What is the Concordance Between the Medical Record and Patient Self-Report as Data Sources for Ambulatory Care?

Diana M. Tisnado, PhD, John L. Adams, PhD,† Honghu Liu, PhD,* Cheryl L. Damberg, PhD,†‡ Wen-Pin Chen, MS,* Fang Ashlee Hu, MD,* David M. Carlisle, MD, PhD,*§ Carol M. Mangione, MD,* and Katherine L. Kahn, MD*†*

Background: The validity of quality of care assessments relies upon data quality, yet little is known about the relative completeness and validity of data sources for evaluating the quality of care.

Objectives: We evaluated concordance between ambulatory medical record and patient survey data. Levels of concordance, variations by type of item, sources of disagreement between data sources, and implications for quality of care assessment efforts are discussed.

Design and Subjects: This was an observational study that included 1270 patients sampled from 39 West Coast medical organizations with at least 1 of the following: diabetes, ischemic heart disease, asthma or chronic obstructive pulmonary disease, or low back pain.

Measures: Items from both data sources were grouped into 4 conceptual domains: *diagnosis*, *clinical services delivered*, *counseling and referral*, and *medication use*. We present total agreement, kappa, sensitivity, and specificity at the item and domain-levels and for all items combined.

Results: We found good concordance between survey and medical records overall, but there was substantial variation within and across domains. The worst concordance was in the *counseling and referrals* domain, the best in the *medication use* domain. Patients were able to report with good sensitivity on memorable items.

Conclusions: Quality ratings are likely to vary in differing directions, depending on the data source used. The most appropriate data source for analyses of components of and overall quality of care must be considered in light of study objectives and resources. We recommend data collection from multiple sources to most accurately portray the patient and provider experience of medical care.

Key Words: health services research, quality measurement, ambulatory care

(*Med Care* 2006;44: 132–140)

From the *University of California at Los Angeles; †RAND, Santa Monica; ‡Pacific Business Group on Health, San Francisco; and §California Office of Statewide Health Planning and Development, Sacramento, California.

Supported by the Agency for Healthcare Research and Quality, American Association of Health Plans, The Pacific Business Group on Health, the Robert Wood Johnson Foundation, and National Research Services Award T32 HS400046 from the Agency for Healthcare Research and Quality.

Reprints: Diana M. Tisnado, PhD, UCLA Department of Medicine, Division of GIM and HSR, 911 Broxton Plaza, Box 951736, Los Angeles, CA 90095-1736. E-mail: dtisnado@mednet.ucla.edu.

ISSN: 0025-7079/06/4402-0132

The validity of quality of care assessments relies upon the quality of the data that are used to produce performance measures. Ambulatory care, accounting for most of the population's contacts with the health care system, is crucial for primary prevention and managing chronic disease. Understanding data quality is fundamental to understanding the validity of ambulatory quality of care measures.

The medical record often is viewed as the preferred data source for measuring patient illness, processes of care, and outcomes. However, medical record review is costly, challenging to implement, and comes with its own sources of measurement error. These include erratic recording of certain topics, such as counseling;[1] failure to include orders, labs, or reports in the chart; delayed recording (leading to physician recall problems); and generally sparse recording in certain health care settings (eg, those with greater time pressure).[2,3] Complete documentation of some aspects of care is dependent upon patients sharing information about symptoms, health behaviors, and nonadherence with recommendations.

Patient self-report data can be subject to error as the result of a variety of factors, including recall, social desirability bias, and patient health knowledge and awareness,[4,5] possibly leading to questionable accuracy in patient report of visits,[6–8] counseling,[9] hypertension, hypercholesterolemia[10] and cancer screening.[11] Yet patient self-report provides an important source of data for monitoring the quality of care from the patient point of view and for patients to express perceptions and experiences with care delivery not routinely documented in the medical record that often is less costly than medical record review.

Important work has compared data from the patient self-report and the medical record, but few recent studies adequately address the question of which is the preferred source for many aspects of care. Several studies have addressed concordance between the 2 data sources, but their usefulness is limited by a focus only on utilization,[6,7,8,12] specific aspects of care, specific diseases[13,14] and non-U.S. settings that may not generalize to U.S. healthcare settings or medical record systems.[15–17] Work by Stange[1] has addressed a comprehensive set of important aspects of ambulatory care and compared medical record and self-report data to direct observation as a gold standard. However, this work was limited to single visits with 138 family practitioners in a relatively small geographic region, with interview data collected immediately after visits.

We studied the concordance between patient self-report and medical record data from a large study of the quality of ambulatory care for patients with chronic illnesses under varying models of managed care in urban and rural areas of 3 West Coast states.[18] We examined data from medical records associated with all encounters with all physicians of key specialty types for the diseases under study, and patients' self-reports about all care received during the study period.

Recognizing that neither the medical record nor the patient perfectly represent "truth," we hypothesized that the best source of data would depend upon the domain of care under consideration. We expected medical records to be a better source for technical aspects of care, such as diagnoses and clinical services delivered. Although evidence suggests that medical records and patients are both imperfect reporters of counseling,[3,9] we expected patients to be better reporters of talking aspects of care than medical records. On the basis of available evidence, it is unclear whether the medical record or the patient should be the preferred data source for current medication use.[13,14,19] Levels of concordance, how concordance varies by the type of item collected, possible sources of disagreement between data sources, and implications for quality of care assessment efforts are discussed.

METHODS

Data were collected as part of the Pacific Business Group on Health (PBGH) Physician Value Check Survey and UCLA Validation project, an observational study evaluating quality of care and reasons for changes in outcomes across 2 years for a cohort of older managed care patients enrolled in physician organizations (POs) located in 3 West Coast states. The study was approved by the UCLA Institutional Review Board. Study design and survey results are described elsewhere.[18]

For this study of concordance across data sources, we examined data from a 1998 patient survey and medical record review of all visits that took place within 30 months before the survey. We selected equivalent items from both data sources from a pool of items that had been used to construct process of care measures. Items selected addressed a range of disease-specific and general issues across a spectrum of care.

Items were grouped into 4 domains conceptualized as part of the larger study to represent important components of the process of care: *diagnosis*, *clinical services delivered*, *counseling and referrals*, and *medication use*. The *diagnosis* domain includes patient history of diagnoses or medical conditions. *Clinical services delivered* include health services patients receive such as physical examination, surgical procedures or special tests. *Counseling and referrals* includes (1) the provider talking with the patient about ways to prevent disease or manage their chronic condition or (2) recommending that the patient consult with another provider. *Medication use* represents medications the patient was using at the time of the 1998 survey.

The analysis included 50 items representing the 4 domains. Items were included only if both the medical record and patient survey instruments recorded patient-level data in comparable time periods. Because of difficulties assessing concordance when prevalence is very low or very high,[20] items were included only if the prevalence of the item as measured by both data sources was between 10% and 90%.

Data Collection Methods

In 1996, the PBGH collected survey data from 30,308 adults from California, Washington, and Oregon who received care in the previous year from 1 of 60 POs (medical groups and independent practice associations [IPAs]). In 1998, we surveyed 3656 patients who had responded to the baseline survey in 1996 (response rate 63%). The mailed, self-administered survey queried patients about diagnoses and health care services received during a 2-year period. Each survey included a disease-specific section to assess processes of care for chronic conditions (ischemic heart disease, diabetes, asthma or chronic obstructive pulmonary disease, or chronic low back pain) reported at baseline. Along with the 1998 mailing, subjects also received an invitation to participate in medical record abstraction and institutional research board-approved consent materials (response rate 54%).

We developed a medical record abstraction tool to collect items representing the aspects of care under study and guidelines with explicit criteria to code items. Nurse abstractors experienced in medical record abstraction and clinical practice successfully completed an intensive training and passed abstraction tests at the end of the training period and throughout the fieldwork.

Abstractors pursued all visits with all primary care providers and key specialty types that took place within 30 months before the survey for consenting patients noted in claims/encounter data provided by participating POs or discovered during abstraction. Study patients had a relatively high volume of clinical encounters and a mean of 5 physicians per patient. Complete medical records were abstracted for 1270 patient survey respondents. A total of 698 patient records were not abstracted or were only partially abstracted because of the inability to locate records, PO closure, and/or study withdrawal. To assess inter-rater reliability, we compared the performance of 11 pairs of abstractors who abstracted components of process measures from the records of 54 unique patients. Concordance between abstractors was excellent with no significant difference noted in overall process scores and with an aggregate 0.87 kappa score across process measures.

Analysis

We calculated the prevalence of each medical condition or service using survey only, medical record only, or either as the data source. To analyze concordance between the 2 data sources, we calculated both the percent total agreement (percent agreement on positives plus negatives) and the kappa statistic. Total agreement may reflect extremes in prevalence and chance agreement. The kappa statistic corrects for chance agreement, but is subject to criticism due to its sensitivity to extremes in prevalence and unbalanced margin totals, which can influence the correction factor.[21]

We also examined the sensitivity (% true positives detected) and specificity (% true negatives detected) of each of the 2 data sources. On the basis of the hypothesis that the patient survey is a better data source for certain items whereas the medical record is a better source for others, we calculated sensitivity and specificity of the patient survey using the medical record as the gold standard, and of the medical record using the patient survey as the gold standard.

We calculated concordance, sensitivity, and specificity at the item level, the domain level, and overall for all items combined. Item-level analyses were based on unique item-patient dyads, classifying agreement and disagreement based upon what was documented by the 2 data sources for each individual item with each unique patient as the unit of analysis. For domain-level and overall analyses, we combined patient-item dyads, using the dyad as the unit of analysis. Since patients may be eligible for multiple items per domain, they could be represented multiple times in these analyses. We calculated 95% confidence intervals (CIs) around domain-level and overall sensitivity and specificity estimates using bootstrap calculations of 2000 replicates in SAS8.[22] The bootstrap sample was drawn at the patient level to account for the correlations induced by individuals contributing multiple observations to each concordance metric.[23]

Items pertinent to all patients regardless of study disease were analyzed including all 1270 patients. However, some item-level concordance analyses were restricted only to those patients eligible for each item. For example, the analysis of concordance on diabetic foot examination was restricted to patients known to have diabetes. Item-level concordance involving laboratory data was calculated only for 1147 patients who had laboratory data available from the medical record. The time period pertinent to each item was comparable for survey and medical record data: typically a period of 1 year, 2 years, or the patient's entire history. For some analyses, an expanding medical record time window allowed evaluation of concordance using an additional 3 to 6 months to account for variations in patients' patterns of clinician visits and timing of survey completion. For example, for survey items that queried about events in the last year, we examined both a 12- and a 15-month medical record time window. For items that queried about the last 2 years, we examined both a 24- and 30-month window. Due to the similarity in results we present only the results using the expanded time windows.

Concordance, sensitivity, and specificity results are categorized as excellent, good, fair, and poor. Cutoffs used to classify kappa are based on Streiner et al[24] and Altman;[25] classification of the other indicators was developed for ease of viewing and making comparisons for this manuscript.

Descriptive analyses were performed to examine the characteristics of the patient sample. On the basis of responses to the baseline patient survey in 1996, respondents were characterized in terms of age, gender, race/ethnicity, education, income, health status (as measured by the physical health components of the SF-12),[26] and type of medical organization from which they were sampled (medical group or IPA).

RESULTS

In Table 1 we present baseline characteristics of the 1270 patients with complete data from the medical record and both the 1996 and 1998 patient self-report surveys. In Table 2, we present item-level prevalence (by medical record, survey, and either data source), concordance (total agreement and kappa) and sensitivity and specificity. Items are presented by domain, in descending order of total agreement. In Table 3, we present measures of concordance, sensitivity, and specificity at the domain-level and for all items combined.

Patient Characteristics

Patient age ranged from 18 to 70 years (mean, 60; SD, 9); 54% were women. The sample was predominantly non-Hispanic white, with greater than 12 years of education and annual income greater than $30,000. Patient self-reported health status was classified as high for those with SF-12 physical scores equal to or greater than the 75th percentile for the sample (52), and low for those who scored below. More patients were from medical groups (69%) than IPAs.

Concordance, Sensitivity, and Specificity

Overall, we found good concordance between survey and medical records, but there was substantial variation in results both within and across domains. Concordance was lowest in the *counseling and referrals* domains, and highest

TABLE 1. Sample Characteristics

	n = 1270	Percent
Age group		
<50	130	10
50–64	659	52
65+	481	38
Gender		
Male	583	46
Female	687	54
Race		
White	1008	79
Black	39	3
Asian	71	6
Hispanic	103	8
Missing race	27	2
Other	22	2
Education		
<High school	562	44
≥High school	708	56
Income		
Low (<$30,000)	371	29
High (>$30,000)	899	71
SF-12 Category		
Low: <75th percentile score (<52)	952	75
High: >75th percentile score (>52)	318	25
Medical organization type		
IPA	390	31
Medical Group	880	69

TABLE 2. Measures of Concordance for Items by Domain

	Prevalence by Data Source				Measures of Concordance		MR = Gold Standard		PSR = Gold Standard	
Domain Item	MR Only	PSR Only	MR-PSR	MR, PSR, or Both	% Total Agreement	Kappa	SE	SP	SE	SP
Diagnoses										
History of acute myocardial infarction	11	13	−2	16	93*	0.7†	78‡	95*	64	97*
History of cancer	12	10	2	15	92*	0.6†	59	97*	72‡	95*
History of diabetes	34	31	3	37	92*	0.8*	84†	97*	93*	92*
Obesity	33	29	4	35	92*	0.8*	83†	92*	94*	92*
History of asthma	16	23	−7	24	91*	0.7†	93*	91*	67	99*
Smoking	20	24	−4	28	88†	0.7†	80†	68	46	98*
Foot ulcers	7	12	−5	15	88†	0.3‡	54	91*	29	97*
Congestive heart failure	13	9	4	18	86†	0.3‡	31	94*	44	90*
History of diabetic retinopathy	23	25	−2	35	79‡	0.4‡	58	85†	54	87†
History of high blood pressure	70	73	−3	85	74‡	0.4‡	84†	52	80†	58
Depressed mood	17	27	−10	35	73‡	0.2	50	78‡	32	89†
History of high cholesterol	63	72	−9	84	69	0.3‡	82†	45	72‡	60
Shortness of breath	72	66	6	85	67	0.2	73‡	53	80†	44
History of arthritis	34	53	−19	61	65	0.3‡	77‡	59	49	84†
Angina/chest pain	49	53	−4	68	65	0.3‡	68	62	63	67
Clinical services delivered										
History of coronary artery bypass or angioplasty	15	17	−2	18	96*	0.9*	94*	97*	83†	99*
Cardiac catheterization	11	17	−6	21	87†	0.5‡	69	89†	44	96*
Treadmill/stress test	48	44	4	58	78‡	0.6†	73‡	83†	80†	76‡
Diabetic foot examination	61	68	−7	81	68	0.3‡	79‡	50	71‡	61
Radiograph of back/spine	34	41	−7	50	66	0.3‡	60	70‡	51	77‡
Echocardiogram	62	38	24	73	55	0.1	44	72‡	72‡	44
Counseling and Referrals										
Saw diabetic nurse educator	7	19	−12	22	83†	0.3‡	68	85†	25	97*
Referred to back pain program or class	11	18	−7	26	78‡	0.1	31	83†	19	91*
Advised to see cardiologist	16	26	−10	33	76‡	0.3‡	56	80†	35	91*
Discussed worsening of angina	33	30	3	45	73‡	0.4‡	54	82†	60	78‡
Counseled or referred for weight loss	23	39	−16	45	72‡	0.4‡	75‡	71‡	43	91*
Referred to orthopedist	21	30	−9	40	71‡	0.2	53	76‡	36	86†
Counseled or referred for depressive symptoms (among antidepressant users)	23	36	−13	45	70‡	0.3‡	62	72‡	40	86†
Counseled about exacerbating factors for shortness of breath	29	49	−20	57	64	0.3‡	73‡	61	43	85†
Counseled about diet/nutrition	49	34	15	62	60	0.2	44	75‡	62	58
Counseled about exercise	45	60	−15	74	57	0.2	69	48	52	65
Medication use										
Theophylline	22	17	5	23	93*	0.8*	71‡	99*	94*	93*
Antidepressant	12	13	−1	17	92*	0.6†	70‡	95*	66‡	96*
Glitazone	15	10	5	17	91*	0.6†	53	98*	84†	92*
Statin or other lipid-lowering drug	30	30	0	35	90*	0.8*	83†	92*	82†	93*
Long-acting nitrate	20	10	10	25	89†	0.6†	47	100*	100*	88†
Beta blocker	37	31	6	45	89†	0.8*	78‡	96*	92*	88†
Long-acting beta agonist	25	18	7	27	88†	0.6†	61	97*	86†	88†
Inhaled steroid	48	48	0	54	87†	0.7†	87†	87†	86†	88†
Angiotensin-converting enzyme inhibitor	31	26	5	35	87†	0.7†	69	94*	85†	87†
Ipratropium	25	16	9	28	86†	0.6†	53	97*	84†	86†
Insulin	15	26	−11	28	85†	0.6†	87†	85†	52	97*
Calcium channel blocker	41	29	12	48	85†	0.7†	67	98*	96*	81†

(*Continued*)

TABLE 2. (*Continued*)

	Prevalence by Data Source				Measures of Concordance		MR = Gold Standard		PSR = Gold Standard	
Domain Item	MR Only	PSR Only	MR-PSR	MR, PSR, or Both	% Total Agreement	Kappa	SE	SP	SE	SP
Narcotic	15	17	−2	24	84†	0.4‡	51	90*	46	91*
Sulfonylurea	59	51	8	64	82†	0.6†	78‡	87†	90*	73‡
Short-acting beta agonist	61	65	−4	73	81†	0.6†	88†	70‡	82†	78‡
Metformin	41	33	8	47	81†	0.6†	67	90*	82†	80†
Hormone-replacement therapy (women >50)	42	43	−1	54	79‡	0.6†	76‡	80†	74‡	82†
Short-acting nitrate	28	6	22	36	75‡	0.2	16	98*	79‡	75‡
NSAID	25	20	5	35	74‡	0.3‡	38	86†	47	81†

n = number of eligible patients included in item-level analysis.
*Excellent agreement (Kappa ≥0.9; SE or SP ≥90).
†Good agreement (Kappa <0.9 and ≥0.6; SE or SP ≥80).
‡Fair agreement (Kappa <0.6 and ≥0.3; SE or SP ≥70).
No symbol indicates Poor Agreement (Kappa <0.3; SE or SP <70).
MR indicates medical record; PSR, patient self-report; SE, sensitivity; SP, specificity.

in the *medication use* domain. Results varied most among items within the domains of *diagnosis* and *clinical services delivered*, whereas results were most consistent within the *counseling and referrals* and the *medication use* domains. Prevalence by either data source was generally higher than by each individual data source, indicating disagreement between the 2 data sources because 1 or both sources failed to report information documented by the other source. Results are described in further detail herein.

Diagnosis

As shown in Table 2, within the *diagnosis* domain, prevalence estimates varied by data source and by item. Survey prevalence estimates were higher than those by medical record for 10 of 15 items, by 2 to 20%. Concordance varied widely by type of diagnosis, with total agreement ranging from 93% for acute myocardial infarction to 65% for a diagnosis of angina. Kappa ranged from 0.8 for diabetes and obesity to 0.2 for shortness of breath and depressed mood. Of 15 diagnoses evaluated by both medical record and survey, 8 had good to excellent total agreement (ie, greater than 80%), all with good to excellent kappa statistics. Four items were associated with poor total agreement and kappa. For all diagnosis items combined (Table 3) concordance was good, with total agreement = 82%. Kappa = 0.6 (95% CI = 0.6–0.6). Survey sensitivity was 78% (95% CI = 77–79%), with specificity somewhat higher at 84% (95% CI = 83–85%).

Clinical Services Delivered

Within the domain *clinical services delivered*, concordance varied by item from excellent to poor. Half of these items showed poor concordance. For all *clinical services delivered* items combined, total agreement and kappa were good (82% and 0.6; 95% CI = 0.5–0.6). Survey sensitivity was fair (72%; 95% CI = 69–75%), and specificity was good at 87% (95% CI = 85–88%).

Counseling and Referrals

Within the *counseling and referrals* domain, prevalence estimates based on survey were higher than those by medical record for 8 of 10 items. Concordance was generally the lowest for items in the *counseling and referrals* domain compared with the other domains. Nine of 10 items were associated with fair to poor total agreement, and all 10 exhibited fair to poor kappa.

TABLE 3. Measures of Concordance by Domain and Overall Across Domains (all items combined)

	Prevalence by Data Source				Measures of Concordance		MR = Gold Standard		PSR = Gold Standard	
Domain	MR Only	PSR Only	MR-PSR	MR, PSR, or Both	% Total Agreement	Kappa (95% CI)	SE (95% CI)	SP (95% CI)	SE (95% CI)	SP (95% CI)
Diagnoses	32	36	−4	42	82	0.6 (0.6–0.6)	78 (77–79)	84 (83–85)	69 (68–71)	89 (89–90)
Clinical services delivered	31	31	0	40	82	0.6 (0.5–0.6)	72 (69–75)	87 (85–88)	71 (68–74)	87 (86–89)
Counseling and referrals	32	39	−7	52	67	0.3 (0.3–0.3)	59 (57–62)	70 (69–72)	48 (46–51)	79 (77–80)
Medication use	32	27	5	37	85	0.6 (0.6–0.7)	68 (66–70)	93 (92–94)	82 (80–84)	86 (85–87)
Overall across domains	32	33	−1	42	80	0.5 (0.5–0.6)	71 (70–72)	84 (84–85)	68 (67–69)	86 (86–87)

Eight of 10 items were associated with poor sensitivity by both survey and medical record, reflecting the disagreement about positives and negatives between the 2 data sources.

At the domain level, concordance was poor, with total agreement = 67% and kappa = 0.3 (95% CI = 0.3–0.3). Taking the patient as the gold standard, medical record sensitivity was poor at 48% (95% CI = 46–51%), and specificity was fair at 79% (95% CI = 77–80%).

Medication Use

Medication use items generally showed the highest levels of concordance across domains. More items exhibited fair to excellent sensitivity by medical record than by patient survey. The use of nonsteroidal anti-inflammatory drugs (NSAIDs) and narcotics was associated with poor sensitivity by both data sources, reflecting disagreement between data sources on both positives and negatives. For all *medication use* items combined, concordance was good: total agreement was 85%, and kappa was 0.6 (95% CI = 0.6–0.7). Survey sensitivity was fair at 68% (95% CI = 66–70%), whereas specificity was excellent at 93% (95% CI = 92–94%). In contrast, medical record sensitivity and specificity were good at 82% (95% CI = 80–84%) and 86% (95% CI = 85–87%), respectively.

Overall

For all items combined across domains, concordance was good according to total agreement = 80% and fair according to kappa = 0.5 (95% CI = 0.5–0.6). Survey sensitivity was fair (71%; 95% CI = 70–72%)), whereas a slightly lower medical record sensitivity fell into the poor range at 68% (95% CI = 67–69%). Specificity was good by both survey (84%; 95% CI = 84–85%) and medical record (86%; 95% CI = 86–87%).

DISCUSSION

These results indicate overall concordance was fair to good. The patient survey was associated with slightly higher prevalence and sensitivity compared with the medical record, and the medical record performed somewhat better in terms of specificity as compared with the patient survey. We found good concordance for the *diagnosis*, *clinical services*, and *medication use* domains, with the highest patient self-report sensitivity for the *diagnosis* domain, and highest medical record sensitivity for the domain of *medication use*.

Despite many consistencies across data sources, domain and item-level analyses show substantial variation in results both across and within domains of medical care. The item-level is of particular importance as it represents the typical level of measurement for components of quality of care indicators. We found excellent concordance at the item-level on diabetes, obesity, coronary artery bypass or angioplasty, and use of statins and beta blockers. Good concordance was also noted for acute myocardial infarction, asthma, smoking, and use of many other medications, including inhaled steroids, angiotensin-converting enzyme inhibitors, and calcium channel blockers. Poor concordance and poor documentation by either data source was noted for angina, counseling about diet/nutrition, and exercise.

At the domain-level, we found the best concordance associated with the *medication use* domain and the worst concordance associated with the *counseling and referrals* domain. The greatest item-level variation was found within the *diagnosis* and *clinical services delivered* domains, whereas item-level results were most consistent within the *counseling and referrals* and the *medication use* domains. These results are consistent with those of previous studies that have demonstrated variations in concordance by domain of medical care.

Our findings highlight the limitations of both the medical record and patient survey as tools for tracking the performance of important interventions in chronic disease care: assessment of health behaviors, education and counseling, and follow-up about behavior change and compliance with recommendations. Previous studies have shown that counseling and referrals are underreported in the medical record as compared with patient survey, indicating that physicians do not consistently record these interventions,[1,3,27] and others have shown underreporting by patients,[9] which may reflect time pressure on physicians, undervaluing the "talking" aspects of medical care, for which physicians cannot routinely bill, a mismatch between physician and patient perceptions of when counseling and referral recommendations have taken place, patient recall bias, patient desire to portray their physician favorably, or some combination of the above. Although total agreement and kappa help us to understand level of concordance, assessing sensitivity and specificity of each data source relative to the other sheds some light on the relative performance of each data source. These findings suggest that a quality assessment effort relying upon medical record data alone could miss diagnoses or conditions that the patient could have reported. This might result in missing important information about, for example, smoking, diabetic foot ulcers, or depressed mood. Conversely, a quality assessment effort relying solely on patient report could miss important information regarding use of short-acting nitrates, long-acting nitrates, ipratropium, or glitazone. Our results also identify instances in which disagreement occurred in both directions, such as in the case of counseling about exercise, suggesting that neither the patient report nor the medical record should be taken as a complete or accurate representation of what took place.

Sensitivity and specificity must be considered in context. The acceptable level and optimal balance of sensitivity and specificity must be determined by the clinical or research objectives and potential applications, the prevalence in the population, and whether, for the purpose at hand, it is preferable to minimize the risk of false positives or false negatives.[28] For Health Plan Employer Data and Information Set (HEDIS) evaluations, health care organizations are driven to minimize false-negative results in terms of services delivered. Quality assurance activities screening care for possible medical errors might purposefully err in the opposite direction, identifying all potential events to evaluate them in more depth.

Potential sources of discrepancies between patient survey and medical records and examples are given in Table 4.

TABLE 4. Possible Sources of Disagreement

Issue	Survey Item	Medical Record Item	Comment
Telescoping	Cardiac catheterization: "During the last 2 years, have you had the following procedures either inside or outside the hospital? cardiac catheterization (heart study where dye is injected into your coronary arteries)"	"Based on provider notes and hospital admission/discharge summaries, give the dates of the specified procedures performed during the data collection period. If study date is not given, provide the date of the provider note describing the study: Cardiac catheterization, angiography, CA" (Code "Yes" for concordance analysis if cardiac catheterization occurred on any date during the study period).	Overreported by patients as compared with MR. This could have been due to patient not understanding item, but item explained procedure clearly. Telescoping, or recalling a memorable event as having occurred more recently than it did is a more likely explanation. 30-month time period shows MR prevalence closer to patient report and better concordance than 24 months, also suggesting telescoping.
	Radiograph of back or spine: "During the last 2 years, have you had the following, either in or outside of the hospital? X-ray of your back or spine?"	"For back pain patients only, based on diagnostic study reports or provider notes describing the studies, list the dates and specify the types of imaging studies performed during the entire data collection period. If study date is not given, provide the date of the provider note describing the study: Plain x-ray" (Code "Yes" for concordance analysis if plain x-ray occurred on any date during the study period).	30-month time window shows MR prevalence closer to patient report and better agreement and Kappa than 24-month results. Trend indicates possible telescoping of patient recall.
Sensitive topics	BMI: derived from self reported weight and height	BMI: derived from medical record weight and height	Due to social stigma, individuals may tend to underestimate weight and to overestimate height
Physicians and patients have different understandings of definitions.	Arthritis: "Do you now have any of the following? Please mark all that apply: Arthritis or any kind of rheumatism?"	"Based on the provider notes during the entire data collection period, did the patient ever have a history of any of the following either before or during the data collection period? Osteoarthritis (DJD)"	Survey question was perhaps more general than a physician's definition for arthritis documented in the medical record, resulting in much higher prevalence by survey.
Problems with item specification	NSAID use: "Are you currently using any prescription medicines? If yes, please gather all of the prescription medicines you are currently using. Write their names below" (Code "Yes" if any NSAID class medication was listed).	"Medication management: Code, dose, and frequency for NSAID medication being used at the beginning of visit or prescribed at the end of visit" (Code "Yes" for concordance analysis if any NSAID use or prescription documented in the last 6 mo of the study period).	Respondents may have been confused about reporting prescribed medications that are available OTC.
	Angina/chest pain: "During the last 2 yr, did you discuss your angina, chest pain, or chest pressure with a doctor or health professional?"	Based on this provider note, did the patient have chest pain? If yes, exertional angina, angina equivalent? Non-cardiac chest pain or chest pain resulting from gastrointestinal or musculoskeletal problems? (Code "Yes" for concordance analysis if any chest pain documented at any visit during the study period).	Survey item may have included patients with non-cardiac chest pain or pressure. We included non-cardiac chest pain in MR definition to improve match, but may have introduced too much imprecision to definitions for good concordance results.
Patient health knowledge	CHF: "During the last 2 yr, have you had congestive heart failure or heart failure?"	"Based on this provider note, did the patient have CHF?"; (Code "Yes" for concordance analysis if CHF documented at any visit during the study period)	Patients may not have been familiar with the term, and may not have known that they had this diagnosis.
	Echocardiogram: "During the last 2 yr, have you had the following procedures either inside or outside the hospital?" echocardiogram (procedure done in the doctor's office or lab where sound waves are bounced off of your heart)"	"Based on the cardiac study reports or provider note describing the studies, provide the dates of the specified information. If study date not given, provide the date of the provider report describing the study" stress echo resting echo" (Code "Yes" for concordance analysis if echocardiogram occurred on any date during the study period).	Although the survey item clearly describes the procedure in layman's terms, patients might not have known what it is or recognized it when they received it

(*Continued*)

TABLE 4. (*Continued*)

Issue	Survey Item	Medical Record Item	Comment
Time period	High cholesterol: "Has a doctor ever said that you have had high cholesterol?"	"Based on the provider notes during the entire data collection period, did the patient ever have a history of any of the following cardiovascular disease problems either before or during the data collection period? hypercholesterolemia, hyperlipidemia" OR Code "high LDL" or "high total cholesterol" for concordance analysis based on clinically detailed criteria using laboratory test results documented during the study period.	"Ever" time period may have introduced patient recall problems, and may have made it difficult to pinpoint documentation of a remote event in the medical record.
	NSAID use: "Are you currently using any prescription medicines? If yes, please gather all of the prescription medicines you are currently using. Write their names below."	Medication management: Code, dose, and frequency for medication being used at the beginning of the visit or prescribed at the end of the visit (collected at each visit). (Code "Yes" for concordance analysis if any NSAID class medication was listed)	NSAID, Narcotic use for low back pain: Survey item asked about current use, and pain relievers may have only been used for a very short duration.

Sources of measurement error pertinent to the patient and patient survey methods include: (1) recall bias with respect to time period (ie, telescoping, or recalling an event as happening more recently than it did);[11,29] (2) response bias due to sensitive items (eg, weight);[5] (3) patient definitions or perceptions of health issues that differ from those of the physician or the researcher (eg, arthritis).[27,30] Either patient survey or medical record review might present errors from (4) problematic item specification (eg, use of NSAIDs: patient confusion about reporting recommended medications that are available over the counter) or lack of understanding of the health issues queried about in the survey (eg, congestive heart failure, echocardiogram) resulting from patient lack of knowledge and/or poor physician-patient communication.[19,31,32]

Careful item design may help to address some of these issues and improve concordance results. Techniques such as bounding and judicious wording may aid patient recall and understanding of the information being requested, and to soften potentially threatening items. Extending the medical record data time window for comparison has been used to address telescoping. Varying the medical record abstraction window can affect dramatically the rates of self-reported screening mammograms and Pap smears that can be validated.[29,33] However, we found that the variations in the time window did not change concordance results substantially. Therefore, it appears that the choice of medical record abstraction window is not a major source of disagreement in this study.

Response bias might limit the generalizability of these findings. Survey nonresponders in 1998 who responded in 1996 were more likely to be nonwhite, less educated, and to have lower SF-12 physical and mental scores than responders,[18] characteristics potentially associated with less accurate self report.

Fragmented medical records for patients across multiple providers present significant challenges to investigators seeking to collect comprehensive medical record data. This is compounded by poor continuity and poor coordination for patients with multiple providers.[34] The difficulty of collecting complete medical record data is not only a measurement problem but is itself symptomatic of quality problems in our health care system. Other sources of error can arise specifically from abstraction procedures. Ambiguous abstraction criteria can contribute to error in medical record data. In some medical organizations, certain data elements such as laboratory data are stored separately from the medical record. Events that occurred remotely pose challenges in terms of locating data elements that may be recorded in archived medical record volumes. To address challenges posed by use of medical record data, we recommend that researchers carefully consider the nature of the data elements sought for abstraction and develop protocols for locating and accessing specific data elements that may be stored separately from the main record.

Considering these challenges, we believe this analysis supports the practice of collecting data from multiple sources to most accurately portray the quality of the patient and physician experience of medical care. We recognize that this is not always feasible and that even when using both data sources, error may still be introduced from sources discussed in this paper or from patient and/or medical organization characteristics. Moreover, we recognize the desirability of making a general recommendation for the use of 1 data source versus the other. However, we do not believe these data support such a recommendation. This analysis suggests that patients are able to report with good to excellent sensitivity on memorable diagnoses and clinical services, and on many medications. With reliance on the medical record alone, we may miss certain events such as those in the counseling domain and even certain medical conditions such as depressive symptoms, arthritis pain, and diabetic foot ulcers.

Measurement is a necessary first step toward change and improvement in health care delivery but, to be effective, our tools must identify true inappropriate variation in care or we risk wasting resources responding to random or systematic variation introduced by imperfect data and measurement methods. The stakes are high as calls have been made for public disclosure of quality data, with reorganization and pay-for-performance based on quality indicators. To effectively and appropriately re-engineer processes, align provider incentives, and inform patients we must measure correctly. The costs of failure could be enormous, both financially and in terms of credibility and buy-in of health care professionals, organizations, and consumers.

ACKNOWLEDGMENTS

We would like to acknowledge the assistance provided by Corinna Koehnenkamp and Sarah Gee in preparing the manuscript.

REFERENCES

1. Stange KC, Zyzanski SJ, Fedirko Smith T, et al. How valid are medical records and patient questionnaires for physician profiling and health services research? A comparison with direct observation of patient visits. *Med Care*. 1998;36:851–867.
2. Peabody JW, Luck J, Glassman P, et al. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA*. 2000;283:1715–1722.
3. Luck J, Peabody JW, Dresselhaus TR, et al. How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. *Am J Med*. 2000;108:642–649.
4. Sudman S, Bradburn NM. *Response Effects in Surveys*. Hawthorne: Adeline; 1974.
5. Andersen RM, Kasper J, Frankel MR, et al. *Total Survey Error: Applications to Improve Health Surveys*. San Francisco: Jossey Bass; 1979.
6. Rozario PA, Morrow-Howell N, Proctor E. Comparing the congruency of self-report and provider records of depressed elders' service use by provider type. *Med Care*. 2004;42:952–959.
7. Ritter PL, Stewart AL, Kaymaz H, et al. Self-reports of health care utilization compared to provider records. *J Clin Epidemiol*. 2001;54:136–141.
8. Wallihan DB, Stump TE, Callahan CM. Accuracy of self-reported health services use and patterns of care among urban older adults. *Med Care*. 1999;37:662–670.
9. Flocke SA, Stange KC. Direct observation and patient recall of health behavior advice. *Prev Med*. 2004;38:343–349.
10. Bowlin SJ, Morill BD, Nafziger AN, et al. Reliability and changes in validity of self-reported cardiovascular disease risk factors using dual response: The Behavioral Risk Factor Survey. *J Clin Epidemiol*. 1996;49: 511–517.
11. Champion VL, Menon U, Hollinden D, et al. Validity of self-reported mammography in low income African American Women. *Am J Prev Med*. 1998;14:111–117.
12. Fowles JB, Fowler EJ, Craft C. Validation of claims diagnoses and self-reported conditions compared with medical records for selected chronic diseases. *J Ambul Care Manage*. 1998;21:24–34.
13. Fowles JB, Rosheim K, Fowler EJ. The validity of self-reported diabetes quality of care measures. *Int J Qual Health Care*. 1999;11:407–412.
14. Kwon A, Bungay KM, Pei Y, et al. Antidepressant use: concordance between self-report and claims records. *Med Care*. 2003;41:368–374.
15. Dendukuri N, McCusker J, Bellavance F, et al. Comparing the validity of different sources of information on emergency department visits: a latent class analysis. *Med Care*. 2005;43:266–275.
16. Raina P, Torrance-Rynard V, Wong M, et al. Agreement between self-reported and routinely collected health-care utilization data among seniors. *Health Serv Res*. 2002;37:751–774.
17. Ungar WJ, Coyte PC, Pharmacy Medication Monitoring Program Advisory Board. Health services utilization reporting in respiratory patients. *J Clin Epidemiol.* 1998;51:1335–1342.
18. Kahn KL, Liu H, Adams JL, et al. Methodological challenges associated with patient responses to follow-up longitudinal surveys regarding quality of care. *Health Serv Res*. 2003;38:1579–1598.
19. Gerbert B, Stone G, Stulbarg M, et al. Agreement among physician assessment methods: searching for the truth among fallible methods. *Med Care*. 1988;26:519–535.
20. Shrout PE, Spitzer RL, Fleiss JL. Quantification of agreement in psychiatric diagnosis revisited. *Arch Gen Psychiatry*. 1987;44:172–177.
21. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol*. 1987;126:161–169.
22. SAS Institute Inc. SAS, Version 9.1. Cary, NC: SAS Institute, 2004.
23. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. London: Chapman & Hall; 1993.
24. Streiner DL, Norman GR. *Health Measurement Scales*, 4th ed. Oxford: Oxford University Press; 1994.
25. Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall; 1991.
26. Ware J, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med Care*. 1996;34:220–233.
27. Rohrbaugh M, Rogers JC. What did the doctor do? When physicians and patients disagree. *Arch Fam Med*. 1994;3:125–129.
28. Sox HC. Probability theory and the interpretation of diagnostic tests. In: Sox H, ed. *Common Diagnostic Tests: Use and Interpretation*. Philadelphia, PA: American College of Physicians; 1990:17–33.
29. 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;84:107–109.
30. Colditz GA, Martin P, Stampfer MJ, et al. Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. *Am J Epidemiol*. 1986;123:894–900.
31. Zuckerman AE, Starfield B, Hochreiter C, et al. Validating the content of pediatric outpatient medical records by means of tape-recording doctor-patient encounters. *Pediatrics*. 1975;56:407–411.
32. Romm FJ, Putnam SM. The validity of the medical record. *Med Care*. 1981;19:310–315.
33. Sawyer JA, Earp J, Fletcher RH, et al. Accuracy of women's self report of their last Pap smear. *Am J Public Health*. 1989;79:1036–1037.
34. Starfield B, Simborg D, Johns C, et al. Coordination of care and its relationship to continuity and medical records. *Med Care*. 1977;15:929–938.