Published in final edited form as:
J Clin Psychiatry. 2012 December ; 73(12): 1503–1508. doi:10.4088/JCP.12m07883.

Response to past depression treatments is not accurately recalled

Gregory E. Simon, MD, MPH, **Carolyn M. Rutter, PhD**, **Christine Stewart, PhD**, **Chester Pabiniak, MS**, and **Linda Wehnes, MA**
Group Health Research Institute, Seattle, WA

Abstract

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

Methods—Electronic medical records from a large health system identified 1878 patients with two or more episodes for clinician-diagnosed depression between 2005 and 2009. 578 of those completed a survey including structured recall of response to each prior treatment – both global improvement during treatment and improvement specifically attributed to treatment. For 269 of these survey participants, at least one treatment episode could be unambiguously linked to both pre- and post-treatment PHQ9 depression scores in electronic medical records. Analyses examined agreement between patients recall of treatment response and improvement in PHQ9 scores from medical records.

Results—Agreement with medical records was poor both for recall of overall improvement following treatment (kappa = 0.10, 95% CI 0.00–0.19) and for recall of 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 her/his ability to recall. Agreement reached a fair level only for episodes in the six months prior to the survey (kappa = 0.23 for overall improvement, kappa = 0.36 for improvement attributed to treatment).

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.

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 guideline for treatment of major depressive disorder (7)

Corresponding author address: Gregory Simon, Group Health Research Institute, 1730 Minor Ave., #1600, Seattle, WA 98101, Phone: 206-287-2979, Fax: 206-287-2872, simon.g@ghc.org.

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

The analytic dataset is available from the first author on request.

calls for “a psychiatric history that particularly notes current treatments and responses to previous treatments”. This recommendation depends on two 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 (8) 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 to past depression treatment.

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

METHODS

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

Electronic medical records were used to identify adult members who had experienced at least two 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 of 2009. An episode of antidepressant treatment was defined by a filled prescription for an antidepressant, an associated diagnosis of major depression or dysthymic disorder, 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 depression and no psychotherapy visit in the prior 180 days. Those 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.

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 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, trouble sleeping, or having thoughts of death or suicide.”

Response to each medication was assessed using three 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 seven-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 five-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 four-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 was followed by three questions regarding each episode of therapy, parallel to those described above regarding antidepressant treatment episodes (details available on request).

For all potential participants invited to complete the survey, computerized medical records were used to compare survey respondents and non-respondents in terms of age, sex, treatment history, and imputed race and ethnicity based on 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 or PHQ9 depression severity questionnaire for initial assessment of depression and at all depression follow-up visits. The PHQ9 has been a valid and sensitive measure of depression severity across a wide range of patient populations and clinical settings (10–12). PHQ9 scores 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 PHQ9 scores for each treatment episode. The eligibility period for a baseline PHQ9 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 one eligible baseline PHQ9 score was identified, then the score closest to the episode start date was selected. The eligibility period for an outcome PHQ9 score extended from 60 days to 120 days after the episode start date. If more than one eligible outcome PHQ9 score was identified, then the score closest to 90 days after the index date was selected. Approximately half of episodes with baseline and outcome scores were excluded because PHQ9 scores could not be linked to a single treatment (i.e. the period between the two scores included exposure to more than one treatment simultaneously or sequentially). The sample was further limited to episodes with baseline PHQ9 score of 5 or greater.

A positive treatment response according to the PHQ9 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 two 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 particpants’ recall (details available on request).

Data analyses proceeded in three steps. The initial step compared eligible patients who did and did not participate in the online survey to assess possible bias due to non-response. The second step considered all survey respondents, comparing those for whom PHQ9 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 PHQ9 data were available, examining agreement between these two sources in identifying positive treatment response. The kappa statistic (13) indicated the degree to which agreement exceeded that expected by chance. Traditional criteria for consider kappa values less than 0.2 to indicate minimal agreement, values of 0.2 to 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 1838 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.

Linkage of survey data to medical records identified 269 respondents (46%) with adequate records data to assess treatment response for at least one prior treatment episode (i.e., both baseline and follow-up PHQ9 scores were recorded, PHQ9 scores could be attached to a single treatment, and baseline PHQ9 score was 5 or greater). As shown in Table 2, those for whom adequate PHQ9 score data were or were not available did not differ significantly in demographic characteristics or treatment history.

Based on PHQ9 scores, 172 of 255 treatment episodes (or 67%) had a favorable response. This compares to a 86/250 (or 34%) with a positive treatment response by recall of global improvement and 174/251 (or 69%) by recall of benefit from treatment. For both measures, agreement between recalled response (by either measure) and depression scores from medical records was generally poor. Table 3a shows agreement between participants’ recall of global improvement and response according to PHQ9 scores. For these two measures, kappa statistic (chance corrected agreement) was 0.10 (95% Conf. Interv. = 0.00 to 0.19). Table 3b shows agreement between records and participants’ recall of improvement specifically attributed to treatment benefit. For these two measures, kappa statistic was 0.12 (95% Conf. Interv = 0.00 to 0.25).

Given the relatively poor agreement observed in the entire sample, post hoc analyses examined agreement in subgroups where we might expect either more accurate recall or more accurate assessment of outcome by the PHQ9. 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 more than three months, 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 recalling treatment within the last six 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 three (rather than top two) categories, the proportion classified as responders increased from 34% to 68%. Using this more lenient classification, kappa statistics regarding agreement with medical records data were essentially unchanged from those in Table 4 (details available on request). When response on the improvement attributed to treatment rating was defined by the top (rather than the top two) categories, the proportion classified as responders decreased from 69% from 25%. Using this stricter classification, kappa statistics regarding agreement with medical records data were generally lower (i.e. poorer agreement) than those in Table 4 (details available on request).

DISCUSSION

We find that recall of response to past depression treatments was generally poor when compared to depression questionnaires from electronic medical records. Overall agreement between patients' recall and PHQ9 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, limiting to those who continued treatment for at least three months, limiting to patients 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 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 (or 71%) experienced a 50% or greater improvement in PHQ9 depression score. The remaining 50 (or 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 (or 42%) did not experience a 50% or greater decrease in PHQ9 depression score. The remaining 45 (or 58%) did experience a 50% or greater improvement and would have been incorrectly classified as non-responders.

Interpretation of these findings should consider 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 non-participants 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 for whom PHQ9 scores were available reported higher confidence in recall of depression outcome but were otherwise similar to those without usable PHQ9 data. The most important determinant of the availability of PHQ9 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 PHQ9 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 PHQ9 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 for 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 under-reporting of past depression, and under-reporting is more likely among those not depressed at the time of recall (15, 16). While we find poor overall agreement between recall and medical records in assessing treatment response, agreement approached a moderate level for treatment episodes in the last six months.

Our findings are not consistent with those previously reported by Posternak and colleagues (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 non-standardized care in community practice.

Response rates according to PHQ9 scores were higher than generally reported for community depression treatment and higher than in previous samples from this health system (19), but this probably reflects the nature of this sample. We included only patients 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 those 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 receiving more regular follow-up care, a group like 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 five years. Personalized cues included the name of the prescribing physician and images of medication received. Furthermore, we limit our sample to episodes involving a single treatment for which outcome was documented in the medical record. We believe that our findings reflect accuracy of recall under ideal conditions, and probably over-estimate 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 (e.g. 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 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 records is recommended.

Acknowledgments

Supported by grant R01 R01MH085930 from the National Institute of Mental Health

References

1. Lam RW, Kennedy SH, Grigoriadis S, McIntyre RS, Milev R, Ramasubbu R, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) clinical guidelines for the management of major depressive disorder in adults. III. Pharmacotherapy. J Affect Disord. 2009 Oct; 117(Suppl 1):S26–43. [PubMed: 19674794]
2. Parikh SV, Segal ZV, Grigoriadis S, Ravindran AV, Kennedy SH, Lam RW, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) clinical guidelines for the management of major depressive disorder in adults. II. Psychotherapy alone or in combination with antidepressant medication. J Affect Disord. 2009 Oct; 117(Suppl 1):S15–25. [PubMed: 19682749]
3. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet. 2009 Feb 28; 373(9665):746–58. [PubMed: 19185342]
4. Gartlehner G, Gaynes BN, Hansen RA, Thieda P, DeVeaugh-Geiss A, Krebs EE, et al. Comparative benefits and harms of second-generation antidepressants: background paper for the American College of Physicians. Ann Intern Med. 2008 Nov 18; 149(10):734–50. [PubMed: 19017592]
5. Simon GE, Perlis RH. Personalized medicine for depression: can we match patients with treatments? Am J Psychiatry. 2010 Dec; 167(12):1445–55. [PubMed: 20843873]
6. Papakostas GI, Fava M. Predictors, moderators, and mediators (correlates) of treatment outcome in major depressive disorder. Dialogues Clin Neurosci. 2008; 10(4):439–51. [PubMed: 19170401]
7. American Psychiatric Association. Practice guideline for the treatment of patients with major depressive disorder (revision). Am J Psychiatry. 2000; 157:s1–s45.
8. Posternak M, Zimmerman M. How accurate are patients in reporting their antidepressant treatment history? J Affect Disord. 2003; 75:115–24. [PubMed: 12798251]
9. Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the Census Bureau’s Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. Health Services and Outcomes Research Methodology. 2009; 9:69–83.
10. Kroenke K, Spitzer R, Williams J. The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med. 2001; 16:606–13. [PubMed: 11556941]
11. Kroenke K, Spitzer RL, Williams JB, Lowe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen Hosp Psychiatry. 2010 Jul-Aug;32(4): 345–59. [PubMed: 20633738]
12. Lowe B, Kroenke K, Herzog W, Grafe K. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). J Affect Disord. 2004; 81:61–6. [PubMed: 15183601]
13. Cohen J. A coefficient of agreement of nominal scales. Educational and Psychological Measurement. 1960; 30:37–46.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar; 33(1):159–74. [PubMed: 843571]

15. Rutter C, Simon G. A Bayesian method for estimating the accuracy of recalled depression. Applied Statistics. 2004; 53:341–53.

16. Simon G, MV. Recall of psychiatric history in cross-sectional surveys: Implications for epidemiologic research. Epidemiol Rev. 1995; 17:221–7. [PubMed: 8521941]

17. Wells JE, Horwood LJ. How accurate is recall of key symptoms of depression? A comparison of recall and longitudinal reports. Psychol Med. 2004 Aug; 34(6):1001–11. [PubMed: 15554571]

18. Patten SB, Williams JV, Lavorato DH, Bulloch AG, D'Arcy C, Streiner DL. Recall of recent and more remote depressive episodes in a prospective cohort study. Soc Psychiatry Psychiatr Epidemiol. 2011 May 1.

19. Simon G, Von Korff M, Rutter C, Peterson D. Treatment process and outcomes for managed care patients receiving new antidepressant prescriptions from psychiatrists and primary care physicians. Arch Gen Psychiatry. 2001; 58:395–401. [PubMed: 11296101]

20. Schwarz, N.; Sudman, S., editors. Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer-Verlag; 1994.

21. Means B, Nigam A, Zarrow M, Loftus E, Donaldson MS. Autobiographical memory for health-related events. Vital Health Stat 1989. 1989; 6(2)

22. Bhandari A, Wagner T. Self-reported utilization of health care services: improving measurement and accuracy. Med Care Res Rev. 2006 Apr; 63(2):217–35. [PubMed: 16595412]

CLINICAL POINTS

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

Table 1

Comparison of survey responders and non-responders[a].

	Responders (n=578) 31%	**Non-Responders (n=1260) 69%**
Female	424 (73%)	913 (73%)
Predicted Minority Race/Ethnicity[b]	99 (17%)	221 (18%)
Age		
18–39	209 (36%)	402 (32%)
40–59	234 (41%)	572 (45%)
60+	135 (23%)	286 (23%)
# of Medication Treatment Episodes		
0	7 (1%)	10 (1%)
1	94 (16%)	195 (16%)
2+	477 (83%)	1055 (84%)
# of Prescriptions in Most Recent Medication Episode (n = 1789[c])		
1	*302 (53%)*	*666 (55%)*
2	*148 (26%)*	*315 (26%)*
3+	*120 (21%)*	*231 (19%)*
# of Psychotherapy Treatment Episodes		
0	187 (32%)	453 (36%)
1	263 (46%)	561 (45%)
2+	128 (22%)	246 (19%)
# of Visits in Most Recent Psychotherapy Episode (n=1178[d])		
1	*153 (39%)*	*278 (35%)*
2	*102 (26%)*	*196 (25%)*
3+	*138 (35%)*	*311 (40%)*

Notes:

[a] Groups did not differ significantly on any measure (based on a chi-square test of independence.

[b] Number (proportion) with ethnicity other than non-Hispanic White, imputed from census block of residence

[c] Limited to those with at least one medication treatment episode

[d] Limited to those with at least one psychotherapy treatment episode

Table 2

Comparison of survey responders with and without usable PHQ9 data for at least one prior treatment episode. Groups did not differ significantly on any measure (based a chi-square test of independence).

	Adequate PHQ9 Data (n=269)	No Adequate PHQ9 Data (n=309)
Female	190 (71%)	234 (76%)
Minority race or ethnicity	49 (18%)	55 (18%)
Age		
18–39	103 (38%)	106 (34%)
40–59	102 (38%)	132 (43%)
60+	64 (25%)	71 (23%)
Characteristics of most recent episode with adequate PHQ9 data		
Treatment received		
Medication	185 (69%)	214 (72%)
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 (60%)
Didn’t 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%)

Table 3

Overall Agreement Between Recalled Effects of Treatment and PHQ9 Depression scores from medical records

a) Recalled Self-Rated Global Improvement

	Response by Recalled Global Improvement		
Response by 50% Improvement in PHQ9	**YES**	**NO**	**TOTAL**
YES	65	107	172
NO	21	62	83
TOTAL	86	169	255

b) Recalled Self-Assessed Benefit of Treatment

	Response by Recalled Benefit from Treatment		
Response by 50% Improvement in PHQ9	**YES**	**NO**	**TOTAL**
YES	124	45	166
NO	50	32	82
TOTAL	174	77	251

Table 4

Agreement between recalled treatment response and response from PHQ9 depression questionnaires in specific participant subgroups.

	Recalled Self-Rated Improvement				Recalled Benefit From Treatment			
	Number	**Response from records**	**Recalled Response**	**Agreement Kappa (95% CI)**	**Number**	**Response from records**	**Recalled response**	**Agreement Kappa (95% CI)**
All Episodes	255	172	86	0.10 (0.00, 0.19)	251	169	174	0.12 (−0.00, 0.25)
Psychotherapy Episodes	79	53	23	0.07 (−0.06, 0.23)	77	51	59	−0.00 (−0.22, 0.21)
Medication Episodes	176	119	63	0.11 (−0.00, 0.23)	174	118	115	0.18 (0.03, 0.33)
Used Medication >= 3 mos	114	85	53	0.08 (−0.07, 0.24)	113	85	93	0.21 (0.01, 0.41)
Baseline PHQ9 >=10	205	147	70	0.08 (−0.02, 0.18)	201	145	139	0.09 (−0.05, 0.23)
“Very Sure” of Recall	150	103	64	0.10 (−0.03, 0.24)	150	103	108	0.03 (−0.13, 0.30)
< 6 mos Since Episode	73	50	22	0.23 (0.08, 0.39)	72	49	54	0.36 (0.12, 0.59)