Title

Methods for Assessing Patient—Clinician Communication about Depression in Primary Care: What You See Depends on How You Look

Journal

Health Services Research, 49(5)

ISSN

0017-9124

Authors

Henry, Stephen G
Feng, Bo
Franks, Peter
et al.

Publication Date

2014-10-01

DOI

10.1111/1475-6773.12187

Peer reviewed

HSR Health Services Research

DOI: 10.1111/1475-6773.12187
METHODS ARTICLE

Methods for Assessing Patient–Clinician Communication about Depression in Primary Care: What You See Depends on How You Look

Stephen G. Henry, Bo Feng, Peter Franks, Robert A. Bell, Daniel J. Tancredi, Dustin Gottfeld, and Richard L. Kravitz

Objective. To advance research on depression communication and treatment by comparing assessments of communication about depression from patient report, clinician report, and chart review to assessments from transcripts.

Data. One hundred sixty-four primary care visits from seven health care systems (2010–2011).

Study Design. Presence or absence of discussion about depressive symptoms, treatment recommendations, and follow-up was measured using patient and clinician post-visit questionnaires, chart review, and coding of audio transcripts. Sensitivity and specificity of indirect measures compared to transcripts were calculated.

Principal Findings. Patient report was sensitive for mood (83 percent) and sleep (83 percent) but not suicide (55 percent). Patient report was specific for suicide (86 percent) but not for other symptoms (44–75 percent). Clinician report was sensitive for all symptoms (83–98 percent) and specific for sleep, memory, and suicide (80–87 percent), but not for other symptoms (45–48 percent). Chart review was not sensitive for symptoms (50–73 percent), but it was specific for sleep, memory, and suicide (88–96 percent). All indirect measures had low sensitivity for treatment recommendations (patient report: 24–42 percent, clinician report 38–50 percent, chart review 49–67 percent) but high specificity (89–96 percent). For definite follow-up plans, all three indirect measures were sensitive (82–96 percent) but not specific (40–57 percent).

Conclusions. Clinician report and chart review generally had the most favorable sensitivity and specificity for measuring discussion of depressive symptoms and treatment recommendations, respectively.

Key Words. Patient–clinician communication, depression, primary care, quality

Primary care clinicians diagnose and manage most of the approximately 13 million Americans affected by major depression each year (Kessler, Merikangas, and Wang 2007). Improving the process of depression care delivered during primary care visits is thus an important part of improving depression

care overall (Donabedian 1979). However, the concordance of different methods for assessing processes of care in depression has received scant attention.

Evaluation and management of depression depend largely on communication-oriented (rather than procedure-oriented) activities that occur during primary care visits. Diagnostic accuracy depends on the history that clinicians obtain and is an important part of quality in depression care (Peabody et al. 2004). Managing depression requires evaluating patients' depressive symptoms and recommending treatment (e.g., medication, mental health referral, or close follow-up) when appropriate (Trangle et al. 2012). Accurately assessing the content and process of communication about depression is important for evaluating diagnostic accuracy, measuring the appropriateness of care, understanding how patients and clinicians make decisions about depression, and improving patient–clinician communication about depression.

Direct observation (usually from audio or video recordings) of patient-clinician communication has limitations (Henry and Fetters 2012) but is generally considered the most accurate method for assessing the content of communication during clinic visits (Donabedian 1988; Hrisos et al. 2009). However, researchers usually rely on one of three indirect measures of communication —patient report, clinician report, or medical chart review—because direct observation is resource intensive and may be considered invasive (Makary 2013). Patient and clinician reports can be collected using postvisit questionnaires and are optimal for assessing participants' subjective experiences. However, these reports are subject to bias and the limitations of human memory (Hertwig, Fanselow, and Hoffrage 2003; Smith, Brown, and Ubel 2008). Medical charts can be examined retrospectively but contain only data that clinicians chose to document, which may be influenced by billing requirements.

A recent systematic review found that, compared to direct observation, patient report is more accurate than chart review for assessing communication

---

Address correspondence to Stephen G. Henry, M.D., Division of General Medicine, Geriatrics, and Bioethics, 4150 V Street, Suite 2400, Sacramento, CA 95817; e-mail: sghenry@ucdavis.edu. Bo Feng, Ph.D., is with the Department of Communication, University of California Davis, Davis, CA. Peter Franks, M.D., is with the Department of Family and Community Medicine, Center for Healthcare Policy and Research, University of California Davis, Sacramento, CA. Robert A. Bell, Ph.D., is with the Departments of Communication and Public Health Sciences, Center for Healthcare Policy and Research, University of California Davis, Davis, CA. Daniel J. Tancredi, Ph.D., is with the Department of Pediatrics, Center for Healthcare Policy and Research, University of California Davis, Sacramento, CA. Dustin Gottfeld, B.S., is with the Center for Healthcare Policy and Research, University of California Davis, Sacramento, CA. Richard L. Kravitz, M.D., M.S.P.H., is with the Division of General Medicine, Geriatrics, and Bioethics, Center for Healthcare Policy and Research, University of California Davis, Sacramento, CA.

behaviors (e.g., dietary counseling), chart review is more accurate than patient report for assessing procedural aspects of care (e.g., medication dosing), and clinician report is understudied (Hrisos et al. 2009). Almost all published studies in this area have focused on clinician behaviors related to preventive services and behavioral counseling (Gilchrist et al. 2004; Shaikh et al. 2012; Stange et al. 1998b). We know of no prior studies that have compared direct and indirect measures of communication about depression in primary care, though one study found that discussion of depression (measured by audio recording) differed by patient race (Ghods et al. 2008). We found only one prior study that simultaneously compared data from direct observation against all three indirect measures (patient report, clinician report, and chart review); that study focused on management of chronic obstructive pulmonary disease (Gerbert and Hargreaves 1986).

In this study we compared patient report, clinician report, and chart review against direct observation for three aspects of depression care: discussion of symptoms, treatment recommendations, and follow-up plans. Assessing depressive symptoms and treatment recommendations are important for evaluating diagnostic accuracy and evidence-based treatment, respectively. Scheduled follow-up visits are important for assessing the effectiveness of new depression treatments and for monitoring patients when the need for treatment is uncertain. We analyzed data collected during a clinical trial in primary care that included patients with a wide range of depressive symptoms. We included patients with minimal depressive symptoms because evaluating patients with potential depression comprises a substantial component of overall depression care in primary care settings and because both over diagnosis and under diagnosis of depression are potential threats to high-quality care.

Data on the sensitivity and specificity of indirect measures of communication about depression are particularly important for researchers seeking valid ways to assess depression communication and treatment. Our results are also likely to be useful for administrators and quality improvement experts interested in selecting methods for assessing the quality and appropriateness of depression treatment.

METHODS

Study Design

Data were taken from a multisite clinical trial comparing two interventions to enhance depression-related patient engagement (a depression engagement

video and an interactive multimedia computer program) and a control arm (a video on sleep hygiene). Primary care providers and adult patients were recruited from seven clinical sites in Northern California. Sites included one health maintenance organization, two academic primary care clinics, two multispecialty group practices, and two Veterans Affairs service areas. Patient inclusion criteria included the ability to speak English, ability to use a computer, and having a scheduled appointment with a participating primary care provider. Patients taking medications for depression were excluded. One-to-two weeks before their scheduled visit, patients were screened for depressive symptoms by telephone using the PHQ-8 scale, which comprises all but the suicidal ideation item from the Patient Health Questionnaire (PHQ-9). The PHQ-8 and PHQ-9 have similar psychometric properties and are both well-established scales for measuring depression severity (Kroenke, Spitzer, and Williams 2001; Kroenke and Spitzer 2002; Kroenke et al. 2009). The study protocol over-sampled patients with clinically significant depressive symptoms. Nondepressed patients were included to evaluate the possibility of intervention-induced overtreatment. Full details of the parent study, which was approved by the institutional review boards of all participating institutions, have been previously published (Kravitz et al. 2013; Tancredi et al. 2013).

Immediately before their appointment, patients completed a computer-based questionnaire that included demographic questions and the PHQ-9. Patients were then randomized and received either the depression engagement video, the interactive multimedia computer program, or the sleep hygiene video (control) prior to their appointment. Clinicians were told that the study goal was to improve communication about common symptoms in primary care and were blind to patient arm assignment. Clinicians and patients were asked for permission to audio record their visits. When both agreed, a research assistant placed a digital audio recorder in the exam room at the start of the visit and then waited outside until the visit ended.

Patient and Clinician Report

Patients and clinicians both completed postvisit questionnaires after the visit that included questions about whether the patient and clinician discussed the following five depressive symptoms: (a) quantity or quality of sleep; (b) difficulty performing usual activities; (c) trouble with memory or concentration; (d) patient's mood, emotions, or feelings; and (e) thoughts of suicide or self-harm. We included these criteria because other criteria were either very rarely documented by clinicians (i.e., guilt, psychomotor changes) or insufficiently

specific to depression (i.e., fatigue, appetite change) (Feldman et al. 2006). These questionnaires also asked whether the clinician recommended (a) medication for depression, (b) referral to a mental health professional, and/or (c) some other depression treatment (patients only). Finally, both questionnaires asked whether the clinician recommended a definite follow-up visit.

Chart Review

Investigators developed a chart abstraction tool to identify depression-related processes of care documented in the medical chart. Two investigators blind to the results of transcript coding, patient report, and clinician report (RLK and DG) independently applied the abstraction tool to 30 charts and then discussed any discrepancies to clarify definitions and abstraction procedures. One investigator (DG) then abstracted data from the remaining charts. Another physician independently over-read and confirmed data abstracted from each chart.

The abstraction tool identified documentation of whether the clinician and patient discussed any of the same five depressive symptoms and three treatment recommendations that were measured by patient and clinician report. It also identified whether the clinician recommended a definite follow-up visit. At the time of data collection, all but one of the seven clinic sites used electronic medical records.

Transcript Coding

Recorded visits were transcribed verbatim, and a coding scheme was developed to collect the same variables that were collected by patient report, clinician report, and chart review. The scheme comprised a series of items that were coded as 1 when the behavior was present in the transcript and 0 when absent. The complete coding scheme is available on request. Two research assistants blind to the study design underwent approximately 20 hours of training to learn the coding scheme.

To facilitate comparison of sensitivity and specificity among patients with different levels of depressive symptoms, we first stratified patients into three groups based on their PHQ-8 screening scores: patients with no depressive symptoms (score 0–4), patients with mild symptoms (score 5–9), and patients with moderate-to-severe symptoms (score 10–24). Two coders blind to the study purpose then applied the transcript coding scheme to a random sample of 55 audio-recorded visits from each group. Coders then coded the

transcripts in three stages. First, both coders independently coded a random sample of 35 transcripts. Next, they independently coded 45 additional transcripts, of which 25 were coded by both. Finally, the remaining transcripts were split between the two coders. This three-stage process allowed for coders and investigators to evaluate inter-coder reliability and discuss discrepancies after the first and second stages to reinforce coder training and answer coders' questions. This process also prevented low reliability due to attrition of coder skill. We calculated final intercoder reliability using the 60 transcripts coded by both coders. Intercoder reliability was high (item–level interclass correlation coefficients ranged from 0.79 to 0.95). Disagreements were resolved through discussion. Coding was performed with NVivo 8 (QSR International, Doncaster, Vic., Australia). One transcript involved a married couple seen together and so was dropped from our sample.

The exact wording used to assess each variable across all four measurement sources is available online (Appendix SA1).

Data Analysis

We analyzed data from the 164 visits to which the validation coding scheme was applied. Analysis focused on five variables measuring discussion of depressive symptoms (i.e., sleep, activity performance, memory/concentration, mood/emotions/feelings, and suicide), three variables measuring treatment recommendations (i.e., medication prescription, mental health referral, and some other treatment), and one variable measuring whether a definite follow-up visit was recommended. All variables were analyzed as dichotomous variables indicating whether the specific topic or recommendation was present or absent during the visit. The prevalences of specific types of depression treatment other than medication and mental health referral were too low to analyze independently. Therefore, instead of specifically analyzing these other types of treatments, we created a composite variable indicating whether the clinician recommended any depression treatment (anti-depressant medication, mental health referral, and/or some other depression treatment). Our main analysis thus included three variables related to treatment: recommendations for medication, for mental health referral, and for any depression treatment.

We calculated the prevalence, sensitivity, specificity, positive predictive value, and negative predictive value for patient report, clinician report, and chart review using transcript coding as the reference (Dickinson et al. 2010). Sensitivity indicates the percentage of transcripts in which a topic is coded as present that are correctly identified as present by an indirect measure.

Specificity indicates the percentage of transcripts in which a topic is coded as absent that are correctly identified as absent by an indirect measure. Positive and negative predictive values indicate the probability that the presence or absence, respectively, of a topic as assessed by an indirect measure agrees with the transcript coding. Our analysis followed the recommendations from a recent review of statistical methods in this area (Dickinson et al. 2010). We calculated 95 percent confidence intervals using robust standard errors to account for patients being clustered within clinicians (Genders et al. 2012).

We performed several exploratory subgroup analyses to investigate whether any of the following factors influenced the sensitivity or specificity of indirect measures: the severity of depressive symptoms (i.e., none, mild, or moderate-to-severe) as measured by patients' PHQ-9 scores, patient race (i.e., white versus nonwhite patients), patient–clinician racial concordance (i.e., visits in which the patient and clinician reported the same versus different race from among the five categories listed in Table 1), and patients' randomization assignment. For each subgroup, we calculated sensitivity and specificity for the three variables we considered to be most critical in managing depression: discussion of mood or emotions, discussion of suicide, and recommendation for any depression treatment.

Clinician report contained a small amount of missing data (6–9 percent) for each question, which we did not impute. Analyses were performed using Stata 12.1 (StataCorp, College Station, TX, USA).

Results

The sample for this study comprised 164 visits involving 54 physicians and two nurse practitioners. The mean number of patients per clinician was 2.9 (median = 3, range 1–9). Table 1 shows participant characteristics. There were no meaningful differences in sensitivity or specificity among the two intervention and control groups; randomization assignment will thus not be discussed further.

Table 2 compares the prevalence of different topics across the four measurement sources. The prevalence of discussions about mood and suicide was similar across all four sources. In contrast, discussions about sleep were substantially less prevalent when measured by chart review (40 percent) compared to the other three sources (66–68 percent). The prevalence of discussions about both memory/concentration and activity performance was substantially lower when measured by chart review or transcript compared to

Table 1: Participant Demographics

	Patients (n = 164)	*Clinicians (n = 56)*
Mean age, years (SD)	52.8 (11.3)	45.3 (9.0)
Female gender, %	54.3	55.8
Race, %		
White, non-Hispanic	67.1	55.8
Black, non-Hispanic	12.2	1.9
Hispanic	10.4	7.7
Asian/Pacific Islander	6.7	30.8
Other/mixed race	3.7	3.9
Annual household income, %		
<$20,000	18.9	
$20,000–35,000	13.4	
$35,000–75,000	27.4	
$75,000–125,000	22.6	
>$125,000	17.7	
Education, %		
High school diploma/GED	12.8	
Some technical school	0.6	
Graduated technical school	3.1	
Some college	40.2	
College graduate	22.0	
Postgraduate degree	21.3	
Depressive symptom severity, *n* (%)		
None (PHQ-9 score 0–4)	53 (32.3)	
Mild (PHQ-9 score 5–9)	55 (33.5)	
Moderate to severe (PHQ-9 score >9)	56 (34.2)	
Clinic type, *n* (%)		
Health maintenance organization	16 (9.8)	6 (10.7)
Multispecialty group practice	73 (44.5)	18 (32.1)
Academic primary care clinic	53 (32.3)	26 (46.3)
Veterans Affairs health system	22 (13.4)	6 (10.7)
Years in current practice, median (interquartile range)		11 (5–16)

patient and clinician report. Recommendations for specific treatments were similar across all four sources, though the sample prevalence was consistently highest when measured by transcript coding and lowest when measured by clinician or patient report. The prevalence of whether a definite follow-up visit was recommended was much lower when measured by transcript coding (48 percent) compared to other sources (62–78 percent).

Patient Report versus Transcript

Table 3 shows the sensitivity, specificity, positive predictive value, and negative predictive value of variables measured by patient report compared to

Table 2: Prevalence of Specific Topics by Measurement Source; % (95% Confidence Intervals)

Topic	*Transcript (n = 164)*	*Patient Report (n = 164)*	*Physician Report (n = 155)*	*Chart Review (n = 164)*
Sleep	65.9 (56.5, 74.1)	67.7 (60.8, 73.9)	67.1 (58.3, 74.9)	40.2 (30.8, 50.5)
Mood	64.0 (54.6, 72.4)	73.2 (65.8, 79.4)	82.1 (71.5, 89.3)	61.6 (50.5, 71.6)
Memory/ concentration	13.4 (9.0, 19.5)	29.9 (23.8, 36.8)	28.0 (19.4, 38.6)	10.4 (6.2, 16.8)
Activity performance	30.5 (22.9, 39.3)	52.4 (45.0, 59.8)	61.7 (50.9, 71.4)	33.5 (26.2, 41.7)
Thoughts of suicide	26.8 (20.2, 34.8)	25.0 (18.7, 32.5)	35.9 (27.6, 45.2)	25.0 (18.0, 33.6)
Recommended medication	22.6 (16.8, 29.6)	11.0 (7.0, 16.8)	14.1 (9.3, 20.7)	17.7 (12.4, 24.5)
Recommended referral	26.2 (20.8, 32.5)	17.7 (12.0, 25.3)	16.7 (11.2, 24.0)	19.5 (14.3, 26.1)
Recommended any depression treatment	36.6 (29.5, 44.3)	22.0 (15.7, 29.8)	25.0 (18.5, 32.8)	31.1 (24.2, 38.9)
Recommended definite follow-up visit	48.2 (39.0, 57.5)	69.5 (61.1, 76.8)	77.6 (68.5, 84.6)	62.2 (52.8, 70.7)

Table 3: Sensitivity, Specificity, and Positive and Negative Predictive Values of Patient Report Versus Transcript

Topic	*n*	*Sensitivity % (95% CI)*	*Specificity % (95% CI)*	*Positive Predictive Value*	*Negative Predictive Value*
Sleep	164	83.3 (73.6, 90.0)	62.5 (50.9, 72.8)	81.1	66.0
Mood	164	82.9 (74.3, 89.0)	44.1 (32.0, 56.9)	72.5	59.1
Memory/concentration	164	59.1 (37.7, 77.5)	74.6 (66.8, 81.1)	26.5	92.2
Activity performance	164	72.0 (58.4, 82.5)	56.1 (47.6, 64.3)	41.9	82.1
Thoughts of suicide	164	54.5 (38.1, 70.1)	85.8 (77.6, 91.4)	58.5	83.7
Recommended medication	164	24.3 (12.6, 41.7)	92.9 (86.5, 96.4)	50.0	80.8
Recommended referral	164	41.9 (26.3, 59.2)	90.9 (84.0, 95.0)	62.1	81.5
Recommended any depression treatment	164	40.0 (27.0, 54.5)	88.5 (81.4, 93.1)	66.7	71.9
Recommended definite follow-up visit	164	87.3 (76.6, 93.6)	47.1 (36.4, 58.0)	60.5	80.0

transcript coding. For variables measuring depressive symptoms, patient report had good sensitivity for mood (83 percent), sleep (83 percent), and activity performance (72 percent) but lower specificity (44–63 percent). For

Table 4: Sensitivity, Specificity, and Positive and Negative Predictive Values of Clinician Report Versus Transcript

Topic	*n*	*Sensitivity % (95% CI)*	*Specificity % (95% CI)*	*Positive Predictive Value*	*Negative Predictive Value*
Sleep	155	93.1 (83.4, 97.3)	81.5 (67.5, 90.3)	90.4	86.3
Mood	156	97.0 (91.6, 99.0)	44.6 (27.9, 62.7)	75.8	89.3
Memory/concentration	150	84.2 (62.2, 94.5)	80.2 (71.0, 87.0)	38.1	97.2
Activity performance	154	86.4 (66.7, 95.2)	48.2 (36.7, 59.8)	40.0	89.8
Thoughts of suicide	153	97.6 (84.7, 99.7)	87.4 (79.0, 92.7)	74.5	99.0
Recommended medication	156	38.2 (22.8, 56.5)	92.6 (87.3, 95.8)	59.1	84.3
Recommended referral	156	50.0 (33.8, 66.2)	95.6 (89.8, 98.2)	80.8	83.8
Recommended any depression treatment	156	50.0 (35.8, 64.2)	89.0 (81.4, 93.7)	71.8	76.1
Recommended definite follow-up visit	156	96.0 (88.7, 98.7)	39.5 (28.2, 52.0)	59.5	91.4

discussions of memory/concentration and suicide, patient report had lower sensitivity (55–59 percent) but higher specificity (75–86 percent). Patient report of treatment recommendations also had low sensitivity (24–42 percent) and high specificity (89–93 percent). Patient report of whether a definite follow-up visit was recommended had high sensitivity (87 percent) but low specificity (47 percent).

Clinician Report versus Transcript

Table 4 shows the results of our main analysis for clinician report. For variables measuring depressive symptoms, clinician report had high sensitivity (84–98 percent) across all five topics. Specificity was high for discussions of sleep (82 percent), memory/concentration (80 percent), and suicide (87 percent) but lower for mood (45 percent) and activity performance (48 percent). Clinician report of treatment recommendations had high specificity (89–96 percent) and low sensitivity (38–50 percent). Clinician report of whether they recommended a definite follow-up visit also had high sensitivity (96 percent) and low specificity (40 percent).

Chart Review versus Transcript

Table 5 shows the results of our main analysis for chart review. For depressive symptoms, chart review had relatively low sensitivity (50–73 percent) across

Table 5: Sensitivity, Specificity, and Positive and Negative Predictive Values of Chart Review Versus Transcript

Topic	*n*	*Sensitivity % (95% CI)*	*Specificity % (95% CI)*	*Positive Predictive Value*	*Negative Predictive Value*
Sleep	164	54.6 (42.9, 65.8)	87.5 (76.2, 93.9)	89.4	50.0
Mood	164	73.3 (62.8, 81.7)	59.3 (41.8, 74.8)	76.2	55.6
Memory/concentration	164	50.0 (31.4, 68.6)	95.8 (89.9, 98.3)	64.7	92.5
Activity performance	164	50.0 (36.2, 63.8)	73.7 (63.3, 82.0)	45.5	77.1
Thoughts of suicide	164	68.2 (51.0, 81.5)	90.8 (83.6, 95.1)	73.2	88.6
Recommended medication	164	48.6 (34.2, 63.3)	91.3 (84.0, 95.5)	62.1	85.9
Recommended referral	164	58.1 (44.5, 70.6)	94.2 (88.5, 97.2)	78.1	86.4
Recommended any depression treatment	164	66.7 (55.8, 76.0)	89.4 (81.9, 94.1)	78.4	82.3
Recommended definite follow-up visit	164	82.3 (71.6, 89.5)	56.5 (44.0, 68.2)	63.7	77.4

all five topics. Chart review had relatively high specificity for discussions of sleep (88 percent), concentration (96 percent), and suicide (91 percent) but it was less specific for discussions of mood (59 percent) and activity performance (74 percent). For treatment recommendations, chart review had high specificity (89–94 percent) but low sensitivity (49–67 percent). Documentation of plans for a definite follow-up visit had high sensitivity (82 percent) but low specificity (57 percent).

Subgroup Analyses

We found no meaningful differences in sensitivity or specificity across subgroups defined by depressive symptom severity levels. Estimates for patients with no symptoms were imprecise because very few clinicians recommended treatments for this group. Similarly, we found no meaningful differences in sensitivity or specificity for white versus nonwhite patients or for racially concordant versus racially discordant patient–clinician dyads. Results of these subgroup analyses are shown in Online Appendix SA2.

DISCUSSION

In this study, we compared measurement of communication about depression using patient report, clinician report, and chart review against direct

observation as coded from written transcripts. Results differed for variables measuring depressive symptoms, treatment recommendations, and follow-up.

For variables measuring discussion of *symptoms*, clinician report consistently had the highest sensitivity relative to coded transcripts and chart review had the lowest. Sensitivity and specificity also differed among the five symptoms we evaluated. Indirect measures of discussion of memory/concentration and activity performance had lower sensitivity and specificity than indirect measures for discussion of sleep, mood, and suicide. One possible explanation for this finding is that patients and clinicians are more likely to perceive sleep, mood, and suicide as medically relevant and/or salient and so are more likely to recall or document discussion of these symptoms accurately. However, patients substantially under-reported discussions about suicide compared to other indirect sources, perhaps because suicide is a stigmatized topic often associated with awkward communication (Vannoy et al. 2011). In contrast, clinician report for discussions about suicide had high sensitivity and specificity, likely due to the high saliency of suicide for clinicians. These overall results for measuring discussion of symptoms are consistent with findings from prior studies that found chart review to be less sensitive than patient report for measuring clinician counseling behavior (Shaikh et al. 2012; Stange et al. 1998b; Wilson and McDonald 1994). Our finding that clinician report has high sensitivity and specificity is also consistent with the sparse literature in this area (Gerbert and Hargreaves 1986; Gilchrist et al. 2004).

For variables measuring *treatment recommendations*, all three indirect measures had high specificity (89–96 percent), suggesting that over-reporting is rare. Chart review had the highest sensitivity, while patient report had the lowest. Slight differences in wording among questionnaires may have contributed to the observed low sensitivity. Transcript coders considered a treatment recommendation present if the clinician "recommended or suggested" a treatment. In contrast, patients were asked whether the clinician "recommended" a treatment, and clinicians were asked whether they "prescribed" a medication or "referred" the patient (see Online Appendix SA1). Few prior studies have specifically examined the accuracy of indirect measures of treatment recommendations outside of preventive services; in two prior studies that examined clinician referrals, all three indirect measures had high specificity and low sensitivity (Gilchrist et al. 2004; Stange et al. 1998a).

In contrast with treatment, all three indirect measures of whether a *definite follow-up visit* was recommended had relatively high sensitivity (82–96 percent) and low specificity (40–57 percent) compared to transcript coding. One likely explanation for this unexpected finding is that patients and

clinicians sometimes discuss or schedule follow-up visits after the patient leaves the exam room. In these cases, using a transcript as the reference would underestimate the proportion of visits in which a definite follow-up visit was recommended.

Our study advances the existing literature on measuring patient–clinician communication in several ways. Most prior studies of sensitivity and specificity have focused on clinician behavioral counseling and/or provision of preventive services (Hrisos et al. 2009). Our study is the first to examine the accuracy of indirect measures of communication about depression care. In addition, few prior studies have examined the accuracy of clinician report or compared all three indirect measures against direct observation.

Researchers studying communication about depression can use our findings to guide choices about the optimal method for measuring communication. Researchers who want to assess patient–clinician communication should consider using clinician report when direct observation is not feasible. Compared to other indirect measures, clinician report has reasonably high sensitivity for measuring the topics discussed during visits. Clinician report also has high specificity and only slightly lower sensitivity than chart review for measuring treatment recommendations. One limitation of using clinician report is that clinicians may be unwilling to complete lengthy questionnaires during busy clinics. Researchers who are primarily interested in assessing treatment recommendations should consider using chart review. Our findings can also inform efforts to assess and improve the quality and appropriateness of depression treatment. Additional research is needed to investigate the extent to which our findings about depression generalize to communication about other common problems in primary care. One implication of our study is that patients and clinicians should be encouraged to participate in research involving audio recording for topics (such as depression) for which content and processes of communication are critical to the quality and appropriateness of care.

Our study has several limitations. We did not measure the processes of depression communication (e.g., affective dimension of communication, supportive comments, patient–clinician negotiation) which may be as important as communication content for assessing the quality of depression care. Our findings may be subject to selection bias if sensitivity or specificity differs for participants who agreed to be recorded compared to those who did not. We were also unable to assess the effect of clinic site given the large number of clinics relative to our sample size. Finally, we did not examine whether the electronic medical records at different sites may have influenced the accuracy

of chart review. Features of the electronic medical record such as automatic templates related to depression or "cut-and-paste" behaviors have the potential to both increase the sensitivity and decrease the specificity of documentation of depression care relative to handwritten charts (Shaikh et al. 2012). Evaluating how electronic medical records influence sensitivity and specificity would require detailed site-specific information, including the kinds of depression-related templates available, whether and how clinicians use these templates, and the prevalence of "cut-and-paste" behaviors.

Direct observation is generally considered to be the most accurate method for measuring patient–clinician communication and communication-related processes of care. However, our findings related to follow-up visits suggest some limitations of direct observation. Furthermore, our measure of direct observation used transcripts, which lack the vocalic and nonverbal cues that are preserved when communication is coded directly from audio or video recordings, respectively. In addition, talk is sometimes inaudible, messy, or inarticulate (Sidnell 2010; Waitzkin 1990), which contributes to discrepancies among what patients or clinicians say, what they think they are saying, what their interaction partners hear, and what third-party coders infer from transcripts. The relative sensitivity and specificity of different indirect measures of communication may vary in contexts other than depression, and researchers may also have specific reasons to prefer one or another data source depending on the outcome of interest. For example, patient report will play an important role in studies that involve patient-centered outcomes or patient behaviors (e.g., treatment adherence) even when patient report of communication content is not highly sensitive or specific relative to written transcripts.

Despite these considerations, valid assessment of patient–clinician communication about depression remains an important first step for researchers who want to measure and improve the quality of depression care. Accurate assessments of communication are necessary for evaluating the completeness of diagnostic appraisal and the extent to which treatment recommendations are evidence-based. Valid information about diagnostic accuracy and the appropriateness of treatment recommendations is, in turn, necessary for health care policy makers and payers to link reimbursement to quality of care in ways that are both related to the quality measure of interest and acceptable to clinicians and patients. Findings in this study should be useful not only for researchers studying the quality of depression care but also for researchers assessing and evaluating patient–clinician communication and quality for other common clinical problems in primary care.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This study was funded by NIMH grant R01 MH079387 to Dr. Kravitz.

Disclosures: None.

Disclaimer: None.

REFERENCES

Dickinson, H. O., S. Hrisos, M. P. Eccles, J. Francis, and M. Johnston. 2010. "Statistical Considerations in a Systematic Review of Proxy Measures of Clinical Behaviour." *Implementation Science* 5: 20

Donabedian, A. 1979. "The Quality of Medical Care: A Concept in Search of a Definition." *Journal of Family Practice* 9 (2): 277–84.

——— 1988. "The Quality of Care: How Can It Be Assessed?" *Journal of the American Medical Association* 260 (12): 1743–8.

Feldman, M. D., P. Franks, R. M. Epstein, C. E. Franz, and R. L. Kravitz. 2006. "Do Patient Requests for Antidepressants Enhance or Hinder Physicians' Evaluation of Depression? A Randomized Controlled Trial." *Medical Care* 44 (12): 1107–13.

Genders, T. S., S. Spronk, T. Stijnen, E. W. Steyerberg, E. Lesaffre, and M. G. Hunink. 2012. "Methods for Calculating Sensitivity and Specificity of Clustered Data: A Tutorial." *Radiology* 265 (3): 910–6.

Gerbert, B., and W. A. Hargreaves. 1986. "Measuring Physician Behavior." *Medical Care* 24 (9): 838–47.

Ghods, B. K., D. L. Roter, D. E. Ford, S. Larson, J. J. Arbelaez, and L. A. Cooper. 2008. "Patient-Physician Communication in the Primary Care Visits of African Americans and Whites with Depression." *Journal of General Internal Medicine* 23 (5): 600–6.

Gilchrist, V. J., K. C. Stange, S. A. Flocke, G. McCord, and C. C. Bourguet. 2004. "A Comparison of the National Ambulatory Medical Care Survey (NAMCS) Measurement Approach with Direct Observation of Outpatient Visits." *Medical Care* 42 (3): 276–80.

Henry, S. G., and M. D. Fetters. 2012. "Video Elicitation Interviews: A Qualitative Research Method for Investigating Physician-Patient Interactions." *Annals of Family Medicine* 10 (2): 118–25.

Hertwig, R., C. Fanselow, and U. Hoffrage. 2003. "Hindsight Bias: How Knowledge and Heuristics Affect our Reconstruction of the Past." *Memory* 11 (4/5): 357–77.

Hrisos, S., M. P. Eccles, J. J. Francis, H. O. Dickinson, E. F. S. Kaner, F. Beyer, and M. Johnston. 2009. "Are There Valid Proxy Measures of Clinical Behaviour? A Systematic Review." *Implementation Science* 4: 37.

Kessler, R. C., K. R. Merikangas, and P. S. Wang. 2007. "Prevalence, Comorbidity, and Service Utilization for Mood Disorders in the United States at the Beginning of the Twenty-First Century." *Annual Review of Clinical Psychology* 3: 137–58.

Kravitz, R. L., P. Franks, M. D. Feldman, D. J. Tancredi, C. A. Slee, R. M. Epstein, P. R. Duberstein, R. A. Bell, M. Jackson-Triche, D. A. Paterniti, C. Cipri, A. M. Iosif, S. Olson, S. Kelly-Reif, A. Hudnut, S. Dvorak, C. Turner, and A. Jerant. 2013. "Patient Engagement Programs for Recognition and Initial Treatment of Depression in Primary Care: A Randomized Trial." *Journal of the American Medical Association* 310 (17): 1818–28.

Kroenke, K., and R. L. Spitzer. 2002. "The PHQ-9: A New Depression Diagnostic and Severity Measure." *Psychiatric Annals* 32 (9): 509–15.

Kroenke, K., R. L. Spitzer, and J. B. Williams. 2001. "The PHQ-9: Validity of a Brief Depression Severity Measure." *Journal of General Internal Medicine* 16 (9): 606–13.

Kroenke, K., T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad. 2009. "The PHQ-8 as a Measure of Current Depression in the General Population." *Journal of Affective Disorders* 114 (1–3): 163–73.

Makary, M. A. 2013. "The Power of Video Recording: Taking Quality to the Next Level." *Journal of the American Medical Association* 309 (15): 1591–2.

Peabody, J. W., J. Luck, S. Jain, D. Bertenthal, and P. Glassman. 2004. "Assessing the Accuracy of Administrative Data in Health Information Systems." *Medical Care* 42 (11): 1066–72.

Shaikh, U., J. Nettiksimmons, R. A. Bell, D. Tancredi, and P. S. Romano. 2012. "Accuracy of Parental Report and Electronic Health Record Documentation as Measures of Diet and Physical Activity Counseling." *Academic Pediatric* 12 (2): 81–7.

Sidnell, J. 2010. *Conversation Analysis: An Introduction.* Chichester, UK/Malden, MA: Wiley-Blackwell.

Smith, D. M., S. L. Brown, and P. A. Ubel. 2008. "Mispredictions and Misrecollections: Challenges for Subjective Outcome Measurement." *Disability and Rehabilitation* 30 (6): 418–24.

Stange, K. C., S. J. Zyzanski, C. R. Jaen, E. J. Callahan, R. B. Kelly, W. R. Gillanders, J. C. Shank, J. Chao, J. H. Medalie, W. L. Miller, B. F. Crabtree, S. A. Flocke, V. J. Gilchrist, D. M. Langa, and M. A. Goodwin. 1998a. "Illuminating the 'Black Box' - A Description of 4454 Patient Visits to 138 Family Physicians." *Journal of Family Practice* 46 (5): 377–89.

Stange, K. C., S. J. Zyzanski, T. F. Smith, R. Kelly, D. M. Langa, S. A. Flocke, and C. R. Jaen. 1998b. "How Valid are Medical Records and Patient Questionnaires for Physician Profiling and Health Services Research? A Comparison with Direct Observation of Patient Visits." *Medical Care* 36 (6): 851–67.

Tancredi, D. J., C. K. Slee, A. Jerant, P. Franks, J. Nettiksimmons, C. Cipri, D. Gottfeld, J. Huerta, M. D. Feldman, M. Jackson-Triche, S. Kelly-Reif, A. Hudnut, S. Olson, J. Shelton, and R. L. Kravitz. 2013. "Targeted Versus Tailored Multimedia Patient Engagement to Enhance Depression Recognition and Treatment in Primary Care: Randomized Controlled Trial Protocol for the AMEP2 Study." *BMC Health Services Research* 13 (1): 141.

Trangle, M., B. Dieperink, T. Gabert, B. Haight, B. Lindvall, J. Mitchell, H. Novak, D. Rich, D. Rossmiller, L. Setterlund, and K. Somers. 2012. *Major Depression in Adults in Primary Care*, pp 119. Bloomington, MN: Institute for Clinical Systems Improvement.

Vannoy, S. D., M. Tai-Seale, P. Duberstein, L. J. Eaton, and M. A. Cook. 2011. "Now What Should I do? Primary Care Physicians' Responses to Older Adults Expressing Thoughts of Suicide." *Journal of General Internal Medicine* 26 (9): 1005–11.

Waitzkin, H. 1990. "On Studying the Discourse of Medical Encounters: A Critique of Quantitative and Qualitative Methods and a Proposal for Reasonable Compromise." *Medical Care* 28 (6): 473–88.

Wilson, A., and P. McDonald. 1994. "Comparison of Patient Questionnaire, Medical Record, and Audio Tape in Assessment of Health Promotion in General Practice Consultations." *British Medical Journal* 309 (6967): 1483–5.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Comparison of Question Wording across Sources.
Appendix SA2: Results of Exploratory Subgroup Analyses.
Appendix SA3: Author Matrix.