Published in final edited form as:
J Affect Disord. 2008 April ; 107(1-3): 271–274.

ACCURACY OF RECALL FOR MANIA SYMPTOMS USING A THREE MONTH TIMELINE FOLLOW-BACK INTERVIEW

Gregory E Simon, MD MPH and **Carolyn M. Rutter, PhD**
Group Health Cooperative Center for Health Studies, Seattle, WA

Abstract

Objective—Little research has examined the accuracy of recall for mood symptoms using retrospective timelines or life charts. We examined accuracy of recall for mania symptoms over a period of 3 months.

Methods—Data were collected from a sample of 392 patients enrolled in a randomized trial of a psychoeducation and care management program for bipolar disorder. Every three months, participants completed in-person assessments including the Longitudinal Interval Follow-Up Examination, a timeline follow-back interview assessing mood symptoms during each week since the previous assessment. Brief telephone assessments of mood symptom severity were performed at a randomly selected point between in-person interviews. Mania symptoms recalled at the in-person assessment were compared to those reported in the previous telephone interview.

Results—The proportions of weeks with full or subthreshold mania symptoms recalled at the in-person interview were similar to those detected by telephone assessments. When compared to telephone assessments, sensitivity of recall for detecting subthreshold or greater symptoms of mania was 63% (95% CI 57% to 69%). Specificity for detecting absence of significant mania symptoms was 76% (95% CI 71% to 80%).

Limitations—Validation of recall was based on brief telephone assessments rather than detailed in-person interviews. Our results may not apply to recall over longer time periods.

Conclusions—A timeline follow-back interview demonstrated acceptable sensitivity and specificity for detecting symptoms of mania during a specific week in the prior three months.

Keywords

bipolar disorder; assessment; mania; recall; validity; accuracy

Bipolar disorder is characterized by recurring symptoms of depression and mania occurring in almost every conceivable pattern. Severity of mood symptom may remain stable for weeks or change rapidly over days or even hours (Judd et al. 2002). In many cases, bipolar disorder is not truly bipolar, with complex combinations of depression and mania symptoms more the norm than the exception (Bauer et al. 2005).

Clinical research in bipolar disorder depends on accurately measuring changes in mood symptoms over time. Clinical assessments separated by weeks or months may not capture

Corresponding Author: Gregory Simon MD MPH Center for Health Studies 1730 Minor Ave. #1600 Seattle, WA 98101 Phone: 206 −287−2979 Fax: 206−287−2871 Email: simon.g@ghc.org.

important changes in mood during the period between assessments. The need to assess interval symptoms is especially important in studies of long-term or maintenance treatment where the goal is to prevent or reduce the severity of relapses over months or years. Consequently, such studies typically rely on recall of symptoms between assessments using a timeline follow-back method (Keller et al. 1987).

Surprisingly little previous research has examined the validity of recalled mood symptoms in timeline follow-back interviews. Warshaw and colleagues (Warshaw et al. 2001) used telephone interviews to assess the accuracy of recalled anxiety symptoms, finding generally good agreement between a telephone interview and symptoms recalled up to 3 months later. We have previously reported on the accuracy of recalled severity of depression over a 3-month period (Rutter and Simon, 2004). We found that recalled severity of depression was, on average, slightly higher than severity assessed by telephone several weeks previously. Accuracy of recall was not affected by elapsed time (up to a maximum 3 months), but patients who were depressed at the time of the assessment tended to “over recall” depression during previous weeks.

Here we examine the accuracy of recall of mania symptoms over a 3-month period. Given that symptoms of mania are less frequent and less stable than are symptoms of depression, we anticipated that accuracy of recall for mania might be poorer.

METHODS

Study methods are described in detail elsewhere (Simon et al. 2002; Simon et al. 2006) and will be summarized here. The primary objective of the study was to evaluate the effectiveness and cost-effectiveness of a psychoeducation and care management program for bipolar disorder. We describe here a sub-study designed to assess the accuracy of the timeline follow-back interviews for assessment of mood symptoms over the prior three months.

Participants

Participants were recruited from mental health clinics of Group Health Cooperative, a prepaid health plan in Washington state and Idaho. Computerized records were used to identify all patients seen at participating clinics in the prior 12 months with a visit diagnosis of bipolar disorder (Type 1 or Type 2). Diagnosis was confirmed at subsequent structured interview.

All potential participants were invited to attend an in-person baseline assessment that included (among other measures) the current depression, past depression, current mania, and past mania modules of the Structured Clinical Interview for DSM-IV (First et al. 1997) or SCID. Eligibility for enrollment required confirmation of bipolar disorder either by SCID interview or by unambiguous medical record documentation of manic episode (for diagnosis of Bipolar disorder Type 1) or depressive and hypomanic episode (for diagnosis of Bipolar disorder Type 2).

Measures

All participants were invited to return for in-person research assessments scheduled every 3 months for 24 months. Each assessment used the Longitudinal Interval Follow-Up Evaluation or LIFE (Keller et al. 1987) to generate week-by-week timeline follow-back ratings of depression and mania/hypomania symptoms since the last in-person assessment. The LIFE yields separate ratings of depression and mania severity for each week using the 6-point Psychiatric Status Rating (PSR) scale. Ratings of 1 or 2 on this scale represent remission or minimal symptoms, ratings of 3 or 4 represent clinically significant subthreshold symptoms,

a rating of 5 represents a current episode of hypomania or moderate major depression, and a rating of 6 represents a current episode of mania or severe depression.

During the first two follow-up periods (between the baseline and 3-month interviews and between the 3- and 6-month interviews) each participant was contacted by telephone for a brief assessment mood symptoms. The SCID current depression and current mania modules were used to assess symptoms during the week prior to the telephone call. Ratings were classified using the same 6-point Psychiatric Status Rating (PSR) scale. Dates for telephone validation interviews were selected at random and were evenly distributed over the interval between assessments.

Telephone interviewers were blinded to the results of the prior in-person assessment. At each in-person interview, the interviewer was aware of the results of the previous in-person assessment but was blinded to the results of any interval telephone assessment.

Interviewers were mental health clinicians with at least one year of clinical experience in the assessment of mood disorders. Details of the interviewer training have been described previously (Simon et al. 2002; Simon et al. 2005).

Data analysis

Each participant contributed up to two telephone assessments: one per 3-month interview period. Each of these assessments was linked to data from the subsequent in-person interview to identify recalled mood ratings for the corresponding one-week period.

To facilitate interpretation, the 6-point PSR scales were analyzed as dichotomous outcomes using two different thresholds: meeting criteria for manic or hypomanic episode (mania PSR >=5) and a lower criterion of subthreshold or greater symptoms (mania PSR >=3).

True positive or sensitivity rates were calculated as the proportion of weeks with mania symptoms as reported during the telephone assessment (denominator) that were also reported to include mania symptoms at the subsequent recall interview (numerator). Sensitivity analyses examined a window of recall (up to +/− 3 weeks before or after the telephone interview) and allowed positive recall anywhere in this window to count as a true positive.

True negative or specificity rates were calculated as the proportion of weeks without mania symptoms as reported during the telephone interview (denominator) that were also reported to be free of mania symptoms at the subsequent recall interview (numerator). Sensitivity analyses examined a window of recall (up to +/− 3 weeks before or after the telephone interview) and required negative recall throughout this window to count as a true negative.

Most (74%) participants contribute information from 2 telephone interviews to estimates of true positive and true negative rates. To account for within patient correlation, we used bootstrap resampling of individuals with 10,000 replicates to calculate percentile-based confidence limits for true positive and true negative rates.

RESULTS

As described in previous reports, the sample of 441 participants enrolled was 68% female with a mean age of 44 years (range 18 to 81). At baseline assessment, 42% met criteria for a current mood episode (major depression, hypomania, or mania) and an additional 39% reported significant subthreshold symptoms. 10% had one or more psychiatric hospitalizations in the prior year and 31% reported significant disability due to bipolar disorder (unable to work or manage household responsibilities for at least 30 of the last 90 days). 392 participants (88%

of the sample) completed at least one telephone assessment and one recall interview and were include in these analyses. In that group, 44 met criteria for a current manic episode during at least one telephone validation interview and an additional 151 reported significant subthreshold mania symptoms during at least one telephone interview.

As shown in Table 1, the distributions of mania severity ratings were similar for telephone assessments and for the subsequent recall interviews: approximately 60% of weeks without significant mania symptoms, approximately 30% of weeks with subthreshold mania symptoms, and fewer than 10% of weeks meeting criteria for hypomania or mania.

The center portion of Table 2 displays true positive rates and true negative rates for detecting weeks with symptoms of mania or hypomania (PSR >=5). The estimated true positive or sensitivity rate increased from approximately 44% to approximately 56% as the recall window expanded to include up to 3 weeks before or after the telephone interview. In the same sensitivity analysis, the estimated true negative or specificity rate decreased from 95% to 89% as the recall window was expanded.

The right portion of Table 2 displays true positive rates and true negative rates for detecting weeks with subthreshold or greater symptoms of mania (PSR >=3. Using this lower threshold, true positive rates were higher and true negative rates were lower than for the stricter threshold of manic or hypomanic episode. As the recall window was expanded to include the three weeks before and after the telephone interview, true positive rates increased modestly and true negative rates decreased modestly.

Using a threshold of PSR>=3 and allowing a recall window of +/− 2 weeks, the kappa statistic for agreement between recall and telephone validation was 0.52.

DISCUSSION

We find reasonably good agreement between ratings of mania symptoms based on a telephone assessment of current symptoms or a timeline follow-back interview several weeks later. When compared to telephone assessments of current symptoms, recall interviews up to three months later produced similar estimates of the overall frequency of mania symptoms. True positive rates were moderate for detection of more severe symptoms and good for detection of subthreshold or greater symptoms. True negative rates were high for detecting the absence of more severe symptoms and moderate for detection the absence of subthreshold symptoms.

While agreement between recall interviews and previous telephone assessments was acceptable, error rates were not trivial. Even after allowing a 3 week window around the time of the telephone assessment, the recall interview failed to identify 23% of weeks with significant mania symptoms. We are not aware of any direct comparisons, but more intensive methods such as daily mood records (Denicoff et al. 2000; Bauer et al. 2004) would almost certainly prove more accurate. Collateral information from family members or other close contacts would also increase sensitivity. Those more intensive methods, however, might not be practical or acceptable in some research settings or patient populations.

We should acknowledge some limitations to the validity and generalizability of these findings. First, we compare recall to a brief telephone assessment, certainly an imperfect gold standard. While frequent clinical assessments would be a more accurate criterion, such assessments might bias results by improving the accuracy of subsequent recall. Available data do support the accuracy of telephone ratings of current mood symptoms in people treated for bipolar disorder (Revicki et al. 1997). Second, our study enrolled outpatients with moderate severity of illness, and recall might be less accurate in patients with more severe or unstable illness. Third, our data do not allow us to examine accuracy of recall over periods longer than three

months. Fourth, the prevalence of mania symptoms was relatively low, so we are not able to examine whether accuracy of recall varied with elapsed time or mood state at the time of the in-person interview.

These results are consistent with those we have reported earlier regarding recall of depression symptoms (Rutter and Simon, 2004). Timeline follow-back ratings showed acceptable agreement with telephone assessments up to 3 months earlier. While recall is only moderately accurate for determining the specific timing of mania symptoms, we do not find significant bias toward over- or underestimating the severity of depression or mania. In combination, these findings generally support the use of timeline follow-back interviews to assess severity of bipolar disorder in longitudinal studies.

Acknowledgements

Supported by grants R01 MH059125 and P20 MH068572 from the National Institute of Mental Health. The NIMH had no further role in study design, data collection, data analysis, or interpretation.

REFERENCES

Bauer M, Grof P, Gyulai L, Ragson N, Glenn T, Whybrow P. Using technology to improve longitudinal studies: self-reporting with ChronoRecord in bipolar disorder. Bipolar Disord 2004;6:67–74. [PubMed: 14996143]

Bauer M, Simon G, Ludman E, Unutzer J. 'Bipolarity' in bipolar disorder: distribution of manic and depressive symptoms in a treated population. Br J Psychiatry 2005;187:87–88. [PubMed: 15994577]

Denicoff K, Leverich G, Nolen W, Rush A, McElroy S, Keck P, Suppes T, Altshuler L, Kupka R, Frye M, Hatef J, Brotman M, Post R. Validation of the prospective NIMH-Life-Chart Method (NIMH-LCM-p) for longitudinal assessment of bipolar illness. Psychol Med 2000;30:1391–1397. [PubMed: 11097079]

First, M.; Spitzer, R.; Gibbon, M.; Williams, J. First, M.; Spitzer, R.; Gibbon, M.; Williams, J., editors. American Psychiatric Press; Washington: 1997.

Judd L, Akiskal H, Schettler P, Endicott J, Maser J, Solomon D, Leon A, Rice J, Keller M. The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch Gen Psychiatry 2002;59:530–537. [PubMed: 12044195]

Keller M, Lavori PW, Friedman B, Nielsen E, Endicott J, McDonald-Scott P, Andreasen NC. The Longitudinal Interval Follow-up Evaluation: A comprehensive method for assessing outcome in prospective longitudinal studies. Arch Gen Psychiat 1987;44:540–548. [PubMed: 3579500]

Revicki D, Tohen M, Gyulai L, Thompson C, Pike S, Davis-Vogel A, Zarate C. Telephone versus in-person clinical and health status assessment interviews in patients with bipolar disorder. Harvard Rev Psychiatry 1997;5:75–81.

Rutter C, Simon G. A Bayesian method for estimating the accuracy of recalled depression. Applied Statistics 2004;53:341–353.

Simon G, Ludman E, Unutzer J, Bauer M. Design and implementation of a randomized trial evaluating systematic care for bipolar disorder. Bipolar Disord 2002;4:226–236. [PubMed: 12190711]

Simon G, Ludman E, Bauer M, Unutzer J, Operskalski B. Long-term effectiveness and cost of a systematic care program for bipolar disorder. Arch Gen Psychiatry 2006;63:500–508. [PubMed: 16651507]

Simon G, Ludman E, Unutzer J, Bauer M, Operskalski B, Rutter C. Randomized trial of a population-based care program for people with bipolar disorder. Psychol Med 2005;35:13–24. [PubMed: 15842025]

Warshaw M, Dyck I, Allsworth J, Stout R, Keller M. Maintaining reliability in a long-term psychiatric study: an ongoing inter-rater reliability monitoring program using the longitudinal interval follow-up evaluation. J Psychiatr Res 2001;35:297–305. [PubMed: 11591433]

Table 1

Proportion of weeks with symptoms of mania according to telephone assessments and recall during subsequent in-person assessment.

	Telephone Assessment	Subsequent Recall
Remission	60.3%	65.2%
Subthreshold Symptoms	32.1%	28.2%
Hypomanic or Manic Episode	7.5%	6.5%

Table 2

Sensitivity (True Positive) and Specificity (True Negative) rates for recall interview when measured against prior telephone assessment.

	Mania or Hypomania (PSR >=5)		**Subthreshold or Greater Symptoms (PSR >=3)**	
	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Exact Match	43.7% (28.3% to 58.5%)	94.5% (92.6% to 96.3%)	63.2% (57.4% to 68.9%)	76.1% (71.7% to 80.3%)
+/− One Week	47.9% (32.0% to 63.3%)	93.4% (91.3% to 95.3%)	68.2% (62.5% to 73.8%)	73.0% (68.6% to 77.4%)
+/− Two Weeks	47.9% (32.0% to 63.3%)	91.0% (88.6% to 93.2%)	75.1% (69.4% to 80.6%)	70.7% (66.0% to 75.5%)
+/− Three Weeks	56.2% (40.8% to 70.5%)	89.3% (86.6% to 91.7%)	77.4% (71.8% to 82.7%)	66.7% (61.8% to 71.4%)