Longitudinal diagnostic efficiency of DSM-IV criteria for borderline personality disorder: a 2-year prospective study

Carlos M. Grilo
Charles A. Sanislow
Andrew E. Skodol
John G. Gunderson, Harvard Medical School
Robert L. Stout, et al.
Longitudinal Diagnostic Efficiency of DSM-IV Criteria for Borderline Personality Disorder: A 2-Year Prospective Study

Carlos M Grilo, PhD¹, Charles A Sanislow, PhD², Andrew E Skodol, MD³, John G Gunderson, MD⁴, Robert L Stout, PhD⁵, Donna S Bender, PhD⁶, Shirley Yen, PhD⁷, M Tracie Shea, PhD⁸, Les C Morey, PhD⁹, Mary C Zanarini, EdD¹⁰, Thomas H McGlashan, MD¹

Objective: To examine the longitudinal diagnostic efficiency of the DSM-IV criteria for borderline personality disorder (BPD).

Methods: At baseline, we used semistructured diagnostic interviews to determine criteria and diagnoses; blinded assessments were performed 24 months later with 550 participants. Diagnostic efficiency indices (specifically, conditional probabilities, total predictive power, and kappa) were calculated for each criterion determined at baseline, with the independent BPD diagnosis at follow-up used as the standard.

Results: Longitudinal diagnostic efficiencies for the BPD criteria varied, with the criteria of suicidality or self-injury and unstable relationships demonstrating the most predictive utility.

Conclusions: BPD criteria differ in their predictive utility for the diagnosis of BPD when considered longitudinally. These findings have implications both for clinicians who are considering diagnoses and for researchers concerned with forthcoming revisions of our nosological system.

(Can J Psychiatry 2007;52:357–362)

Information on funding and support and author affiliations appears at the end of the article.

Clinical Implications

- BPD criteria differ in their utility for predicting the presence of BPD diagnosis 2 years later.
- The criteria of suicidality and unstable relationships had the highest predictive utility.
- Although it is not advisable to derive diagnoses from only partial criteria, the findings have relevance for future nosological refinements.

Limitations

- The diagnostic criteria were not randomly assessed, and this might artificially inflate the degree to which they appear interrelated.
- The findings may not generalize to the community or to groups not seeking treatment.
- The findings may not generalize to individuals younger or to older than our patient group aged 18 to 45 years.

Key Words: borderline personality disorder, diagnosis, diagnostic efficiency, longitudinal follow-along, suicidality, impulsivity, unstable relationships
Borderline personality disorder is a prevalent disorder associated with considerable psychiatric comorbidity, functional impairment, and use of treatment. The current DSM-IV BPD diagnostic construct follows from the seminal contributions of Gunderson and Singer and Spitzer and colleagues. Although the original BPD construct evolved from clinical observations and empirical research, one concern has been the availability of sufficient empirical justification for the revisions across the DSM editions. Indeed, the DSM-IV Task Force and Work Group aimed to minimize arbitrary revisions.

In terms of psychometric approaches to classification, diagnostic efficiency is one fundamental approach to addressing questions about the construct validity of BPD. Diagnostic efficiency refers to the extent to which criteria are able to discriminate individuals with a given disorder from those without that disorder, as determined by the application of conditional probabilities. This approach has helped to refine certain diagnoses, and such data influenced, to some degree, the DSM-IV Work Group for BPD. For the DSM-IV, 2 BPD criteria were unchanged, 6 criteria had minor revisions, and 1 new criterion was added (paranoia or dissociation).

Although the DSM-IV Work Group reviewed available performance data for BPD criteria, some of these data were collected without the benefit of semistructured diagnostic interviews. We are aware of only 4 published studies of BPD diagnostic efficiency that used semistructured interviews—one for DSM-III-R criteria and 3 for the DSM-IV.

The study of the 8 DSM-III-R BPD criteria found that they differed little in their predictive capacity in a study group of psychiatric inpatients. In contrast, the 3 DSM-IV studies reported substantial variability in the efficiency of the 9 criteria for predicting the BPD diagnosis. Farmer and Chapman found that 6 of the 9 BPD criteria had PPP of less than 50/50 chance (mean 0.45, range 0.35 to 0.53) in a heterogeneous community sample (n = 149) of respondents to advertise-ments. It is worth noting, however, that Farmer and Chapman used as their criterion standard the BPD diagnosis derived solely from the diagnostic interview; thus, their findings might be best viewed as a type of item-total correlation rather than as conditional probabilities. Findings from our initial report based on a large, heterogeneous patient group (n = 668) suggested that all BPD criteria at baseline performed better than 50/50 chance for predicting the BPD diagnosis, with a mean PPP of 0.71 and a range from 0.61 (for the criterion of chronic emptiness) to 0.82 (for the criterion of suicidality or self-injury). More recently, Grilo and colleagues examined this issue in a patient group of 130 monolingual Hispanic outpatients and found that 8 of the 9 BPD criteria performed better than 50/50 chance, with a mean PPP of 0.61 and a range from 0.49 (for affective instability) to 0.82 (for suicidality or self-injury). Taking both positive and negative predictive capacity into account, the criterion of suicidality or self-injury demonstrated the greatest overall utility in correctly diagnosing BPD. In both of these studies, the criterion of affective instability performed best for exclusionary purposes (that is, it had the highest NPP).

This study aimed to extend our initial analysis by examining the longitudinal diagnostic efficiency of the DSM-IV criteria for BPD. We used data from the CLPS, an ongoing prospective repeated-measures study designed to examine disease course and outcome among patients with PDs. Specifically, we analyze the diagnostic efficiency of the BPD criteria for predicting the presence and (or) absence of the BPD diagnosis determined independently 2 years later. We refer to this as “longitudinal” diagnostic efficiency and it represents—to our knowledge—the first application of such conditional probabilities to the diagnosis of BPD. This approach is particularly relevant to the study of PDs because their current definition is based in part on their presumed stable and enduring nature.

Recent empirical findings, however, suggest that the stability of BPD above the diagnostic threshold over time is modest. Therefore, it seems relevant to use a criterion of the presence or absence of the BPD diagnosis assessed independently 2 years later for understanding the predictive utility of the BPD criteria.

**Methods**

**Participants**

The overall study aims, methods, and characteristics of participants are described elsewhere. Most participants were patients recruited from in- and outpatient clinical programs at

---

**Abbreviations used in this article**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPD</td>
<td>borderline personality disorder</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>CLPS</td>
<td>Collaborative Longitudinal Personality Disorders Study</td>
</tr>
<tr>
<td>DIPD-IV</td>
<td>Diagnostic Interview for DSM-IV Personality Disorders</td>
</tr>
<tr>
<td>GAF</td>
<td>Global Assessment of Functioning</td>
</tr>
<tr>
<td>NPP</td>
<td>negative predictive power</td>
</tr>
<tr>
<td>OR</td>
<td>odds ratio</td>
</tr>
<tr>
<td>PD</td>
<td>personality disorder</td>
</tr>
<tr>
<td>PPP</td>
<td>positive predictive power</td>
</tr>
<tr>
<td>SD</td>
<td>standard deviation</td>
</tr>
<tr>
<td>TPV</td>
<td>total predictive value</td>
</tr>
</tbody>
</table>
each of the 4 CLPS recruitment sites, with the intent to obtain a diverse and clinically representative study group. Advertising was also used to recruit participants with current or past psychiatric treatment. The current report is based on 550 subjects from the overall CLPS study group (n = 668), for whom complete data through the 24-month follow-up assessment were available at the time of analysis. Participants in this study for whom follow-up data were available did not differ from those who were not included in the analyses in age, sex, ethnicity, or PD group including BPD, nor did they differ in the frequency of any of the 9 BPD criteria (P levels for continuity-adjusted chi-square tests ranged from 0.20 to 0.97).

Of the 550 participants in this study, 352 (64%) were women, and most were white (n = 419, 76%). The participants’ mean age was 32.9 years, SD 8.3, and the mean GAF score at baseline for the past month was 51.5, SD 9.7. As detailed elsewhere,20 cooccurring Axis I and Axis II disorders were common and comparable to findings in other clinical samples.21

Procedures
All participants provided written informed consent following a full description of study procedures. The study was approved by each collaborating site’s institutional review board. Participants were interviewed in person by experienced interviewers who had undergone extensive standardized training to achieve reliability in the administration of the diagnostic measures. Interviewers were also monitored and received regular, ongoing supervision by the investigators at each site; as well, there was regular supervision across sites to maintain reliability and prevent drift over time.

The DIPD-IV,22 a semistructured interview, was administered to assess all PDs (that is, the 10 formal and 2 research criteria of the DSM-IV). Each criterion for all disorders is assessed with one or more questions, which are then rated on a 3-point scale (0 = not present; 1 = present but of uncertain clinical significance; 2 = present and clinically significant). The time frame covered is the prior 2 years, but to be counted toward a diagnosis, traits or behaviors must be characteristic of the person for most of his or her adult life.

DIPD-IV Reliability. Interrater reliability kappa coefficients were based on 84 pairs of raters using taped interviews and ranged from 0.58 to 1.00 (median 0.68) for all PDs; for BPD, \( \kappa = 0.68 \).23 One-week test–retest reliability for BPD (based on 52 cases) ranged from \( \kappa = 0.69 \) to \( \kappa = 0.74 \).23 For the specific BPD criteria, kappa coefficients for interrater and test–retest reliability, respectively, were as follows: 0.84 and 0.55 (abandonment fears), 0.70 and 0.57 (unstable relationships), 0.56 and 0.61 (identity disturbance), 0.64 and 0.55 (impulsivity), 0.75 and 0.62 (suicidality or self-injury), 0.66 and 0.60 (affective instability), 0.73 and 0.65 (feelings of emptiness), 0.67 and 0.65 (inappropriate anger), and 0.60 and 0.56 (paranoia or dissociation).

Follow-Up Evaluations. Participants were reinterviewed at 6, 12, and 24 months after the baseline assessment. The 24-month assessment included a separate, independent administration of the DIPD-IV to assess all PDs. This DIPD-IV was administered by an interviewer blind to baseline and all interim assessments.

Statistical Analysis
We considered longitudinal diagnostic efficiency of the BPD criteria at baseline entry into the study against the BPD diagnosis obtained at 2-year follow-up. We examined efficiency of the criterion (that is, a DIPD-IV rating of 2), using 4 types of conditional probabilities. The first, sensitivity (true-positive rate), is the proportion of patients with the diagnosis who meet the criterion—or the probability of meeting the criterion, given that one has the diagnosis. The second, specificity (true-negative rate), is the probability of not meeting the criterion, given that one does not have the diagnosis. The third, PPP, is the probability of having the diagnosis, given that one meets the criterion. Lastly, NPP is the probability that one does not have the diagnosis, given that one does not meet the criterion. In addition, TPV, a measure of percentage agreement (not a conditional probability), was calculated as an overall indicator of a criterion’s utility in making a correct diagnosis. We also calculated the kappa coefficient, which corrects for chance agreement.

PPP indicates whether a criterion will have utility as an inclusion criterion, and the relative values of the PPPs for the various criteria of a diagnosis can provide information about which criteria are the best predictors of the disorder. All else being equal, PPP increases and NPP decreases with increasing diagnosis base rates.24,25 TPV has an advantage over PPP and NPP because it relates to both inclusion and exclusion. While the kappa coefficient provides information similar to the TPV, it has an important advantage in that its scale does not inflate a criterion’s prediction of diagnosis.

Results
Table 1 shows the base rates of the 9 BPD criteria at baseline. Affective instability was the most frequently found BPD criterion, occurring in 54% of all participants. It was followed by inappropriate anger (43%) and impulsivity (46%). The least frequently found criterion was suicidality or self-injury (26%).

At 24-month follow-up, 119 of 550 participants (22%) received the BPD diagnosis. Table 1 also shows, for each criterion, the longitudinal diagnostic efficiency indices for predicting the diagnosis of BPD at the 24-month blinded DIPD-IV assessment. Each of the efficiency indices was quite
variable across the criteria. For predicting BPD at 24 months, all PPPs are less than 0.50 (which reflects a 50/50 chance). Of course, given the BPD base rate of 0.22, all 9 criteria performed better than chance, suggesting that they do possess some degree of predictive utility. In this regard, the criterion of suicidality or self-injury had the highest PPP (0.48), followed by the criteria of unstable relationships (0.43) and abandonment fears (0.43). Affective instability had the highest NPP (0.95), suggesting its efficiency for exclusionary purposes. Taking both PPP and NPP into account, the TPVs indicate that suicidality or self-injury (0.77) had the most overall utility in correctly diagnosing BPD. Notably, the kappa findings, which take into account chance, highlight the overall utility of 2 criteria—suicidality or self-injury and unstable relationships—for predicting the diagnosis of BPD 2 years later.

Our longitudinal diagnostic efficiency findings have particular relevance, given emerging questions about the course of PDs. Recent findings suggest that the stability of BPD at diagnostic threshold levels over time is modest. The findings presented here can be considered to address the critical question of which BPD criteria are most strongly predictive of primarily stable cases of BPD. Thus our findings speak to specific aspects of the predictive validity of the DSM-IV BPD criteria, which have been substantially revised from earlier versions.

We used a logistic regression analysis (stepwise forward method) to examine the joint contributions of the BPD criteria at baseline for predicting the BPD diagnosis at 24-month follow-up. The overall model was significant ($\chi^2 = 168.10, P < 0.001$) and revealed 5 criteria as making significant independent contributions. These criteria, and their ORs and CIs, are as follows: unstable relationships (OR 3.86; 95%CI, 2.27 to 6.57; $P < 0.001$), affective instability (OR 3.82; 95%CI, 1.90 to 7.67; $P < 0.001$), emptiness (OR 2.34; 95%CI, 1.40 to 3.91; $P < 0.001$), suicidality or self-injury (OR 2.22; 95%CI, 1.32 to 3.74; $P < 0.004$), and paranoia or dissociation (OR 1.76; 95%CI, 1.05 to 2.96; $P < 0.04$).

Discussion

This study, using data obtained with reliably administered semistructured research interviews, examined the longitudinal diagnostic efficiency of the DSM-IV criteria for BPD for prospectively predicting the BPD diagnosis at 24-month follow-up. The best criterion for predicting the presence of BPD assessed 2 years later (that is, having the highest PPP) was suicidality or self-injury, which was also overall the most predictive criterion (having the highest TPV). The best criterion for predicting absence of BPD was affective instability. Kappa coefficients, which take into account chance, highlight the overall utility of 2 criteria—suicidality or self-injury and unstable relationships—for predicting the diagnosis of BPD 2 years later.

Our longitudinal diagnostic findings here are consistent with our baseline findings and with those reported for a separate group of Hispanic patients. In these studies, suicidality or self-injury had the highest PPP, and affective instability had the highest NPP. These striking convergences support this aspect of the construct validity of BPD across these 2 cultures.
editions of the DSM. This study also underscores the importance of studying the criteria sets for PDs longitudinally.

We note several limitations. The diagnostic interviews did not randomly assess the BPD criteria; thus, a “halo effect” is possible, whereby interviewers could potentially rate criteria similarly within a given diagnosis. This might artificially inflate the degree to which criteria appear to be interrelated and make them appear to be more similar in their efficiency. Alternatively, this method of assessing criteria grouped by disorder is more clinically ecologic because criteria are evaluated within the context of the disorder. Assessments were performed by trained and monitored interviewers. The variable reliability for the specific BPD criteria at baseline did not appear related to the patterns of diagnostic efficiency indices.

The study recruited patients and treatment-seeking participants, and our findings may not generalize to the community or to groups not seeking treatment. The pattern of diagnoses, however, resembles those reported for other clinical samples. Our sampling across diverse clinical settings affiliated with universities in 4 northeastern urban settings produced a heterogeneous adult study group that approximates US norms for ethnicity. Since we targeted individuals aged 18 to 45 years, our findings may not generalize to patient groups of different ages. Indeed, studies of BPD in younger adolescent patients have found that, despite some surface similarities in signs, patterns, or cutting scores. Psychol Bull. 1955;52:194–216.

Our findings highlight the particular relevance of the presence of 2 criteria (suicidality or self-injury and unstable relationships) or the absence of the criterion of affective instability. Of course, it is important to emphasize that deriving diagnoses only from partial criteria is not advisable and may prove to be erroneous. Indeed, Morey and Ochoa documented that incorrect BPD diagnoses were associated with clinicians’ frequently relying on the presence of depressed mood and impulsive suicidal behaviours. We offer these findings regarding the predictive utility of criteria within this important cautionary context, with the hope they inform future revisions of diagnostic schemes.

**Funding and Support**

This research was supported by the following research grants from the National Institutes of Health: MH 50837, 50838, 50839, 50840, 50850, K05 MH 01645, and MH 073708. There are no conflicts of interest or commercial products related to this study.

**References**


Manuscript received September 2006, revised, and accepted November 2006.
1Professor, Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut.
2Assistant Professor, Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut.
3Professor, New York State Psychiatric Institute, Columbia University, New York, New York.
Résumé : Efficacité diagnostique longitudinale des critères du DSM-IV pour le trouble de personnalité limite : une étude prospective de 2 ans

Objectif : Examiner l’efficacité diagnostique longitudinale des critères du DSM-IV pour le trouble de personnalité limite (TPL).

Méthode : Au départ, nous avons utilisé des entrevues diagnostiques semistructurées pour déterminer les critères et les diagnostics ; des évaluations en aveugle ont été effectuées 24 mois plus tard auprès de 550 participants. Les indices d’efficacité diagnostique (spécifiquement, les probabilités conditionnelles, la capacité prédictive totale, et l’indice kappa) ont été calculés pour chaque critère déterminé au départ, et le diagnostic de TPL indépendant au suivi servait de norme.

Résultats : L’efficacité diagnostique longitudinale des critères du TPL variait, et les critères de suicidabilité ou d’automutilation et de relations instables démontraient la plus grande utilité prédictive.

Conclusion : Les critères du TPL diffèrent d’utilité prédictive pour le diagnostic du TPL quand on les considère longitudinalment. Ces résultats ont des implications tant pour les cliniciens qui posent des diagnostics que pour les chercheurs qui se préoccupent des révisions à venir de notre système nosologique.