Psychological Medicine, 1998, **28**, 857–870. Printed in the United Kingdom

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

D. L. FOLEY, M. C. NEALE AND K. S. KENDLER[1]

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 ($\kappa = 0{\cdot}43$). 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)

[1] Address for correspondence: Dr Kenneth Kendler, Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, PO Box 980710, Richmond, VA 23298-0710, USA.

(Spitzer *et al.* 1975) was good ($\kappa = 0{\cdot}61$). 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 ($\kappa = 0{\cdot}34$). 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 ($h^2 = 0{\cdot}70$ for reliably diagnosed MD *versus* $h^2 = 0{\cdot}49$ for time 1 (self-report) MD and $h^2 = 0{\cdot}35$ 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 (h^2) 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);

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

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 $Vg = Vg + Ve + Vk = 1{\cdot}00$ (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

$$h^2 = Vg/Vp, \quad \text{where } Vp = Vg + Ve \quad (2)$$

and Vk ('error') was estimated separately. As the denominator in (2) is smaller than the denominator in (1), by a factor equivalent to Vk, 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 Vk 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 Vk 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 (S.D. = 4·0) after receipt of the questionnaire (index lifetime history interview). Of these 2163 twins, 2002 (93%) individuals twins from 938 twin pairs were interviewed over the telephone an average of 17 months later (S.D. = 3·7) about onsets that occurred during the preceding year. Of these 2002 individuals, 1895 (95%) individuals from 849 twin pairs completed another lifetime history interview and average of 44·3 months later (S.D. = 3·9) over the telephone (follow-up lifetime history interview). All interviews were conducted blind to the status of the co-twin by trained field staff who held a Master's degree in Social Work or had at least 2 years of clinical experience. Informed consent was obtained in writing prior to the personal interview, and verbally prior to the telephone interviews.

The data analysed here are for the 847 twin pairs who completed the depression module of both the index and follow-up lifetime history interviews: 496 monozygotic (MZ) and 351 dizygotic (DZ) twin pairs. Zygosity was determined by the twin's responses to standard questions regarding their physical similarity to their co-twin, the frequency with which they were confused as children, photographs and DNA typing (Spence *et al.* 1988; Eaves *et al.* 1989). These 847 twin pairs were aged between 17 and 55 years at index ($\bar{x} = 30$; 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 1 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

MZ twins share all their genes identical by descent, while DZ twins share on average only half their genes); (*ii*) family or 'common' environment (those familial factors which make twins similar in their liability to MD) (C), which contributes equally to the correlation in MZ and DZ twins; and (*iii*) individual specific environment (E), which reflects environmental experiences not shared by both members of a twin pair and therefore contribute to differences between them in their liability to MD. We have previously examined the equal environment assumption for MD (that the exposure to familial environmental risk factors for MD is approximately equal in MZ and DZ twins) in this sample of twins and found no evidence to reject it (Kendler *et al.* 1994).

The twin model for MD utilizes both our index and follow-up diagnostic data. As pictured in Fig. 1 *a*, the model assumes that there is a true latent liability to MD. Each of our two assessments of LTH are considered to be fallible indices 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 index and LTH of MD at follow-up (k_1 and k_2, respectively) represent transient or 'occasion-specific' influences on the individual assessments of LTH of MD, including measurement error. By definition, $\lambda^2 + k^2 = 1{\cdot}0$. 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 index latent liability to MD (λ). 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 (κ) are correlated between twins because of occasion-specific genetic (κa) or familial environmental influences (κc) 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 Rice *et al.* (1992). The backward prediction sample comprises the 416 twins diagnosed with a LTH of MD at follow-up, and the predictor variables are the number of symptoms, the number of episodes, help-seeking, impairment, worst episode duration and age at onset reported at follow-up. This analysis is comparable to that employed by Kendler *et al.* (1993). 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

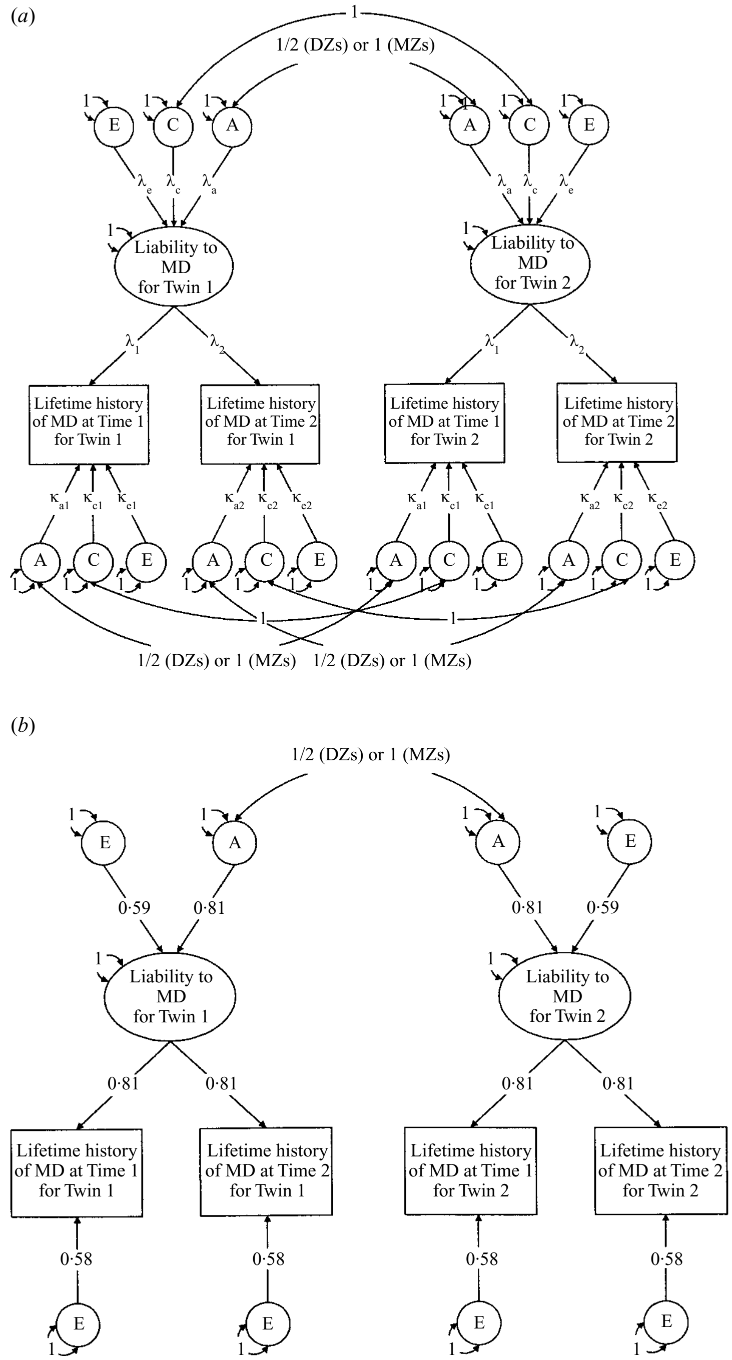

FIG. 1. (*a*) A twin model for the heritability of liability to a lifetime history (LTH) of major depression (MD) including transient influences on the recall and diagnosis of a LTH of MD. This model assumes that there is a true (latent) liability to a LTH of MD, which is indexed by two assessments, at time 1 and 2. The paths λ_1 and λ_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 (κ_a, κ_c and κ_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 (κ_{a1}) or time 2 (κ_{a2}), shared or common environmental influences on the recall and diagnosis of MD at time 1 (κ_{c1}) or time 2 (κ_{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 $P < 0{\cdot}05$ 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, $Z = 0{\cdot}67$, $P = 0{\cdot}50$). 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, $\chi^2 = 52{\cdot}5$ df = 1, $P < 0{\cdot}001$).

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.

Table 1. *Temporal stability over 5 years for a diagnosis of a lifetime history of major depression*

	Lifetime history of major depression at follow-up	
Lifetime history of major depression at index	Present	Absent
Present	286 (A)	276 (B)
Absent	130 (C)	1002 (D)

Twin models

The first twin model, Model 1 (Fig. 1*a*, the 'full model'), fit the data very well ($\chi^2 = 1{\cdot}95$, df = 6, $P = 0{\cdot}92$, AIC = $-10{\cdot}05$). Five paths in this model were estimated at zero: familial (shared) environmental influences on liability to reliably diagnosed MD (λc), and the occasion-specific genetic and familial environmental influences on a MD diagnosis at index (κa1 and κc1) and follow-up (κa2 and κc2). These parameters could, therefore, be set to zero in Model 2 with no change in model fit ($\chi^2 = 1{\cdot}95$, df = 11, $P = 0{\cdot}99$, AIC = $-20{\cdot}05$). 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 (λa), the environmental liability path (λe) was set to zero in Model 3. Compared to Model 2, Model 3 provided a significantly worse fit to the data ($\chi^2 = 27{\cdot}22$, df = 12, $P < 0{\cdot}01$, 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

recall and diagnosis of MD at time 1 (κ_{e1}) or time 2 (κ_{e2}). The model is constrained such that $\lambda^2 + (\kappa a + \kappa c + \kappa e)^2 = 1$. Both the true liability for, and the transient influences on the recall and diagnosis of, a LTH of MD are modelled in a standard twin design with the sources of variance in each divided between additive genetic (A), common (C) environmental, and individual-specific (E) environmental factors. 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 (MZ) twins and correlated 0·50 in dizygotic (DZ) twins. Lower case letters (a, c, 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 variables is equal to the square of the connecting path. Heritability of reliably recalled and diagnosed LTH of MD for example equals λa^2. Observed variables are depicted in boxes and latent variables in circles and ellipses. (*b*) Parameter estimates for the best fitting model (model 5). Parameter estimates are constrained to be equal for twin 1 and twin 2. No evidence was found for common environmental influences on liability to a LTH of MD. No evidence was found for transient genetic or common environmental influences on the recall and diagnosis of a LTH of MD. Heritability of true liability to a LTH of MD = $0{\cdot}81^2 = 66\%$.

Table 2. *Prediction of reliable reporting of lifetime history of major depression: I. Bivariate analyses*

	Bivariate predictors of reliable recall of LTH of MD							
	I. Forward prediction Characteristics reported at index that predict a LTH of MD is recalled at follow-up				II. Backward prediction Characteristics reported at follow-up that predict a LTH of MD was previously recalled at index			
Predictors	β	χ^2	$P <$	OR (95% CI)	β	χ^2	$P <$	OR (95% CI)
No. of symptoms	0·41	31·62	0·01	1·50 (1·30, 1·73)	0·31	11·74	0·01	1·37 (1·14, 1·64)
Help-seeking	1·06	26·31	0·01	2·88 (1·92, 4·31)	0·34	2·60	0·11	1·41 (0·93, 2·15)
No. of episodes	0·14	1·84	0·18	1·15 (0·94, 1·34)	0·31	6·37	0·01	1·36 (1·07, 1·73)
Impairment	0·39	10·44	0·01	1·47 (1·16, 1·85)	0·20	2·33	0·13	1·22 (0·94, 1·59)
Duration	0·07	0·71	0·40	1·07 (0·91, 1·26)	0·34	9·26	0·01	1·41 (1·13, 1·73)
Age at onset	−0·01	0·08	0·77	0·99 (0·97, 1·02)	−0·01	0·06	0·81	0·99 (0·97, 1·02)

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

	Multivariate predictors of reliable recall of LTH of MD							
	I. Forward prediction Characteristics reported at index that predict a LTH of MD is recalled at follow-up				II. Backward prediction Characteristics reported at follow-up that predict a LTH of MD was previously recalled at index			
Predictors	β	χ^2	$P <$	OR (95% CI)	β	χ^2	$P <$	OR (95% CI)
Help-seeking	0·89	17·13	0·01	2·44 (1·60, 3·72)	*	—	—	—
No. of symptoms	0·38	19·35	0·01	1·46 (1·23, 1·73)	0·26	7·92	0·01	1·30 (1·08, 1·57)
No. of episodes	*	—	—	—	0·30	5·77	0·02	1·35 (1·06, 1·73)
Impairment	*	—	—	—	*	—	—	—
Duration	*	—	—	—	0·30	6·82	0·01	1·35 (1·08, 1·70)
Age at onset	*	—	—	—	*	—	—	—

* Variable excluded in multiple logistic regression using forward selection.

the genetic liability path to zero ($\lambda a = 0$) in Model 4. This led to an even worse model fit ($\chi^2 = 99{\cdot}18$, df = 12, $P < 0{\cdot}01$, 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 ($\lambda_1 = \lambda_2$), but was 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 ($\chi^2 = 2{\cdot}51$, df = 12, $P = 0{\cdot}998$, 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. 1*b*), estimating that liability to MD in this unselected sample of twins reflects both genetic ($\lambda a^2 = 66\%$, CI = 53%, 78%) and environmental risk factors unshared by co-twins ($\lambda e^2 = 34\%$, 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 $Vg = \lambda a^2$ and $Vp = \lambda a^2 + \lambda e^2$. 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 $Vg = \lambda a^2$ and $Vp = \lambda a^2 + \lambda e^2 + \kappa e^2$.

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

Table 4. *Prediction of the temporal order of inconsistent recall of a LTH of MD: what characteristics distinguish subjects who recall a LTH of MD only at index from those who report a LTH of MD only at follow-up?*

	Characteristics of inconsistent recall of a LTH of MD									
	I. Bivariate regression Characteristics that predict an inconsistently recalled LTH of MD will be recalled only at index, and not at follow-up					II. Multivariate regression Characteristics that predict an inconsistently recalled LTH of MD will be recalled only at index, and not at follow-up				
Predictors	β	χ^2	S.E.	$P <$	OR (95% CI)	β	χ^2	S.E.	$P <$	OR (95% CI)
Help-seeking	−0·84	12·42	0·24	0·01	0·43 (0·27, 0·69)	−0·93	13·94	0·25	0·01	0·39 (0·24, 0·64)
No. of symptoms	0·27	7·42	0·10	0·01	1·31 (1·08, 1·59)	0·35	10·63	0·11	0·01	1·42 (1·15, 1·75)
No. of episodes	0·03	0·06	0·12	0·81	1·03 (0·81, 1·31)	*	—	—	—	—
Impairment	0·13	0·84	0·15	0·36	1·14 (0·86, 1·52)	*	—	—	—	—
Duration	−0·22	3·98	0·11	0·05	0·81 (0·65, 0·97)	*	—	—	—	—
Age at onset	−0·02	2·86	0·01	0·09	0·98 (0·95, 1·01)	−0·04	6·04	0·01	0·01	0·97 (0·94, 0·99)

* Variable excluded in multiple logistic regression.

Table 5. *Does co-morbidity affected the odds that major depression will be reliably diagnosed*?

	LTH of MD recalled once or twice *v.* never					LTH of MD recalled twice *v.* once				
Risk factor at index	β	S.E.	χ^2	$P <$	Odds ratio (CI)	β	S.E.	χ^2	$P <$	Odds ratio (CI)
LTH of GAD	2·23	0·28	65·49	0·0001	9·33 (5·43, 16·03)	2·39	0·21	125·43	0·0001	10·95 (7·20, 16·65)
LTH of panic disorder	1·61	0·28	32·57	0·0001	5·00 (2·88, 8·70)	0·91	0·29	9·81	0·002	2·47 (1·40, 4·36)

LTH, lifetime history; GAD, generalized anxiety disorder; MD, major depression.

diagnosis are help seeking (OR = 2·9), an increasing number of symptoms (OR = 1·5) and an increasing level of impairment (OR = 1·5, Table 2). The multivariate predictors are help seeking (OR = 2·4) and an increasing number of symptoms (OR = 1·5, Table 3).

Backward prediction: what depressive features at time 2 predict a previous diagnosis of MD at time 1?

Using a diagnosis of MD at follow-up to designate a positive case, the bivariate predictors of a reliable diagnosis are help seeking (OR = 1·4), duration of worst episode (OR = 1·4), and an increasing number of symptoms (OR = 1·4, Table 2). The multivariate predictors are duration of worst episode (OR = 1·3), the number of lifetime episodes (OR = 1·3) and an increasing number of symptoms (OR = 1·3, Table 3). Although the predictors of reliability differ depending on how we designate a positive case, inspection of the 95% confidence intervals for the estimated odds ratios indicates that the prediction of reliability does not differ significantly across paradigms.

The temporal order of unreliable diagnosis: what distinguishes a diagnosis given only at index from one given only at follow-up?

Compared with a diagnosis of LTH of MD that is made only at follow-up, a diagnosis of LTH of MD that is made only at index is characterized by an increasing number of symptoms (OR = 1·31), less help-seeking (OR = 0·4), and a shorter duration of illness (OR = 0·8) 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 = 1·4), less help-seeking (OR = 0·4) and a slightly younger age at onset (OR = 0·97, 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 = 9·3, 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 = 0·14, 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 ($h^2 = 0·66$ *versus* $h^2 = 0·43$ 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.* (1993*a*), 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*. 1992*b*; 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 ($\kappa = 0{\cdot}43$) 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 ($\kappa = 0{\cdot}49$) (Williams *et al*. 1992) and the 18 month reliability of RDC MD in an epidemiological sample of 391 women ($\kappa = 0{\cdot}41$) (Bromet *et al*. 1986). Kappa ranges between 0·21 and 0·75 for a LTH of MD in other unselected samples retested 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 suggests 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*. 1992*a*). 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

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* **52**, 317–332.

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (*3rd edition, revised*) (*DSM-III-R*). APA: Washington, DC.

Andrade, L., Eaton, W. W. & Chilcoat, H. (1994). Lifetime comorbidity of panic attacks and major depression in a population based study. Symptom profiles. *British Journal of Psychiatry* **165**, 363–369.

Aneshensel, C. S., Estrada, A. L., Hansell, M. J. & Clark, V. A. (1987). Social psychological aspects of reporting behavior: lifetime depressive episode report. *Journal of Health and Social Behavior* **28**, 232–246.

Angst, J., Dobler-Mikola, A. & Binder, J. (1984). The Zurich Study – a prospective epidemiological study of depressive, neurotic and psychosomatic syndromes. I. Problem, Methodology. *European Archives of Psychiatric and Neurological Science* **234**, 13–20.

Bowlby, J. (1980). *Attachment and Loss: Volume III: Loss. Sadness and Depression*. Basic Books: London.

Bromet, E. J., Dunn, L. O., Connell, M. M., Dew, M. A. & Schulberg, H. C. (1986). Long-term reliability of diagnosing lifetime major depression in a community sample. *Archives of General Psychiatry* **43**, 435–440.

Brown, C., Schulberg, H. C., Madonia, M. J., Shear, M. K. & Houck, P. R. (1996). Treatment outcomes for primary care patients with major depression and lifetime anxiety disorders. *American Journal of Psychiatry* **153**, 1293–1300.

Brown, G. W. & Harris, T. O. (1978). *Social Origins of Depression: A Study of Psychiatric Disorder in Women*. Tavistock: London.

Cannell, C. F. & Fowler, F. J. (1963). A comparison of a self-enumerative procedure and a personal interview: a validity study. *Public Opinion Quarterly* **27**, 250–264.

Carey, G. & Gottesman, I. I. (1978). Reliability and validity of binary ratings: areas of common misunderstanding in diagnosis and symptom ratings. *Archives of General Psychiatry* **35**, 1454–1459.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46.

Eaton, W. W., Kramer, M., Anthony, J. C., Dryman, A., Shapiro, S. & Locke, B. Z. (1989). The incidence of specific DIS/DSM-III mental disorders: data from the NIMH Epidemiological Catchment Area Program. *Acta Psychiatrica Scandinavica* **79**, 163–178.

Eaves, L. J., Eysenck, H. J. & Martin, N. G. (1989). *Genes, Culture and Personality: An Empirical Approach*. Academic Press: San Diego.

Falconer, D. S. (1967). *Introduction to Quantitative Genetics*. Oliver & Boyd: London.

Fendrich, M., Weissman, M. M., Warner, V. & Mufson, L. (1990). Two-year recall of lifetime diagnoses in offspring at high and low risk for major depression: the stability of offspring reports. *Archives of General Psychiatry* **47**, 1121–1127.

Goethe, J. W., Fischer, E. H. & Wright, J. S. (1993). Severity as a key construct in depression. *Journal of Nervous and Mental Disease* **181**, 718–724.

Helzer, J. E., Robins, L. N., Groughan, J. & Welner, A. (1981). Renard Diagnostic Interview: its reliability and procedural validity with physicians and lay interviewers. *Archives of General Psychiatry* **38**, 393–398.

Jorm, A. F., Duncan-Jones, P. & Scott, R. (1989). An analysis of the re-test artefact in longitudinal studies of psychiatric symptoms and personality. *Psychological Medicine* **19**, 487–493.

Keller, M. B., Klein, D. N. & Hirschfeld, M. A. (1995). Results of the DSM-IV mood disorders field trial. *American Journal of Psychiatry* **152**, 843–849.

Kendler, K. S. & Gardner, Jr., C. O. (1998). The boundaries of major depression: an evaluation of DSM-IV criteria. *American Journal of Psychiatry* **155**, 172–177.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. (1992*a*). Childhood parental loss and adult psychopathology in women: a twin study perspective. *Archives of General Psychiatry* **49**, 109–116.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. (1992*b*). Major depression and generalized anxiety disorder. Same genes, (partly) different environments? *Archives of General Psychiatry* **49**, 716–722.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. (1993*a*). The lifetime history of major depression. Reliability of diagnosis and heritability. *Archives of General Psychiatry* **50**, 863–870.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. (1994). Parental treatment and the equal environment assumption in twin studies of psychiatric illness. *Psychological Medicine* **24**, 579–590.

MacClean, C. J. (1988). Assessing changes in risk factor effect over multiple levels of severity. *American Journal of Epidemiology* **127**, 663–673.

McGuffin, P., Katz, R., Watkins, S. & Rutherford, J. (1996). A hospital based twin register of the heritability of DSM-IV unipolar depression. *Archives of General Psychiatry* **53**, 129–136.

Neale, M. C. (1997). *Mx: Statistical Modeling, Fourth Edition*. Department of Psychiatry, Box 710, MCV, Richmond, VA 23298.

Neale, M. C. & Cardon, L. R. (1992). *Methodology for Genetic Studies of Twins and Families*. Kluwer Academic Publishers: The Netherlands.

Neale, M. C. & Kendler, K. S. (1995). Models of comorbidity for multifactorial disorders. *American Journal of Human Genetics* **57**, 935–953.

Neale, M. C., Eaves, L. J. & Kendler, K. S. (1994). The power of the classical twin study to resolve variation in threshold traits. *Behavior Genetics* **24**, 239–258.

Pini, S., Goldstein, R. B., Wickramaratne, P. J. & Weissman, M. M. (1994). Phenomenology of panic disorder and major depression in a family study. *Journal of Affective Disorders* **30**, 257–72.

Prusoff, B. A., Merikangas, K. R. & Weissman, M. M. (1988). Lifetime prevalence and age of onset of psychiatric disorders: recall 4 years later. *Journal of Psychiatric Research* **22**, 107–117.

Reich, J., Warshaw, M., Peterson, L. G., White, K., Keller, M., Lavori, P. & Yonkers, K. A. (1993). Comorbidity of panic and major depressive disorder. *Journal of Psychiatric Research* **27**, (suppl. 1), 23–33.

Rice, J. P., Reich, T., Anderson, N. C., Lavori, P. W., Endicott, J., Clayton, P. J., Keller, M. B., Hirschfeld, R. M. A. & Klerman, G. L. (1984). Sex-related differences in depression. *Journal of Affective Disorders* **71**, 199–210.

Rice, J. P., Rochberg, N., Endicott, J., Lavori, P. W., Miller, C. (1992). Stability of psychiatric diagnoses: an application to the affective disorders. *Archives of General Psychiatry* **49**, 824–830.

Robins, E. & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. *American Journal of Psychiatry* **126**, 107–111.

Roy, A., Neale, M. C., Pedersen, N. L., Mathe, A. A. & Kendler, K. S. (1995). A twin study of generalized anxiety disorder and major depression. *Psychological Medicine* **25**, 1037–1049.

Rubio-Stipec, M., Freeman, Jr D. H., Robins, L., Shrout, P., Canino, G. & Bravo, M. (1992). Response error and the estimation of lifetime prevalence and incidence of alcoholism: experience in a community survey. *International Journal of Methods in Psychiatric Research* **2**, 217–224.

SAS Institute Inc. (1990). *SAS/STAT Users Guide, Version 6, Fourth Edition*. SAS Institute Inc.: Cary, NC.

Sobin, C., Weissman, M. M., Goldstein, R. B., Adams, P., Wickramaratne, P., Warner, V. & Lish, J. D. (1993). Diagnostic interviewing for family studies: comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. *Psychiatric Genetics* **3**, 227–233.

Spence, J. E., Corey, L. A., Nance, W. E., Marazita, M. L., Kendler, K. S. & Schieken, R. M. (1988). Molecular analysis of twin zygosity using VNTR DNA probes. *American Journal of Human Genetics* **43**, A159 (Abstract).

Spitzer, R. L., Williams, J. B. & Gibbon, M. (1987). *Structured Clinical Interview for DSM-III-R*. Biometrics Research Division, New York State Psychiatric Institute: New York.

Spitzer, R. L., Endicott, J. & Robins, E. (1975). *Research Diagnostic Criteria (RDC) for a Selected Group of Functional Disorders*. Biometrics Research Division, New York State Psychiatric Institute: New York.

Wainwright, N. W. J., Surtees, P. G. & Gilks, W. R. (1997). Diagnostic boundaries, reasoning and depressive disorder, I. Development of a probabilistic morbidity model for public health psychiatry. *Psychological Medicine* **27**, 835–845.

Weeks, M. F., Kulka, R. A., Lessler, J. T. & Whitmore, R. W. (1983). Personal versus telephone surveys for collecting household health data at the local level. *American Journal of Public Health* **73**, 1389–1394.

Weissman, M. M., Bruce, M. L., Leaf, P. J., Florio, L. P. & Holzer, C. (1991). Affective disorders. In *Psychiatric Disorders in America: the Epidemiologic Catchment Area Study* (ed. L. N. Robins, and D. A. Regier), pp. 53–80. Free Press: New York.

Wells, K. B., Burnam, M. A., Leake, B. & Robins, L. N. (1988).

Agreement between face-to-face and telephone-administered versions of the depression section of the NIMH Diagnostic Interview Schedule. *Journal of Psychiatric Research* **22**, 207–220.
Wilhelm, K. & Parker, G. (1994). Sex differences in lifetime depression rates: fact or artefact? *Psychological Medicine* **24**, 97–111.
Wilhelm, K., Parker, G. & Hadzi-Pavlovic, D. (1997). Fifteen years on: evolving ideas in researching sex differences in depression. *Psychological Medicine* **27**, 875–883.
Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, Jr. H. G., Rounsaville, B. & Wittchen, H-U. (1992). The structured clinical interview for DSM-III-R (SCID) 11. Multisite test–retest reliability. *Archives of General Psychiatry* **49**, 630–636.
Williams, L. J. & Holahan, P. J. (1994). Parsimony-based fit indices for multiple indicator models: do they work? *Structural Equation Modeling* **1**, 161–189.