of Videotapes on Which Service Was Observed
Heart examination	1.00	8
Smoking cessation	1.00	7
Urinalysis	1.00	2
Diet advice	1.00	3
Back examination	0.73	3
Extremity examination	0.73	7
Neurological examination	0.69	5
Head/neck examination	0.69	12
Lung examination	0.68	9
Abdomen	0.67	6
Skin examination	0.65	3
Exercise advice	0.52	4
Family history	0.31	3
Social history	0.17	13

NA, not applicable.
*Research nurse assessment of components of the visit that led to the assignment of a Current Procedures Technology (CPT) Code for visit.[44]

son for visit that was overt and recorded by the research nurse.

Discussion

This description of the methods of the Direct Observation of Primary Care Practice Study demonstrates the feasibility of carrying out a large multimethod observational study in busy nonacademic practice sites. The physician sample represented recent demographic trends toward increasing numbers of female and residency-trained practitioners. The patient sample was highly representative of patients seeing family physicians, although questionnaire responders showed similar selection factors to other survey research. The measurement of important variables from multiple vantage points is important for studies of variables for which there is no clear gold standard. In addition, the concurrent use of both quantitative and qualitative methods allows simultaneous testing of a priori hypotheses and generation of new hypotheses.[39,60]

Before this study, measurement of the content of the ambulatory care visit rarely had been compared to direct observation. Direct observation is expensive and potentially intrusive, but provides a gold standard for assessing the validity of more easily performed nonobservational methods such as medical record review and patient exit questionnaire. This study directly observed 4,454 patient visits to 138 physicians, providing a solid basis for making judgments about the validity and reliability of nonobservational methods.

Some areas of service delivery were measured validly by both medical record review and self-report compared to direct observation. These include Pap smears, breast, pelvic and rectal examinations, and influenza immunization. Our findings are consistent with those of Montaño and Phillips,[23] who found a high degree of correlation between rates of cancer preventive service delivery measured by medical record review and patient survey. For these services, researchers can confidently choose either method of measurement, making the decision on the basis of other factors, such as ease of access to data from each source.

Delivery of some services were more sensitively ascertained by medical record review, including most physical examination items, laboratory tests,

TABLE 3. Concordance of Medical Record Review With Direct Observation

Service (sex/age group eligible)	*n* Eligible	*n* Performed	Sensitivity (%)	Specificity (%)	Kappa
Counseling services					
Alcohol history (≥13)	3,700	350	57	96	0.54
Tobacco history (≥13)	3,700	371	53	88	0.33
Estrogen prescription (F ≥ 18)	2,230	101	49	98	0.49
Contraception (≥13)	3,700	187	47	99	0.54
Drug abuse history (≥13)	3,700	100	45	98	0.40
Aspirin for myocardial infarction prophylaxis (≥18)	3,475	81	44	99	0.46
Seat belt use	4,432	47	43	99	0.30
Smoking cessation	4,432	242	41	99	0.52
Exercise	4,432	858	38	96	0.41
Social history	4,432	2825	36	83	0.16
Alcohol counseling (≥13)	3,700	58	33	100	0.39
Estrogen discussion (F ≥ 18)	2,230	126	31	100	0.43
Back Pain prevention (≥13)	3,700	62	29	100	0.38
Family history	4,432	1080	28	96	0.29
Other sexually transmitted disease prevention (≥13)	3,700	40	28	100	0.34
Diet advice	4,432	525	27	96	0.28
Breast self-examination (F ≥ 18)	2,231	150	27	100	0.39
Accident prevention	4,432	55	27	100	0.34
Diet advice about sodium	4,432	111	25	100	0.36
Dental health	4,432	121	25	100	0.34
Diet advice about cholesterol/fat	4,432	258	24	99	0.32
Diet advice about calories	4,432	218	23	99	0.28
Distribution of educational materials	4,432	236	19	99	0.25
Other injury prevention	4,432	49	16	100	0.23
Passive tabacco exposure	4,432	56	13	100	0.20
Avoidance of sun exposure	4,432	34	12	100	0.20
Diet advice about calcium	4,432	37	8	100	0.12
Physical examination					
Bimanual pelvic (F ≥ 13)	2,367	226	90	99	0.89
Lung examination	4,432	2538	86	87	0.72
Breast examination (F ≥ 13)	2,367	260	86	99	0.87
Head/neck examination	4,432	2627	81	87	0.67
Rectal examination	4,432	236	78	99	0.80
Abdominal examination	4,432	1274	77	91	0.68
Heart examination	4,432	2160	76	86	0.62
Gonorrhea culture (F ≥ 13)	2,367	38	76	100	0.79
Extremity examination	4,432	1579	70	89	0.59
Chlamydia culture (F ≥ 13)	2,367	40	68	100	0.73
Testicular examination (M)	1,699	173	67	98	0.68
Neurological examination	4,432	566	64	93	0.54
Skin examination	4,432	803	61	89	0.48
Back examination	4,432	685	49	96	0.50

(*Continues*)

TABLE 3. (Continued)

Service (sex/age group eligible)	*n* Eligible	*n* Performed	Sensitivity (%)	Specificity (%)	Kappa
Screening services					
PAP test (F ≥ 13)	2,367	205	90	99	0.89
In office fecal occult blood test (≥18)	3,475	114	70	99	0.67
Home fecal occult blood test (≥18)	3,475	94	67	99	0.64
Sigmoidoscopy (≥18)	3,475	34	65	99	0.57
Mammogram (F ≥ 18)	2,231	134	62	99	0.68
Lab tests					
EKG	4,432	119	92	99	0.84
TB skin test	4,432	67	82	100	0.81
Chest x-ray	4,432	99	81	99	0.77
Urinalysis	4,432	233	80	94	0.53
Prostate specific antigen (M ≥ 18)	1,235	41	78	98	0.61
Hemotrocrit/hemoglobin	4,432	309	76	95	0.59
Thyroid tests	4,432	165	73	97	0.60
Chemistry panel	4,432	365	72	96	0.63
Glucose test	4,432	322	67	93	0.49
Cholesterol	4,432	288	64	96	0.55
Adult immunization					
Flu shot	4,432	115	83	99	0.78
Tetanus booster	4,432	62	81	100	0.79
Hepatitis B vaccine	4,432	35	80	100	0.78
Children ≤ 12 yr of age only					
Polio immunization	713	55	86	99	0.86
DPT immunization	713	65	86	98	0.83
HIB immunization	713	60	85	99	0.85
Genital examination	713	139	68	96	0.69
Breast feeding advice	713	30	67	99	0.75
Nutrient intake dietary advice	713	184	54	95	0.54
Eye examination	713	47	37	98	0.38
MMR immunization	713	47	30	99	0.41
Referral					
Another physician	4,276	327	61	98	0.62
Nonphysician out-of-office	4,238	104	31	99	0.35
Nonphysician in-office	4,276	69	25	100	0.32
Any	4,454	455	58	97	0.60
Reason for visit					
Prenatal	4,432	47	92	99	0.92
Acute visit (initial and follow-up)	4,454	2582	83	81	0.63
Well visit	4,454	536	81	98	0.79
Chronic disease (routine and flare-up)	4,454	1040	68	87	0.54
Chronic problem, routine	4,432	784	67	92	0.57
Counseling/advice visit	4,432	61	33	99	0.33
Other reason for visit	4,432	96	29	99	0.35
Administrative purposes	4,432	43	23	99	0.26

n Eligible, number of patients eligible for this service by the age and gender criteria in parentheses; *n* Performed, number of times this service was performed during visits to eligible patients, assessed by direct observation.

TABLE 4. Concordance of Patient Exit Questionnaire With Direct Observation

Service(sex/age group eligible)	*n* Eligible	*n* Performed	Sensitivity (%)	Specificity (%)	Kappa
Counseling services					
Breast self-examination (F ≥ 18)	1,734	119	83	97	0.74
Smoking cessation	3,283	155	72	96	0.53
Seat belt use	3,283	32	72	98	0.34
Alcohol counseling (≥13)	2,813	38	63	97	0.31
Diet	3,283	157	58	87	0.22
Back pain prevention (≥13)	2,813	44	57	97	0.32
Estrogen discussion (F ≥ 18)	1,734	111	52	97	0.51
Aspirin for myocardial infarction prophylaxis (≥18)	2,670	64	47	98	0.37
Educational materials given	3,282	166	47	96	0.39
Exercise	3,283	678	42	94	0.42
Contraception	2,813	142	39	99	0.48
Birth control	2,813	133	38	99	0.49
Dental health	3,283	86	37	99	0.43
Family history	3,282	776	33	93	0.30
Passive tobacco exposure	3,283	34	27	97	0.12
Accident prevention	3,283	44	25	99	0.21
Physical examination					
Breast examination (F ≥ 13)	1,820	215	87	99	0.86
Bimanual pelvic examination	1,820	172	83	98	0.80
Pelvic examination	1,820	172	83	98	0.80
Rectal examination	3,283	194	82	98	0.74
Heart or lung	3,283	2,024	75	88	0.60
Abdominal examination	3,283	943	64	93	0.60
Back examination	3,283	531	47	93	0.42
Testicular examination (M)	1,222	124	46	99	0.54
Extremity examination	3,283	1,216	45	93	0.41
Skin examination	3,283	617	40	95	0.40
Head/neck examination	3,283	1,944	40	93	0.29
Screening services					
PAP (F ≥ 13)	1,820	159	89	98	0.86
In office fecal occult blood test (≥18)	2,670	95	50	98	0.44
Home fecal occult blood test (≥18)	2,670	86	47	98	0.46
Mammogram (F ≥ 18)	1,734	115	17	99	0.25

(*Continues*)

and immunizations. Thus, medical record review is preferable to patient questionnaire for variables for which there are limitations in patient understanding or knowledge of what the physician did during the encounter. Efforts to improve the accuracy of medical record charting, particularly better recording of health habit counseling, would increase the utility of the medical record for research and for providing quality patient care.

For other areas of service delivery, patient report is a more valid measure than medical record review. Most health habit counseling is poorly documented in the medical record, but is reliably reported by the patient. Others also have found

TABLE 4. (Continued)

Service(sex/age group eligible)	*n* Eligible	*n* Performed	Sensitivity (%)	Specificity (%)	Kappa
Lab tests					
EKG	3,283	92	74	99	0.67
Urinalysis	3,283	169	74	94	0.48
Prostate specific antigen (M ≥ 18)	929	35	60	95	0.40
Chest x-ray	3,283	75	55	99	0.55
Cholesterol	3,283	227	53	95	0.44
Glucose test	3,283	248	43	94	0.35
Thyroid function tests	3,283	133	41	98	0.43
Hematocrit/hemoglobin	3,283	233	28	97	0.28
Adult immunization					
Flu shot	3,283	98	85	98	0.64
Tetanus booster	3,283	50	74	99	0.67
For children ≤12 yr of age only					
Polio immunization	452	32	81	98	0.77
Diphtheria pertussis tetanus immunization	452	40	75	98	0.74
Genital examination	452	85	68	97	0.70
Hemophilus influenza B immunization	452	39	49	99	0.60
Referral					
Another physician	2,918	226	72	96	0.62
Nonphysician out-of-office	2,860	70	36	96	0.21
Nonphysician in-office	2,908	35	34	85	0.03
Any	2,908	455	50	86	0.28
Reason for visit					
Chronic illness	1,776	999	51	91	0.40
Acute illness	1,776	447	76	64	0.31
Well visit	1,776	235	57	94	0.52
Other reason for visit	1,776	82	56	83	0.16

n Eligible, number of patients eligible for this service by the age and gender criteria in parentheses; *n* Performed, number of times this service was performed during visits to eligible patients, assessed by direct observation.

that the medical record typically severely underreports the provision of patient counseling about health habits.[27] Among samples with predominantly minority patients, receipt of mammography and Pap smears were significantly overreported by patients.[25,26,32] It is likely that socioeconomic and cultural factors may affect the accuracy of patient report of receipt of medical services. One telephone survey of women who received a mammogram at a mobile van found an 82% accuracy rate for patient report of having had a mammogram within the past year.[35] The accuracy of patient recall declined with the duration of time since having had the procedure. The findings of our patient exit questionnaire, which was completed shortly after the end of the visit, may not extend to patient surveys conducted much longer after the patient visit.

Some items were poorly measured by both medical record review and patient exit questionnaire. Thus, for items such as physician counseling about passive tobacco exposure or accident prevention, more in-depth questioning of patients may be necessary than was provided by our single-item questions about whether or not the physician provided this advice to the patient during the observed visit. Although both the medical record and the patient questionnaire were moder-

ately accurate for measuring if a referral was made to a physician, both measurement methods showed poor sensitivity for referrals to nonphysicians. Direct observation or other methods are necessary to accurately measure this variable.

These data provide insights into the best nonobservational method for measuring various aspects of the content of the ambulatory patient visit. This information will be useful for readers of the medical literature in understanding previously published work by calling into question studies that draw conclusions based on measurement methods that our study has shown to be insensitive to the delivery of particular services. In addition, Tables 3 and 4 will be useful for quality managers and for researchers who need to choose the best nonobservational method for measuring the delivery of particular medical services.

For example, in part because of incentives involved in managed care, physicians and health care plans are increasingly subject to performance and quality assessment.[61] "Physician profiles are already being used in decisions about hiring, firing, disciplining, and paying physicians."[5] Because of limitations in the scope of claims data, patient surveys and medical record reviews increasingly are being used to create these profiles despite the lack of validity data on these surrogate measures of the actual delivery of services to patients.[4,6,9,10,16,62] Our data validate some of the decisions made by the National Committee on Quality Assurance (NCQA) in choosing measures for their revised HEDIS 3.0 performance measures.[62] For example, because of the higher sensitivity of the patient questionnaire compared with medical record review for ascertaining delivery of smoking cessation advice, patient survey is the appropriate choice for measuring delivery of this preventive service. The data also showed that although medical record review is accurate for ascertaining delivery of influenza vaccine to patients, the patient questionnaire was nearly as sensitive and specific. Thus, the patient survey could be used to measure this, potentially with considerable cost savings. As the National Committee on Quality Assurance and others move to expand the scope of measurement of quality of care, the data in this report will be useful in choosing the most appropriate measure for different services.

Acknowledgments

The authors thank the physician members of the Research Association of Practicing Physicians (RAPP) and the office staffs and patients without whose participation this study would not have been possible. Members of the Direct Observation of Primary Care Practice Study Team contributed to the development of the methods reported in this article: Edward Callahan, PhD, Jason Chao, MD, Benjamin Crabtree, PhD, Daniel Dunn, PhD, William Gillanders, PhD, Jack Medalie, MD, MPH, William Miller, MD, MS, and J. Christopher Shank, MD.

References

1. **Clancy C, Gold M, Wall E.** Primary care and health care reform: The next 100 days. J Fam Pract 1993;36:233.

2. **Starfield B, Simpson L.** Primary care as part of US health services reform. JAMA 1993;269:3136–3139.

3. **Langsley DG.** Medical competence and performance assessment: A new era. JAMA 1991:977.

4. **Dauphinee WD.** Assessing clinical performance: Where do we stand and what might we expect? JAMA 1995;274:741.

5. **Kassirer JP.** The use and abuse of practice profiles. N Engl J Med 1994;330:634.

6. **Brand DA, Quam L, Leatherman S.** Medical practice profiling: Concepts and caveats. Med Care Res Rev 1995;52:223.

7. **McAuliffe WE.** Studies of process-outcome correlations in medical care evaluations: A critique. Med Care 1978;16:907.

8. **Brook RH, Williams KN, Avery AD.** Quality assurance today and tomorrow: Forecast for the future. Ann Intern Med 1976;85:809.

9. **Wingert TD, Krawlewski JE, Lindquist TJ, Knutson DJ.** Constructing episodes of care from encounter and claims data: Some methodological issues. Inquiry 1995;32:430.

10. **Weiner JP, Powe NR, Steinwachs DM, Dent G.** Applying insurance claims data to assess quality of care: A compilation of potential indicators. Qual Rev Bull 1990;16:424.

11. **Weiner JP, Parente ST, Garnick DW, et al.** Variation in office-based quality: A claims-based profile of care provided to Medicare patients. JAMA 1995;273: 1503.

12. **Fisher ES, Whaley FS, Krushat M, et al.** The accuracy of Medicare's hospital claims data: Progress has been made, but problems remain. Am J Public Health 1992;82:243.

13. **Lawthers AG, Palmer RH, Edwards JE, Fowles J, Garnick DW, Weiner JP.** Developing and evaluating performance measures for ambulatory care quality: A preliminary report of the DEMPAQ project. Joint Commission Journal on Quality Improvement 1993;19:552.

14. **Romano PS, Roos LL, Luft HS, et al.** A comparison of administrative versus clinical data: Coronary artery bypass surgery as an example. J Clin Epidemiol 1994;47:249.

15. **Romano PS, Mark DH.** Bias in the coding of hospital discharge data and its implications for quality assessment. Med Care 1994;32:81.

16. **Garnick DW, Fowles J, Lawthers AG, Weiner JP, Parente ST, Palmer RH.** Focus on quality: Profiling physicians' practice patterns. J Ambulatory Care Manage 1994;17:44.

17. **Parente ST, Weiner JP, Garnick DW, et al.** Developing a quality improvement database using health insurance data: A guided tour with application to Medicare's National Claims History file. Am J Med Qual 1995;10:162.

18. **Fowles JB, Weiner JP, Knutson D, Fowler E, Tucker AM, Ireland M.** Taking health status into account when setting capitation rates: A comparison of risk-adjustment methods. JAMA 1996;276:1316.

19. **US Department of Health and Human Services.** Report to Congress: The Feasibility of Linking Research-Related Data Bases to Federal and Non-Federal Medical Administrative Data Bases. Rockville, MD: US DHHS, Public Health Service, Agency for Health Care Policy and Research, 1991.

20. **Stange KC.** Primary care research: Barriers and opportunities. J Fam Pract 1996;42:192.

21. **McPhee SJ, Richard RJ, Solkowitz SN.** Performance of cancer screening in a university general internal medicine practice: Comparison with the 1980 American Cancer Society Guidelines. J Gen Intern Med 1986;1:275.

22. **Woo B, Woo B, Cook EF, Weisberg M, Goldman L.** Screening procedures in the asymptomatic adult: A comparison of physicians' recommendations, patients' desires, published guidelines, and actual practice. JAMA 1985;254:1480.

23. **Montaño DE, Phillips WR.** Cancer screening by primary care physicians: A comparison of rates obtained from physician self-report, patient survey, and chart audit. Am J Public Health 1995;85:795.

24. **Palmer RH, Hargraves JL.** The ambulatory care medical audit demonstration project: Research design. Med Care 1996;34:SS12.

25. **Starfield B, Steinwachs D, Morris I, Bause G, Siebert S, Westin C.** Concordance between medical records and observations regarding information on coordination of care. Med Care 1979;17:758.

26. **Zuckerman AE, Starfield B, Hochreiter C, Kovasznay B.** Validating the content of pediatric outpatient medical records by means of tape-recording doctor-patient encounters. Pediatrics 1975;56:407.

27. **Callahan EJ, Bertakis KD.** Development and validation of the Davis Observation Code. Fam Med 1991;23:19.

28. **Suarez L, Goldman DA, Weiss NS.** Validity of Pap smear and mammogram: Self-reports in a low-income Hispanic population. Am J Prev Med 1995;11:94.

29. **McKenna MT, Speers M, Mallin K, Warnecke R.** Agreement between patient self-reports and medical records for Pap smear histories. Am J Prev Med 1992;8:287.

30. **Sawyer JA, Earp JA, Fletcher RH, Daye FF, Wynn TM.** Accuracy of women's self-report of their last Pap smear. Am J Public Health 1989;79:1036.

31. **Gerbert B, Stone G, Stulbarg M, Gullion DS, Greenfield S.** Agreement among physician assessment methods: Searching for the truth among fallible methods. Med Care 1988;26:519.

32. **Johnson CS, Archer J, Campos-Outcalt D.** Accuracy of Pap smear and mammogram self-reports in a southwestern native American tribe. Am Prev Med 1995;11:360.

33. **Silagy C, Muir J, Coulter A, Thorogood M, Yudkin P, Roe L.** Lifestyle advice in general practice: Rates recalled by patients. BMJ 1992;305:871.

34. **Harlow SD, Linet MS.** Agreement between questionnaire data and medical records: The evidence for accuracy of recall. Am J Epidemiol 1989;129:233.

35. **Etzi S, Lane DS, Grimson R.** The use of mammography vans by low-income women: The accuracy of self-reports. Am J Public Health 1994;84:107.

36. **Degnan D, Harris R, Ranney J, Quade D, Earp JA, Gonzalez J.** Measuring the use of mammography: Two methods compared. Am J Public Health 1992;82:1386.

37. **Zapka JG, Bigelow C, Hurley T, et al.** Mammography use among sociodemographically diverse women: The accuracy of self-report. Am J Public Health 1996;86:1016.

38. **Rohrbaugh M, Rogers JC.** What did the doctor do? When physicians and patients disagree. Arch Fam Med 1994;3:125.

39. **Stange KC, Miller WL, Crabtree BF, O'Connor PJ, Zyzanski SJ.** Multimethod research: Approaches for integrating qualitative and quantitative methods. J Gen Int Med 1994;9:278.

40. **Stange KC.** Practice-based research networks: Their current level of validity, generalizability, and potential for wider application. Arch Fam Med 1993;2:921.

41. **Stange KC.** "One size doesn't fit all." Multimethod research yields new insights into interventions to improve preventive service delivery in family practice. J Fam Pract 1996;43:358.

42. **Stange KC, Zyzanski SJ, Jaén CR, et al.** Illuminating the black box: A description of 4454 patient visits to 138 family physicians. J Fam Pract 1998;46(5). In press.

43. **Ware J, Nelson E, Sherboume C, Stewart A.** Preliminary tests of a 6-item general health survey: A patient application. In: Ware ASJ, ed. Measuring functioning and well-being. Durham: Duke University Press, 1992.

44. **American Medical Association**. Physicians'current procedural terminology: CPT '95. Chicago, IL: AMA, 1995.

45. St. Anthony's ICD-9-CM: Code book for physician payment. Alexandria, VA: St. Anthony Publishing, Inc. 1994:533;410.

46. **Bogdewic SP.** Participant observation. In: Crabtree BF, Miller WL, eds. Doing qualitative research: Multiple strategies. Newbury Park, CA: Sage Publications, 1992.

47. **Schneider D, Appleton L, McLemore T.** A reason for visit classification for ambulatory care. Vital Health Stat 2 1979;78:1.

48. **Schappert SM.** National ambulatory medical care survey: 1991 summary. Vital Health Stat 13 1993;116:1.

49. **American Academy of Family Physicians.** Facts about family practice. Kansas City, MO: American Academy of Family Physicians, 1996.

50. **Schappert SM.** National ambulatory medical care survey: 1994 summary. Advance data from vital and health statistics. Adv Data 1996;273:1.

51. **Cogner AJ.** Integration and generalization of kappas for multiple raters. Psych Bull 1980;88:322.

52. **Strube MJ.** A general program for the calculation of the kappa coefficient. Behavior Research Methods, Instruments & Computers 1989;21:643.

53. **Fleiss JL.** The use of mammography vans by low-income women: The accuracy of self-reports. Am J Public Health 1971;76:378.

54. **Maclure M, Willett WC.** Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126: 161.

55. **Uebersax JS.** A generalized kappa coefficient. Educ Psych Meas 1982;42:151.

56. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. Biometrics 1977;33:159.

57. **Feinstein AR, Cicchetti DV.** High agreement but low Kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543.

58. **Cicchetti DV, Feinstein AR**. High agreement but low Kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551.

59. **Brown JB, Adams ME.** Patients as reliable reporters of medical care process: Recall of ambulatory encounter events. Med Care 1992;30:400.

60. **Stange KC, Zyzanski SJ.** Integrating quantitative and qualitative research methods. Fam Med 1989;22:183.

61. **Iglehart JK.** The National Committee for Quality Assurance. N Engl J Med 1996;335:995.

62. **National Committee for Quality Assurance.** HEDIS 3.0 Health Plan Employer Data & Info. Washington, DC: NCQA, 1996.