30-Month Stability of Personality Disorder Diagnoses in Depressed Outpatients

Tova Ferro, Ph.D., Daniel N. Klein, Ph.D., Joseph E. Schwartz, Ph.D., Karen L. Kasch, M.A., and Julie B. Leader, M.A.

Objective: This study examined the 30-month stability of axis II conditions. Method: One hundred eight depressed outpatients received comprehensive, semistructured personality disorder assessments at baseline and at follow-up. Results: The diagnostic stability of personality disorders ranged from low to moderate at the categorical level and was generally moderate at the dimensional level. Most disorders exhibited good discriminant validity, in that the association between a disorder at baseline and at follow-up was greater than the associations between that disorder at baseline and the other 11 axis II disorders at follow-up. Two variables, sex and lifetime history of substance abuse or dependence, were significantly related to change in level of personality disorder features over time. Conclusions: Personality disorders have low to moderate stability over a 30-month period in depressed outpatients.

(Am J Psychiatry 1998; 155:653–659)

In DSM-IV, personality disorders are defined as stable and enduring. However, little empirical evidence exists to substantiate this defining characteristic of personality disorders (1, 2). Stability is examined by determining the concordance between diagnoses in the same individual at two points in time. When the interval between diagnostic assessments is brief, much of the variance is due to differences in diagnosticians and is best conceptualized as an examination of interrater reliability (3). With longer time intervals (generally greater than 12 months) between assessments, presumably more of the variance is due to diagnostic/subject change over time, and the level of concordance can be viewed as an index of diagnostic stability (3).

Diagnostic stability has long been viewed as integral to establishing the validity of axis I disorders (4). However, there are relatively few empirical examinations of stability. Moreover, most axis I "stability" studies have examined consistency of recall for the same time period (generally lifetime up to the occasion of the subject's first assessment), rather than diagnostic stability versus change over time. Loranger et al. (5) summarized the stability of axis I diagnoses derived with the Schedule for Affective Disorders and Schizophrenia—Lifetime Version (6) in four studies with follow-up intervals of greater than 1 year. Stability coefficients, measured with kappa, for the major affective, anxiety, and substance abuse disorders ranged from 0.09 to 0.73 (median=0.51).

The diagnostic stability of axis II disorders has only recently been examined in the literature. Furthermore, most studies examining the stability of personality disorders are fraught with methodological limitations. For example, most studies focus on only a single personality disorder, the samples are small, chart reviews or self-reports are employed rather than structured diagnostic interviews, raters at the time of reassessment are not blind to index diagnoses, and data on interrater reliability are not reported.

Only four studies examining the diagnostic stability of personality disorders have assessed multiple axis II disorders and have had a follow-up of 12 months or longer (7–10). In the earliest report, Barasch et al. (7) examined the stability of borderline personality disorder and "other personality disorder" over a 3-year interval in 30 outpatients. DSM-III diagnoses of borderline personality disorder were stable in 60% of subjects, and "other personality disorders" were stable in 80% of subjects. In addition to the small sample size, limitations of the Barasch et al. study (7) include its reliance on chart reviews to make initial diagnoses and a high (61%) attrition rate.

In the second report, Bernstein et al. (8) followed up 724 adolescents 2 years after their initial personality disorder assessment. Stability of diagnosis was 6%–32% in subjects with moderate levels of disorder and 7%–43%

Received Aug. 5, 1997; revision received Nov. 18, 1997; accepted Dec. 4, 1997. From the State University of New York at Stony Brook; and New York State Psychiatric Institute, Columbia University. Address reprint requests to Dr. Ferro, Department of Child Psychiatry, Unit 78, New York State Psychiatric Institute, Columbia University, 722 West 168th St., New York, NY 10032.

Supported in part by NIMH grants MH-45757, MH-16434, and MH-43878.

The authors thank L. Klein and K. Norden for their help with data collection and processing.

TABLE 1. Demographic and Clinical Characteristics of Depressed Outpatients (N=108) in a Study of Personality Disorder Diagnoses

Characteristic	Mean	SD	N	%
Baseline				
Age (years)	31.6	8.8		
Female gender			78	72.2
Caucasian race			97	89.8
Education (years)	13.7	2.2		
Marital status				
Single			48	44.4
Married			35	32.4
Separated or divorced			23	21.3
Widowed			2	1.9
Socioeconomic status	37.0	14.0		
Global Assessment of Functioning Scale score	57.1	9.6		
Hamilton Depression Rating Scale score	26.8	9.6		
Diagnoses				
Lifetime major depression			93	86.1
Current major depression			76	70.4
Current dysthymia			73	67.6
Current anxiety disorder			30	27.8
Lifetime substance abuse or dependence			46	42.6
Follow-up				
Hamilton Depression Rating Scale score	14.7	10.7		
Treatment status				
Weekly psychotherapy			23	21.3
Biweekly psychotherapy			13	12.0
Receiving adequate dose of antidepressant medication			18	16.7

in subjects with severe levels of disorder. The limitations of the Bernstein et al. study (8) include the following: 1) subjects were adolescents, who may be in flux with regard to personality development; 2) items from many instruments, not designed for the assessment of personality disorders, were employed to approximate axis II diagnoses; and 3) not all DSM criteria could be assessed.

In the third study, Vaglum et al. (10) assessed 73 inpatients with axis II diagnoses 1.6–4.9 years after initial evaluation. Patients with cluster A and B diagnoses were combined into a "severe" group, while patients with cluster C and no personality disorder diagnosis were combined into a nonsevere group. Seventy percent of patients in the severe personality disorder group remained diagnostically stable (kappa=0.65). As with the Barasch et al. study (7), Vaglum et al. (10) relied on chart review, not structured interviews, to make initial diagnoses.

The most recent study assessed the 2-year stability of the full range of personality disorders as part of a study of 118 homosexual, HIV-positive and -negative men (9). Johnson et al. (9) employed a semistructured diagnostic interview, the Structured Clinical Interview for DSM-III-R Personality Disorders (11), to make axis II evaluations. However, their raters were not blind at reassessment, 19% of subjects were reinterviewed by the same rater, and all raters had access to baseline diagnostic information about all subjects. Kappa for the stability of any personality disorder was 0.29. Kappas for the three clusters ranged from 0.12 to 0.32. While the rates of specific personality disorders were too low to analyze, the stability of dimensional scores for individual disorders ranged from 0.08 to 0.70, with a median intraclass correlation coefficient (ICC) of 0.43. Changes in personality disorder symptoms were associated with changes in psychological distress but not HIV status. Unfortunately, the approach of Johnson et al. to analyzing change underestimates the effects of initial distress, and insufficient data were presented to determine whether the effects for distress at follow-up really represented change or primarily reflected cross-sectional associations.

In the present study, we report on the diagnostic stability of DSM-III-R personality disorders over a 30-month interval in 108 depressed outpatients. Diagnoses were based on semistructured interviews assessing the full range of axis II disorders, and follow-up evaluations were conducted by raters who were blind to the patients' initial diagnostic status.

METHOD

Subjects

The patients and method at the index evaluation have been described in detail in prior publications (12–14). In brief, the original study group consisted of 97 outpatients diagnosed with DSM-III-R primary, early-onset dysthymia and 45 outpatients with DSM-III-R nonchronic (episodic) major depression. Subjects were between the ages of 18 and 60, were English speaking, were not currently psychotic, were never psychotic outside of a major depressive episode, and had knowledge of at least one first-degree relative. In addition, subjects with episodic major depression were required to have an onset before age 35, and the depression could not be due to another axis I or chronic medical condition. The majority of subjects were selected from consecutive admissions to the State University of New York at Stony Brook Outpatient Psychiatry Department and Psychological Center. Several subjects were referred from a community mental health center and the State University of New York at Stony Brook University Counseling Center. All subjects were given a complete description of the study, and written informed consent was obtained.

Complete follow-up evaluations of personality disorders were available for 108 patients (76.1%) (75.3% of subjects with early-onset dysthymia and 77.8% of subjects with episodic major depression). All follow-up assessments were limited to the previous 30 months, which was the period since the baseline evaluation. Follow-up assessments were conducted a median 31 months (range=29–45) after the baseline evaluation. Descriptive characteristics of the patients are presented in table 1. Patients with axis II data at follow-up did not differ from the 34 patients for whom axis II data at follow-up were not available on sex, age, race, marital status, socioeconomic status according to the Hollingshead index (15), the Global Assessment of Functioning Scale from the Structured Clinical Interview for DSM-III-R (SCID) (16), the modified Hamilton Depression Rating Scale (17), lifetime and current major depression, current dysthymia, current anxiety disorders, lifetime substance abuse or dependence, any personality disorder, and any cluster A, B, or C personality disorder at baseline. However, patients with complete axis II evaluations had significantly more education (mean=13.7 years, SD=2.2) than those who did not have axis II data at follow-up (mean=12.8, SD=2.1) ($t=2.13$, $df=140$, $p<0.04$).

Measures

At entry into the study, subjects were administered the SCID (16), 24-item Hamilton depression scale (17), and Personality Disorder Examination (18) and completed the Eysenck Personality Questionnaire (19). As described elsewhere, the interrater reliability of our baseline diagnoses was good to excellent (12).

The follow-up assessment included the Longitudinal Interval Follow-

Up Evaluation (20), 24-item Hamilton depression scale, Personality Disorder Examination, and Eysenck Personality Questionnaire. The Longitudinal Interval Follow-Up Evaluation is a semistructured interview assessing the longitudinal course of axis I disorders and treatment through the follow-up period. The Personality Disorder Examination is a semistructured interview for the assessment of the 11 personality disorders included in DSM-III-R, with the addition of self-defeating and sadistic personality disorders and a category of personality disorder not otherwise specified. Diagnoses were made by using both narrow (definite only) and broad (definite or probable) thresholds. Probable Personality Disorder Examination diagnoses are made when patients are one symptom shy of meeting full criteria. The Personality Disorder Examination also yields dimensional scores, which consist of the summed ratings for all items within the diagnostic category. Because the follow-up assessments were based on the previous 30 months, the child conduct disorder items for antisocial personality disorder were not reassessed. Therefore, in the present study, follow-up diagnoses of antisocial personality disorder were based on the adult portion of the criteria only. To maintain parity, the child conduct items were excluded from the baseline antisocial personality dimensional scores.

The Eysenck Personality Questionnaire is a widely used, broad band, personality inventory. It includes three scales: extraversion (versus introversion), neuroticism (versus emotional stability), and psychoticism (which taps primarily impulsivity and antisocial behavior).

Follow-up interviews were conducted by a master's level psychiatric social worker with several years of research diagnostic experience, a doctoral level clinical psychology research fellow, and three advanced graduate students in clinical psychology with previous training and experience in diagnostic interviewing. To guard against biases in the evaluation of diagnostic stability, different interviewers conducted baseline and follow-up assessments with all patients, and follow-up interviewers were blind to all baseline data.

In order to assess interrater reliability, one rater independently rated audiotapes of several randomly selected Hamilton depression scale interviews conducted by each of the other interviewers in the study (total N=13). The ICC (case 1) (21) was 0.95. Interrater reliability of the Personality Disorder Examination was assessed through independent evaluations of 20 videotaped Personality Disorder Examination interviews. Interrater reliability, expressed with kappa, was 0.80 for any personality disorder and 0.44, 0.76, and 0.88 for clusters A, B, and C, respectively. Kappas for the three disorders with prevalence rates over 5% were 0.73 for avoidant personality and 0.69 for both borderline and histrionic personality. ICCs (case 1) (21) for the dimensional scores ranged from 0.76 to 0.90 (median=0.84).

Data Analysis

Although patients with early-onset dysthymia and episodic major depression differed on rates of personality disorders (14), they had similar levels of stability and hence were combined for all analyses. The degree of association between initial and follow-up axis II data was calculated with kappa for diagnostic categories and the ICC (case 1) (21) for dimensional ratings. Paired t tests were employed to measure change in dimensional scores between time 1 and time 2. The discriminant validity of axis II disorders over time was examined with Pearson product-moment correlations and hierarchical multiple linear regression analyses. Finally, predictors of change in total Personality Disorder Examination dimensional scores were examined with repeated measures analysis of variance; total dimensional score at baseline and follow-up was used as the within-subjects measure (time), categorical predictors were the between-subjects measure, and dimensional predictors were used as covariates. Significant predictors of change were identified by a significant predictor-by-time interaction.

RESULTS

Diagnostic Stability

Data on the 2½-year stability of DSM-III-R personality disorder diagnoses are presented in table 2. Kappas were calculated for all diagnoses with a prevalence of greater than 5% at both assessments. The data were analyzed by using both narrow (definite only) and broad (probable or definite) thresholds. Data on sadistic personality disorder are not presented, since no patients received this diagnosis at either evaluation.

The level of diagnostic stability varied across disorders and thresholds. The kappas for any personality disorder were 0.41 and 0.22 when narrow and broad thresholds, respectively, were used. The kappas for the three clusters ranged from 0.24 to 0.44 when a narrow threshold was used and from 0.28 to 0.60 with a broad threshold. When the narrow threshold was used, kappas for the two specific disorders with sufficient prevalence rates, borderline personality and avoidant personality, were 0.54 and 0.24, respectively. With the broad threshold, kappas for the five specific disorders with adequate base rates ranged from 0.33 to 0.73, with a median of 0.48.

Of patients with any definite personality disorder diagnosis at entry into the study, 51% continued to meet criteria for a definite personality disorder, and another 22% met criteria for a probable personality disorder at follow-up. Of patients entering the study with a definite or probable personality disorder, 59% met criteria for a definite or probable personality disorder at follow-up.

As can be seen in table 2, personality disorder dimensional scores were moderately stable over 2½ years. The ICC between total dimensional scores at baseline and follow-up was 0.52. ICCs for dimensional scores for the three clusters ranged from 0.44 to 0.55. The ICCs for individual disorders were highest for histrionic (0.65) and avoidant (0.52), and lowest for narcissistic (0.22) and antisocial (0.27), with a median of 0.48 across all disorders.

These data probably underestimate stability, since they are attenuated by imperfect interrater reliability. In the last column of table 2, we present the ICCs corrected for attenuation, using the paired rater-interrater reliability videotape data summarized earlier. Corrected for attenuation, the stability of dimensional scores for specific personality disorders ranged from 0.25 to 0.72, with a median of 0.58. These figures should be regarded as conservative, since the paired rater design provides an upper-bound estimate of interrater reliability, thereby minimizing the degree of correction for attenuation.

For comparative purposes, we also examined the stability of the Eysenck Personality Questionnaire. The ICCs for extraversion, neuroticism, and psychoticism were 0.73, 0.52, and 0.70, respectively.

We also examined whether mean levels of dimensional scores changed over time. As can be seen from table 2, there was a significant decrease over time in total dimensional scores and dimensional scores for cluster B, cluster C, and schizotypal, antisocial, borderline, histrionic, dependent, avoidant, obsessive-compulsive, passive-aggressive, and self-defeating personality disorders.

TABLE 2. Concordance Between Baseline and 30-Month Follow-Up Assessments of Personality Disorder in Depressed Outpatients (N=108)

	Patients With Definite Diagnoses							Patients With Definite/Probable Diagnoses						
	Baseline		Follow-Up		Retained[a]			Baseline		Follow-Up		Retained[a]		
Personality Disorder	N	%	N	%	N	%	Kappa[b]	N	%	N	%	N	%	Kappa[b]
Any disorder	37	34.3	28	25.9	19	51.4	0.41	70	64.8	54	50.0	41	58.6	0.22
Cluster A	9	8.3	9	8.3	3	33.3	0.27	19	17.6	17	15.7	12	63.2	0.60
Cluster B	19	17.6	9	8.3	7	36.8	0.44	28	25.9	21	19.4	12	42.9	0.34
Cluster C	20	18.5	10	9.3	5	25.0	0.24	35	32.4	27	25.0	15	42.9	0.28
Paranoid	8	7.4	5	4.6	2	25.0		16	14.8	12	11.1	9	56.3	0.59
Schizoid	1	0.9	4	3.7	1	100.0		3	2.8	6	5.6	2	66.7	
Schizotypal	3	2.8	0	0.0	0	0.0		6	5.6	1	0.9	0	0.0	
Antisocial	4	3.7	1	0.9	0	0.0		6	5.6	4	3.7	1	16.7	
Borderline	16	14.8	8	7.4	7	43.8	0.54	22	20.4	11	10.2	8	36.4	0.40
Histrionic	8	7.4	3	2.8	1	12.5		14	13.0	15	13.9	8	57.1	0.48
Narcissistic	4	3.7	0	0.0	0	0.0		4	3.7	1	0.9	0	0.0	
Dependent	7	6.5	0	0.0	0	0.0		17	15.7	4	3.7	2	11.8	
Avoidant	13	12.0	7	6.5	3	23.1	0.24	22	20.4	18	16.7	9	40.9	0.33
Obsessive-compulsive	3	2.8	2	1.9	0	0.0		8	7.4	5	4.6	1	12.5	
Passive-aggressive	1	0.9	2	1.9	0	0.0		4	3.7	6	5.6	1	25.0	
Self-defeating	7	6.5	4	3.7	3	42.9		9	8.3	7	6.5	6	66.7	0.73
Not otherwise specified	9	8.3	8	7.4	1	11.1	0.04	27	25.0	27	25.0	8	29.6	0.06

[a] Patients with the diagnosis at baseline who retained the diagnosis at follow-up.
[b] Calculated only for disorders with a prevalence of more than 5% in both assessments.
*p<0.10. **p<0.05. ***p<0.01. † p<0.005. †† p<0.001.

Discriminant Validity

The next set of analyses examined the discriminant validity of axis II disorders over time. In other words, is the relationship between a personality disorder at entry into the study and the same disorder at follow-up stronger than the relationships between that personality disorder at baseline and the other disorders at follow-up? These analyses were conducted at the dimensional level only because of the low rates of some diagnoses. Because of space limitations, these data are not presented in detail here; however, they can be obtained from Dr. Ferro upon request.

The correlations between the baseline and follow-up dimensional scores for the same disorder were higher than the highest correlations between the baseline dimensional score for the index disorder and the follow-up dimensional scores for all the other disorders, with the exceptions of narcissistic and self-defeating personality. With only one exception (narcissistic), the correlations between the baseline and follow-up dimensional scores for the same disorder were higher than the median correlations between the index disorder at baseline and all the other disorders. Narcissistic personality at baseline was more highly correlated with eight other disorders than with itself at follow-up, while self-defeating personality at baseline was more highly correlated with borderline and dependent personality than with itself at follow-up.

The differences between the correlation of the index disorder at baseline with the same disorder at follow-up and the correlation of the index disorder at baseline with the most highly correlated other disorder at follow-up ranged from –0.21 to 0.19, with a median of 0.09. The differences between the correlation of the index disorder at baseline with the same disorder at follow-up and the median correlation of the index disorder at baseline with the other 11 disorders at follow-up ranged from –0.04 to 0.44, with a median of 0.31.

We also conducted a series of hierarchical multiple regression analyses in which we examined the unique association between each disorder at baseline and the same disorder at follow-up after controlling for all 11 other disorders at baseline, and the independent contribution of the 11 other disorders at baseline to predicting each disorder at follow-up after controlling for that same disorder at baseline. With the exception of antisocial and narcissistic personality, the baseline values of each disorder were significantly associated with the same disorder at follow-up even after we controlled for the baseline values of all of the other disorders. In addition, only paranoid and antisocial personality at follow-up were significantly predicted by the set of all other disorders at baseline after we controlled for paranoid and antisocial personality, respectively, at baseline.

Predictors of Change

In light of the decrease in personality disorder features over time, we explored whether individual differences in change were predicted by a variety of demographic and clinical features. Because the temporal decline was a fairly general phenomenon, characterizing most axis II disorders, we focused on predictors of change in total dimensional scores.

We examined 17 demographic and clinical predictors: age; sex; race; marital status; education; socioeconomic status; number of months between the initial and follow-up assessments; current major depression, dys-

Dimensional Score						
Baseline		Follow-Up		t		Corrected
Mean	SD	Mean	SD	(df=107)	ICC	ICC
38.6	23.5	28.5	20.3	5.35††	0.52	0.56
7.0	6.5	6.0	5.6	1.77*	0.55	0.63
12.6	9.9	8.8	8.1	4.77††	0.50	0.53
15.2	9.0	10.8	7.8	5.60††	0.44	0.50
2.9	3.1	2.4	2.7	1.59	0.48	0.63
1.4	2.0	1.7	2.3	1.11	0.50	0.60
2.6	2.9	1.9	2.0	3.30††	0.50	0.60
2.1	2.9	1.3	2.1	2.69***	0.27	0.30
5.4	3.9	3.4	3.5	6.15††	0.48	0.55
2.9	3.1	2.2	2.7	2.85†	0.65	0.72
2.3	2.7	1.8	2.0	1.76*	0.22	0.25
4.4	3.5	2.5	2.5	6.51††	0.34	0.45
3.8	3.4	3.0	2.9	2.98†	0.52	0.68
3.7	3.1	2.8	2.4	3.35††	0.44	0.56
3.2	2.7	2.6	2.9	2.19**	0.43	0.52
3.3	3.2	2.5	2.9	2.73***	0.50	0.59

thymia, and anxiety disorder at baseline; lifetime history of substance abuse or dependence; lifetime number of major depressive episodes; baseline scores on the Global Assessment of Functioning Scale and Hamilton depression scale; whether the patient was receiving psychotherapy or medication at follow-up; and recovery from depression at follow-up. Two definitions of recovery were employed: a Hamilton depression scale score of 8 or less at follow-up, and a 50% decrease in Hamilton depression scores between the baseline and follow-up evaluations.

Only two variables exhibited significant interactions with time: sex (F=7.01, df=1, 106, p=0.009) and a lifetime history of substance abuse or dependence (F=7.72, df=1, 106, p=0.006). The decrease in total dimensional scores over time was significantly greater for men (mean= 18.0, SD=23.7) than for women (mean=7.1, SD=17.2). In addition, patients with a lifetime history of substance abuse or dependence exhibited a significantly greater decrease in total dimensional scores (mean=16.1, SD=23.2) than patients without a history of substance abuse or dependence (mean=5.7, SD=15.4). When both variables were included in the same analysis simultaneously, both interactions were significant, indicating that they had independent effects.

DISCUSSION

A core, defining feature of personality disorders is stability. However, only limited data on the stability of personality disorders are available (1, 2). We examined the stability of DSM-III-R personality disorders at both the categorical and dimensional levels over a 2½-year period. At the diagnostic level, stability varied across disorders and diagnostic thresholds. The kappas for any personality disorder were 0.41 with a narrow, and 0.22 with a broad, threshold. When a narrow threshold was used, kappas for the two specific disorders with sufficient prevalences were 0.54 and 0.24; when a broad threshold was used, kappas for the five disorders with adequate base rates ranged from 0.33 to 0.73, with a median of 0.48. Overall, of patients with a definite personality disorder at entry into the study, 73% met criteria for a definite or probable personality disorder at follow-up.

At the dimensional level, most disorders were moderately stable. The ICCs for specific disorders ranged from 0.22 to 0.65, with a median of 0.48. Corrected for attenuation due to imperfect interrater reliability, the ICCs ranged from 0.25 to 0.72, with a median of 0.58. As noted earlier, paired rater-interrater reliability data provide a very conservative correction for attenuation. The corrected ICCs probably would have been considerably higher if test-retest reliability data had been employed.

We also compared the stability of personality disorder features to that of normal-range personality traits. Dimensional scores on the Personality Disorder Examination were not as stable as Eysenck Personality Questionnaire extraversion and psychoticism scores, but the stability estimates for a number of disorders were similar to Eysenck Personality Questionnaire neuroticism.

Borderline personality, with a narrow threshold, and paranoid and self-defeating personality, with a broad threshold, were the most stable disorders at the categorical level. The personality disorder not otherwise specified category was particularly unstable, with kappas of 0.04 and 0.06 when narrow and broad thresholds, respectively, were used. This is not surprising, since personality disorder not otherwise specified lacks specific defining features and, at least as operationalized by the Personality Disorder Examination, is a residual category for patients with a number of personality disorder features who do not meet criteria for any other category.

At the dimensional level, the most stable disorders were histrionic and avoidant personality, while narcissistic and antisocial personality were the least stable. The poor stability of antisocial personality may have been due, at least in part, to the restriction in range necessitated by considering only the adult portion of the criteria.

It is difficult to compare these results to previous reports of the stability of personality disorders because of the many methodological differences between studies. However, the present findings fall roughly within the mid-range of the stability estimates from other studies (1, 7, 9, 10). Overall, these data indicate that the stability of personality disorders is lower than expected given the nature of the construct but is broadly comparable to that of many of the better established axis I disorders (5).

Previous studies have not addressed the discriminant validity of personality disorders over time. We found that with the exception of narcissistic personality, the personality disorders had fairly good discriminant validity. In general, the initial assessments of each disorder correlated more highly with the follow-up assessments of that same disorder than with the follow-up assessments of the other disorders. Further, the baseline values of the other disorders as a set did not contribute significant variance over and above the baseline value of the index disorder in predicting the level of the index disorder at follow-up.

The rates and levels of personality disorders tended to decrease over time. This finding is consistent with the study of Loranger et al. (22) and may be due to regression to the mean or to state effects. However, it should be noted that similar declines in measures of psychopathology and personality are frequently reported in community samples, in which such factors are less likely to operate (23).

We were able to identify two variables that significantly predicted individual differences in change in personality disorder features over time: both men and patients with a lifetime history of substance abuse or dependence exhibited a significantly greater decrease in axis II symptoms between the initial and follow-up evaluations. The effect for sex may reflect a tendency for women to overreport maladaptive personality traits when they are at the peak of their distress and entering treatment; however, it may also reflect a tendency for men to underreport personality disturbance when they are not motivated by being in treatment or experiencing a high level of distress. The greater decline in personality disorder features over time among patients with a history of substance abuse or dependence may be due to the fact that substance abuse can mimic personality disorder. Although less than 20% of patients with a lifetime history of substance abuse or dependence met criteria within the month before the baseline evaluation, most patients with such a history had abused substances within the 5-year period used by the Personality Disorder Examination to assess personality disorders. With an additional 30 months of follow-up, many of the symptoms that had been attributed to personality disorder at entry into the study may have been recognized as effects of prior substance abuse or dependence or were no longer evident within the period of the assessment.

Recovery from depression was not associated with change in axis II features. It is important to note, however, that this does not rule out the possibility that a reduction in depression contributed to the mean decrease in Personality Disorder Examination dimensional scores for the study group as a whole. Rather, it indicates that recovery from depression was not associated with individual differences in the decrease in axis II symptoms over time.

This study has a number of strengths, including the use of comprehensive, structured assessments at both baseline and follow-up, follow-up interviewers who were blind to the initial diagnoses, and the use of both categorical and dimensional approaches. However, the study also has several significant limitations. First, for many axis II categories, the number of patients with the diagnosis was small, particularly when a narrow diagnostic threshold was employed. Second, because the study was initiated before DSM-IV, the diagnoses were based on DSM-III-R criteria. Hence, the stability of DSM-IV-defined personality disorders may differ slightly from the data reported here. Third, some of the instability of personality disorders in this study may reflect the limitations of the use of semistructured interviews for personality disorders (24). Thus, it would be important to employ other assessment techniques and methodologies in future research. Finally, the study was limited to patients with depressive disorders. While mood disorders are probably the most common diagnoses in outpatient practice, and the majority of patients with personality disorders have comorbid mood disorders (25, 26), it will be important to conduct further studies of diagnostic stability with unselected samples.

REFERENCES

1. McDavid JD, Pilkonis PA: The stability of personality disorder diagnoses. J Personality Disorders 1996; 10:1–15
2. Perry JC: Longitudinal studies of personality disorders. J Personality Disorders 1993; 7(suppl):63–85
3. Zimmerman M: Diagnosing personality disorders: a review of issues and research methods. Arch Gen Psychiatry 1994; 51: 225–245
4. Robins E, Guze SB: Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. Am J Psychiatry 1970; 126:983–987
5. Loranger AW, Sartorius N, Andreoli A, Berger P, Buchheim P, Channabasavanna SM, Coid B, Dahl A, Diekstra RFW, Ferguson B, Jacobsberg LB, Mombour W, Pull C, Ono Y, Regier DA: The International Personality Disorder Examination: the World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration International Pilot Study of Personality Disorders. Arch Gen Psychiatry 1994; 51:215–224
6. Endicott J, Spitzer RL: A diagnostic interview: the Schedule for Affective Disorders and Schizophrenia. Arch Gen Psychiatry 1978; 35:837–844
7. Barasch A, Frances A, Hurt S, Clarkin J, Cohen S: Stability and distinctness of borderline personality disorder. Am J Psychiatry 1985; 142:1484–1486
8. Bernstein DP, Cohen P, Velez CN, Schwab-Stone M, Siever LJ, Shinsato L: Prevalence and stability of the DSM-III-R personality disorders in a community-based survey of adolescents. Am J Psychiatry 1993; 150:1237–1243
9. Johnson JG, Williams JBW, Goetz RR, Rabkin JG, Lipsitz JD, Remien RH: Stability and change in personality disorder symptomatology: findings from a longitudinal study of HIV+ and HIV- men. J Abnorm Psychol 1997; 106:154–158
10. Vaglum P, Friis S, Karterud S, Mehlum L, Vaglum S: Stability of the severe personality disorder diagnosis: a 2- to 5-year prospective study. J Personality Disorders 1993; 7:348–353
11. Spitzer RL, Williams JBW, Gibbon M: Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II). New York, New York State Psychiatric Institute, Biometrics Research, 1987
12. Klein DN, Ouimette PC, Kelly HS, Ferro T, Riso LP: Test-retest reliability of team consensus best-estimate diagnoses of axis I and II disorders in a family study. Am J Psychiatry 1994; 151: 1043–1047
13. Klein DN, Riso LP, Donaldson SK, Schwartz JE, Anderson RL, Ouimette PC, Lizardi H, Aronson TA: Family study of early-onset dysthymia: mood and personality disorders in relatives of

outpatients with dysthymia and episodic major depression and normal controls. Arch Gen Psychiatry 1995; 52:487–496

14. Pepper CM, Klein DN, Anderson RL, Riso LP, Ouimette PC, Lizardi H: DSM-III-R axis II comorbidity in dysthymia and major depression. Am J Psychiatry 1995; 152:239–247
15. Hollingshead AB: Four Factor Index of Social Position. New Haven, Conn, Department of Sociology, Yale University, 1975
16. Spitzer RL, Williams JBW, Gibbon M, First MB: User's Guide for the Structured Clinical Interview for DSM-III-R (SCID). Washington, DC, American Psychiatric Press, 1990
17. Miller IW, Bishop S, Norman WH, Maddever H: The Modified Hamilton Rating Scale for Depression: reliability and validity. Psychiatry Res 1985; 14:131–142
18. Loranger AW, Susman VL, Oldham JM, Russakoff M: The Personality Disorder Examination (PDE) Manual. Yonkers, NY, DV Communications, 1988
19. Eysenck SBG, Eysenck HJ, Barrett P: A revised version of the psychoticism scale. Pers Indiv Diff 1985; 6:21–29
20. Keller MB, 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 Psychiatry 1987; 44:540–548
21. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing interrater reliability. Psychol Bull 1979; 86:420–428
22. Loranger AW, Lenzenweger MF, Gartner AF, Susman VL, Herzig J, Zammit GK, Gartner JD, Abrams RC, Young RC: Trait-state artifacts and the diagnosis of personality disorders. Arch Gen Psychiatry 1991; 48:720–728
23. Jorm AF, Duncan-Jones P, Scott R: An analysis of the re-test artifact in longitudinal studies of psychiatric symptoms and personality. Psychol Med 1989; 19:487–493
24. Westen D: Divergences between clinical and research methods for assessing personality disorders: implications for research and the evolution of axis II. Am J Psychiatry 1997; 154:895–903
25. Farmer R, Nelson-Gray RO: Personality disorders and depression: hypothetical relations, empirical findings, and methodological considerations. Clin Psychol Rev 1990; 10:453–476
26. Shea MT, Widiger TA, Klein MH: Comorbidity of personality disorders and depression: implications for treatment. J Consult Clin Psychol 1992; 60:857–868