Reliability of a lifetime history of major depression: implications for heritability and co-morbidity

D. L. FOLEY, M. C. NEALE AND K. S. KENDLER

From the Departments of Human Genetics and Psychiatry at the Virginia Institute for Psychiatric and Behavioral Genetics, Medical College of Virginia/Virginia Commonwealth University, Virginia, USA

ABSTRACT

Background. In unselected samples, the diagnosis of major depression (MD) is not highly reliable. It is not known if occasion-specific influences on reliability index familial risk factors for MD, or how reliability is associated with risk for co-morbid anxiety disorders.

Methods. An unselected sample of 847 female twin pairs was interviewed twice, 5 years apart, about their lifetime history (LTH) of MD, generalized anxiety disorder (GAD) and panic disorder (PD). Familial influences on reliability were examined using structural equation models. Logistic regression was used to identify clinical features that predict reliable diagnosis. Co-morbidity was characterized using the continuation ratio test.

Results. The reliability of a LTH of MD over 5 years was fair (κ = 0.43). There was no evidence for occasion-specific familial influences on reliability, and heritability of reliably diagnosed MD was estimated at 66%. Subjects with unreliably diagnosed MD reported fewer symptoms and, if diagnosed with MD only at the first interview, less impairment and help seeking, or, if diagnosed with MD only at the second interview, fewer episodes and a longer illness. A history of co-morbid GAD or PD is more prevalent among subjects with reliably diagnosed MD.

Conclusions. A diagnosis of MD based on a single psychiatric interview incorporates a substantial amount of measurement error but there is no evidence that transient influences on recall and diagnosis index familial risk for MD. Quantitative indices of risk for MD based on multiple interviews should reflect both the characteristics of MD and the temporal order of positive diagnoses.

INTRODUCTION

Epidemiological studies commonly assess a subject’s risk for major depression (MD) on the basis of a single psychiatric interview. Given that the accuracy of diagnostic assignments predicate the accurate characterization of familial–epidemiological risk factors for MD, the reliability with which non-clinical subject’s recall past episodes of MD is of considerable importance. In community samples, kappa (κ) for the test–retest reliability of a lifetime history (LTH) of MD ranges between 0·21 and 0·87 (Prusoff et al. 1988; Williams et al. 1992; Keller et al. 1995). In samples comprising patients and their relative, κ ranges between 0·61 and 0·85 (Prusoff et al. 1988; Fendrich et al. 1990; Rice et al. 1992). These reliability estimates indicate that the diagnosis of a LTH of MD is not highly reliable, especially in epidemiological samples.

Two previous investigators have examined the characteristics of reliably diagnosed MD within a genetically informative framework. Rice et al. (1992) conducted two personal interviews with a selected sample of 1629 first-degree relatives of 187 bipolar, 78 bipolar II and 331 MD probands. Six year stability of the Schedule for Affective Disorders and Schizophrenia (SADS) depressive disorder diagnosed following Research Diagnostic Criteria (RDC)
(Spitzer et al. 1975) was good (κ = 0.61). Covariates of stable diagnostic assignments were assessed using a forward prediction paradigm that used a diagnosis of depression at time 1 to identify a positive case. The depressive features at time 1 that predicted that depression was diagnosed at time 2 were an increasing number of depressive symptoms and hospitalization or treatment with medication or electroconvulsive therapy.

Kendler et al. (1993) subsequently examined the reliability of a LTH of MD in an unselected sample of 1721 female twins. Twins were initially surveyed by a mailed questionnaire that included five of the nine DSM-III-R (American Psychiatric Association (APA), 1987) MD symptoms and a LTH of MD was assigned if any three symptoms, including depressed mood, had co-occurred for at least 2 weeks. Approximately 1 year later, a LTH of MD was assessed again using a face-to-face interview based on the Structured Clinical Interview for DSM-III-R (SCID, Spitzer et al. 1987). The reliability of these two LTH diagnoses was poor (κ = 0.34). Kendler et al. (1993) examined the generality of the covariates of stable diagnosis reported by Rice et al. (1992) by using a backward prediction paradigm that used the depressive features reported at time 2 to predict a diagnosis of MD was previously made at time 1. This analytical approach, therefore, assumes that a diagnosis at time 2 identifies a positive case. The number of depressive symptoms again predicted a reliable diagnosis, as did impairment and help seeking associated with the worst episode, and the number of depressive episodes experienced over the lifetime. Reliably diagnosed MD was more heritable than MD assessed and diagnosed at only one occasion (h² = 0.70 for reliably diagnosed MD versus h² = 0.49 for time 1 (self-report) MD and h² = 0.35 for time 2 (SCID) MD), although Kendler et al. (1993) noted that the questionnaire and interview were not statistically equivalent indices of liability for MD.

On the basis of these findings both Rice and Kendler have argued for the adoption of a quantitative risk index, derived from the depressive features that predict reliable diagnosis, rather than a continued reliance on the more error-prone and less heritable RDC and DSM diagnoses. Two important issues remain to be addressed however.

First, it is necessary to demonstrate that unreliable diagnosis indexes only measurement error and non-genetic influences on liability to MD. A variety of transient or ‘occasion-specific’ influences on the recall of depressive episodes have been characterized, including internal symptom cues (current mood state) and external stress cues (recent salient life events). Aneshensel et al. (1987) reported that such effects are not random, but follow orderly patterns. The possibility that such effects may index genetic or other familial influences on true liability for MD has not, however, been formally tested in models that characterize the heritability of reliably diagnosed MD.

When only cross-sectional data are available, heritability (h²) is calculated as the ratio of the estimated genetic variance (Vg) to the total observed (phenotypic) variance (Vp) – with the genetic variance estimated from the pattern of resemblance between different groups of relatives (Neale & Cardon, 1992). Vp incorporates both genetic (Vg) and environmental variance (Ve), including that which is occasion-specific or ‘unreliable’ (Vk);

\[ h^2 = \frac{Vg}{Vp}, \quad \text{where} \quad Vp = Vg + Ve + Vk \quad (1). \]

In cross-sectional data, Ve and Vk are confounded and the estimated heritability therefore cannot exceed the reliability of the measured trait. If a trait is 100% heritable and measured with perfectly reliability then \( Vg = Vg + Ve + Vk = 1.00 \) (i.e. \( Ve = 0 \) and \( Vk = 0 \)). If, however, the trait is 100% heritable but measurement of the trait is unreliable because of measurement error (\( Ve = 0 \) but \( Vk > 0 \)), then the genetic variance will be estimated as \( < 100\% \) (i.e. \( Vg < Vg + Ve + Vk = 1 \)). A method for estimating Vk using longitudinal data was developed by one of us (MCN), and this method was subsequently used to estimate the heritability of reliably diagnosed MD (Kendler et al. 1993). In that study, Vk was assumed to reflect errors of measurement and occasion-specific influences on the subject’s recall of MD that are uncorrelated between relatives. Heritability of reliably diagnosed MD was calculated as

\[ h^2 = \frac{Vg}{Vp}, \quad \text{where} \quad Vp = Vg + Ve \quad (2) \]

and Vk (‘error’) was estimated separately. As the denominator in (2) is smaller than the
Reliability of major depression

859

denominator in (1), by a factor equivalent to \( V_k \), \( h^2 \) for reliably diagnosed MD is larger than \( h^2 \) for MD assessed and diagnosed at only one occasion.

Kendler et al. (1993) did not formally test if \( V_k \) only indexes errors of measurement and occasion-specific influences on the subject’s recall of MD that are uncorrelated between relatives. Multiple measurements that index both genetic and environmental influences, including measurement error, on occasion-specific variance have been reported previously, and the impact of such variance on the estimation of heritability discussed in detail elsewhere (Falconer, 1967). We therefore cannot assume a priori that the estimated heritability for reliably diagnosed MD captures all the relevant genetic influences on liability to MD. It is possible that \( V_k \) is correlated between relatives, reflecting, for example, occasion-specific genetic influences on recent mood that are correlated with consistency of recall and true liability to MD.

The second issue raised by the findings reported by Rice et al. (1992) and Kendler et al. (1993) concerns the somewhat different correlates of reliably diagnosed MD reported by these two investigators. A variety of factors may account for the discrepant findings. First, each study assigned a positive case differently. Rice used the index diagnosis and a forward prediction paradigm to characterize stability over time, whereas Kendler used the follow-up diagnosis and a backward prediction paradigm. It is not clear, however, if a depressive history endorsed at index is equivalent in liability to one endorsed at follow-up given the lower re-test prevalence of negative affective states surveyed by rating scales (Jorm et al. 1989) and interviews (Helzer et al. 1981; Bromet et al. 1986, Eaton et al. 1989). Secondly, Kendler used different diagnostic criteria to assign a LTH of MD at time 1 and 2. At time 1, MD was diagnosed on the basis of a subset of self-rated DSM-III-R criteria. At time 2, MD was diagnosed using the full DSM-III-R criteria evaluated at personal interview. Thirdly, Rice used RDC criteria and Kendler used (variable) DSM-III-R criteria. Fourthly, the depressive features used to characterize stable diagnosis are likely to be highly correlated and the pattern of correlations may differ in selected and unselected samples. Fifthly, there may be differences in the correlates of consistent recall in community versus selected samples given the latter are likely to include a greater proportion of milder cases. Before the latter explanation can be assumed to account for all cross-study differences, and the data reported by Rice used as a basis for assigning different liability weight to subjects in both community and selected samples (Rice et al. 1992), points 1 to 4 require further consideration.

Lastly, temporal stability is just one test of validity (Rice et al. 1992). Robins & Guze (1970) proposed five other criteria to establish diagnostic validity, including delimitation from other diagnoses. Does a reliable diagnosis of MD attenuate the association between MD and other disorders? If diagnostic reliability indexes severity of liability for MD (Rice et al. 1992; Kendler et al. 1993) and if co-morbidity between MD and GAD (Goethe et al. 1993; Brown et al. 1996) and between MD and panic disorder (Reich et al. 1993; Andrade et al. 1994; Pini et al. 1994) reflects a more severe depression, then reliably diagnosed MD may be associated with a higher risk for co-morbid anxiety disorder over the lifetime. If we are to understand the implications of reliable diagnoses we need to broaden our investigation to include a consideration of the multivariate pattern of risk associated with a reliable diagnosis.

Since the publication of Kendler et al. (1993), a LTH of DSM-III-R MD in this same sample of twins was surveyed again an average of 5 years after administration of the index interview. With these two wave interview data in hand, we are now able to address the following questions:

1 Do multiple surveys that cover the same risk period index occasion-specific genetic effects on liability to major depression?

2 Does the assignment of a positive case based on a diagnosis of MD at index versus follow-up affect the characterization of reliable diagnosis? Do the characteristics of an index diagnosis of MD differ from those reported at follow-up? What are the implications for the derivation of a quantitative risk index for depression?

3 How does diagnostic reliability affect the characterization of multivariate patterns of risk? What does this imply about the
validity of reliably diagnosed MD and the boundary between MD and anxiety disorders?

METHOD

Sample
The sample of Caucasian female twins who are the subject of the present report are a subset of those registered with the population-based Virginia Twin Register (VTR). The VTR was formed from a systematic review of birth records in the Commonwealth of Virginia from 1915 onwards. Twins were initially mailed a self-report questionnaire to which 64% of individuals responded. The true cooperation rate is likely to be higher than this figure suggests, however, because an unknown proportion of non-responding twins never received the questionnaire due to incorrect mailing addresses, incorrect forwarding of mail etc. Of the 2352 individual twins from 1176 twin pairs who returned the questionnaire, 2163 (92%) twins from 1033 twin pairs were interviewed about their lifetime history of psychiatric disorder an average of 12 ± 3 months (x = 30; s.d. = 7).

Measures
A LTH of MD was assessed at index and follow-up with an adapted version of the SCID (Spitzer et al. 1987) following DSM-III-R criteria (APA, 1987). At follow-up, only lifetime episodes reported to precede the index interview were analysed here in an effort to distinguish new onsets from unreliable recall. This may underestimate diagnostic reliability, however, because some subjects who recall previously denied symptoms at re-test may erroneously move the onset of these symptoms forward (Angst et al. 1984; Rubio-Stipec et al. 1992). Twenty subjects who reported a LTH of MD at both interviews reported an onset at follow-up that was later than their age at index, and these subjects are rated here as having a LTH of MD only at index. Characteristics of MD assessed at both interviews included the number of depressive symptoms, degree of reported impairment, help seeking, age at onset, length of worst episode and the number of lifetime episodes.

A LTH of GAD was assessed at index using Criterion D from DSM-III-R. Criterion A (unrealistic or excessive anxiety and worry (apprehensive expectation) about two or more life circumstances) and criterion B (the focus of the anxiety and worry in A is not attributable to another Axis I disorder) were not surveyed, and no diagnostic hierarchy or exclusion criteria were applied (criterion C and E).

A LTH of panic disorder was assessed at index using criteria A(1), B, C and D from DSM-III-R. Criterion A(2) – panic attacks were not triggered by situations in which the person was the focus of others’ attention – was not surveyed, and criterion E – it cannot be established that an organic factor initiated and maintained the disturbance – was not applied. All diagnoses were assigned by computer algorithm.

Twin analyses
The twin model used here is based on a liability-threshold model that divides the variation in liability to MD into three classes: (i) additive genetic (A), which contribute twice as much to the correlation in MZ twins as DZ twins (because...
The twin model for MD utilizes both our index and follow-up diagnostic data. As pictured in Fig. 1a, the model assumes that there is a true latent liability to MD. Each of our two assessments of LTH of MD at index and follow-up (k1 and k2, respectively) represent transient or ‘occasion-specific’ influences on the individual assessments of LTH of MD, including measurement error. By definition, \( \lambda^2 + k^2 = 1.0 \). The latent liability to lifetime MD and the unreliable or ‘occasion-specific’ influences on recall/diagnosis at each measurement occasion are then modelled 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.

It is important to emphasize two critical differences between this model and the standard twin model. First, this model estimates occasion-specific influences on recall that include errors of measurement (k). Secondly, it provides a direct estimate of the degree to which the individual assessments of LTH of MD at index latent liability to MD (\( \lambda \)). Lastly, this model differs from the model used by Kendler et al. (1993) in one important way. It has been extended to test whether occasion-specific influences on recall (\( \kappa \)) are correlated between twins because of occasion-specific genetic (\( \kappa_a \)) or familial environmental influences (\( \kappa_c \)) on liability for MD.

### Characteristics of reliably diagnosed major depression (MD)

The MD features at index that predict a diagnosis of MD at follow-up (forward prediction), and the MD features at follow-up that predict a diagnosis of MD at index (backward prediction) are characterized using bivariate and multiple logistic regression. The forward prediction sample comprises the 562 twins diagnosed with a LTH of MD at index, and the predictor variables are the number of symptoms, number of episodes, help seeking, impairment, worst episode duration and age at onset reported at index. This approach is comparable to that employed by Kendler et al. (1994). To compare the depressive features of subjects diagnosed as having a LTH of MD only at index or only at follow-up (\( N = 406 \)), bivariate and multiple logistic regression is used to model the depressive features that predicted the diagnosis was assigned only at index (versus follow-up).

### Reliable diagnosis and co-morbidity

To determine if co-morbidity significantly influences the odds of a reliable diagnosis for a LTH of MD, the logistic regression continuation ratio test (CRT, MacClean, 1988) was used to compare the odds that MD was diagnosed once or twice (CRT 1), versus twice (CRT 2), given a lifetime history of GAD or panic disorder at index. If the slope of the regression line is constant across CRT 1 and CRT 2, then the continuation ratio test statistic (D) will be insignificantly different from 0. If D is not insignificantly different from zero this indicates that the odds that MD is reliably diagnosed increase linearly if the subject has a LTH of GAD or panic disorder at index. If D is significantly greater than 0 then this indicates that the odds of a reliable diagnosis increase in...
Fig. 1. (a) A twin model for the heritability of liability to a lifetime history (LTH) of major depression (MD) including transient influences on the recall and diagnosis of a LTH of MD. This model assumes that there is a true (latent) liability to a LTH of MD, which is indexed by two assessments, at time 1 and 2. The paths $\lambda_1$ and $\lambda_2$ represent the degree to which these assessments reflect true liability to LTH of MD. The square of these paths is a measure of the reliability of these assessments. The other paths to a LTH of MD ($\kappa_a$, $\kappa_c$, and $\kappa_e$) represent transient influences on each assessment of a LTH of MD, which may reflect genetic effects on the recall and diagnosis of MD at time 1 ($\kappa_{a1}$) or time 2 ($\kappa_{a2}$), shared or common environmental influences on the recall and diagnosis of MD at time 1 ($\kappa_{c1}$) or time 2 ($\kappa_{c2}$), or unshared or individual-specific environmental influences and/or measurement error on the
a non-linear fashion when the subject has a LTH of GAD or panic disorder at index, and if D is significantly less than 0 then this indicates that the odds of a reliable diagnosis of MD decrease when the subject has a LTH of GAD or panic disorder at index.

**Software**

The twin modelling is performed using Mx (Neale, 1997) with the best-fitting model in our analyses selected using Akaike’s Information Criterion (AIC) (Akaike, 1987; Williams & Holahan, 1994). Logistic regression is performed using forward selection, using a \( P < 0.05 \) significance level to enter variables in the model (SAS Institute Inc., 1990). The continuation ratio test is calculated using a SAS program written by Dr Charles Gardner.

**RESULTS**

Prevalence and reliability of a lifetime history of major depression

The index and follow-up LTH interviews were conducted an average of 62.6 months apart (s.d. = 4.9, range = 46–92). The retest interval is not significantly different for MZ and DZ twin pairs (Wilcoxin two-sample test, \( Z = 0.67, P = 0.50 \)). At index, the prevalence of a LTH of MD is 33.2% (562/1694). At follow-up, the prevalence of MD for the lifetime preceding the index interview is 24.6% (416/1694). This difference is statistically significant (McNemar Test, \( \chi^2 = 52.5, df = 1, P < 0.001 \)).

If a diagnosis of MD at index is defined as the criterion variable, and a diagnosis of MD at follow-up is defined as the ‘test’ variable, then the sensitivity of our diagnoses is 51% (A/A + B in Table 1). The specificity is 89% (D/C + D), the positive predictive value is 69% (A/A + C), and the negative predictive value is 78% (D/B + D). Kappa (Cohen, 1960) for LTH of MD is 0.43 (95% confidence interval 0.38–0.48), indicating only fair agreement across time.

**Twin models**

The first twin model, Model 1 (Fig. 1a, the ‘full model’), fit the data very well (\( \chi^2 = 1.95, df = 6, P = 0.92, AIC = -10.05 \)). Five paths in this model were estimated at zero: familial (shared) environmental influences on liability to reliably diagnosed MD (\( \lambda_c \)), and the occasion-specific genetic and familial environmental influences on a MD diagnosis at index (\( \lambda_a+c1 \) and \( \kappa_c1 \)) and follow-up (\( \lambda_a+c2 \) and \( \kappa_c2 \)). These parameters could, therefore, be set to zero in Model 2 with no change in model fit (\( \chi^2 = 1.95, df = 11, P = 0.99, AIC = -20.05 \)). Model 2 therefore suggests that familial environmental factors and occasion-specific genetic effects do not influence liability to MD. To test if liability to reliably diagnosed MD can be attributed exclusively to genetic effects (\( \lambda_a \)), the environmental liability path (\( \lambda_e \)) was set to zero in Model 3. Compared to Model 2, Model 3 provided a significantly worse fit to the data (\( \chi^2 = 27.22, df = 12, P < 0.01, AIC = 3.22 \)) suggesting that experiences that are not shared by twins significantly contribute to their liability for MD. We next tested if liability to MD can be attributed exclusively to such experiences and therefore set

<table>
<thead>
<tr>
<th>Lifetime history of major depression at follow-up</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lifetime history of major depression at index</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>Present</td>
<td>286 (A)</td>
<td>276 (B)</td>
</tr>
<tr>
<td>Absent</td>
<td>130 (C)</td>
<td>1002 (D)</td>
</tr>
</tbody>
</table>
Table 2. Prediction of reliable reporting of lifetime history of major depression: I. Bivariate analyses

<table>
<thead>
<tr>
<th>Predictors</th>
<th>β</th>
<th>χ²</th>
<th>P</th>
<th>OR (95% CI)</th>
<th>β</th>
<th>χ²</th>
<th>P</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Forward prediction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Characteristics reported at index that predict a LTH of MD is recalled at follow-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of symptoms</td>
<td>0.41</td>
<td>31.62</td>
<td>0.01</td>
<td>1.50 (1.30, 1.73)</td>
<td>0.31</td>
<td>11.74</td>
<td>0.01</td>
<td>1.37 (1.14, 1.64)</td>
</tr>
<tr>
<td>Help-seeking</td>
<td>1.06</td>
<td>26.31</td>
<td>0.01</td>
<td>2.88 (1.92, 4.31)</td>
<td>0.34</td>
<td>2.60</td>
<td>0.11</td>
<td>1.41 (0.93, 2.15)</td>
</tr>
<tr>
<td>No. of episodes</td>
<td>0.14</td>
<td>1.34</td>
<td>0.18</td>
<td>1.15 (0.94, 1.34)</td>
<td>0.31</td>
<td>6.37</td>
<td>0.01</td>
<td>1.36 (1.07, 1.73)</td>
</tr>
<tr>
<td>Impairment</td>
<td>0.39</td>
<td>10.44</td>
<td>0.01</td>
<td>1.47 (1.16, 1.85)</td>
<td>0.20</td>
<td>2.33</td>
<td>0.13</td>
<td>1.22 (0.94, 1.59)</td>
</tr>
<tr>
<td>Duration</td>
<td>0.07</td>
<td>0.71</td>
<td>0.40</td>
<td>1.07 (0.91, 1.26)</td>
<td>0.34</td>
<td>9.26</td>
<td>0.01</td>
<td>1.41 (1.13, 1.73)</td>
</tr>
<tr>
<td>Age at onset</td>
<td>−0.01</td>
<td>0.08</td>
<td>0.77</td>
<td>0.99 (0.97, 1.02)</td>
<td>−0.01</td>
<td>0.06</td>
<td>0.81</td>
<td>0.99 (0.97, 1.02)</td>
</tr>
<tr>
<td>II. Backward prediction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Characteristics reported at follow-up that predict a LTH of MD was previously recalled at index</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Prediction of reliable reporting of lifetime history of major depression: do the same characteristics predict forward and backward discordance? II. Multivariate analyses

<table>
<thead>
<tr>
<th>Predictors</th>
<th>β</th>
<th>χ²</th>
<th>P</th>
<th>OR (95% CI)</th>
<th>β</th>
<th>χ²</th>
<th>P</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Forward prediction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Characteristics reported at index that predict a LTH of MD is recalled at follow-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Help-seeking</td>
<td>0.89</td>
<td>17.13</td>
<td>0.01</td>
<td>2.44 (1.60, 3.72)</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of symptoms</td>
<td>0.38</td>
<td>19.35</td>
<td>0.01</td>
<td>1.46 (1.23, 1.73)</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of episodes</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impairment</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at onset</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Variable excluded in multiple logistic regression using forward selection.

the genetic liability path to zero (λa = 0) in Model 4. This led to an even worse model fit (χ² = 99.18, df = 12, P < 0.01, AIC = 75.18) indicating that both genetic and non-familial environmental risk factors significantly contribute to liability for MD. The final model fit to these data (Model 5) equated the reliability of our index and follow-up interviews (λr = λa), but in all other respects identical to Model 2. In model 5, these two reliability paths could be equated with no significant reduction in model fit (χ² = 2.51, df = 12, P = 0.998, AIC = −21.48) which suggests that the index and follow-up diagnoses are not significantly different indices of liability for MD. Model 5 provides the best fit to these longitudinal diagnostic data (Fig. 1b), estimating that liability to MD in this unselected sample of twins reflects both genetic (λa² = 66%, CI = 53%, 78%) and environmental risk factors unshared by co-twins (λe² = 34%, CI = 22%, 47%). Occasion-specific influences on the recall and diagnosis of MD at onset and follow-up reflect environmental effects unshared by relatives and/or measurement error, and together these account for 34% (CI = 22%, 47%) of the total variance in the diagnosis of LTH of MD at each interview. The estimated heritability of a reliably diagnosed LTH of MD is calculated as Vg/Vp = 0.66 where Vg = λa² and Vp = λa² + λe². The estimated heritability of LTH of MD assessed and diagnosed on the basis of a single psychiatric interview is calculated as Vg/Vp = 0.43 where Vg = λa² and Vp = λa² + λe² + ke².

Characteristics of reliably diagnosed major depression

Forward prediction: what depressive features at time 1 predict a second lifetime diagnosis of major depression at time 2?

Using a diagnosis of MD at index to designate a positive case, the bivariate predictors of a reliable
The lifetime prevalence of GAD at index is 9%, duration of worst episode (OR = 31), less help-seeking (OR = 4), and a shorter duration of illness (OR = 8) in bivariate regressions (Table 4). Using multiple regression, a diagnosis that is made only at index is characterized by an increasing number of symptoms (OR = 1), less help-seeking (OR = 4) and a slightly younger age at onset (OR = 7, Table 4).

**Co-morbidity associated with a reliably diagnoses lifetime history of major depression**

The lifetime prevalence of GAD at index is 6.3% (N = 107/1694). GAD is significantly associated with MD in this sample (OR = 93, Table 5), and the odds of a reliable diagnosis of MD increase when the subject has a history of GAD diagnosed at index (D = 0.54, variance
D = 0.10, Z test = 1.71, P = 0.08, Table 5). The prevalence of GAD among subjects diagnosed with a LTH of MD at both index and follow-up, only at index, only at follow-up, or at neither index or follow-up is 23.8% (N = 68/286), 7.7% (N = 20/276), 2.3% (N = 3/130) and 1.6% (N = 16/1002) respectively.

The lifetime prevalence of panic disorder at index is 2.2% (N = 38/1694). Panic disorder is significantly associated with MD in this sample (OR = 5.0, Table 5) and the odds of a reliable diagnosis of MD increase when a history of panic disorder is diagnosed at index (D = −0.43, variance D = 0.44, Z test = 1.16, P = 0.25, Table 5). The prevalence of panic disorder among subjects diagnosed with a LTH of MD at both index and follow-up, only at index, only at follow-up, or at neither index or follow-up is 11.9% (34/286), 5.4% (15/276), 4.6% (6/130) and 1.7% (17/1002) respectively.

DISCUSSION

An unreliable diagnosis of LTH of MD in this epidemiological sample of adult female twins indexes only measurement error and/or transient non-familial influences on liability to MD. We found no evidence for genetic influences on transient or ‘occasion-specific’ effects on our index and follow-up assessments of LTH of MD. This finding provides further support for the utility of a quantitative risk index derived from the features of reliably recalled and diagnosed depressive histories (Rice et al. 1992; Kendler et al. 1993). A reliably diagnosed LTH of MD in this sample of unselected twins is 50% more heritable than a LTH of MD surveyed and diagnosed at only one occasion (h² = 0.66 versus h² = 0.43 respectively). This underscores the impact of measurement error and other occasion-specific influences on the recall and rating of MD, and the higher mean genetic liability of subjects who consistently report a LTH of MD in epidemiological surveys. Liability to reliably diagnosed MD also reflects the cumulative impact of experiences unshared by relatives, compatible with other aetiological models of MD (Brown & Harris, 1978; Bowlby, 1980).

These results confirm those previously reported by Kendler et al. (1993a), using a self-report measure of MD and a follow-up SCID-based interview, and accord well with the recent findings of McGuffin et al. (1996). McGuffin and colleagues estimated the heritability of DSM-IV major depression in 177 pairs of twins ascertained from the Maudsley Hospital Twin Register. These twins were administered a PSE-based personal interview if their first contact diagnosis met DSM-III criteria for affective disorder. Diagnosis of DSM-IV unipolar depression, therefore, reflects diagnostic agreement for a LTH of MD (albeit variably defined) over two occasions, which is broadly compatible with our concept of reliably diagnosed MD. Assuming a population risk of 16.6% in women, the heritability of MD in the Maudsley twins is estimated at 75%. The prevalence of a reliably recalled LTH of MD in the VTR twins is 16.9%, and heritability is estimated here at 66% (95% CI 53–78%). The comparability of these findings is important because the Maudsley sample represents the largest systematically ascertained clinical sample of twins with unipolar depression.

The time 1 depressive features that predicted a LTH of MD was diagnosed again at time 2 are an increasing number of symptoms and help-seeking. These findings accord with those previously reported by Rice et al. (1992) as help-seeking in the VTR twins subsumes treatment and hospitalization as a result of seeking the help of a medical professional. This suggests that the features of a reliably diagnosed LTH of MD do not vary for DSM-III-R and RDC definitions of depression, and that the correlates of reliable recall do not differ in community and selected samples. If a diagnosis of LTH of MD at follow-up is used to designate a positive case, the depressive features that predict a LTH of MD was previously diagnosed at index are an increasing number of symptoms, an increasing number of lifetime episodes and a longer (worst) episode of illness. These findings partly replicate those reported by Kendler et al. (1993). In that study episode duration was not a significant predictor of the questionnaire-based diagnosis, and help-seeking and impairment were more strongly (and significantly) associated with a reliable diagnosis. The differences between the present study and that conducted by Kendler et al. (1993) are likely to reflect the impact of criterion variance in Kendler’s study, whereas the differences between the findings reported by Kendler et al. (1993) and Rice et al. (1992) are
likely to reflect both criterion variance in Kendler et al. (1993) and differences between subjects diagnosed with a LTH of MD only at index or only at follow-up. In the present study, the latter report relatively fewer symptoms, a longer (worst) episode of illness and more help seeking. It is of interest that McGuffin et al. (1996) reported a significantly higher MZ to DZ concordance ratio associated with depressive episodes of <13 months, which they suggest may reflect a greater genetic loading for depressions of shorter duration. Furthermore, Kendler & Gardner (1998) report that an increasing number of MD symptoms predict an increased risk for future depressive episodes and a heightened co-twin risk for MD. Taken together, these data suggest that a reliably diagnosed LTH of MD indexes the memorability of depression, help-seeking and severity of liability, with a LTH diagnosed only at follow-up likely to index a slightly lower mean liability to MD than a LTH recalled only at index. Quantitative caseness indices of MD may, therefore, need to incorporate the temporal sequence of LTH diagnoses to accurately characterize weighted risk estimates in longitudinal surveys.

McGuffin et al. (1996) have emphasized that estimates of heritability depend, in part, on the (estimated) population frequency of MD, but suggest that differences in prevalence across surveys (as a function of variable threshold placement) are unlikely to affect the overall pattern of results. Although this may be true for univariate model fitting results, the positioning of the diagnostic (or reliability) threshold may have important implications for the characterization of multivariate patterns of risk.

In the VTR twins, a history of GAD is strongly associated with a reliably diagnosed LTH of MD. The genetic correlation between GAD and MD diagnosed on the basis of a single interview has been estimated at unity (Kendler et al. 1992b; Roy et al. 1995), suggesting that a common set of genes underlie the familial component of risk for both disorders. The present findings further suggest that a history of chronic (>6 month) GAD indexes a higher mean liability for MD, consistent with the finding that co-morbidity between MD and GAD is associated with a greater severity of depression in clinical samples (Goethe et al. 1993; Brown et al. 1996). A history of panic disorder also predicts reliably diagnosed MD, supporting the suggestion that co-morbidity reflects a more severe illness (Reich et al. 1993; Andrade et al. 1994; Pini et al. 1994). In the analyses reported here, we have assumed that risk for MD reflects a normally distributed multifactorial liability and that a reliable diagnosis of MD indexes a higher mean liability. Although we consider it very unlikely that a higher liability to MD would not subsume the risk factors for a lower liability to MD (a cumulative risk model), this does not preclude the possibility that there are risk factors that operate only at higher liability levels. For example, if co-morbidity reflects an epiphenomena of illness severity then co-morbidity will characterize more severe depressions.

A history of GAD or panic disorder increases the odds that MD will be reliably diagnosed over a 5-year period. This suggests, first, that more reliable diagnostic assignments may not serve to validate the existing diagnostic boundaries that have been drawn between disorders and, second, that models which formally test if co-morbidity reflects epiphenomena associated with severity of liability of the focal disorder should be more widely utilized (Neale & Kendler, 1995).

The present data also highlight the modest 4 to 8 year reliability ($\kappa = 0.43$) of a SCID-based LTH of MD in this population based sample. This is, however, very similar to the 1 day to 2 week reliability of the SCID in 202 non-patients ascertained via community advertisements ($\kappa = 0.49$) (Williams et al. 1992) and the 18 month reliability of RDC MD in an epidemiological sample of 391 women ($\kappa = 0.41$) (Bromet et al. 1986). Kappa ranges between 0.21 and 0.75 for a LTH of MD in other unselected samples re-tested after 5 days to 7 years using a variety of different interviews and diagnostic criteria (Prusoff et al. 1988).

Cannell & Fowler (1963) suggested that follow-up interviews may provide progressively less accurate information due to lowered subject motivation over time. In the present study, the (personal) index and (telephone) follow-up interviews are statistically equivalent indices of liability to MD. The significantly lower lifetime prevalence of MD estimated at follow-up here may reflect a `re-test artefact', an explanation
previously invoked to account for the lower re-test prevalence of (negative) affective states surveyed by rating scales (Jorm et al. 1989) or at interview (Helzer et al. 1981; Bromet et al. 1986; Eaton et al. 1989). It is also possible that the different memory tasks required at each occasion may be partly responsible for the difference in lifetime prevalence estimated by each interview. At index, subjects are asked to recall any previous episode of MD. At follow-up, subjects are asked to recall any previous MD episode that occurred prior to the index interview. Given that only 20 subjects who reported a history of MD at onset reported that their histories of MD post-dated the onset interview at follow-up, this explanation is unlikely, however, to account for all the variance in prevalence across time.

Wells et al. (1988) concluded that disagreement in LTH assignments over time reflected unreliability and a re-test artefact that affected the reporting of certain depressive symptoms, rather than the method of administration. Weeks et al. (1983) found that subjects interviewed over the telephone were less likely than subjects interviewed face-to-face to report conditions for which they had been hospitalized, but, when they did report such conditions, they did so more accurately. These findings suggest that a substantial amount of time (Sobin et al. 1993) and money (Weeks et al. 1983) could be saved by assessing MD over the telephone, although longitudinal data are likely to incorporate a re-test artefact associated with lower endorsement rates of negative affective states.

Limitations
The results presented here should be interpreted in light of the following limitations. First, reliability is assessed by kappa, and modelled using dichotomous diagnostic assignments, which may exaggerate the cross-time disagreement in the reporting of MD symptoms/duration (Wainwright et al. 1997). Secondly, our twin model of diagnostic reliability assumes that error is random across subjects and is not informative regarding the type of misclassification (Carey & Gottesman, 1978). Thirdly, the sample comprises only women. Although Rice et al. (1992) reported no impact of gender on the stability of MD diagnoses in relatives of patients, Angst et al. (1984) reported that men forgot certain symptoms of depression more often than women did and Wilhelm & Parker (1994) found that women were more likely than men to ‘remember’ episodes of depression that had not previously reached case criteria. Fourthly, given the sex differences in prevalence and familial transmission of MD (Rice et al. 1984; Weissman et al. 1991; Wilhelm et al. 1997), the present findings may not be replicated with a male sample. Finally, the findings derived from latent variable models and logistic regression are informative in so far as the model assumptions are supportable or their violation has a negligible effect upon the parameter estimates obtained. For example, no significant common familial environmental influences on liability were identified here although separation from parents, due to factors such as divorce or parental illness prior to age 17, has been previously shown to have a small (1.6%) but significant effect on risk of MD when separation is modelled as a specified index of the common familial environment of co-twins (Kendler et al. 1992a). Although a comprehensive treatise of putative environmental risk factors evaluated in a similar manner was beyond the scope of the present study, readers should note the relatively low power of latent variable models for identifying any but very sizeable effects of the familial environment shared by twins (Neale et al. 1994).

This work was supported by grants MH-45268, MH-40829, MH-49492, MH-01277 and RR-08123 from the National Institutes of Health, Bethesda, MD and grants from the Carman Trust and the MacArthur Network on Psychopathology.

REFERENCES


