Title

Does the concordance between medical records and patient self-report vary with patient characteristics?

Journal

Health Services and Outcomes Research Methodology, 6(3-4)

ISSN

1387-3741

Authors

Tisnado, Diana M
Adams, John L
Liu, Honghu
et al.

Publication Date

2006-12-01

DOI

10.1007/s10742-006-0012-1

Peer reviewed

Health Serv Outcomes Res Method (2006) 6:157–175
DOI 10.1007/s10742-006-0012-1

Does the concordance between medical records and patient self-report vary with patient characteristics?

**Diana M. Tisnado · John L. Adams · Honghu Liu
Cheryl L. Damberg · Fang Ashlee Hu · Wen-Pin Chen
David M. Carlisle · Carol M. Mangione · Katherine L. Kahn**

Received: 13 January 2006 / Revised: 2 September 2006 /
Accepted: 5 September 2006 /
Published online: 23 November 2006

Abstract Few studies of the concordance between patient self-report and medical record data have examined how concordance varies with patient characteristics, and results of such studies have been mixed. Given discrepancies in the quality of care

D. M. Tisnado (✉) · H. Liu · F. A. Hu · W.-P. Chen · D. M. Carlisle · C. M. Mangione · K. L. Kahn
Division of GIM and HSR, Department of Medicine, University of California at Los Angeles, 911 Broxton Plaza, Box 951736, Los Angeles, CA 90095-1736, USA
e-mail: dtisnado@mednet.ucla.edu

H. Liu
e-mail: hhliu@mednet.ucla.edu

F. A. Hu
e-mail: fhu@mednet.ucla.edu

W.-P. Chen
e-mail: wenpinc@uci.edu

C. M. Mangione
e-mail: cmangione@mednet.ucla.edu

K. L. Kahn
e-mail: kkahn@mednet.ucla.edu

J. L. Adams · C. L. Damberg · K. L. Kahn
RAND, 1776 Main Street, Santa Monica, CA 90407-2138, USA
e-mail: John_Adams@rand.org

C. L. Damberg
Pacific Business Group on Health, San Francisco, CA, USA
e-mail: Cheryl_Damberg@rand.org

D. M. Carlisle
California Office of Statewide Health Planning and Development, 1600 Ninth St., Rm, 433, Sacramento, CA 95814, USA
e-mail: dcarlisl@oshpd.state.ca.us

received across patient cohorts, it is important to understand the degree to which concordance metrics are robust across patient characteristics.We hypothesized that concordance between ambulatory medical record and patient survey data varies by patient demographic characteristics, especially education, income, and race/ethnicity. We present the results of bivariate and multivariate analyses including data from 1,270 patients with at least one of: diabetes, ischemic heart disease, asthma or COPD, or low back pain sampled from 39 West Coast medical organizations.We present total agreement, kappa, and survey sensitivity and specificity, stratified by patient demographic and health status characteristics. We conducted logistic regressions to test the impact of patient demographic characteristics, domain of medical care, and health status on these three measures of concordance. Survey sensitivity varied significantly by race/ethnicity in bivariate analyses, but this effect was erased in multivariate analyses. Our findings do not support the hypothesis that patient education, income, or race/ethnicity have an independent effect on concordance when controlling for other factors. However, concordance varied significantly by patient health status. The medical record and patient self-report do not measure quality comparably across patient cohorts. We recommend continued efforts to improve survey data collection across different patient populations and to improve the quality of clinical data.

Keywords Health services research · Quality measurement · Ambulatory care

1 Introduction

Quality of care measurement must be accurate for purposes of public reporting, pay-for-performance, and quality improvement. The data source used to inform quality is one of the key determinants of quality measurement. The ideal data source would allow: assessments of under-use as well as over-use in a target population; comparisons of care that can be delivered in different settings or by different provider types in different organizations; assessments of technical and interpersonal aspects of care; assessments of patient-centered measures such as barriers to access and satisfaction; and adequate risk adjustment (Siu et al. 1991). Studies of concordance across data sources indicate that no one data source has all of the characteristics desired to inform quality measurement.

Although the medical record is frequently viewed as the preferred data source, it is generally viewed as too costly for routine quality assessment. Patient self-report data, on the other hand, is more economical to collect and may provide data on experiences and perspectives not routinely captured by the medical record, but is subject to error due to problems with recall, social desirability bias, and patient health knowledge (Andersen et al. 1979; Sudman and Bradburn 1974).

In order to provide accurate survey data, respondents must understand what information is being asked of them, be able to recall the information, and be motivated to report it accurately (Cannell 1965), which may be affected by the salience of the issue for the patient (Madow 1967). Patient performance of these tasks could be expected to vary according to demographic and health status characteristics due to differences in cognitive function, salience of the health topics, and health knowledge by patient age, socioeconomic status, and health status.

Given the vast discrepancies in quality care received across patient cohorts, one important metric is the degree to which concordance metrics are robust across patient characteristics, including age, health status, race/ethnicity, education and income.

In a review of the literature from 1986 to 2006, 12 of 30 US articles on the topic of the concordance between different data sources examined the influence of one or more patient characteristics (age, gender, race/ethnicity, education, income, health status, utilization) on concordance, with mixed results (Bush et al. 1989; Flocke and Stange 2004; Klein et al. 1986; Ritter et al. 2001; Roberts et al. 1996; Rohrbaugh and Rogers 1994; Rozario et al. 2004; Sawyer et al. 1989; Wallihan et al. 1999; Brown and Adams 1992; Katz et al. 1996; Linet et al. 1989).

Concordance between data sources has been found by some studies to vary by patient age (Roberts et al. 1996; Wallihan et al. 1999; Brown and Adams 1992), gender (Brown and Adams 1992; Linet et al. 1989), race/ethnicity (Rohrbaugh and Rogers 1994; Linet et al. 1989), education (Katz et al. 1996), income level (Roberts et al. 1996), health status or severity of study condition (Klein et al. 1986; Rozario et al. 2004), and utilization rates (Roberts et al. 1996; Rozario et al. 2004; Wallihan et al. 1999). Other studies have found no associations between concordance and patient age (Flocke and Stange 2004; Ritter et al. 2001; Rozario et al. 2004; Sawyer et al. 1989; Katz et al. 1996), gender (Bush et al. 1989; Flocke and Stange 2004; Ritter et al. 2001; Rozario et al. 2004), race/ethnicity (Rozario et al. 2004), education (Flocke and Stange 2004; Ritter et al. 2001; Rozario et al. 2004; Sawyer et al. 1989), income level (Sawyer et al. 1989; Wallihan et al. 1999), or health status (Flocke and Stange 2004; Ritter et al. 2001; Wallihan et al. 1999).

Most of these studies are subject to important limitations. Some (Ritter et al. 2001; Roberts et al. 1996; Rozario et al. 2004; Wallihan et al. 1999) examined only concordance on utilization (numbers of health care system encounters), one based in a single HMO with electronic medical records (Ritter et al. 2001), and one based on a convenience sample of depressed elders from a single clinic (Rozario et al. 2004). Others were limited to patient cohorts with single conditions, such as diabetic retinopathy (Klein et al. 1986), depression (Rozario et al. 2004), conditions affecting urinary function (Roberts et al. 1996), and chronic lymphocytic leukemia (Linet et al. 1989). Many were based on small, convenience samples associated with one site of care (Bush et al. 1989; Rohrbaugh and Rogers 1994; Rozario et al. 2004; Katz et al. 1996).

This study uses a clinically detailed data set from patients with one of four common chronic diseases associated with 39 medical organizations in three West Coast states to (1) apply a method for aggregating data from multiple items across the spectrum of medical care to calculate overall measures of concordance, sensitivity and specificity, and (2) assess how the overall concordance between patient self-report and medical record data is influenced by patient health status and demographic characteristics. We hypothesized concordance between patient self-report and medical record data would vary according to patient characteristics, with patient health education, income, and White race being positively associated with concordance. We present results of bivariate and multivariate analyses, and implications for quality of care assessment efforts.

2 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 managed care patients with diabetes, asthma or COPD, or ischemic heart disease enrolled in physician organizations (POs) located in three West Coast states. The study was approved by the UCLA Institutional Review Board (IRB). Study design and survey results are described elsewhere (Kahn et al. 2003).

For this work evaluating the concordance between data sources, we examined data from a 1998 patient survey and medical record review of all visits that took place within 30 months prior to the survey. We selected equivalent items from both data sources from a pool of items that had been used to construct explicit process of care measures as part of the larger study. Items selected addressed a range of disease-specific and generic topics across the spectrum of care pertinent to patients with chronic disease.

We analyzed the presence or absence of 50 items representing diagnoses, clinical services, counseling and referrals, and medication use in each data source. Items were included only if both the medical record and patient survey instruments recorded patient-level data in comparable time periods. Due to difficulties assessing concordance when prevalence is very low or very high (Shrout et al. 1987), items were included only if the prevalence of the item as measured by both data sources was between 10 and 90%.

2.1 Data collection methods

In 1996, the PBGH collected survey data from 30,308 adults from California, Washington and Oregon who received care in the prior year from one of 60 physician organizations. In 1998, we surveyed 3,656 patients who had responded to the baseline survey in 1996 indicating that they had at least one of the four study conditions (response rate 63%). The mailed, self-administered survey queried patients about diagnoses and health care services received over a 2-year period. Each survey included a disease-specific section to assess processes of care for chronic conditions reported at baseline. Along with the 1998 mailing, subjects also received an invitation to participate in medical record abstraction and IRB-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. Nurses 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 records of all visits with all key health care providers, including records of primary care providers and key specialists for the study conditions noted in the claims/encounter data provided by participating medical organizations. Records of encounters newly discovered during abstraction though not previously noted by claims/encounter data were also located and abstracted.

In all, complete medical records were abstracted for 1,270 patient survey respondents. A total of 698 patients' records were not abstracted or were only partially abstracted due to medical practice closures, inability to locate records, 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 medical 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.

2.2 Analyses

Items were grouped according to type of medical care or service into four conceptual domains: *diagnosis*, *clinical services delivered*, *medication use*, and *counseling and referrals*. The *diagnosis* domain comprises items that represent patient history of diagnoses or medical conditions. *Clinical services delivered* includes health services patients receive such as physical examination, surgical procedures or special tests. *Medication use* represents medications the patient was using at the time of the 1998 survey. *Counseling and referrals* refers to items representing (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. Due to poor performance of both the patient self-report and medical record as data sources for counseling and referrals (Tisnado et al. 2006), neither data source could be considered a gold standard for this domain. Therefore, 10 items from this domain were excluded from this analysis.

To analyze concordance between the two data sources, we calculated both the percent total agreement (percent agreement on positives plus negatives) and the kappa statistic, which corrects for chance agreement, to evaluate agreement between data sources at the item-level and overall. Based on the hypothesis that the medical record is the gold standard data source for patient diagnoses, clinical services received, and medications used, we calculated the sensitivity (% true positives detected) and specificity (% true negatives detected) of the patient survey using the medical record 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 two 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. In other words, we aggregated all dyads into a single 2×2 table to calculate the overall concordance, sensitivity and specificity metrics. Since patients may be eligible for multiple items per domain, unique patients could be represented multiple times in these analyses. Item-level and domain-level analyses have been reported elsewhere (Tisnado et al. 2006).

Independent variables of interest included variables representing patient demographics (age, gender, education, income, race/ethnicity), domain of medical care (diagnoses, clinical services delivered, medication use), and six measures of health status (self-reported health status, comorbidity count, study disease severity, medication count, body mass index (BMI), and visit count).

We determined age group, gender, race/ethnicity, education level, income level, and self-reported health status from 1996 patient self-report. Self-reported health status was measured using the SF-12 (Ware et al. 1996). The SF-12 variable was divided by 10 in multivariate analyses to predict the effect of a 10 point change in SF score on the outcome variables. Comorbidity count was based on a count of up to 39 patient comorbidities noted by either patient self-report or medical record review. Items eligible for scoring a point in the comorbidity index include: cardiovascular problems; cerebrovascular disease; cancer; diabetes; chronic lung disease; common ambulatory problems; depression; measures of functional impairment; and habits associated with medical problems. Scores representing the severity of the patient's study disease (coronary heart disease, lung disease (asthma or emphysema), or

diabetes) were calculated in a disease-specific manner defined to be independent of use of services. To test the validity of the comorbidity and staging systems, we checked the relationships between the comorbidity and staging scores and the construct of burden of illness as measured by the number of drug categories the patient used (Kahn et al. in press). Medication count was determined from a count of medications the patient reported using. BMI was calculated using the medical record report of patient weight and height. Visit count was determined from a count of outpatient visits with a clinician at which vital signs were documented in the medical record.

The dichotomous outcome variables were in the following way. To calculate agreement, the cohort was defined to include all patient-item dyads. Each patient-item dyad was associated with a patient self-report and a medical record report. Agreement was classified as 1 if the patient and medical record both reported Yes or both reported No (Yes–Yes or No–No). To define sensitivity of the patient self-report, the cohort was defined to include all dyads for whom the medical record reported Yes. If the patient also reported Yes, the outcome variable was coded as 1 (true positive). For the model predicting specificity, the cohort was defined to include those dyads for whom the medical record reported No or No data. Of those, the dyads for whom the patient also reported No or No data were classified as 1, (true negative). This approach allows us to utilize stratified Chi Squares to test for bivariate associations, and logistic regression modeling to predict the odds of the patient and medical record being in agreement, and the odds of the patient providing a true positive and a true negative. For example, if the medical record indicated that an eligible patient had a diabetic foot exam during the study time window, this would be classified as a 1 (true positive) if the patient self-report agreed, or 0 (false negative) if the patient disagreed.

We used the Test for Equal Kappa Coefficients to test for differences in kappa, and stratified Chi Squares to test for differences in sensitivity and specificity by patient characteristics (age, gender, income, education, race/ethnicity, health status, number of visits) and domain in medical care (SAS Institute Inc. 2004).

We conducted multivariate logistic regression analyses to test the impact of patient characteristics and domain of medical care on three measures of concordance: agreement, survey sensitivity (self-report of true positive), and survey specificity (self-report of true negative), controlling for variables representing patient demographics (age, gender, education, income, race/ethnicity) domain of medical care, and six measures representing health status (self-reported health status, comorbidity count, study disease severity, medication count, BMI, and visit count). Because our previous work indicates that concordance varies with the domain of medical care in question (Tisnado et al. 2006), dummy variables representing domain of care were included as control variables in the regression models. Regression analyses were adjusted for clustering of observations within patient using the Huber correction (Statcorp. 2003).

3 Results

Table 1 presents the characteristics of the 1,270 patients.

Bivariate results are presented in Table 2, with total agreement, kappa, sensitivity and specificity for all items combined, stratified by variables representing patient demographics, domain of medical care, and health status.

Table 1 Sample characteristics

	N	%
Age group		
< 65	789	62.1
65+	481	37.9
Gender		
Male	583	45.9
Female	687	54.1
Race/ethnicity		
White	1,008	79.4
Black	39	3.1
Asian	71	5.6
Hispanic	103	8.1
Other/missing race	49	3.9
Education		
< High school	562	44.3
≥ High school	708	55.8
Income		
≤ $30,000	371	29.2
> $30,000	899	70.8
Body mass index		
Not obese (BMI < 30)	838	66.0
Obese (BMI ≥ 30)	432	34.0
	Mean (SD)	Range
SF-12	41.7 (11.4)	13.8–65.6
Comorbidity count	7.6 (3.4)	1–26
Severity index	0.51 (0.34)	0–1
Medication count	4.1 (3.0)	0–21
Visit count	7.3 (6.4)	0–57

Overall, total agreement was 83%, and kappa was 0.6. In bivariate analyses, no significant associations were found by gender, education or income. We found kappa was significantly and positively associated with three measures of health status: self-reported health status, medication count, and visit count ($P < 0.01$). Sensitivity was positively associated with White and Asian patient race (as compared with Black, Hispanic, or Other). Sensitivity was also positively associated with lower disease severity, obesity, higher medication count, and lower visit count.

Specificity was positively associated with higher age, better self-reported health status, lower comorbidity count, lower medication count, and higher visit count.

3.1 Multivariate results

Table 3 presents the results of the multivariate models for each of the three outcome variables: agreement, sensitivity (odds of reporting a true positive) and specificity (odds of reporting a true negative). We controlled for patient demographic characteristics (age, gender, education, income, race/ethnicity), domain of medical care, and six measures representing health status, in all three models.

Table 2 Measures of concordance for all items combined, by patient and medical organization characteristics

Item	% Agreement	Kappa	MR as gold standard	
			Sensitivity	Specificity
All items combined	83.2	0.61	73.9	87.5
Demographics				
Age				*
Age < 65	83.1	0.60	74.2	86.9
Age 65+	83.3	0.63	73.4	88.5
Gender				
Female	83.2	0.61	73.6	87.6
Male	83.2	0.62	74.2	87.5
Education				
≤ 12 years	82.6	0.61	72.7	87.6
> 12 years	83.7	0.62	75.0	87.4
Income				
≤ $30K	82.2	0.60	72.3	87.2
> $30K	83.6	0.62	74.6	87.7
Race/ethnicity			*	
Black	81.8	0.58	69.8	87.6
Hispanic	82.3	0.58	68.0	88.8
Asian	85.4	0.63	75.3	89.2
Other/missing	81.4	0.56	68.8	86.8
White	83.3	0.62	74.8	87.3
Health status				
SF-12		*		*
Lowest quartile	80.7	0.59	73.0	85.3
2nd quartile	82.1	0.59	73.0	86.3
3rd quartile	85.1	0.65	75.7	89.2
Highest quartile	85.0	0.63	74.2	89.2
Comorbidity count				*
Lowest quartile	85.8	0.61	75.2	88.9
2nd quartile	84.2	0.62	75.3	87.9
3rd quartile	82.4	0.61	73.7	86.8
Highest quartile	80.3	0.59	72.3	85.9
Severity index			*	
Lowest quartile	84.8	0.59	73.9	88.0
2nd quartile	83.6	0.62	76.9	86.5
3rd quartile	82.4	0.62	73.3	87.6
Highest quartile	81.7	0.61	71.5	88.1
Medication count		*	*	*
Lowest quartile	83.0	0.49	59.6	89.5
2nd quartile	84.1	0.60	72.4	88.3
3rd quartile	83.0	0.62	74.8	87.3
Highest quartile	82.6	0.64	78.6	85.4
BMI			*	
Not obese	83.6	0.60	72.8	87.9
Obese	82.6	0.62	75.4	86.8
Visit count		*	*	*
Lowest quartile	83.9	0.56	76.5	85.9
2nd quartile	84.1	0.61	75.1	87.5
3rd quartile	83.1	0.62	74.2	87.8
Highest quartile	82.0	0.62	71.8	88.9
Domain of care		*	*	*
Clinical services	82.3	0.59	72.0	86.9
Diagnoses	85.0	0.64	68.1	93.0
Medication use	82.2	0.60	78.1	84.1

*$P < 0.01$

Table 3 Multivariate analyses predicting total agreement, sensitivity and specificity of self-report

Predictors	1. % Agreement		2. PSR sensitivity (MR = standard)		3. PSR specificity (MR = standard)	
	n = 25,801 (Area under ROC curve = 0.56)		*n* = 8,159 (Area under ROC curve = 0.62)		*n* = 17,642 (Area under ROC curve = 0.63)	
	Odds ratio	*P*-value	Odds ratio	*P*-value	Odds ratio	*P*-value
Demographics						
Age 65+	1.09	0.02	1.00	0.86	1.23	0.00
Female	1.01	0.88	0.92	0.14	1.02	0.68
Black	0.95	0.59	0.85	0.29	1.01	0.96
Hispanic	0.92	0.23	0.82	0.05	1.06	0.52
Asian	1.08	0.32	0.97	0.82	1.08	0.46
Other or missing race	0.90	0.21	0.81	0.09	0.95	0.64
Education (HS Grad)	1.01	0.78	1.04	0.53	0.97	0.52
Income 30 (inc > 30K)	1.07	0.11	1.10	0.13	1.03	0.59
Health status						
Self-reported health	1.07	0.00	1.02	0.55	1.13	0.00
Comorbidity count	0.96	0.00	0.96	0.00	0.98	0.00
Severity score	0.91	0.10	0.88	0.20	1.11	0.19
Medication count	1.01	0.08	1.13	0.00	0.96	0.00
Obesity	1.01	0.88	1.11	0.06	0.95	0.36
Visit count	1.00	0.36	0.98	0.00	1.03	0.00
Domain of care						
Diagnosis domain	0.96	0.42	1.36	0.00	0.77	0.00
Medication domain	1.20	0.00	0.77	0.01	1.97	0.00

Taken together, the variables representing patient demographics were not significant in the prediction of agreement in any of the multivariate models. However, the set of variables representing health status made a significant contribution to the prediction of agreement in all three models (Wald Chi2 test. $P < 0.0001$) (Stata-corp. 2003). Results for each model are described in further detail below. The statistic representing the area under the ROC curve, and the adjusted odds ratios and *P*-values associated with each independent variable are shown for each model in Table 3.

3.1.1 Modeling survey and medical record agreement

Column 1 presents the model predicting agreement (versus lack of agreement) between items measured by patient self-report and medical records.

Two individual measures of health status: patient self-reported health status and comorbidity count, were significantly associated with the likelihood that the two data sources would agree. Individuals with better self-reported health status had higher odds of agreement with the medical record as compared with those with lower self-reported health (OR = 1.07, $P < 0.001$). Individuals with higher comorbidity counts had lower odds of agreement with the medical record (OR = 0.96, $P < 0.001$).

In addition, self-report of items from the *medication use* domain was associated with higher odds of agreement as compared with the reference domain (*clinical services delivered)* (OR = 1.20, $P < 0.001$).

3.1.2 Modeling survey sensitivity

We modeled survey sensitivity, or the odds of the patient self-reporting a true positive, with the medical record report as the gold standard (Column 2). An odds ratio of less than one indicates patient under-reporting, whereas an odds ratio of greater than one indicates higher odds of the patient reporting a positive consistent with the medical record (i.e., the presence of a diagnosis, a procedure having taken place, according to the medical record) as compared with the reference group.

No patient demographic characteristics were found to be significant.

However, three of the six measures of health status were significant predictors. Individuals with a higher comorbidity count were found to be slightly more likely to under-report as compared with those with lower comorbidity counts (OR = 0.96, $P < 0.001$). Those with more visits were similarly more likely to under-report items as compared with those with fewer visits (OR = 0.98, $P < 0.001$). Individuals using more medications were more likely to report medications documented in the medical record than those with fewer numbers of medications (OR = 1.13, $P < 0.001$).

Significant associations were also found with the domain of care. Patients were more likely correctly report positives matching the medical record on items associated with the *diagnosis* domain as compared with the clinical services domain (OR = 1.36, $P < 0.001$). Items from the *medication* domain were associated with patient under-report (OR = 0.77, $P < 0.01$).

Although patients of Hispanic ethnicity or black race were observed in the unadjusted analyses to be more likely to under-report as compared with Whites, these effects were erased in the multivariate analyses controlling for all other factors including enriched measures of health status.

3.1.3 Modeling survey specificity

We modeled survey specificity, or the likelihood of the patient self-reporting a true negative, with the medical record as the gold standard (Column 3). An odds ratio of less than one indicates patient over-reporting, whereas an odds ratio of greater than one indicates higher odds of the patient reporting a negative consistent with the medical record (i.e., documentation of ''no'' or no data regarding diagnoses or services according to the medical record) as compared with the reference group.

We found four of the six measures of health status were associated with survey specificity. Better self-reported health status was associated with higher odds of the patient reporting a negative consistent with the medical record (OR = 1.13, $P < 0.001$). Similarly, higher comorbidity count (OR = 0.98, $P < 0.001$) and medication count (OR = 0.96, $P < 0.001$) were associated with patient over-report (lower

odds of specificity). More visits was associated with slightly higher odds of the patient report matching a medical record negative.

In addition, we found that patient age greater than 65 was associated with higher odds of the patient self-report correctly matching a medical record negative as compared with younger age (OR = 1.23, $P < 0.001$).

Domain of medical care was again significantly associated with the outcome, with patients more likely to over-report items from the diagnosis domain as compared with the clinical services domain (OR = 0.78, $P < 0.001$). Patient self-reports were more likely to match a medical record negative for items from the medication domain as compared with the clinical services domain (OR = 1.97, $P < 0.001$).

4 Discussion

Few studies have examined how concordance varies with patient characteristics. Our main hypothesis was that concordance would vary by patient demographic characteristics. After controlling for patient demographics including age, gender, education, income, and race/ethnicity, as well as for health status and domain of medical care, we found no significant differences in the odds of agreement, patient over-report or a patient under-report by patient demographics. However, we did find significant differences by patient health status.

Patient education and income were not significantly associated with any of the measures of concordance in bivariate or multivariate analyses. Differences in sensitivity by race/ethnicity were observed in unadjusted, bivariate analyses. However, these differences were erased when we controlled for six measures representing health status in the multivariate analysis. Four of these six measures were significant in one or more of the three multivariate models, suggesting that differences in sensitivity of the patient self-report by race/ethnicity are explained by differences in health status.

Of the few previous studies to address the topic in the last 20 years, there have been mixed results regarding the relationship between health status and concordance (Flocke and Stange 2004; Klein et al. 1986; Ritter et al. 2001; Rozario et al. 2004; Wallihan et al. 1999). More serious conditions have been found to be more accurately reported by patients (Katz et al. 1996), supporting the hypothesis that greater salience is associated with more accurate self-report. Two other studies involved specific diseases: one found more severe diabetes to be positively associated with the sensitivity of patient self-report of diabetic eye exam (Klein et al. 1986); the other found lower concordance with greater severity of depression (Rozario et al. 2004), likely due to issues of cognitive functioning and motivation. Others have found no relationship (Flocke and Stange 2004; Ritter et al. 2001; Wallihan et al. 1999).

Health status has been postulated to affect concordance in several ways. Sicker patients may find that their health conditions and encounters with the healthcare system have greater salience to them, facilitating recall (Madow 1967). However, in the extreme, poor health could be associated with confusion if there are too many diagnoses, services, and medications to remember, as well as cognitive difficulties and fatigue affecting respondent burden and thus recall (Andersen et al. 1979). At

the other extreme, the very healthy may have fewer issues to track and therefore fewer to get wrong.

Simultaneously controlling for six measures representing health status, we found that patients with better self-assessed health status had higher odds of agreement with the medical record, and higher odds of accurate self-report of true negatives (i.e., less over-report). Higher comorbidity count was associated with less patient accuracy in terms of both under-report and over-report. In contrast, a higher medication count was associated with a more accurate report of positive events but also with a higher likelihood of over-reporting. The difference may have to do with the salience to patients of conditions for which one must take medication as compared with a count of all comorbidities, each of which may have varying degrees of impact on the patient.

Patients with higher visit counts were less likely to accurately report a true positive (more likely to under-report) and were more likely to accurately report a true negative (less over-report), possibly due to difficulties recalling individual events as numbers of visits and event increases.

Motivation has also been hypothesized to affect accurate reporting. Motivation may be related to the degree of embarrassment or emotional threat posed by an item (Cannell 1965). Cannell (1965) proposed a set of conditions hypothesized to be associated with a high degree of threat, including conditions such as cancer, mental illness, STDs, and issues affecting the prostate or breast. This study included few items classified by Cannell as highly threatening. Potentially threatening items in this study include cancer, depressive symptoms, anti-depressant use, weight, and possibly narcotic use.

These hypotheses focus on the effects of health status on the accuracy of the patient self-report. In addition, patient health status likely interacts with the quality of medical record data. Physicians of patients with multiple medical problems may be more likely to focus their recording in the chart on acute issues while omitting details about other, perhaps more long-standing problems. Patients and providers may hold different definitions of health conditions, and may differ in their perceptions of the most salient issues, which could result in patient under-or over-report compared with the medical record.

This work is subject to some limitations. Each of the models presented was analyzed with a different sample size due to the differing number of eligible events for each. Differing sample sizes may explain lack of consistency in significance levels of predictors across the three models (i.e., age).

Calculations of patient over-report may be biased upward if records of medical encounters were missed during medical record abstraction. Medical record abstractors pursued records of all visits with all key health care providers, including records of primary care providers and key specialists for the study conditions noted in claims/encounter data provided by participating medical organizations. In addition, we used a snowballing approach, whereby evidence of encounters not previously noted by claims/encounter data but newly discovered during abstraction triggered additional chart pursuit and abstraction. Following this strategy allowed us to minimize the chances of missing records of care obtained outside of a particular medical organization.

Non-response bias may limit the generalizability of these findings. Non-responders in 1998 who had responded in 1996 were more likely to be non-white, less educated, and to have lower self-reported health status than

responders. In addition, the self-administered survey was only available in English. Therefore, these findings may not be representative of limited-English speaking individuals.

Although the medical record is frequently viewed as the preferred data source, it is subject to error due to sparse recording of certain topics such as counseling (Stange et al. 1998), failure to include orders, laboratory and procedure reports in the chart, and delayed recording resulting in physician recall problems (Luck et al. 2000). Moreover, medical record errors are not only errors of omission. Luck found that the medical record both under-reported *and over-reported* care delivered. Medical record accuracy may also be affected by the setting of care, with differences in time pressures, continuity, and coordination of care, and systems such as integrated medical records or electronic medical records (Luck et al. 2000). However, data for this study were drawn from the commonly used data sources of patient self-report and medical records pertinent to thousands of encounters with physicians across three states, making use of a ''true'' gold standard such as direct observation impracticable.

The absolute magnitude of the differences in concordance are significant but small. It is not clear what effect differences in kappa, sensitivity, and specificity of this magnitude would have on the results of large group-level assessments of quality of care.

The effect these differences in sensitivity and specificity would have on actual quality of care scores would depend on the prevalence of the disease state or indicated medical service in question. The resulting net under-report of medical services delivered (i.e., misclassification of numerators) could artificially depress quality of care scores while a net over-report could have the opposite effect. Net under-report of conditions triggering patient eligibility for quality assessment (i.e., misclassification of denominators) could affect quality of care scores in either direction, while a net over-report could artificially depress quality of care scores.

In an example of satisfying a quality indicator, such as beta blocker use by patients with coronary disease, consider an HMO with a true rate of 98% (as reported by a Los Angeles health plan using a hybrid method of both medical records and self-report for case identification). Our concordance findings suggest that self-report would result in a rate of 76.5%. But all health plans using self-report would be similarly biased downward, not necessarily changing rankings significantly. However the findings in our analysis suggest that an HMO with patients clustered around the 3rd quartile of medication use would have a self-report rate of approximately 75.7% while a plan with a population clustering around the 1st quartile of medication use would have a self-report rate of approximately 66.9% despite the plans' true rates being equal. In a consumer report card, this would put the plan with the latter population at a significant disadvantage. The California Office of the Patient Advocate presents performance measures for California HMO's on their web site (http://www.opa.ca.gov) by market for consumer use. Although the beta blocker after heart attack HEDIS measure reported there is not a precise match for the beta blocker measure studied here, it does give a range for comparison. The difference in self-report scores calculated here (75.7–66.9 = 8.8) would be substantially larger than the range of true beta blocker scores for Los Angeles HMOs. A plan performing as well as the best plan in Los

Angeles despite a much different population could easily have the worst score in Los Angeles by self-report.

5 Conclusion

In our previous work, we concluded that neither the survey nor the medical record alone provides sufficiently complete data across all aspects of care for optimal quality of care measurement, although the adequacy of a given data source ultimately depends on the research topic in question. Based on the findings presented in this paper, we conclude that small but significant variation in concordance exists across patients of different health status. We caution consumers, payors, as well as researchers collecting data from populations with wide variations in burden of illness, that concordance between patient self-report and medical records documentation varies with health status sufficient to impact quality of care scores.

To truly understand disparities in health status and quality of care, more robust measures are needed. The true sources of the variations in concordance by patient characteristics, and the proportions attributable to true differences as compared with survey or medical record measurement error, cannot be determined from these data. Some of the potential sources of inconsistencies between data sources were mentioned above.

Many have pointed out how automated, electronic clinical information systems could improve the quality of health care delivery as well as our ability to measure quality through many mechanisms including prompts and reminder systems; enhanced information sharing, communications, and coordination; and data capture across multiple providers and settings of care (Committee on Quality Healthcare in America, Institutes of Medicine 2001; Schneider et al. 1999).

However, lack of concordance between the medical record and patient self-report due to true differences in doctor and patient understanding, perspective, or poor doctor–patient communication, will persist despite attempts to perfect data collection methods. It is possible that only research using alternative, gold standard data collection methods (e.g., direct observation) can provide information on which research can base adjustments to quality scores.

If quality assessment and accountability efforts are to succeed in achieving health care system improvement, robust measurement is essential. Scientists involved in quality measurement must make every effort to maximize the accuracy of our methods. The stakes are high financially and politically, in terms of credibility and buy-in of clinicians, health care organizations, consumers, and policymakers, and in terms of achieving the ultimate goal of realizing meaningful improvements in quality of care.

Acknowledgements This work was 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.

Appendix I Measures of concordance for items by domain

Domain	Prevalence by data source				Measures of concordance		MR = Gold standard		PSR = Gold standard	
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 exam ϕ	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
Medication use										
Theophylline*	22	17	5	23	93	0.8	71.4	99	94	93

Appendix I continued

Domain	Prevalence by data source				Measures of concordance		MR = Gold standard		PSR = Gold standard	
Item	MR only	PSR only	MR–PSR	MR, PSR, or both	% Total agreement	Kappa	SE	SP	SE	SP
Anti-depressant*	12	13	–1	17	92	0.6	69.5	95	66	96
Glitazone*	15	10	5	17	91	0.6	52.5	98	84	92
Statin or other lipid lowering*	30	30	0	35	90	0.8	83.2	92	82	93
Long-acting nitrate*	20	10	10	25	89	0.6	47.1	100	100	88
Beta blocker*	37	31	6	45	89	0.8	78.2	96	92	88
Long-acting beta agonist*	25	18	7	27	88	0.6	60.8	97	86	88
Inhaled steroid*	48	48	0	54	87	0.7	86.9	87	86	88
Angiotensin-converting enzyme inhibitor*	31	26	5	35	87	0.7	69.4	94	85	87
Atrovent*	25	16	9	28	86	0.6	52.5	97	84	86
Insulin*	15	26	–11	28	85	0.6	86.7	85	52	97
Calcium channel blocker*	41	29	12	48	85	0.7	67.1	98	96	81
Narcotic*	15	17	–2	24	84	0.4	51.2	90	46	91
Sulfonylurea*	59	51	8	64	82	0.6	77.5	87	90	73
Short-acting beta agonist*	61	65	–4	73	81	0.6	87.8	70	82	78
Metformin*	41	33	8	47	81	0.6	66.9	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	15.6	98	79	75
NSAID*	25	20	5	35	74	0.3	37.5	86	47	81
Counseling and referrals[α]										
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

Appendix I continued

Domain	Prevalence by data source				Measures of concordance		MR = Gold standard		PSR = Gold standard	
Item	MR only	PSR only	MR–PSR	MR, PSR, or both	% Total agreement	Kappa	SE	SP	SE	SP
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

MR, medical record; PSR, patient self-report; SE, sensitivity; SP, specificity

*Survey item queried about current time period (do you now have or are you currently using...)

**Survey item queried about last 12 months

***Survey item queried about last 2 years or ''ever''

ϕ indicates that survey item asked about each of several time periods, which were lumped together to represent the last 2 years

α Note that items from the counseling and referrals domain were excluded from the current analysis

Reproduced from Tisnado et al. 2006

References

Andersen, R.M., Kasper, J., Frankel, M.R., et al.: Total Survey Error: Applications to Improve Health Surveys. Jossey Bass, San Francisco (1979)

Brown, J.B., Adams, M.E.: Patients as reliable reporters of medical care process. Med. Care **30**(5), 400–411 (1992)

Bush, T.L., Miller, S.R., Golden, A.L., et al.: Self reported and medical record report agreement of selected medical conditions in the elderly. Am. J. Public Health **79**(11), 1554–1556 (1989)

Cannell, C.F.: Reporting hospitalization for the Health Interview Survey. National Center for Health Statistics, Series 2, No. 6. US Department of Health, Education and Welfare, Washington, D.C (1965)

Committee on Quality Healthcare in America, Institutes of Medicine: Crossing the Quality Chasm: A New Health System for the 21st Century. National Academy Press, Washington, D.C (2001)

Flocke, S.A., Stange, K.C.: Direct observation and patient recall of health behavior advice. Prev. Med. **38**, 343–349 (2004)

Gerbert, B., Stone, G., Stulbarg, M., et al.: Agreement among physician assessment methods: searching for the truth among fallible methods. Med. Care **26**, 519–535 (1988)

Kahn, K.L., Tisnado, D.M., Adams, J.L., Liu, H., Chen, W.P., Hu, F.A., Mangione, C.M., Hays, R.D., Damberg, C.L.: Does ambulatory process of care predict health-related quality of life outcomes? Health Serv. Res. (in press)

Kahn, K.L., Liu, H., Adams, J.L., et al.: Methodological challenges associated with patient responses to follow-up longitudinal surveys regarding quality of care. Health Serv. Res. **38**, 1579–1598 (2003)

Katz, N., Chang, L.C., Sangha, O., Fossel, A.H., Bates, D.W.: Can comorbidity be measured by questionnaire rather than medical record review? Med. Care **34**(1), 73–84 (1996)

Klein, R., Klein, B., Moss, S.E., et al.: The validity of a survey question to study diabetic retinopathy. Am. J. Epidemiol. **124**, 104–110 (1986)

Linet, M.S., Harlow, S.D., Mclaughlin, J.K., McCaffrey, L.D.: A comparison of interview data and medical records for previous medical conditions and surgery. J. Clin. Epidemiol. **42**(12), 1207–1213 (1989)

Luck, J., Peabody, J.W., Dresselhaus, T.R., et al.: How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. Am. J. Med. **108**, 642–649 (2000)

Madow, W.G.: Interview data on chronic conditions compared with information derived from medical records. National Center for Health Statistics, Series 2, No. 23. US Department of Health, Education and Welfare, Washington, D.C (1967)

Ritter, P.L., Stewart, A.L., Kaymaz, H., et al.: Self-reports of health care utilization compared to provider records. J. Clin. Epidemiol. **54**, 136–141 (2001)

Roberts, R., Bergstralh, E.J., Schmidt, L., et al.: Comparison of self-reported and medical record health care utilization measures. J. Clin. Epidemiol. **49**(9), 989–995 (1996)

Rohrbaugh, M., Rogers, J.C.: What did the doctor do? When physicians and patients disagree. Arch. Fam. Med. **3**, 125–129 (1994)

Rozario, P.A., Morrow-Howell, N., Proctor, E.: Comparing the congruency of self-report and provider records of depressed elders' service use by provider type. Med. Care **42**, 952–959 (2004)

SAS Institute Inc.: SAS/STAT 9.1 User's Guide, pp. 1431–1557. SAS Institute Inc., Gary, NC (2004)

Sawyer, J.A., Earp, J., Fletcher, R.H., et al.: Accuracy of women's self report of their last Pap smear. Am. J. Public Health **79**, 1036–1037 (1989)

Schneider, E.C., Riehl, V., Courte-Wienecke, S., et al.: Enhancing performance measurement: NCQA's road map for a health information framework. JAMA **282**(12), 1184–1190 (1999)

Shrout, P.E., Spitzer, R.L., Fleiss, J.L.: Quantification of agreement in psychiatric diagnosis revisited. Arch. Gen. Psychiatry **44**, 172–177 (1987)

Siu, A.L., McGlynn, E.A., Morgenstern, H., et al.: A fair approach to comparing quality of care. Health Aff. **10**(1), 62–75 (1991)

Stange, K.C., Zyzanski, S.J., 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 **36**, 851–867 (1998)

Statacorp.: STATA 8 Reference G-M, vol. 2, pp. 298–299. Stata press, College Station, TX (2003)

Statacorp.: STATA 8 User's Guide, pp. 263–267. Stata press, College Station, TX (2003)

Sudman, S., Bradburn, N.M.: Response Effects in Surveys. Adeline, Hawthorne (1974)

Tisnado, D.M., Adams, J.L., Liu, H., Damberg, C., Chen, W.P., Hu, F.A., Carlisle, D.M., Mangione, C.M., Kahn, K.L.: What is the concordance between the medical record and patient self-report as data sources for ambulatory care? Med. Care **44**(2), 132–140 (2006)
Wallihan, D.B., Stump, T.E., Callahan, C.M.: Accuracy of self-reported health services use and patterns of care among urban older adults. Med. Care **37**, 662–670 (1999)
Ware, J., Kosinski, M., Keller, S.D.: A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med. Care **34**, 220–233 (1996)