

Response to Past Depression Treatments Is Not Accurately Recalled: Comparison of Structured Recall and Patient Health Questionnaire Scores in Medical Records

Gregory E. Simon, MD, MPH; Carolyn M. Rutter, PhD;
Christine Stewart, PhD; Chester Pabiniak, MS; and Linda Wehnes, MA

ABSTRACT

Objective: Assessing response to prior depression treatments is common in research and clinical practice, but few data are available regarding accuracy of patient recall. Data from a population-based survey were linked to electronic medical records to examine agreement between patients' recalled treatment response and depression severity scores in the medical records.

Method: Electronic medical records from a large health system identified 1,878 adult patients with 2 or more episodes of clinician-diagnosed major depressive disorder or dysthymia between January 2005 and December 2009 (diagnoses had been recorded using ICD-9 codes). These potential participants were mailed an invitation letter, and, of these, 578 completed an online or mailed survey including structured recall of response to each prior depression treatment, rating both global improvement during treatment and improvement specifically attributed to treatment. For 269 of the survey participants, at least 1 treatment episode could be unambiguously linked to both pretreatment and posttreatment Patient Health Questionnaire (PHQ-9) depression scores in the electronic medical records. Analyses examined the agreement between patients' recall of treatment response and response according to PHQ-9 scores from the medical records.

Results: Agreement between recall and the medical records was poor for both overall improvement following treatment ($\kappa=0.10$; 95% CI, 0.00–0.19) and improvement attributed to treatment ($\kappa=0.12$; 95% CI, 0.00–0.25). Agreement remained poor when the sample was limited to medication treatment episodes, episodes lasting 3 months or more, or episodes for which the participant was "very sure" of his or her ability to recall. Agreement reached a fair level only for episodes in the 6 months prior to the survey, for both overall improvement ($\kappa=0.23$; 95% CI, 0.08–0.39) and improvement attributed to treatment ($\kappa=0.36$; 95% CI, 0.12–0.59).

Conclusions: Patients' recall of response to past depression treatments agrees poorly with data from medical records. Interview assessment of prior treatment response may not be a useful tool for research or clinical practice.

J Clin Psychiatry 2012;73(12):1503–1508

© Copyright 2012 Physicians Postgraduate Press, Inc.

Submitted: May 7, 2012; accepted August 23, 2012
(doi:10.4088/JCP.12m07883).

Corresponding author: Gregory E. Simon, MD, MPH, Group Health Research Institute, 1730 Minor Ave, #1600, Seattle, WA 98101 (simon.g@ghc.org).

When selecting an initial treatment for depression, patients and providers can choose from several medications and specific psychotherapies. While these various treatment choices have, on average, similar likelihood of success,^{1–4} both the benefits and adverse effects of treatments vary from individual to individual. At this time, we have no evidence-based criteria for selecting specific treatments for individual patients.^{5,6}

Absent accurate predictors of individual response, guidelines typically recommend that providers consider each patient's response to prior treatments.^{1,7} For example, the American Psychiatric Association's *Practice Guideline for the Treatment of Patients With Major Depressive Disorder*⁷ calls for "a psychiatric history, including identification of past symptoms of mania, hypomania, or mixed episodes and responses to previous treatments."^{7(p13)} This recommendation depends on 2 assumptions—that past treatment response predicts future response to the same or similar treatment and that patients can accurately recall responses to past treatments.

Surprisingly, few data are available regarding the accuracy of patients' recall of past treatment response. In a study of 73 outpatients receiving antidepressant treatment, Posternak and Zimmerman⁸ compared recall of prior treatment and treatment response with outpatient records. Patients were able to recall 81% of past medication trials, and patients' recall of treatment response showed moderate to substantial agreement ($\kappa=0.56$) with response documented in outpatient records. We are aware of no other published data regarding accuracy of recall of past depression treatment.

Here we use data from a population-based survey to examine accuracy of recall of response to past depression treatments. Survey data were linked to results from standardized depression questionnaires in electronic medical records.

METHOD

Study Setting

Data were collected to evaluate the feasibility of using patient surveys and electronic medical records to identify predictors of response to specific depression treatments. All participants were members of Group Health Cooperative, a member-owned integrated health system providing general medical and mental health care to approximately 650,000 Washington and Idaho residents. Group Health members are enrolled through a combination of employer-sponsored insurance, individually purchased insurance, Medicare, and Medicaid or other subsidized insurance for low-income residents. Members are generally similar to the area population in distribution of age, socioeconomic status, and race/

- Patients' recall of response to past depression treatments agrees poorly with standardized outcome questionnaires they completed at the time of treatment.
- Recall of treatment response is fair for the preceding 6 months and poor for treatments earlier than the last 6 months.
- Accurate assessment of past treatment response will probably require review of medical records.

ethnicity. All study procedures were reviewed and approved by the Group Health Human Subjects Review Committee.

Study Sample

Electronic medical records were used to identify adult members who had experienced at least 2 episodes of depression treatment (either medications or psychotherapy) between January 2005 (when Group Health's outpatient electronic medical records system was fully implemented) and December 2009. An episode of antidepressant treatment was defined by a filled prescription for an antidepressant, an associated diagnosis of major depressive disorder or dysthymic disorder (clinician diagnoses had been recorded using ICD-9 codes), and no filled prescription for any antidepressant in the prior 270 days. An episode of psychotherapy for depression was defined by an initial psychotherapy visit associated with a diagnosis of major depressive disorder and no psychotherapy visit in the prior 180 days. Individuals with a recorded diagnosis of bipolar disorder or schizophrenia spectrum disorder were excluded, but there were no other exclusions for co-occurring psychiatric, general medical, or substance use disorder diagnoses.

Survey Methods

Each potential participant was mailed an invitation letter including a brief description of study procedures and instructions for completing an online survey. Those unable to complete the survey online were offered a paper survey by mail. Both the mail and online surveys began with a complete description of the study purpose, procedures, potential risks, and right to refuse or withdraw. Each participant provided signed consent (electronically or by paper), including consent to link survey responses to electronic medical records regarding depression treatment.

Questions regarding response to past antidepressant medication treatments began with specific prompts to improve recall ("Your Group Health records show that sometime since 2005 you have taken these medications:" followed by list of medications and initial prescriptions dates). This orienting prompt was followed by questions regarding each specific medication, as follows: "Now we will ask some specific questions about [generic name], also known as [brand name]. In [month and year], you filled a prescription for this medication from [name of prescribing physician]. They were

probably [physical description of medication dispensed] that looked like this [color photographic image of medication dispensed]." Physical descriptions and images of medications were derived from National Drug Code (NDC) codes in prescription records. A second prompt concerning specific symptoms of interest read as follows: "Try to remember how you felt before you started taking [brand and generic name of medication] in [month and year of first prescription]. Think about symptoms of depression or stress, like feeling low or depressed, having no interest in things, feeling tired, feeling guilty or worthless, having trouble sleeping, or having thoughts of death or suicide."

Response to each medication was assessed using 3 Likert-type questions. The first question assessed self-rated global improvement, as follows: "Please rate how much those symptoms or problems improved after you started taking [brand and generic name of medication]," followed by a 7-point response scale ranging from "very much worse" to "very much better" (with an additional option for "cannot recall"). The second question assessed improvement specifically attributable to medication, as follows: "Try to remember how much you thought that [brand and generic name of medication] helped you after you started taking it. Please rate whether the medicine helped or made things worse," followed by a 5-point response scale ranging from "made things very much worse" to "helped very much" (with an additional option for "cannot recall"). The third question assessed confidence of recall, as follows: "How sure are you that you can remember how things changed after you started taking [brand and generic name of medication]?" followed by a 4-point response scale ranging from "very sure" to "not at all sure."

Questions regarding past psychotherapy followed a similar structure. An initial prompt listed the date and provider for the initial visit in each episode of psychotherapy. A second prompt concerned specific symptoms of interest. This prompt was followed by 3 questions regarding each episode of therapy, parallel to those described above regarding antidepressant treatment episodes (details available upon request from the authors).

Medical Records Data

For all potential participants invited to complete the survey, computerized medical records were used to compare survey respondents and nonrespondents in terms of age, sex, treatment history, and imputed race and ethnicity from US Census data.⁹

For all survey respondents, computerized medical records were used to assess outcome of past treatment episodes. Since 2006, all Group Health providers have been encouraged to use the Patient Health Questionnaire (PHQ-9) depression severity questionnaire for initial assessment of depression and at all depression follow-up visits. The PHQ-9 has been a valid and sensitive measure of depression severity across a wide range of patient populations and clinical settings.¹⁰⁻¹² Scores for the PHQ-9 are stored in the electronic record of each outpatient encounter. These electronic medical records

data were used to identify baseline and follow-up or outcome PHQ-9 scores for each treatment episode. The eligibility period for a baseline PHQ-9 score extended from 14 days prior to the episode start date (initial prescription or psychotherapy visit) until 3 days after the start date. If more than 1 eligible baseline PHQ-9 score was identified, then the score closest to the episode start date was selected. The eligibility period for an outcome PHQ-9 score extended from 60 days to 120 days after the episode start date. If more than 1 eligible outcome PHQ-9 score was identified, then the score closest to 90 days after the index date was selected. Approximately half of the episodes with baseline and outcome scores were excluded because PHQ-9 scores could not be linked to a single treatment (ie, the period between the 2 scores included exposure to more than 1 treatment simultaneously or sequentially). The sample was further limited to episodes with a baseline PHQ-9 score of 5 or greater.

Classification of Response

A positive treatment response according to the PHQ-9 was defined as a 50% or greater decrease in total score between the baseline and outcome measure. A positive response according to recall was defined by the 2 highest categories for each measure ("very much improved" or "much improved" for recalled global improvement and "helped very much" or "helped some" for recalled benefit from treatment).

While some individuals had both recall and medical records data for multiple episodes, we included only the most recent episode for each individual. Inclusion of all episodes with complete data led to slightly lower estimates of agreement between medical records and participants' recall (details available upon request from the authors).

Data Analyses

Data analyses proceeded in 3 steps. The initial step compared eligible patients who did and did not participate in the online survey to assess possible bias due to nonresponse. The second step considered all survey respondents, comparing those for whom PHQ-9 depression data (both baseline and outcome) were and were not available in the electronic medical record. The third step considered treatment episodes for which both survey and PHQ-9 data were available, examining agreement between these 2 sources in identifying positive treatment response. The κ statistic¹³ indicated the degree to which agreement exceeded that expected by chance. Traditional criteria consider κ values less than 0.2 to indicate minimal agreement, values of 0.2 to less than 0.4 to indicate fair agreement, and values of 0.4 or greater to indicate moderate or better agreement.¹⁴

RESULTS

Invitation letters were mailed to 1,838 potential participants, and 578 (31%) completed the survey. As shown in Table 1, those responding and not responding to the survey did not differ significantly in distribution of demographic characteristics or treatment history.

Table 1. Comparison of Survey Responders and Nonresponders (N = 1,838)^a

Variable	Responders, n = 578 (31%), n (%)	Nonresponders, n = 1,260 (69%), n (%)
Female sex	424 (73)	913 (72)
Predicted minority race or ethnicity ^b	99 (17)	221 (18)
Age, y		
18–39	209 (36)	402 (32)
40–59	234 (40)	572 (45)
60+	135 (23)	286 (23)
No. of medication treatment episodes		
0	7 (1)	10 (1)
1	94 (16)	195 (15)
2+	477 (83)	1,055 (84)
No. of prescriptions in most recent medication episode ^c		
1	302 (53)	666 (53)
2	148 (26)	315 (25)
3+	120 (21)	231 (18)
No. of psychotherapy treatment episodes		
0	187 (32)	453 (36)
1	263 (46)	561 (45)
2+	128 (22)	246 (20)
No. of visits in most recent psychotherapy episode ^d		
1	153 (39)	278 (34)
2	102 (26)	196 (24)
3+	138 (35)	311 (39)

^aGroups did not differ significantly on any measure (based on a χ^2 test of independence).

^bNumber (proportion) with ethnicity other than non-Hispanic white, imputed from census block of residence.

^cLimited to those with at least 1 medication treatment episode (n = 571 for responders; n = 1,250 for nonresponders).

^dLimited to those with at least 1 psychotherapy treatment episode (n = 391 for responders; n = 807 for nonresponders).

Linkage of survey data to medical records identified 269 respondents (47% of 578) with adequate records data to assess treatment response for at least 1 prior treatment episode (ie, both baseline and follow-up PHQ-9 scores were recorded, PHQ-9 scores could be attached to a single treatment, and baseline PHQ-9 score was 5 or greater). As shown in Table 2, those for whom adequate PHQ-9 score data were or were not available did not differ significantly in demographic characteristics or treatment history.

On the basis of PHQ-9 scores, 172 of 255 treatment episodes (67%) had a favorable response. This proportion compares to 86 of 255 (34%) with a positive treatment response by recall of global improvement and 174 of 251 (69%) by recall of benefit from treatment. For both measures, agreement between recalled response (for either measure) and depression scores from medical records was generally poor. Table 3 shows agreement between participants' recall of global improvement and response according to PHQ-9 scores. For these 2 measures, the κ statistic (chance-corrected agreement) was 0.10 (95% CI, 0.00–0.19). Table 3 also shows agreement between records and participants' recall of improvement specifically attributed to treatment benefit. For these 2 measures, the κ statistic was 0.12 (95% CI, 0.00–0.25).

Given the relatively poor agreement observed in the entire sample, post hoc analyses examined agreement in subgroups

Table 2. Comparison of Survey Responders With and Without Adequate PHQ-9 Data for at Least 1 Prior Treatment Episode (N = 578)^a

Variable	Adequate PHQ-9 Data (n = 269), n (%)	Inadequate PHQ-9 Data (n = 309), n (%)
Female sex	190 (71)	234 (76)
Minority race or ethnicity	49 (18)	55 (18)
Age, y		
18–39	103 (38)	106 (34)
40–59	102 (38)	132 (43)
60+	64 (24)	71 (23)
Characteristics of most recent treatment episode		
Treatment received		
Medication	185 (69)	214 (69)
Psychotherapy	84 (31)	95 (31)
Self-rated improvement following treatment		
Very much or much better	173 (64)	190 (61)
A little better	53 (20)	68 (22)
No change or worse	29 (11)	25 (8)
Unable to recall or left blank	14 (5)	26 (8)
Self-rated benefit from treatment		
Helped very much or helped some	174 (65)	187 (61)
Did not help or hurt	53 (20)	77 (25)
Made worse or very much worse	24 (9)	20 (6)
Unable to recall or left blank	18 (7)	25 (8)
Confidence in recall of treatment outcome		
Very sure	152 (57)	161 (52)
Less than very sure	113 (42)	145 (47)

^aGroups did not differ significantly on any measure (based on a χ^2 test of independence).

Abbreviation: PHQ-9 = Patient Health Questionnaire.

in which we might expect either more accurate recall or more accurate assessment of outcome by the PHQ-9. As shown in Table 4, agreement was poor for both medication treatment episodes and psychotherapy episodes, and agreement was not meaningfully improved by limiting analyses to participants who continued medication treatment for 3 months or more, to participants who had at least moderate severity of depression at baseline, or to those who reported high confidence in recall. It was only in the subset of patients recalling treatment within the last 6 months that agreement approached a moderate level.

Additional analyses examined whether findings were sensitive to the thresholds or cut points used to define treatment response from patient surveys. When response on the global improvement rating was defined by the top 3 (rather than top 2) categories, the proportion classified as responders increased from 34% to 68%. Using this more lenient classification, the κ statistics regarding agreement with medical records data were essentially unchanged from those in Table 4 (details available upon request from the authors). When response on improvement attributed to treatment rating was defined by the top category (rather than the top 2 categories), the proportion classified as responders decreased from 69% to 25%. When we used this stricter classification, the κ statistics regarding agreement with medical records data were generally lower (ie, poorer agreement) than those in Table 4 (details available upon request from the authors).

Table 3. Overall Agreement Between Recalled Effects of Treatment and PHQ-9 Depression Scores From Medical Records

Response According to 50% Improvement in PHQ-9 Scores	Response According to Recalled Global Improvement, n			Response According to Recalled Benefit From Treatment, n		
	Yes	No	Total	Yes	No	Total
Yes	65	17	172	124	45	169
No	21	62	83	50	32	82
Total	86	169	255	174	77	251

Abbreviation: PHQ-9 = Patient Health Questionnaire.

DISCUSSION

We found that recall of response to past depression treatments was generally poor when compared to depression questionnaire scores from electronic medical records. Overall agreement between patients' recall and PHQ-9 scores in records was only marginally better than chance. Accuracy of recall was not improved by limiting the sample to patients with more severe depression at baseline, to those who continued treatment for at least 3 months, or to those reporting high confidence in accuracy of recall or by varying the cut points used to define treatment response. Recall was more accurate regarding recent treatment episodes, but agreement with medical records data still did not reach a moderate level.

To illustrate the practical implications of these results, we can examine the proportion of patients who would be correctly or incorrectly classified using recalled benefit of treatment (assuming that PHQ-9 depression scores from medical records are the true indicator of response). Of 174 participants who recalled that a specific treatment "helped some" or "helped very much," 124 (71%) experienced a 50% or greater improvement in PHQ-9 depression score. The remaining 50 (29%) did not and would have been incorrectly classified as responders. Of 77 participants who recalled that a specific treatment "did not help or hurt" or "made things worse," 32 (42%) did not experience a 50% or greater decrease in PHQ-9 depression score. The remaining 45 (58%) did experience a 50% or greater improvement and would have been incorrectly classified as nonresponders.

Interpretation of these findings should give consideration to several important limitations. First, only one-third of potential participants completed the survey. Survey participants did not differ from nonparticipants in any characteristic we were able to measure using available computerized records. It is not clear what any unmeasured differences between participants and nonparticipants might imply about accuracy of recall among those not responding to our survey, but we would not predict that accuracy of recall would be greater among those declining to respond to questions regarding prior depression treatments. Second, appropriate depression outcome data were available in medical records for only 47% of survey participants. Those patients for whom PHQ-9 scores were available reported higher confidence in recall of depression outcome but were

Table 4. Agreement Between Recalled Treatment Response and Response According to PHQ-9 Depression Questionnaires in Specific Participant Subgroups

Variable	Global Improvement				Benefit From Treatment			
	n	Response From Records, n	Recalled Response, n	Agreement, κ (95% CI)	n	Response From Records, n	Recalled Response, n	Agreement, κ (95% CI)
All treatment episodes	255	172	86	0.10 (0.00 to 0.19)	251	169	174	0.12 (−0.00 to 0.25)
Psychotherapy episodes	79	53	23	0.07 (−0.06 to 0.23)	77	51	59	−0.00 (−0.22 to 0.21)
Medication episodes	176	119	63	0.11 (−0.00 to 0.23)	174	118	115	0.18 (0.03 to 0.33)
Episodes with ≥ 3 months of medication use	114	85	53	0.08 (−0.07 to 0.24)	113	85	93	0.21 (0.01 to 0.41)
Episodes with baseline PHQ-9 score ≥ 10	205	147	70	0.08 (−0.02 to 0.18)	201	145	139	0.09 (−0.05 to 0.23)
Episodes for which respondent was “very sure” of recall	150	103	64	0.10 (−0.03 to 0.24)	150	103	108	0.03 (−0.13 to 0.30)
Episodes within 6 months of survey date	73	50	22	0.23 (0.08 to 0.39)	72	49	54	0.36 (0.12 to 0.59)

Abbreviation: PHQ-9 = Patient Health Questionnaire.

otherwise similar to those without usable PHQ-9 data. The most important determinant of the availability of PHQ-9 data is attendance at follow-up visits; only those who make follow-up visits would have scores reported. We would not predict that those failing to attend follow-up visits would more accurately recall outcomes of treatment. Third, a single PHQ-9 score from medical records might not accurately represent change in severity of depression. Patients might recall improvement that occurred before or after the visit at which the PHQ-9 score was recorded. More detailed or more frequent clinical assessments might have yielded a more accurate indicator of treatment response.

These findings are consistent with other research regarding recall of past depressive episodes and depressive symptoms. While little previous research has examined accuracy of recall of response to specific depression treatments, several previous studies have examined recall of depression over periods of several weeks to several years. We have previously reported that recall of prior depression is moderately accurate over several weeks.¹⁵ Studies of recall over periods of a year or more generally find moderate or poor accuracy.^{16–18} Recall errors are more often due to underreporting of past depression, and underreporting is more likely to occur among those not depressed at the time of recall.^{15,16} While we found poor overall agreement between recall and medical records in assessing treatment response, agreement approached a moderate level for treatment episodes in the last 6 months.

Our findings are not consistent with those previously reported by Posternak and Zimmerman.⁸ Our sample included patients treated by community mental health and primary care providers under the conditions of usual practice (diverse treatments, variable adherence, variable frequency of follow-up assessment). Recall of past treatment response may be poorer under these conditions than in specialty or referral clinics. We would expect our results to apply to patients receiving nonstandardized care in community practice.

Response rates according to PHQ-9 scores were higher than generally reported for community depression treatment and higher than in previous samples from this health

system,¹⁹ but this difference probably reflects the nature of this sample. We included only patients who were willing to participate in a survey regarding past depression treatments, limited to those with follow-up depression scores in medical records. We might expect that individuals choosing to participate in the survey would have had more favorable experience with depression treatment. And availability of depression scores in records would be limited to patients who had received more regular follow-up care, a group likely to experience more favorable outcomes.

We examined accuracy of recall in a highly structured research survey, and it is possible that recall might be more accurate during an in-person clinical assessment. Nevertheless, our survey incorporated several proven techniques for improving recall that would not be customary in clinical assessments.^{20–22} Preparation or priming questions oriented participants to the recall task and provided additional time for retrieval of memories regarding past treatment. A personal timeline listed all treatment episodes in the past 5 years. Personalized cues included the name of the prescribing physician and images of the medication received. Furthermore, we limited our sample to episodes involving a single treatment for which the outcome was documented in the medical record. We believe that our findings reflect accuracy of recall under ideal conditions and that they probably overestimate the accuracy of recall under more typical conditions.

Our findings do not support the use of recalled treatment outcome in research to identify individual predictors of treatment response. Our hope was that data regarding response across multiple treatment episodes might help identify groups of patients with especially informative patterns of treatment response (eg, good response to one class of medications and poor response to another, good response to psychotherapy and poor response to medication). Our data indicate that patients' recall regarding past depression treatment is not accurate enough to identify those patterns of response. While it may still be useful to examine response across multiple treatment episodes, doing so will probably require outcome data collected at the time of treatment.

These findings also have significant implications for clinical practice. Practice guidelines recommend assessment

of prior treatment response,^{1,7} and inquiring about past treatment response is common clinical practice. Our data suggest, however, that treatment decisions based on recollections of past treatment response may be no better than chance. If current treatment choices are to be guided by past treatment experience, then review of medical records is recommended.

Author affiliation: The Group Health Research Institute, Seattle, Washington.

Potential conflicts of interest: The authors have no financial or other conflicts of interest to disclose.

Funding/support: Supported by grant R01 MH085930 from the National Institute of Mental Health, Bethesda, Maryland.

Additional information: The analytic dataset is available from the first author upon request.

REFERENCES

1. Lam RW, Kennedy SH, Grigoriadis S, et al; Canadian Network for Mood and Anxiety Treatments (CANMAT). Canadian Network for Mood and Anxiety Treatments (CANMAT) clinical guidelines for the management of major depressive disorder in adults, 3: pharmacotherapy. *J Affect Disord.* 2009;117(suppl 1):S26–S43.
2. Parikh SV, Segal ZV, Grigoriadis S, et al; Canadian Network for Mood and Anxiety Treatments (CANMAT). Canadian Network for Mood and Anxiety Treatments (CANMAT) clinical guidelines for the management of major depressive disorder in adults, 2: psychotherapy alone or in combination with antidepressant medication. *J Affect Disord.* 2009;117(suppl 1):S15–S25.
3. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet.* 2009;373(9665):746–758.
4. Gartlehner G, Gaynes BN, Hansen RA, et al. Comparative benefits and harms of second-generation antidepressants: background paper for the American College of Physicians. *Ann Intern Med.* 2008;149(10):734–750.
5. Simon GE, Perlis RH. Personalized medicine for depression: can we match patients with treatments? *Am J Psychiatry.* 2010;167(12):1445–1455.
6. Papakostas GI, Fava M. Predictors, moderators, and mediators (correlates) of treatment outcome in major depressive disorder. *Dialogues Clin Neurosci.* 2008;10(4):439–451.
7. American Psychiatric Association. *Practice Guideline for the Treatment of Patients With Major Depressive Disorder.* 3rd ed. Arlington, VA: American Psychiatric Association; 2010:1–152.
8. Posternak MA, Zimmerman M. How accurate are patients in reporting their antidepressant treatment history? *J Affect Disord.* 2003;75(2):115–124.
9. Elliott MN, Morrison PA, Fremont A, et al. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcomes Res Methodol.* 2009;9(2):69–83.
10. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* 2001;16(9):606–613.
11. Kroenke K, Spitzer RL, Williams JB, et al. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *Gen Hosp Psychiatry.* 2010;32(4):345–359.
12. Löwe B, Kroenke K, Herzog W, et al. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). *J Affect Disord.* 2004;81(1):61–66.
13. Cohen J. A coefficient of agreement of nominal scales. *Educ Psychol Meas.* 1960;20(1):37–46.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–174.
15. Rutter C, Simon G. A Bayesian method for estimating the accuracy of recalled depression. *Appl Stat.* 2004;53(2):341–353.
16. Simon GE, Von Korff M. Recall of psychiatric history in cross-sectional surveys: implications for epidemiologic research. *Epidemiol Rev.* 1995;17(1):221–227.
17. Wells JE, Horwood LJ. How accurate is recall of key symptoms of depression? a comparison of recall and longitudinal reports. *Psychol Med.* 2004;34(6):1001–1011.
18. Patten SB, Williams JV, Lavorato DH, et al. Recall of recent and more remote depressive episodes in a prospective cohort study. *Soc Psychiatry Psychiatr Epidemiol.* 2012;47(5):691–696.
19. Simon GE, Von Korff M, Rutter CM, et al. Treatment process and outcomes for managed care patients receiving new antidepressant prescriptions from psychiatrists and primary care physicians. *Arch Gen Psychiatry.* 2001;58(4):395–401.
20. Schwarz N, Sudman S, eds. *Autobiographical Memory and the Validity of Retrospective Reports.* New York, NY: Springer-Verlag; 1994.
21. Means B, Nigam A, Zarrow M, et al. *Vital and Health Statistics: Autobiographical Memory for Health-Related Events. Series 6: Cognition and Survey Measurement, No 2. US Dept of Health and Human Services publication (PHS) 89-1077.* Hyattsville, MD: US Dept of Health and Human Services; 1989.
22. Bhandari A, Wagner T. Self-reported utilization of health care services: improving measurement and accuracy. *Med Care Res Rev.* 2006;63(2):217–235.