ORIGINAL ARTICLE

The Lifetime History of Major Depression in Women

Reliability of Diagnosis and Heritability

Kenneth S. Kendler, MD; Michael C. Neale, PhD; Ronald C. Kessler, PhD; Andrew C. Heath, DPhil; Lindon J. Eaves, PhD, DSc

Background: In epidemiologic samples, the assessment of lifetime history (LTH) of major depression (MD) is not highly reliable. In female twins, we previously found that LTH of MD, as assessed at a single personal interview, was moderately heritable (approximately 40%). In that analysis, errors of measurement could not be discriminated from true environmental effects.

Methods: In 1721 female twins from a population-based register, including both members of 742 pairs, LTH of MD, covering approximately the same time period, was obtained twice, once by self-administered questionnaire and once at personal interview.

Results: Reliability of LTH of MD was modest (κ=+.34, tetrachoric *r*=+.56) and was predicted by the number of depressive symptoms, treatment seeking, number of episodes, and degree of impairment. Deriving an "index of caseness" from these predictors, the estimated heritability of LTH of MD was greater for more restrictive definitions. Incorporating error of measurement into a structural equation model including both occasions of measurement, the estimated heritability of the liability to LTH of MD increased substantially (approximately 70%). More than half of what was considered environmental effects when LTH of MD was analyzed on the basis of one assessment appeared, when two assessments were used, to reflect measurement error.

Conclusions: Major depression, as assessed over the lifetime, may be a rather highly heritable disorder of moderate reliability rather than a moderately heritable disorder of high reliability.

(Arch Gen Psychiatry. 1993;50:863-870)

MANY RECENT epidemiologic and genetic studies have assessed the lifetime history (LTH) of psychiatric disorders by a single psychiatric interview. In accord with studies of the accuracy of recall for relatively neutral facts, such as medical symptoms and medication use,[1] previous investigations have found that, especially in nonclinical populations, the assessment of LTH of psychiatric illness is not highly reliable.[2-4] This unreliability of assessment has important implications for psychiatric genetic studies.[5] In particular, for family, twin, or adoption studies that use a one-time assessment of LTH, the degree of familial resemblance or concordance may be substantially attenuated by error of measurement.

On the basis of a single personal interview in a population-based sample of female-female twin pairs, we found that lifetime major depression (MD), as defined by *DSM-III-R*, was moderately heritable.[6] As in most such studies, our analyses included no index of the reliability of our assessment. Therefore, errors of measurement were indistinguishable from true environmental differences between twins. If a pair of monozygotic (MZ) twins were discordant for LTH of MD, we were unable to distinguish whether the discordance resulted from a "true" difference, due to discrepant environmental experiences in the two twins, or from an error in assessment (ie, a false-positive report from the "affected" member or a false-negative report from the "unaffected" member). Put another way, we were unable to distinguish

From the Departments of Psychiatry (Drs Kendler, Neale, and Eaves) and Human Genetics (Drs Kendler and Eaves), Medical College of Virginia/Virginia Commonwealth University, Richmond; the Institute for Social Research, University of Michigan, Ann Arbor (Dr Kessler); and the Department of Psychiatry, Washington University School of Medicine, St Louis, Mo (Dr Heath).

See Subjects and Methods on next page

SUBJECTS AND METHODS

SAMPLE AND DIAGNOSTIC METHODS

As outlined previously,[6] this sample of white female same-sex twins was obtained from the population-based Virginia Twin Register, formed from a systematic review of birth records in the Commonwealth of Virginia from 1915 onward. Twins were eligible to participate if both members of the pair responded to one of two self-report questionnaires entitled "Health and Life-style Survey" and "Health and Personality." The average individual response rate to these two questionnaires was 64%, but the cooperation rate was higher, as a proportion of nonresponding twins never received the questionnaire because of improper addresses, incorrect forwarding of mail, etc. Of the 2352 individuals from 1176 twin pairs who met these criteria, we succeeded in personally interviewing 2163 (92.0%) of them, including both members of 1033 pairs. Zygosity was determined blindly by standard questions,[7] photographs, and, when necessary, DNA.[8] Interviewers were instructed to interview twins a minimum of 12 months after the completion of the mailed questionnaires. Of the completed personal interviews, 89.3% were performed face to face and 10.7% by telephone. Interviews were conducted by trained social workers, unaware of the status of the cotwin.

In most analyses reported herein, we focus on twins who completed both the relevant sections of the "Health and Personality" self-administered questionnaire and the personal interview (n=1721, including both members of 742 pairs, of whom 444 were considered to be MZ, 296 dizygotic [DZ], and two of unknown zygosity). The rate of LTH of MD did not differ in the twins included vs excluded from this subsample ($\chi^2=0.05$, *df*=1, not significant). The "Health and Personality" questionnaire contained a section assessing, by self-report, the LTH of MD, which we call our *time 1 assessment.* Individuals were asked for the lifetime occurrence of five key depressive symptoms chosen from the nine symptomatic criteria for MD in *DSM-III-R:* (1) sad mood, (2) change in appetite, (3) loss of energy, (4) feelings of guilt or worthlessness, and (5) problems in concentration. They were then asked whether any three of these symptoms co-occurred in their life for at least 2 weeks. In this study, individuals who responded positively to this item *and* admitted to sad mood were considered to have reported an LTH of MD.

The personal interviews, based on the Structured Clinical Interview for *DSM-III-R* Diagnosis,[9] assessed the history of MD twice in two separate sections: one section covered the last year and another covered the lifetime *before* the last year. Only the latter section will be examined in this report and will be termed our *time 2 assessment* of LTH of MD. Previous analyses of LTH of MD in this sample have included results from both sections.[6,10] In this case, we examine only LTH before the last year so that the period covered is similar in the time 1 and time 2 assessments. In fact, the periods covered by these two lifetime assessments differed by a mean of only 2.5 months. Diagnoses of MD in our time 2 assessment were assigned, on the basis of *DSM-III-R* criteria, after a "blind" review by one of us (K.S.K.), an experienced psychiatric diagnostician. Additional variables of interest assessed at the time 2 interview included age at onset of MD, number of episodes, duration of the longest episode, degree of impairment (none, moderate, and severe, the last meaning incapacitation), and treatment seeking (defined as "seeking professional help"). Neither hospitalization for MD nor treatment with medication was assessed.

Interrater reliability of the personal interview for LTH of MD was measured among 53 randomly chosen cases assessed at a single interview by two raters with perfect agreement ($\kappa=1.00$).

REGRESSION ANALYSES

To assess predictors of reliability in the diagnosis of LTH of MD, we used a "follow-back" approach because our time 2 assessment of LTH of MD contained far more clinical details than our time 1 assessment. Beginning with twins who met *DSM-III-R* criteria for LTH of MD at time 2 (n=535), we examined the ability of clinical features of MD assessed at time 2 to predict, by logistic regression,[11] those who reported LTH of MD at time 1 (n=313). To maximize comparability, we coded, wherever possible, our predictor variables in the same manner as that used by Rice et al.[5] To adjust for the correlated observations in twin pairs, we previously corrected the SEs upward as a function of the proportion of the sample that were complete twin pairs and the magnitude of the cor-

whether MD was a *highly reliable, moderately heritable disorder* or *a moderately reliable, highly heritable disorder.*

In a large proportion of this twin sample, we had obtained, around 1 year before the personal interview, a self-administered assessment of a lifetime history of MD. Using these data, we address the following questions: (1) How reliable is the LTH of MD in an epidemiologic sample of women? (2) What clinical features of MD predict reliability across two occasions of measurement? (3) Using our two times of assessment to calculate an index of true caseness after Rice et al,[5] will the index of caseness predict heritability of LTH of MD? (4) What impact will the formal inclusion of errors of measurement into a structural equation twin model have on estimates of the heritability of liability to MD?

RESULTS

AGREEMENT ON LTH OF MD AT THE TWO ASSESSMENTS

Examining all individuals in our sample with both a time 1 and time 2 assessment (n=1721), the frequency of LTH of MDD at time 1 and time 2 was 33.1% and 31.1%, respectively. Of the 569 twins who reported an LTH of MD at time 1, 313 (55.0%) reported an LTH of MD at relation of the dependent variable in those twin pairs.[12,13] However, in the current analyses, only 36% of the total sample were members of complete pairs. Furthermore, within these pairs, the correlation in the dependent variable (LTH of MD reported at time 1) was small (+.03). Therefore, no correction was needed. In the stepwise logistic regression, we used a *P* value of .50 for variables to be entered into the analysis and .10 for them to be excluded.[14]

We established an "index of caseness" of MD from the results of these analyses, according to the procedure outlined by Rice et al.[5,15] Covariates that significantly predicted the stability of the LTH of MD across the two times were selected. Cases with the highest possible value of these covariates (which, in these analyses, were individuals with three or more episodes of incapacitating depression, who endorsed all nine symptoms and sought treatment) are assumed to be "true cases" and assigned an index of caseness of unity. The index of all other cases, which varied in value between zero and unity, were assigned on the basis of these covariates so that the higher the index, the greater the probability of "true" caseness. For example, an individual with two episodes of depression with moderate impairment, who never sought treatment and endorsed seven symptoms, would have an index of 0.48.

TWIN ANALYSES

Two kinds of twin models are examined in this report, the first of which is a standard univariate model described previously.[6] All the models used here are based on a liability-threshold model and divide the variation in liability to MD into three classes: (1) additive genetic (A), which contributes twice as much to the correlation in MZ twins as DZ twins (because MZ twins share all their genes identical by descent, while DZ twins share on average only half their genes), (2) family or "common" environment (C), those familial factors that make twins similar in their liability to MD, which contributes equally to the correlation in MZ and DZ twins, and (3) individual-specific environment (E), which reflects environmental experiences not shared by both members of a twin pair and therefore contributes to *differences* between them in their liability to MD. Because previous analyses, as well as analyses of the models used herein, provided no evidence of significant dominance genetic effects on the liability MD, it will not be further considered. We have previously examined, from several perspectives, the equal environment assumption for MD in these data (that the exposure to environmental risk factors for MD is approximately equal in MZ and DZ twins) and found no evidence to reject it.[6,16,17] The best-fit model in our analyses was selected by means of Akaike's information criterion (AIC).[18]

In the first series of twin analyses, we applied a standard twin model to MD, but changed the definition of affection as a function of the index of caseness. For example, if a twin had an index of caseness for MD of 0.21, she would be considered affected if the index of caseness "cutoff" was greater than 0.1 or greater than 0.2, but unaffected when the index was raised to greater than 0.3.

Our second twin model for MD uses simultaneously both our time 1 and time 2 data. As pictured in **Figure 1**, left, the model assumes that there is a true latent liability to LTH of MD. Each of our two assessments of LTH are considered to be fallible indexes of this true latent liability. The paths λ_1 and λ_2 represent the degree to which the assessments of LTH of MD obtained at the two time points reflect this true liability. The square of these paths is one potential measure of the reliability of these assessments. The other paths to LTH of MD at time 1 and LTH of MD at time 2 (k_1 and k_2, respectively) represent error in the individual assessments of LTH of MD. By definition, $\lambda^2+k^2=1.0$. The latent liability to lifetime MD is then modeled in a standard twin design, as outlined above, with the sources of variance in liability divided between additive genetic, common environmental, and individual-specific environmental factors.

Although not pictured in Figure 1, left, further elaborations of this model are possible and relevant. In particular, it is possible that the errors of measurement may be correlated in twin pairs. This is incorporated in the model by adding two paths reflecting correlated errors in twin pairs at time 1 (from MD at time 1 for twin 1 to MD at time 1 for twin 2) and at time 2 (from MD at time 2 for twin 1 to MD at time 2 for twin 2).

It is important to emphasize *two* critical differences between this model and the standard twin model. First, this model provides *separate* estimates for error of measurement (k) and true individual-specific environment (e). Second, it provides a direct estimate of the degree to which the individual assessments of LTH of MD reflect the latent liability (λ).

time 2. Of the 1152 twins who did not report an LTH of MD at time 1, there were 222 (19.3%) who reported an LTH of MD at time 2. While the association between the report of an LTH of MD at the two assessments was highly statistically significant, the degree of agreement was modest (κ=+.34±.02, *P*<.0000; tetrachoric *r*=+.56±.03).

CLINICAL CHARACTERISTICS OF REPORTED MD THAT PREDICTED RELIABILITY OF DIAGNOSIS

In the subsample of twins who reported an LTH of MD at time 2 (n=535), we examined the relationship between the clinical characteristics of MD assessed at time 2 and the probability that they also reported an LTH for MD at time 1. **Table 1** illustrates these results, following, as closely as possible, the presentation by Rice et al.[5] For example, while 59% of all twins who reported a lifetime episode of MD at time 2 also reported one at time 1, this was true for only 48% of those reporting no impairment vs 74% of those with severe impairment. For symptoms, around 45% for those reporting five or six symptoms reported a lifetime depressive episode at time 1, compared with 86% of those reporting all nine symptoms.

Logistic regression analyses of these relationships, taken one at a time, are seen in the left side of **Table 2**. Stability of the diagnosis of lifetime MD was most strongly predicted by number of symptoms, treatment seeking, and level of impairment. Duration of the longest episode and the number of reported episodes were also significant predictors. Only age at onset was unrelated to the reliability of LTH of MD.

We then repeated these analyses by means of stepwise logistic regression (**Table 2**, right side), with somewhat different results. The number of symptoms, treatment seeking, and number of episodes were now robust predictors. Degree of impairment was a weaker but still significant predictor, while duration of the longest episode and age at onset provided no additional predictive power.

INDEX OF CASENESS AND THE HERITABILITY OF LIABILITY TO MD

According to the methods of Rice et al,[5] we developed for each individual who reported an LTH of MD at our second assessment an *index of caseness*. To maximize our power, these analyses were conducted with the use of all twins assessed at our time 2 interview with known zygosity (n=1030 pairs). We then calculated the heritability of liability to lifetime MD, defining affection as a function of the index of caseness. For all analyses, the best-fitting model required only additive genes and individual-specific environment (details available on request). The results are depicted in **Figure 2**. Taking any case meeting *DSM-III-R* criteria, the heritability of liability to LTH of MD was estimated at 33%. As the value of the index of caseness that was required to define caseness increased, heritability tended, with some variation, also to increase. At the highest level of index of caseness at which stable estimates could be obtained (>0.8), the heritability of liability to MD was estimated at 52%.

THE HERITABILITY OF LIABILITY TO LIFETIME MD INCORPORATING UNRELIABILITY OF MEASUREMENT

A second way to investigate the relationship between unreliability of measurement and the heritability of liability to MD is to include formally such unreliability in a structural equation model (Figure 1, left). First, however, we applied standard twin models to LTH of MD as assessed at time 1 and as assessed at time 2. At both occasions of measurement, the ACE model fit well and estimated the common environmental path (c) to be zero. For both time

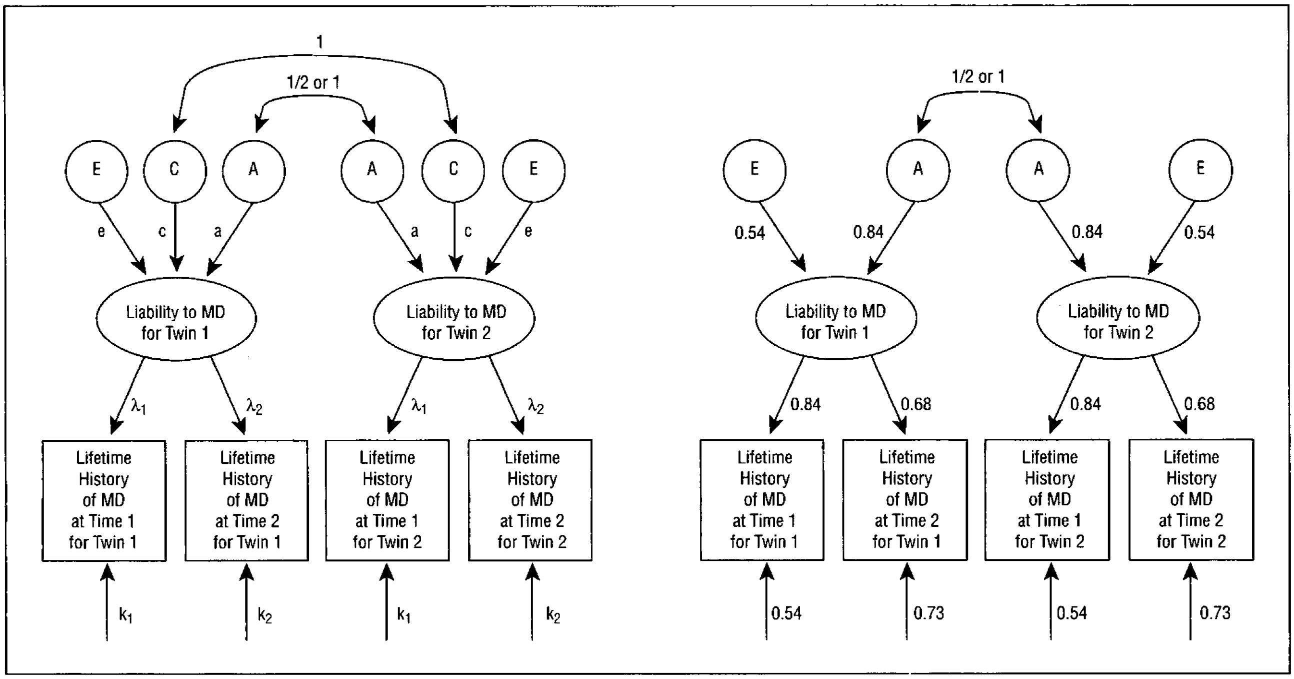
1
1/2 or 1
E
C
A
A
C
E
e
c
a
a
c
e
Liability to MD for Twin 1
Liability to MD for Twin 2
λ1
λ2
λ1
λ2
Lifetime History of MD at Time 1 for Twin 1
Lifetime History of MD at Time 2 for Twin 1
Lifetime History of MD at Time 1 for Twin 2
Lifetime History of MD at Time 2 for Twin 2
k1
k2
k1
k2
1/2 or 1
E
A
A
E
0.54
0.84
0.84
0.54
Liability to MD for Twin 1
Liability to MD for Twin 2
0.84
0.68
0.84
0.68
Lifetime History of MD at Time 1 for Twin 1
Lifetime History of MD at Time 2 for Twin 1
Lifetime History of MD at Time 1 for Twin 2
Lifetime History of MD at Time 2 for Twin 2
0.54
0.73
0.54
0.73

Figure 1. *Left, A twin model for the heritability of liability to lifetime history of major depression (MD) including error of measurement. This model assumes that there is a* true *liability to the lifetime history of MD, which is indexed by two assessments, at time 1 and time 2. The paths λ_1 and λ_2 represent the degree to which these assessments reflect the true liability to MD. The square of these paths is a measure of the reliability of these assessments. The other paths to lifetime history of MD at times 1 and 2 (k_1 and k_2) represent error in the individual assessments of lifetime history of MD. The model is constrained so that for each twin assessment, $\lambda^2+k^2=1.0$. The latent liability to lifetime MD is modeled as in a standard twin design,[6] with the sources of variance in liability divided between additive genetic (A), common (C) environmental, and individual-specific environmental factors (E). By definition, the common environmental components are perfectly correlated in all twins, while the individual-specific environment is uncorrelated. Additive genetic factors are perfectly correlated in monozygotic twins and correlated 0.5 in dizygotic twins. Lowercase letters (a, c, and e) are used to label the paths from these factors. The individual paths represent standardized regression coefficients, so that the proportion of variance in the dependent variables accounted for by the independent variable is equal to the square of the connecting path. Heritability, for example, equals a^2. Observed variables are depicted in boxes and latent variables in circles and ellipses. Right, Parameter estimates from the best-fitting model (model 2). Parameter estimates are constrained to be equal for twin 1 and twin 2. No evidence was found for common environmental factors.*

1 and time 2, the best-fitting models by AIC were the AE models, with estimates of heritability of 49% and 35%, respectively.

We then fit the full model (depicted in Figure 1, left) —or model 1, as we will call it—to the tetrachoric correlation matrixes of the LTH of MD in twin 1 and twin 2 at time 1 and time 2 in MZ and DZ twins (matrixes available on request). This model fit very well (χ^2=7.5, *df*=8, *P*=.28, AIC=−8.5). The common environmental path was estimated at zero and could, for model 2, be set to zero with no change in fit (χ^2=7.5, *df*=9, AIC=−10.5). For model 3, we added correlated errors of measurements for twins at time 1 and time 2. However, both of these paths were estimated at near zero and did not improve the fit of the model (χ^2=7.4, *df*=7, AIC=−6.6). In model 4, we constrained paths λ_1 and λ_2 to be equal, meaning that the time 1 and time 2 assessments reflected, with the same accuracy, the underlying latent liability to MD. However, this resulted in an AIC that was worse than that of model 2 (χ^2=10.6, *df*=10, AIC=−9.4). That is, our time 1 assessment was a significantly better index of the latent liability to LTH of MD than was our time 2 assessment. No other improvements were possible, so that model 2 was the best fit. We also analyzed these data with a standard bivariate twin model[10] that treats LTH of MD assessed at times 1 and 2 as entirely separate phenotypes. The fit of this general model was similar to that obtained herein, suggesting that the strong assumptions inherent in our model are well supported by the data. The parameter estimates of the best-fitting model 2 are seen in Figure 1, right. Corrected for unreliability at two occasions of measurement, the heritability of liability to MD was now estimated at 71%. Individual-specific environment accounted for the remaining 29% of the variation in liability. The paths from our time 1 self-administered and our time 2 personal interview assessments of the LTH of MD to the latent liability to MD were estimated at +0.84 and +0.68, respectively.

Table 1. Predictors of Reliability of Lifetime History of Major Depression

	No. With LTH-MD at Time 2*	No. (%) Who Also Reported LTH-MD at Time 1*
Major depression	535	313 (59)
No. of symptoms		
5	88	40 (46)
6	119	50 (42)
7	152	85 (56)
8	121	91 (75)
9	55	47 (86)
No. of episodes		
1	203	103 (51)
2	113	67 (59)
≥3	216	141 (65)
Treatment		
No	309	148 (48)
Yes	162	126 (78)
Impairment		
None	136	61 (45)
Moderate	234	130 (56)
Severe	164	122 (74)
Duration, wk		
≤2	81	43 (53)
>2-4	114	50 (44)
>4-51	262	162 (62)
>51	78	58 (74)

*LTH-MD indicates lifetime history of major depression.

RESULTS

The goal of this article was to further our understanding of measurement error in the assessment of LTH of MD and its potential impact on genetic-epidemiologic investigations. We will address in turn the four major questions posed above.

THE RELIABILITY OF LTH OF MD

We asked 1721 female twins ascertained from a population-based register twice about their lifetime history of MD. The first occasion was by mailed questionnaire and covered the occurrence of MD at any point in their lives up to that time. The second occasion was by personal interview slightly more

Table 2. Prediction of Diagnostic Stability of Lifetime History of Major Depression by Clinical Covariates

	Covariates Assessed One at a Time			Covariates Assessed by Stepwise Procedure		
Variable	β*	χ^2†	*P*	β*	χ^2†	*P*
No. of symptoms	.50	41.20	.000	.46	23.45	.000
No. of episodes	.30	9.02	.003	.40	11.35	.001
Treatment	1.34	36.72	.000	1.09	19.99	.000
Impairment	.63	26.54	.000	.37	6.99	.008
Duration of longest episode	.36	13.06	.000	. . .‡	. . .	. . .
Age at onset	−.02	2.22	.14	. . .‡	. . .	. . .

*Logistic regression coefficient.
†df=1.
‡Variable excluded in stepwise analysis.

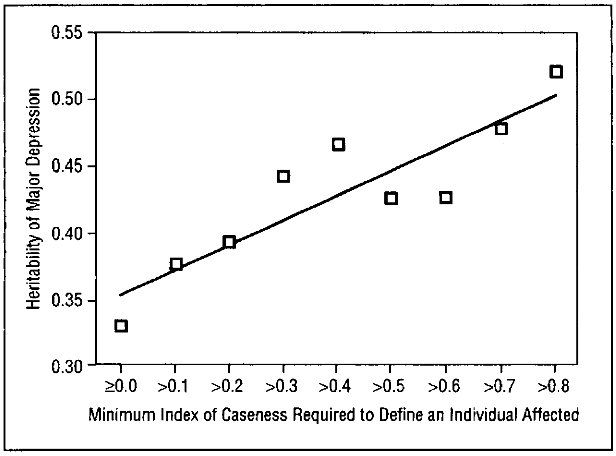

Figure 2. *The relationship between heritability of liability to major depression and the minimal index of caseness required to define an individual as affected. The best-fitting twin model for all of these analyses included only additive genetic and individual-specific environmental factors. The figure also contains an estimated regression line for this relationship.*

than 1 year later in which they were asked for the occurrence of MD any time *before* 1 year before the interview. The agreement between LTH of MD assessed at these two time points was modest (κ=.34, tetrachoric r=+.56).

The most similar previous study of this issue was conducted by Bromet et al.[3] In an epidemiologic sample of 391 women, they conducted an 18-month test-retest study of the reliability of LTH of MD at or before the first interview. Personal interviews and Research Diagnostic Criteria[19] were used on both occasions of measurement. They found the overall reliability of LTH of MD to be slightly greater than we did (κ=.41). Prusoff et al[2] reviewed earlier studies of the reliability of LTH of MD in nonclinical populations and reported their own results. Excluding studies of 1-day test-retest, they noted six different studies with κ values ranging from .21 to .75 and averaging +.46.

Rice et al[5] recently reported on the "stability" of LTH of MD in 2226 first-degree relatives of probands with affective illness during a 6-year period. Their approach differed from ours in one important way. In this study, our two assessments for LTH of MD covered approximately the same period in the respondent's life. However, the second assessment by Rice et al[5] covered 6 more years of exposure to MD than their first assessment did. They did not differentiate onsets of MD reported at the second assessment that occurred before vs after the first assessment. They found that the "stability" of LTH of MD in their sample was relatively high (κ=.61).

Our results are consistent with most previous studies, suggesting that, in nonclinical populations, the assessment of LTH of MD is only moderately reliable.

THE PREDICTION OF RELIABILITY IN LIFETIME MD

We examined the clinical features that predicted consistency of reporting of LTH of MD across two occasions of measurement. Because only our second assessment included these clinical details, we employed a "follow-back" procedure, using clinical characteristics at time 2 to predict the reporting of MD at time 1. Three strong predictors emerged: number of symptoms, treatment-seeking, and number of episodes. Analyzed on its own, the degree of impairment was also a powerful predictor, but it lost much of this power in multivariate analysis.

Rice and colleagues[5] performed a similar analysis attempting to predict *stability* of the lifetime diagnosis of MD in a large sample of relatives of probands with affective illness. The results they obtained were similar to those found herein. Examined individually, they found that the number of symptoms, number of episodes, treatment, and duration of longest episode all significantly predicted stability.[5] Using stepwise regression, they found only two significant predictors, which were also the most powerful predictors in our analyses: number of symptoms and treatment seeking. Bromet et al[3] also reported that treatment seeking strongly predicted consistency in reporting MD over time.

These results suggest that there may be a relatively consistent pattern to the features that make lifetime episodes of MD *memorable*. First, memorable MD tends to be severe. Depressive episodes that were more symptomatic and more disabling appear to be more consistently recalled. Second, frequent recurrence tends to make the depressive experience more memorable, independent of severity. Third, treatment makes MD more memorable independent of severity *and* recurrence. While memorable depressions tend to be severe, clinical severity is apparently not the only factor that influences consistent recall of previous depressive episodes.

INDEX OF CASENESS AND THE HERITABILITY OF MD

Following Rice and colleagues[5] in using consistency of reporting during two occasions as a validating criterion, we constructed an index of "true caseness." The higher this index, the higher the probability that the individual has a "true" case of MD. If unreliability of measurement substantially influences estimates of the heritability of MD, psychometric theory predicts that heritability should increase if the diagnosis of MD is made more reliable. We tested this hypothesis by examining heritability while altering the definition of affection as a function of the index of caseness. The results were as predicted. Heritability tended to increase as the index of caseness required for affection increased. The overall increase was substantial, as the most rigorous levels of index of caseness produced an estimated heritability of MD 58% greater than the lowest possible index. These results suggest that in estimations of the heritability of MD from a single assessment, a significant proportion of what was considered true environmental differences between twins were in fact errors of measurement.

INCORPORATION OF ERRORS OF MEASUREMENT INTO A STRUCTURAL EQUATION TWIN MODEL FOR MD

Using the flexibility of structural equation modeling, we were able to incorporate unreliability of assessment of LTH of MD directly into our twin model. Standard biometric twin models for psychiatric illness, such as that used to analyze MD previously in this sample,[6] assume that the disorder is assessed without error. In these models, unreliability of measurement, if uncorrelated across twin pairs, is entirely confounded with true differences in environmental exposure across twins. However, the twin model used in this report both separates error of measurement from true environmental differences and provides estimates of the reliability of the individual assessments.

Six results from this model fitting are particularly worthy of note. First, as we have found previously with these data,[6,16] common or familial environmental influences appear to have little impact on liability to MD. Second, these models showed that the true liability to lifetime MD is only imperfectly assessed by any one evaluation. Results suggest that at any one time of measurement of LTH of MD, 30% to 50% of the variance in liability is, from a psychometric perspective, "error."

Third, our model suggested that the error in assessment of LTH on one occasion is largely individual specific. We found no evidence that such error was correlated in twin pairs. Fourth, our self-administered assessment of LTH of MD was a *better* index of the latent liability to depression than was that obtained at personal structured interview by a mental health professional. The time 1 self-administered assessment might be superior because respondents were asked there to report any previous episode of MD. However, at the time 2 interview, they were asked about episodes of MD *before* the last year, a more difficult task that may be performed less reliably. Alternatively, the assumption that valid assessment of psychiatric illness must be based on a face-to-face interview, preferably by a trained clinician, may be incorrect. Survey research has suggested that sensitive information may be more reliably obtained by means of more anonymous as opposed to more personal means of assessment.[20,21] Furthermore, the interviewer-respondent interaction may be a source of additional error variance that is missing in self-administered measures.

Fifth, the model provided for the first time an estimate of the true individual environmental contribution to the liability to MD, unfounded with error of measurement. The estimate was around 30%, *lower* than the contributions of error. That is, these results suggest that more than half of what was previously estimated as the environmental contribution to liability to MD[6,16] was error. Finally, when corrected for errors of measurement, the heritability of liability to MD was high, around 70%. Heritabilities in this range have been previously reported for schizophrenia[22] and bipolar illness[23] in studies where assessment almost always included medical records. While MD, which often goes untreated,[24] may be less reliably assessed than schizophrenia or bipolar illness, it may be nearly as heritable.

COMPARISON WITH OUR PREVIOUS LONGITUDINAL TWIN STUDY OF MD

We previously reported, in this twin sample, a longitudinal study of the *1-year* prevalence of MD.[16] It is useful to compare and contrast these two analyses. In our previous analysis,[16] twins reported the occurrence of MD during the preceding 1 year at two separate times. The time periods covered by these two assessments were *completely nonoverlapping.* In this report, twins were asked twice about episodes of MD in their *lifetime,* and the periods covered by these reports were, with minor exceptions, *entirely overlapping.*

Our previous analyses indicated that genetic factors were largely responsible for the temporal stability of risk to MD, while environmental factors were occasion-specific in their effect.[16] This analysis suggests that genetic factors are mostly responsible for the reliable component of the liability to MD, while much of what we previously considered to be environmental effects may be "error." These two findings suggest, in different ways, that there is a lot of "error" or "short-term environmental effects" in a single assessment of MD. When we move beyond this standard cross-sectional approach and obtain assessments on more than one occasion, the impact of genetic factors on the liability to MD becomes substantially greater.

LIMITATIONS

The results presented herein should be interpreted in the context of six potential methodologic limitations. First, this sample was restricted entirely to women. Although Rice et al[5] found no impact of gender on the stability of the diagnosis of MD, the interrelation between error and heritability of MD may differ in men and women.

Second, because our second assessment was not always exactly 1 year after our first assessment, the periods covered by these two measures of LTH of MD often differed by a few months. However, agreement in reporting LTH of MD at these two assessments was not significantly predicted by the length of time by which these two periods differed (χ^2=0.20, *df*=1, not significant).

Third, different methods of assessment and different criteria for MD were used at our two times of measurement. However, two analyses suggested that these differences are unlikely to have a major impact on the results obtained. In the time 2 personal interview data, the criteria for MD used in the time 1 assessment agreed with the *DSM-III-R* criteria closely (κ=+.96±.01). We also reanalyzed our data applying the time 1 criteria for MD to the time 2 data, including fitting the full twin model pictured in Figure 1, left. Results were similar to those found with the use of *DSM-III-R* criteria with the time 2 data. In particular, the agreement in LTH of MD across the two times of assessment did not differ if the *DSM-III-R* or the time 1 criteria for MD were used with the time 2 data (κ=+.363 and +.357, respectively). The low reliability of LTH of MD in our data is unlikely to be due, to a substantial degree, to the use of two different definitions of MD at the two occasions of measurement. However, because of differences in criteria and methods of assessment, the results of this investigation should be regarded as preliminary. Replication is needed, preferably with the use of identical assessment methods and diagnostic criteria at both times of measurement. While overall probably less desirable, the use of two different assessment methods does have one important methodologic strength—it minimizes the possibility of correlated errors on the two occasions of measurement.

Fourth, our structural equation model (Figure 1, left) assumes that on the liability dimension, error is a random variable with a mean of zero and unit variance. That is, error may as likely go in the false-positive as false-negative direction. However, it could be argued that no false-positive results occur in the assessment of the LTH of MD—that every positive report is a "true" report. If this hypothesis were correct, which we consider unlikely, it would suggest a quite different approach to modeling "unreliability" of assessment.

Fifth, we have frequently used the term *error* to reflect unstable influences on the reporting of lifetime MD. While our results suggest that this "error" is neither stable over time nor substantially correlated in relatives, we do not mean to imply by this usage that such error is entirely random or unworthy of study. Such error could, for example, be influenced by short-term state-dependent effects of mood on memory[25] or interviewer-respondent interactions.

Finally, it could be argued that our results can all be explained by the simple axiom that more "severe" MD is more heritable. This is almost certainly, however, an oversimplification. Using a single cross-sectional assessment of LTH of MD, we were unable in this sample to detect any simple relationship between "narrowness" of diagnosis and heritability.[6] Furthermore, controlling for severity (as indexed by number of symptoms *or* degree of impairment), both treatment seeking *and* number of episodes predicted reliability of reporting and hence higher heritability. Our results suggest that it is *memorable* lifetime episodes of MD that have high heritability. They have high heritability not mainly because they are severe, but because they can be recalled on multiple occasions with a low degree of error.

Accepted for publication May 11, 1993.

Reprint requests to Box 710 MCV Station, Richmond, VA 23298-0710 (Dr Kendler).

REFERENCES

1. Harlow SD, Linet MS. Agreement between questionnaire data and medical records: the evidence for accuracy of recall. *Am J Epidemiol.* 1989;129:233-248.
2. Prusoff BA, Merikangas KR, Weissman MM. Lifetime prevalence and age of onset of psychiatric disorders: recall 4 years later. *J Psychiatr Res.* 1988;22:107-117.
3. Bromet EJ, Dunn LO, Connell MM, Dew MA, Schulberg HC. Long-term reliability of diagnosing lifetime major depression in a community sample. *Arch Gen Psychiatry.* 1986;43:435-440.
4. Aneshensel CS, Estrada AL, Hansell MJ, Clark VA. Social psychological aspects of reporting behavior: lifetime depressive episode reports. *J Health Soc Behav.* 1987;28:232-246.
5. Rice JP, Rochberg M, Endicott J, Lavori PW, Miller C. Stability of psychiatric diagnoses: an application to the affective disorders. *Arch Gen Psychiatry.* 1992;49:824-830.
6. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. A population based twin study of major depression in women: the impact of varying definitions of illness. *Arch Gen Psychiatry.* 1992;49:257-266.
7. Eaves LJ, Eysenck HJ, Martin NG, Jardine R, Heath AC, Feingold L, Young PA, Kendler KS. *Genes, Culture and Personality: An Empirical Approach.* London, England: Oxford University Press; 1989.
8. Spence JE, Corey LA, Nance WE, Marazita ML, Kendler KS, Schieken RM. Molecular analysis of twin zygosity using VNTR DNA probes. *Am J Hum Genet.* 1988;43(3):A159. Abstract.
9. Spitzer RL, Williams JB, Gibbon M. *Structured Clinical Interview for* DSM-III-R. New York, NY: Biometrics Research Division, New York State Psychiatric Institute; 1987.
10. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. Major depression and generalized anxiety disorder: same genes, (partly) different environments? *Arch Gen Psychiatry.* 1992;49:716-722.
11. SAS Institute Inc. CATMOD. In: *SAS User's Guide: Statistics, Version 5 Edition.* Cary, NC: SAS Institute Inc; 1985:171-254.
12. Kish L, Frankel MR. Inferences from complex samples. *J R Stat Soc Ser B.* 1974;36:1-37.
13. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. A longitudinal twin study of personality and major depression in women. *Arch Gen Psychiatry.* 1993;50:853-862.
14. SAS Institute Inc. *SAS/STAT User's Guide, Version 6, Fourth Edition, Volumes 1 and 2.* Cary, NC: SAS Institute Inc; 1990.
15. Rice JP, Endicott J, Knesevich MA, Rochberg N. The estimation of diagnostic sensitivity using stability data: an application to major depressive disorder. *J Psychiatr Res.* 1987;21:337-345.
16. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. A longitudinal twin study of one-year prevalence of major depression in women. *Arch Gen Psychiatry.* 1993;50:843-852.
17. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. A test of the equal environment assumption in twin studies of psychiatric illness. *Behav Genet.* 1993;23:21-27.
18. Akaike H. Factor analysis and AIC. *Psychometrika.* 1987;52:317-332.
19. Spitzer RL, Endicott J, Robins E. *Research Diagnostic Criteria for a Selected Group of Functional Disorders.* 2nd ed. New York, NY: New York State Psychiatric Institute; 1975.
20. Rolnick SJ, Gross CR, Garrard J, Gibson R. A comparison of response rate, data quality, and cost in the collection of data on sexual history and personal behaviors: mail survey approaches and in person interview. *Am J Epidemiol.* 1989;129:1052-1061.
21. Siemiatycki J. A comparison of mail, telephone, and home interview stategies for household health surveys. *Am J Public Health.* 1979;69:238-245.
22. Kendler KS. Overview: a current perspective on twin studies of schizophrenia. *Am J Psychiatry.* 1983;140:1413-1425.
23. McGuffin P, Katz R. The genetics of depression and manic-depressive illness. *Br J Psychiatry.* 1989;155:294-304.
24. Shapiro S, Skinner EA, Kessler LG, Von Korff M, German PS, Tischler GL, Leaf PJ, Benham L, Cottler L, Regier DA. Utilization of health and mental health services: three epidemiologic catchment area sites. *Arch Gen Psychiatry.* 1984;41:971-978.
25. Blaney PH. Affect and memory: a review. *Psychol Bull.* 1986;99:229-246.