Diagnostic Reliability of Bipolar II Disorder

Sylvia G. Simpson, MD; Francis J. McMahon, MD; Melvin G. McInnis, MD; Dean F. MacKinnon, MD; David Edwin, PhD; Susan E. Folstein, MD; J. Raymond DePaulo, MD

Background: Although the diagnostic reliability of major depression and mania has been well established, that of hypomania and bipolar II (BPII) disorder has not. This remains an important issue for clinicians, especially for those undertaking genetic studies of BP disorder since bipolar I (BPI) and BPII disorders often cluster in the same families. We have assessed our diagnostic reliability of BP disorders, recurrent unipolar disorder, and their constituent episodes (major depression, mania, and hypomania) using interview and best-estimate diagnostic procedures used in a genetic study of families with BPI disorder.

Methods: Reliability was assessed for (1) co-rated Schedule for Affective Disorders and Schizophrenia–Lifetime version interviews of 37 subjects including 15 with BP disorders; (2) test-retest Schedule for Affective Disorders and Schizophrenia–Lifetime version interviews of 26 subjects including 13 with BP disorders; and (3) best-estimate diagnoses made by 2 non-interviewing psychiatrists on 524 subjects in a genetic linkage study of BPI disorder. Diagnoses were based on Research Diagnostic Criteria for a Selected Group of Functional Disorders, except that recurrent major depression as well as hypomania was required for a diagnosis of BPII disorder.

Results: On co-rated interviews, we observed complete agreement between interviewers for diagnosing major depressive, manic, and hypomanic episodes. For test-retest interviews, the Cohen \( \kappa \) coefficients were 0.83 for manic, 0.72 for hypomanic, and 1.0 for major depressive episodes. At the best-estimate level, the Cohen \( \kappa \) coefficients were 0.99 for BPI, 0.99 for BPII, and 0.98 for recurrent unipolar disorder.

Conclusion: Good interrater reliability for BPII can be achieved when the interviews and best-estimate diagnoses are done by experienced psychiatrists.

Arch Gen Psychiatry. 2002;59:736-740
SUBJECTS AND METHODS

SUBJECTS

In part 1 of the reliability study, we conducted 37 co-rated SADS-L and 26 test-retest SADS-L interviews. The group of 37 co-rated subjects consisted of 12 patients selected from the Johns Hopkins Hospital psychiatric inpatient units and 25 participants from the family study of BPI. Six of the inpatients had BP disorders (3 with BPI and 3 with BPII) compared with 9 of the family subjects (5 of whom had BPI and 4 of whom had BPII). Sixty-one percent of the co-rated subjects were female and 61% had been married (ie, “had been married” includes currently married, separated, divorced, and widowed individuals). Their mean age was 34.5 years and mean years of education was 13. Most of the inpatients were recruited from units other than the affective disorders unit. Subjects were excluded if their primary diagnosis was a substance use disorder but not if this was a comorbid diagnosis. Most of the patients or subjects with BP disorder and RUP had 1 or more comorbid diagnoses.

The 26 test-retest subjects consisted of inpatients, day-hospital patients, and healthy control subjects and included 13 subjects with BP disorders (7 with BPI, and 6 with BPII). Of the 26, 17 (65%) were female and 16 (62%) had been married; their mean age was 40 years and mean years of education was 14. Four of the 6 subjects having the diagnosis of BPII in the test-retest study were young people in their early to mid-20s. This was the first hospitalization for only 1 of the subjects, but for most it was the first time they had been diagnosed as having BPII. These subjects with BPII were quite complex. Five of the 6 had 1 or more substance use disorders, 3 had 1 or more anxiety disorders, and 3 had an eating disorder. Healthy controls were recruited from among friends and acquaintances of the research staff.

The subjects for part 2 of the reliability study, the assessment of agreement between pairs of best-estimate diagnoses, were 524 relatives from 71 families in the Johns Hopkins BPI disorder family study. Fifteen of those who had BPI disorder and RUP had 1 or more comorbid diagnoses. Experienced psychiatrists conduct the interviews in our family study of BPI. We present reliability data at the interview and best-estimate levels for BPI, BPII, and RUP diagnoses, based on Research Diagnostic Criteria for a Selected Group of Functional Disorders (RDC). 13 The interview data were collected specifically as a preparatory study. Interviewers were blind to the subjects’ diagnoses. All pertinent RDC diagnoses, current and lifetime, were made on all subjects. The paired test-retest interviews were done within 72 hours of each other, most of them within a 24-hour period. Interviews were done by the first available psychiatrist, with no formal randomization as to who did the test or retest interviews. Fifteen of the 26 test-retest interviews were done by a more junior psychiatrist (S.G.S.).

The pairs of best-estimate diagnoses were made by the same 5 psychiatrists (other than the interviewing psychiatrist) and a sixth senior psychiatrist (S.E.F.) and were based on family-history data, treatment records, and the direct interview,13 including a narrative summary done by the interviewing psychiatrist. Specific best-estimate diagnoses were assigned for major recurrent affective disorder diagnoses, while a best-estimate diagnosis of “uncertain phenotype” was assigned if a subject had a single episode of major depression or a minor affective diagnosis such as hypomania or minor depression.

After each diagnostian separately assigned a best-estimate diagnosis, the diagnosticians were allowed to discuss the case. In cases where the diagnosticians did not agree, they were encouraged to rectify any incomplete or misreading of family history data, medical records, or SADS-L data, and were also able to question the interviewing psychiatrist to clarify information in the narrative summary or to correct coding discrepancies in the interview.

If the diagnosticians still did not agree after a guided review of the data, we did not encourage much less force agreement. For the purposes of the linkage study, the subject was assigned the more “conservative” of the 2 diagnoses as the final consensus best-estimate diagnosis. For example, if there was disagreement as to whether the subject had an affective disorder, the subject was designated as having an uncertain phenotype. Nearly one quarter of our sample has been so designated; in most cases, this was because they either had a single episode of major depression or hypomania without recurrent major depression. If there was disagreement between 2 affective disorder diagnoses, for example, if one reviewer assigned a diagnosis of BPI and the other BPII, the consensus final best-estimate would be BPII. The final diagnosis of each reviewer, however, is retained in the database as a record of their disagreement.

STATISTICAL ANALYSES

Reliability was assessed using the unweighted Cohen κ statistic18 that measures agreement on nominal categories such as diagnosis and that incorporates a correction for chance agreement. The asymptotic SEs of the κ values and 2-tailed level of statistical significance over a 95% confidence interval were calculated using the SPSS software package.19,20

(Reprinted) Arch Gen Psychiatry/Vol. 59, Aug 2002
WWW.ARCHGENPSYCHIATRY.COM

737

©2002 American Medical Association. All rights reserved.

Downloaded From: http://archpsyc.jamanetwork.com/pdfaccess.ashx?url=/data/journals/psych/5145/ on 09/29/2017
There was agreement on 98% of the affective disorder diagnoses. The interviewers differed on whether the subjects had ever experienced a hypomanic episode. In 2 cases, the interviewers disagreed on the hypomania diagnosis (see Table 1 for agreement on the hypomania diagnosis). There was disagreement over the hypomania diagnosis in 3 of the 26 test-retest subjects. In 2 cases, the interviewers differed on whether the subjects had ever experienced a hypomanic episode. In the third, one interviewer diagnosed mania and the other diagnosed hypomania.

### DIAGNOSTIC RELIABILITY AT THE BEST-ESTIMATE LEVEL

We compared affective disorder diagnoses assigned by pairs of noninvestigating psychiatrists making best-estimate diagnoses on 524 members of families ascertained through probands with BP disorder (Table 2). There was agreement on 98% of the affective disorder diagnoses. The values (SE) for best-estimate diagnoses were 0.99 (0.006) for BPI, 0.99 (0.007) for BPII, and 0.98 (0.014) for RUP. All values were significant at the P<.001 level.

### COMMENT

Our findings from the co-rated and test-retest SADS-L interviews indicate that good interrater diagnostic reliability can be achieved for BPII if experienced clinicians (in this case, psychiatrists) conduct the interviews. These results are consistent with those of Dunner and Tay, but challenge the prevalent view that BPII cannot be diagnosed reliably. Although our sample size for the test-retest study was modest and included mainly inpatients, it included twice as many subjects with BPII as the sample that has been considered the benchmark for reliability of the BPII diagnosis. The prevalent view that the hypomania diagnosis has low reliability was based on a sample of 50 subjects from the National Institute of Mental Health Psychobiology of Depression Study, only 3 of whom were diagnosed as having BPII. Those subjects were interviewed twice, 6 months apart, and at time 2 they were interviewed twice on the same day. There was good diagnostic agreement on hypomania on the same-day interviews, with an intraclass correlation coefficient of 0.6, but poor diagnostic agreement between the interviews done 6 months apart, with an intraclass correlation coefficient of 0.06. In another study of similar size, Maziade and Gershon did test-retest SADS-L interviews 6 months apart and reported relatively good agreement, with an overall κ of 0.79. Of the 3 subjects with hypomania, 2 were diagnosed as hypomanic at both assessments, compared with 6 of 6 subjects diagnosed with mania.

The hypomania diagnosis has been shown to have predictive power even if made on only 1 of 2 interviews. In a group of relatives from the Psychobiology of Depression Study who were interviewed 5 years apart, 9 of 10 subjects who were diagnosed as having hypomania at one interview were not diagnosed with it at the other interview, but all 10 cases predicted a proband with BPI or BPII.

Conclusions regarding the reliability of the best-estimate diagnoses must consider the limited independence of the best-estimate diagnosticians in this study. Our procedures, which were devised to maximize diagnostic validity for genetic studies, allow for some communication between reviewers. We compared our best-estimate reliability study to a recent study by Maziade et al. They compared diagnoses made by psychiatrists in the field based on a summary of all of the clinical data, to diagnoses made by a board of research psychiatrists who were blind to the probands’ and relatives’ diagnoses. Two research psychiatrists reviewed all available clinical data and made best-estimate diagnoses independently, after which they discussed the case. If they agreed, the complete record was presented to the board for a final consensus diagnosis. If they disagreed, the complete record was reviewed by 2 other psychiatrists, who made independent diagnoses and sent the case to the board.

While there are many similarities with our diagnostic procedure, there are also some differences. Our psychiatrist-interviewers make diagnoses based only on the SADS-L interview. We have a panel of 6 psychiatrists (composed of 6 of us) who serve as best-estimate diagnosticians, except on subjects whom they had interviewed. Any 2 psychiatrists separately review all sources of data and make their diagnoses. If there is a difference of opinion between diagnosticians, they are allowed to

### Table 1. Agreement on Hypomania Diagnoses in Test-Retest SADS-L Interviews*

<table>
<thead>
<tr>
<th>Interviewer 1</th>
<th>Hypomania</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypomania</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>17</td>
</tr>
</tbody>
</table>

* SADS-L indicates Schedule for Affective Disorders and Schizophrenia—Lifetime version. The total sample size was 26 subjects.
discuss the case. This process allows correctable errors, such as misreading of the clinical data, to be detected. If the diagnosticians still do not agree, agreement is not forced, nor is a third psychiatrist brought in as a tie-breaker. Instead, for the purpose of assigning a phenotype for genetic study, the subject is given the more "conservative" of the 2 diagnoses. By conservative we mean that if 1 reviewer made the diagnosis of BPII (with recurrent major depression) and the other made the diagnosis of RUP, we would assign RUP as the diagnosis. We, thus, would capture the part of the diagnosis agreed on and would exclude the subject from the narrowest affection status in the genetic analyses.

Applicability of our findings to clinical samples must be qualified as follows. While most of the test-retest subjects and some of the co-rated subjects were drawn from the general psychiatry inpatient units and are, therefore, likely to reflect a representative clinical population, some of the co-rated subjects were relatives of patients but not patients themselves. In addition, the sample for the best-estimate reliability study consisted of patients with BP and their relatives, some of whom were affected and some of whom were not, and may not be representative of clinical samples. Our findings may not be applicable to community-based, nontreatment seeking samples where reliability might be considerably more variable.

In addition to using DSM-IV criteria, we believe it is important to continue to use the RDC in genetic studies of BP disorder so that findings can be compared with those of earlier studies. "Probable hypomania" as defined by RDC requires a minimum of 2 manic symptoms for at least 2 days without associated impairment, while DSM-IV requires 3 symptoms for at least 4 days and an associated change in function that is observable by others. While raising the diagnostic threshold will improve the reliability of the hypomania diagnosis, raising the threshold may also decrease the sensitivity, with milder episodes going undiagnosed.

We have demonstrated that experienced psychiatrists using a semistructured interview such as the SADS-L can reliably diagnose BPII. This has important implications for genetic studies of BP disorders since individuals with BPII are prevalent in families ascertained through probands with BP. Based on clinical data from our family studies, we have proposed that BPII may be genetically less complex than BPI and that identifying the subjects with BPII in these families may be crucial to understanding the genetics of BP disorder generally. Furthermore, misdiagnosing BPII as recurrent unipolar depression decreases the power of the sample to identify genes for BP disorder.

Misdiagnosing BPII as unipolar depression also has important clinical implications, both for training of clinicians and for treatment. Treatment with antidepressant medications alone may lead to a worsening of the course of illness (with the possible development of mixed states or rapid-cycling) and may also deprive the patient of the potential benefits of mood stabilizing medications.

Submitted for publication April 11, 2000; final revision received September 26, 2001; accepted October 1, 2001.

This study was supported by grants from the National Institute of Mental Health, Rockville, Md; the Charles A. Dana Foundation, New York, NY; the National Alliance for Research on Schizophrenia and Depression, Great Neck, NY (Dr Simpson); the Ted and Varda Stanley Foundation, Arlington, Va; and contributors to the Affective Disorders Fund and the George Browne Laboratory Fund at The Johns Hopkins Hospital.

This study was presented as a poster at the 1995 World Congress of Psychiatric Genetics, Cardiff, Wales, August 30, 1995.

We thank the many research assistants, technicians, secretaries, and medical students who have contributed their energies to this study. We also thank the clinicians who referred families for study and the family volunteers and individuals who volunteered, without whose collaboration this research would not have been possible.

Corresponding author and reprints: Sylvia G. Simpson, MD, University of Colorado Health Sciences Center, 4200 E Ninth Ave, Box C268-71, Denver, CO 80262 (e-mail: sylvia.simpson@uchsc.edu).

Table 2. Agreement Between Diagnosticians 1 and 2 on 524 Best-Estimate Diagnoses

<table>
<thead>
<tr>
<th>Diagnostician 1</th>
<th>Unaffected</th>
<th>Uncertain</th>
<th>Bipolar I</th>
<th>Bipolar II</th>
<th>Recurrent Unipolar</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostician 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaffected</td>
<td>144</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>149</td>
</tr>
<tr>
<td>Uncertain</td>
<td>3</td>
<td>123</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>126</td>
</tr>
<tr>
<td>Bipolar I</td>
<td>0</td>
<td>1</td>
<td>104</td>
<td>0</td>
<td>0</td>
<td>105</td>
</tr>
<tr>
<td>Bipolar II</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>86</td>
<td>0</td>
<td>87</td>
</tr>
<tr>
<td>Recurrent unipolar</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>55</td>
<td>57</td>
</tr>
<tr>
<td>Total</td>
<td>147</td>
<td>132</td>
<td>104</td>
<td>86</td>
<td>55</td>
<td>524</td>
</tr>
</tbody>
</table>

REFERENCES


20. SPSS Base 7.0 Applications Guide. Chicago, III: SPSS Inc; 1996.


