Eur Arch Psychiatry Clin Neurosci (1993) 243:158–163

European Archives of Psychiatry and Clinical Neuroscience

Phenotype definition for genetic studies

John P. Rice

Department of Psychiatry (Box 8134), Washington University School of Medicine, 4940 Children's Place, St. Louis, MO 63110, USA

Summary. A difficulty in the interpretation of the reliability/stability of a lifetime diagnosis of mental disorders is the lack of a theoretical perspective. A model expressed in terms of the three unknowns-sensitivity, specificity and true base rate – is problematic due to the lack of a "gold standard", so that only two of these unknowns can be estimated. We extend this model to allow for clinical covariates that increase the likelihood that a positive case at Time 1 will be positive at Time 2. Under the assumption that all observed cases are true cases at the highest covariate values, we obtain a direct estimate of the sensitivity, so that all unknowns can be estimated. Moreover, we then calculate the likelihood that an observed case with given covariate levels is in fact a true case. The implications of diagnostic error for the fitting of genetic models are given.

These methods are applied to stability data collected as part of the NIMH Psychobiology of Depression Program. A total of 1,629 relatives have been assessed with interviews separated by a 6-year interval. A logistic function was used to model the stability in relatives with an initial lifetime diagnosis of affective disorders.

We discuss the use of these techniques in genetic models to increase information by defining an ordinal phenotype, use multiple assessments to minimize the impact of diagnostic error and increase the heritability, and utilize clinical covariates to model the certainty of diagnosis.

Key words: Diagnostic stability – Genetics of affective disorders – Psychiatric genetics

Introduction

The rapid advances in molecular genetics have provided a detailed linkage map of the human genome. This provides an exciting opportunity to use linkage methods to discover genes which contribute to susceptibility to mental disorders. However, success will depend on phenotype definition and appropriate data analytic techniques to detect a particular locus for a complex and heterogeneous trait. In this paper, we concentrate on two issues: the definition of a polychotomus phenotype to increase information for a linkage study, and the use of multiple, blind diagnostic assessments to increase the heritability of the liability for the disorder. This work is a synthesis of prior publications (Rice et al. 1986, 1987a, 1992, Rice and Todorov, in press).

The reliability of psychiatric assessment has received considerable attention since the introduction of structured diagnostic interviews. Studies of reliability have compared diagnoses made by two or more raters for either a single, joint interview or separate interviews spaced a few days apart. Accordingly, these studies use rater agreement to focus on the psychometric properties of a particular instrument and to ensure uniformity among raters. Results are typically reported using the kappa statistic as a yardstick, although there can be difficulties with this approach due to the lack of an appropriate underlying model, especially when the base rate of diagnosis is low (Grove et al. 1981).

Reliability is an indication of the repeatability of a measurement, but does not ensure that the measure is a valid indicator of an underlying construct. Studies of the validity of diagnosis have considered variables associated with course, treatment response, biological and psychosocial correlates, and familial aggregation (Robins and Guze 1970).

A concept related to reliability is that of the temoral stability of diagnosis. Assessment of temporal stability uses interviews at widely separated time points (or successive admissions) and is influenced by sources of disagreement other than those in the usual test-retest comparison. The distinction between instrument error and stability is well illustrated in the Muscatine hyperlipidemia family study (Schrott et al. 1979), and is given since comparable data are not available in psychiatry. The correlation for same-day measures of systolic and diastolic blood pressures were both found to be 0.83 and indicate relatively little instrument or rater error. In contrast, the correlations across time were relatively low: 0.41 and 0.27 after 2 years, and 0.30 and 0.18 after 6 years for systolic and diastolic measures, respectively. The correlations across time may be correted for instrument error (i.e., attenuation) by dividing by 0.83, so that 2-year correlation in "true" diastolic pressure (correcting for measurement error) would be 0.49 rather than 0.41,

Correspondence to: J.P. Rice

and would be the value expected if the instrument were perfectly reliable. Accordingly, we infer that the trait is relatively unstable over time and that the instability of observed measures over time is not accounted for solely by measurement error.

The paradigm is conceptually more complicated when considering diagnoses made using a psychiatric interview. The situation where patients are evaluated for their current clinical state at successive admissions is most directly comparable. For example, we may consider whether the subtyping of episodes of major depression according to the endogenous-nonendogenous dichotomy is stable over episodes and whether instability is accounted for by instrument unreliability at the two assessments. This addresses whether the endogenous subtype is a trait characteristic of the individual or a temporally unstable characteristic of a particular episode.

In family studies and in surveys of the general population, it is necessary to consider lifetime diagnoses based on retrospective information. We consider here the situation where independent lifetime assessments are made at widely separated time points. We distinguish between the reliability of the instrument as measured in a short-term test-retest design and stability of a lifetime diagnosis. For example, when assessment are only a few days apart, subjects will likely remember their first set of responses and be influenced accordingly, whereas it is unlikely that a subject recalls responses from an interview several years earlier. Indeed, it is possible to have perfect short-term test-restest reliability with little stability over longer periods. Stability studies have been less common in psychiatry, and have mostly used patient populations with a positive diagnosis at initial evaluation.

Although the reliability and stability designs appear similar – two independent assessments over time – there are conceptual differences in two areas. Cloninger et al. (1979) used path analysis to model sources of diagnostic disagreement. They included inconsistent histories on the subjects' part; borderline or atypical clinical features of the subject; interviewer procedure; and other sources of error as causal factors. These, together with true clinical status, determine the diagnosis made. In addition, effects due to retrospective recall and the setting or circumstances of the interview may be important. In a short-term reliability study, there is likely to be little variation in many of the error terms between the two interviews, and the focus is on quantification of variability in the interview procedure. Indeed, a videotape reliability study is an excellent approach to control for within-subject variation in these other variables, so that the correlations in error components across time are near unity. In contrast, for a stability study, the assumption will be made that these correlations are negligible if the time period between interviews is great enough. Under this assumption, agreement reflects the component corresponding to true clinical state (however, it is still necessary to consider external correlates of validity as a test of the assumption of uncorrelated errors).

The other conceptual difference is that the true clinical state may change in the time interval between interviews in a stability study. For example, if a unipolar subject has a manic episode between interviews, this is not an error in diagnosis. Rather, if diagnoses are assessed on a lifetime basis, there is a different degree of censoring of the risk period at the two evaluations. Accordingly, in a stability study, it is necessary to consider the true lifetime clinical status and allow for the true clinical status at evaluation to reflect the period of incomplete observation. Moreover, the diagnostic picture before the frist interview may be clearer with a longer period of observation.

In a pioneering investigation, Masserman and Carmichael (1938) studied a series of 100 patients seen at a psychiatric clinic and after a year found that major diagnostic changes were necessary in 40% of their sample. Ødegaard (1966) compared the first and last diagnoses for patients who were admitted to Norwegian mental hospitals for the first time between 1950 and 1954 and who had a subsequent admission. He found many with an initial diagnosis of reactive psychosis to have a subsequent diagnosis of either schizophrenia or manic-depressive illness. Cooper (1967) studied inpatients in England and Wales with four different admissions in 1954–55. Only 54% were given diagnoses in the same broad categories on all four occasions, with only 16% having evidence for a true change in the clinical status that would account for the change in diagnosis. It should be noted that in these studies the base changes and reassessments were not blind to the first assessment, so the results must be interpreted accordingly. More recently, Kendell (1974) used hospital diagnoses, whereas studies by Murphy et al. (1974), Tsuang et al. (1981) and Faravelli and Poli (1982) used the criteria of Feighner et al. (1972). Murphy et al. (1974) started with 115 hospitalized patients and found 40% of those not initially diagnosed as having primary depression to be so diagnosed at follow-up. Tsuang and colleagues (1981) found 63% of unipolar depressions and 56% of bipolar depressions to have the same diagnosis in a 30–40-year follow-up. Faravelli and Poli (1982) found 54% of their primary major depressions to be so diagnosed 4 years later. The magnitude of error we found for major depressive disorder (MDD) using the Schedule for Affective Disorders and Schizophrenia-Lifetime version (SADS-L) is consistent with that reported in the studies of Bromet et al. (1986) and Prusoff et al. (1988) which also used the SADS-L.

Despite the fundamental importance of the validity of diagnosis and the framework set forth by Robins and Guze (1970) and Kendell (1975), there has been a lack of new methods used in psychiatric family studies. The most commonly used statistic in reliability study is Kappa which is usually defined by computational procedures rather than a parametric model for a given population.

In other areas of medicine, some work has been done (Hui and Walter 1980; Diamond et al. 1986; Henkelmen et al. 1990; Schulzer et al. 1991) using multiple populations or more than two reassessments to estimate sensitivity and specificity. These methods have been reviewed by Walter and Irwing (1988) and Uebersax and Grove (1989).

In what follows, we described methods in Rice et al. (1986, 1987a, 1992, in press) and present an application to major depressive disorder.

Methods

Sensitivity, specificity, true base rate, and kappa

Let x and y denote the true state and observed state, with values of 1 for "affected" and 0 for "not affected". If the true state x can be determined (i.e. if a gold standard exists), then the sensitivity (p), specificity (q) and true base rate (K) are defined as

$$p = \text{Prob}(y = 1|x = 1) \quad (1)$$
$$q = \text{Prob}(y = 0|x = 0) \quad (2)$$
$$K = \text{Prob}(x = 1), \quad (3)$$

where Prob( ) denotes the probability of the event in parentheses.

Note the observed rate in the populations is given by

$$\text{Prob}(y = 1) = pK + (1 - q)(1 - K) \quad (4)$$

and consists of true positives and false positives.

A related quantity is the predictive value and is given by

$$PV = \frac{pK}{pK + (1 - q)(1 - K)}. \quad (5)$$

The predictive value is the proportion of individuals who are observed to be positive who are in fact true positives. With a constant sensitivity and specificity, PV decreases as K decreases.

Unfortunately, these unknowns (p, q, K) cannot be estimated in the standard test-retest study, although any subset of two parameters can be.

Kraemer (1979) has derived an elegant formula for the value of kappa (K) as:

$$\kappa = \frac{K(1 - K)}{P(y = 1)\,P(y = 0)}(p + q - 1)^2. \quad (6)$$

From Eq. (6) we see the dependence between Kappa and true base rate when the sensitivity and specificity are held constant.

The use of diagnostic covariates

For time i, $i = 1, 2$, let x_i and y_i denote the true state and observed state at evaluation i.

Let Z be a set of covariates for an observed case at Time 1. The variables Z may include the number of symptoms and other clinical features, the number of reported episodes and treatment-seeking behavior, as well as variables such as sex and age. We will make the assumption at Time 2 that the probability of an observed positive case conditional on the true state at Time 2, does not depend on y_1 or Z. That is, that

$$\text{Prob}(y_2 = 1|x_2 = 1, y_1 = i, Z) = \text{Prob}(y_2 = 1|x_2 = 1) = p \quad (7)$$
$$\text{Prob}(y_2 = 1|x_2 = 0, y_1 = i, Z) = \text{Prob}(y_2 = 1|x_2 = 0) = 1 - q. \quad (8)$$

Thus, if we knew the true state at Time 2, then the Time 1 information would not influence sensitivity or specificity. Under these assumptions, we can derive that

$$I(Z) = \text{Prob}(x_2 = 1|y_1 = 1, Z) = \frac{\text{Prob}(y_2 = 1|y_1 = 1, Z) - (1 - q)}{p + q - 1}. \quad (9)$$

We model $\text{Prob}(y_2 = 1|y_1 = 1, Z)$, using a logistic regression model, with

$$L(Z) = \text{Prob}(y_2 = 1|y_1 = 1, Z) = \frac{\exp(\alpha + \beta' Z)}{1 + \exp(\alpha + \beta' Z)}. \quad (10)$$

If we assume that all observed cases are true cases at the highest covariate value Z_{max}, we have that

$$\text{Prob}(x_2 = 1|y_1 = 1, Z_{max}) = 1,$$

so that

$$p = \text{Prob}(y_2 = 1|y_1 = 1, Z_{max}) = L(Z_{max}), \quad (11)$$

and the full model may be estimated from stability data.

The coefficients α and β may be estimated using standard logistic regression software. This not only permits estimation of the sensitivity p, but yields a parametric form for Eq. (9) that can be used in epidemiologic and genetic analysis. We refer to $I(Z)$ as the index of caseness.

From the estimate of the sensitivity p, we can then derive an estimate of the specificity q as

$$q = 1 - \frac{\text{Prob}(y_1 = 1) - p\,\text{Prob}(y_1 = 1, y_2 = 1)}{p - \text{Prob}(y_2 = 1)}. \quad (12)$$

Moreover, we can derive the probability $I(Z)$ of being a true case at Time 2, conditioned on being positive at Time 1 as

$$I(Z) = (L(Z) - (1 - q))/(p + q - 1). \quad (13)$$

Note that $I(Z)$ depends on p, q, α, β and the covariates from Time 1. Accordingly $I(Z)$ may be defined using the parameter values obtained from a study of stability and applied in a study using a single assessment. We use $I(Z)$ as our index of caseness for an observed case.

The estimation of the sensitivity p requires some model-based statistical assumption to be made. The rationale behind Eq. (11) is that being a true case is related to stability over time and that all individuals with values Z_{max} are true cases. If, for example, predictors are not highly associated with a positive diagnosis at Time 2, then the estimate of p will be low. To the extent that all such cases are not in fact true cases, the p will be underestimated.

Data application

Subjects

The Collaborative Depression Study (CDS) of the National Institute of Mental Health is a naturalistic, prospective investigation of 955 probands and 2,226 relatives of 612 of these probands who participated in a family study (Andreasen et al. 1987; Rice et al. 1987b). One component of this study was a second, "blind" reassessment of all relatives 6 years after initial evaluation using the SADS-L. We use the sample of 1,629 relatives described previously (Rice et al. 1992), and consider the hierarchical lifetime diagnosis as displayed in Table 1 (see also Table 7 of Rice et al. 1992).

Results

We first use logistic regression to determine whether the putative hierarchy in Table 1 reflects an ordinal relation-

Table 1. Hierarchical diagnoses[a] at time 1 and time 2

	SA/M	MAN	HYP	MDD	Minor	Other	NMI	N
SA/M	3	2						5
MAN	1	17	5	4		1	1	29
HYP	1	3	31	34	2	12	14	97
MDD	1	3	24	292	32	39	36	427
Minor			5	30	15	21	32	103
Other		1	9	61	10	123	53	258
NMI		2	10	101	40	97	449	700
N	6	28	84	522	99	293	585	1619

[a] SA/M, Schizoaffective, mania; MAN, mania; HYP, hypomania; MDD, major depression disorder; NMI, never mentally ill

Adapted from Rice and Todorov (in press)

Table 2. Predictors of MDD at time 2

T_1 diagnosis	Categorical analysis		Ordinal analysis
	Odds ratio	χ^2_1	Odds ratio
MDD-A	53.8	141.4	50.2
MDD-B	24.5	102.2	26.1
MDD-C	13.9	101.6	13.6
MDD-D	7.2	90.5	7.1
Minor depression	2.6	14.8	3.7
Other diagnosis	1.9	12.4	1.9
Never mentally ill	1.0	–	1.0

Adapted from Rice and Todorov (in press)

Table 3. Predictors of hypomania at time 2

T_1 diagnosis	Odds ratio	χ^2_1	Effect of hypomania	Ordinal
Hypomania, MDD-A	45.8	54.5		65.9
Hypomania, MDD-B	27.5	24.3		28.5
Hypomania, MDD-C	39.3	27.2	34.3	12.3
Hypomania, MDD-D	38.2	32.1		5.3
Hypomania, no MDD	26.2	38.8		2.3
MDD	4.2	13.8	4.1	1.6
Minor depression	3.5	5.0	3.5	1.4
Other diagnosis	2.5	3.9	2.5	1.0
Never mentally ill	1.0	–	–	–
Model fit (df)			1.2 (4)	62.0 (4)

Adapted from Rice and Todorov (in press)

Table 4. Predictors of mania at time 2

T_1 diagnosis	Odds ratio	χ^2_1
Mania	88.2	87.2
Hypomania	15.1	16.9
MDD	3.3	5.4
Never mentally ill	1.3	–

Adapted from Rice and Todorov (in press)

ship in predicting the various Time 2 diagnoses. We begin at the bottom of the hierarchy.

We used the four levels of severity of MDD based on the index of caseness (Rice et al. 1992) denoted MDD-A, MDD-B, MDD-C and MDD-D. We created six dummy variables corresponding to the seven categories MDD-A, MDD-B, MDD-C, MDD-D, minor depression, other diagnosis and never mentally ill. The resulting odds ratios for MDD represent a gradient of risk (Table 2). The regression coefficient is 0.6527. The resulting odds ratios (third column of Table 2) are very similar to those obtained when ordinality of the categories are not assumed (first column). The hypothesis of ordinality is accepted ($\chi^2_5 = 2.8$, NS).

In contrast to the ordinal relationship for MDD, hypomania in combination with the index of caseness for MDD did not define an ordinal relationship for stability of hypomania. The ordinal model in Table 3 gives a chi square of 62.0 with four degrees of freedom ($P < 0.0001$), whereas the model with an effect of only the presence of hypomania (with or without MDD) gave an odds ratio of 34.3 and a good fit to the data ($\chi^2_4 = 1.2$).

The predictors of mania are given in Table 4. It is interesting that hypomania predicts an intermediate risk between mania and MDD. Compared to other variables, MDD is not a powerful predictor of either mania (OR = 3.3) or hypomania (OR = 4.1).

The polygenic model. According to the polygenic model (Rice et al. 1991), we can partition the liability to develop a disorder (denoted X) as

$$X = gG + cE_c + uE_u + eE_{error}, \tag{14}$$

where G, E_c, E_u, and E_{error} denote the polygenic, shared environmental, unique environmental, and error components, respectively, and where the components are assumed independent. The variance of X is 1, so that $g^2 + c^2 + u^2 + e^2 = 1$. The parent-offspring, sibling, dizygotic, monozygotic and within-person correlations are given by:

$$\begin{aligned} r_{po} &= \tfrac{1}{2}g^2 \\ r_{oo} &= \tfrac{1}{2}g^2 + c^2 \\ r_{dz} &= \tfrac{1}{2}g^2 + c^2 \\ r_{mz} &= g^2 + c^2 \\ r_w &= g^2 + c^2 + u^2. \end{aligned} \tag{15}$$

Polygenic model with repeated diagnostic assessments. As noted by Falconer (1981), the standard methods for multiple measures are useful in dealing with the error term E_{error}. If n independent measurements are made then their average Y can be written as

$$Y = gG + cE_c + uE_u + (e/n)\, E_{error}, \tag{16}$$

so that the heritability estimate h^2 for Y is given by

$$h^2 = g^2/(1 - e^2\,(n^2 - 1)/n^2). \tag{17}$$

Thus, repeated measures reduce the impact of error by a factor of $(n^2 - 1)/n^2$. As noted in Eq. (15), the estimates of familial correlations will be increased when repeated measurements, rather than a single measurement, are used.

Geneticists distinguish between true cases and phenocopies, individuals who have a positive diagnosis which does not reflect the same underlying genetic/familial process as that in the proband (index case). In clinical samples, probands are often ill at the time of ascertainment, have identified themselves as cases and have multiple sources of clinical corroboration. In contrast, the relatives may have milder, untreated forms of illness that must be retrospectively assessed from a lifetime perspective. The methods outlined above allow the use of clinical correlates in the relatives to assign a probability that a relative is a true case rather than a phenocopy.

To assess the impact of diagnostic error on the estimation of parameters in a familial transmission model, we examined the multifactorial model in which an underlying liability distribution is postulated with true cases reflecting individuals whose liability values are

Table 5. The effect of imperfect sensitivity on observed multifactorial parameters

p	q	Probands include false positives			Probands are true case	
		Observed prevalence	Rate in relatives	r	Rate in relatives	r
1.00	1.00	0.100	0.324	0.500	0.324	0.500
0.95	1.00	0.095	0.308	0.484	0.308	0.484
0.90	1.00	0.090	0.292	0.469	0.292	0.469
0.85	1.00	0.085	0.275	0.453	0.275	0.453
1.00	0.95	0.145	0.284	0.326	0.358	0.463
0.95	0.95	0.140	0.270	0.309	0.342	0.445
0.90	0.95	0.135	0.255	0.291	0.325	0.426
0.85	0.95	0.130	0.240	0.272	0.309	0.408
1.00	0.90	0.190	0.285	0.225	0.392	0.438
0.95	0.90	0.185	0.272	0.208	0.375	0.416
0.90	0.90	0.180	0.260	0.193	0.359	0.396
0.85	0.90	0.175	0.247	0.176	0.343	0.376
1.00	0.85	0.235	0.304	0.164	0.425	0.418
0.95	0.85	0.230	0.292	0.148	0.409	0.396
0.90	0.85	0.225	0.281	0.135	0.393	0.374
0.85	0.85	0.220	0.270	0.121	0.377	0.352

Adapted from Rice et al. (1987a)

above a threshold value (Rice and Reich 1985). The joint distribution of liability within a family is assumed multivariate normal, and familial resemblance is quantified in terms of the correlation in liability between family members. We fixed a true prevalence of 10% in the population and a correlation of 0.5 in liability. In this instance, 32% of the relatives of a true case would also be true cases.

In Table 5 we examine the effect of different combinations of sensitivity and specificity on the observed population prevalence and observed proportion of affected relatives of a proband. We compute this latter quantity under the assumption that probands and relatives are sampled with error and under the assumption that a proband is a true case. We also display the value of the tetrachoric correlation that would be obtained from the observed prevalence and proportion of affected relatives if it were assumed that $p = q = 1$.

As expected, the presence of (unrecognized) diagnostic error can considerably lower the estimate of familial resemblance. This is especially true when probands who are false positives are included. If the values of p and q were specifically modelled, then the true base rate and proportions of true cases in relatives could be estimated. Moreover, including covariates in the affected relatives should enhance the power in fitting such familial transmission models.

Discussion

Kendell (1975) has provided as excellent overview of conceptual issues involved in classifying individuals according to discrete categories versus a dimensional system. The above model based on sensitivity and specificity assumes an underlying discrete classification based on being a true case, and in addition, defines a quantitative index of caseness based on stability over time. Note, however, that the caseness index is defined only for people who receive the diagnosis, and is related to the probability of being a true case.

The data presented indicate that there is substantial error in a cross-sectional assessment of this non-clinical sample. Moreover, this error appears to decrease as reported severity increases. It is important to realize that the interpretation of this error depends on the assumptions of the statistical model applied to the data. We have assumed that the error components are uncorrelated between assessements, and that the sensitivity and specificity are constants. Neither of these assumptions are likely to hold exactly. For example, personality attributes, which tend to have high temporal stability, may influence reporting across occasions and may in part determine the probability of diagnosis. In addition, we have made the assumption that individuals with the highest covariate values are true cases. Without an assumption such as this, the sensitivity can not be estimated and the model is not determined. The reader must recognize the limitations inherent in any statistical model and that the conclusions are dependent on the assumptions made. However, anlayses which do not take into account diagnostic error (i.e. try to avoid statistical modelling) in some sense assume that the sensitivity and specificity are perfect ($p = q = 1$); our model is a first attempt to relax these assumptions.

A further key assumption is that the diagnosis arises from an underlying dichotomy. An alternative formulation is that there is an underlying continuum, termed liability to depression, with a threshold value such that cases correspond to individuals above the threshold value. This approach is often used in genetic models of the affective disorders, and is a latent dimensional approach and is that given by Eq. (14). In this setting, the index of caseness may be used to grade affected individuals along the liability scale according to severity. Moreover, this provides a way to redefine who is case (e.g. an MDD diagnosis with $I(Z) \geq 0.9$), or to define multiple liability classes used in genetic analysis.

In many approaches to genetic analysis, for example in most linkage analyses, individuals are classified as "affected", "unaffected", or "unknown". Although false negatives are usually modelled using incomplete penetrance, false positives may appear as recombinant individuals and reduce the power to detect linkage unless a high sporadic rate is used. By allowing for certainty of diagnosis, different classes of affected individuals may have different sets of penetrances to reflect this diagnostic uncertainty.

As noted in Eq. (16), multiple assessments can be used to reduce the effect of measurement error. In the case of oligogenic transmission (i.e. several loci which contribute to liability), the heritability will be a key parameter in determining the power in genetic analysis.

The second area of impact is in disease definition. Reich et al. (1972) note the use of multiple thresholds to

increase information for genetic analysis. The ability to grade affected individuals (e.g. MDD-A through MDD-D) on an ordinal scale is provided by the above methods. Moreover, "spectrum" diagnoses may be modelled and related to the primary diagnosis of interest. The above methods can be used to assign a probability to a given diagnosis. For example, individuals with minor depression give an intermediate odds of 2.6 for MDD (Table 2). Rather than combine individuals with minor depression with those with MDD or those who are never mentally ill, they may be left as an intermediate class.

In summary, the approach outlined above may have application to family study or epidemiologic studies of mental illness. Due to a general lack of a "gold standard", error in diagnosis has not been directly modeled in most studies. Our data show a moderate degree of instability over time and indicate that such error is prominent near the threshold of diagnosis. Using consistency over time as a partial gold standard, we provide one approach to model this error explicitly.

Acknowledgements. This study was supported in part by Public Health Service grants MH-37685, MH-25430, MH31302 and AA08401.

References

Andreasen NC, Rice J, Endicott J, Coryell W, Grove WM, Reich T (1987) Familial rates of affective disorder. Arch Gen Psychiatry 44:461–469

Bromet EJ, Dunn LO, Conell MM, Dew MA, Schulberg HC (1986) Long-term reliability of diagnosing lifetime major depression in a communtiy sample. Arch Gen Psychiatry 43:435–440

Cloninger CR, Miller JP, Wette R, Martin RL, Guze SB (1979) The evaluation of diagnostic concordance in follow-up studies: I. A general model of causal analysis and a methodological critique. J Psychiatr Res 15:85

Cooper JE (1967) Diagnostic change in a longitudinal study of psychiatric patients. Brit J Psychiatry 113:129

Diamond GA, Rozanski A, Forrester JS, Morris D, Pollock BH, Staniloff HM, Berman DS, Swan HJC (1986) A model for assessing the sensitivity and specificity of tests subject to selection bias. J Chronic Dis 39:343–355

Falconer DS (1981) Introduction of Quantitative Genetics Longman, New York

Faravelli C, Poli E (1982) Stability of the diagnosis of primary affective disorder. J Affect Disord 4:35

Feighner JP, Robins E, Guze SB, Woodruff RA, Winokur G (1972) Diagnostic criteria for use in psychiatric research. Arch Gen Psychiatry 26:57

Grove WM, Andreasen NC, McDonald-Scott P, Keller MB, Shapiro RW (1981) Reliability studies of psychiatric diagnosis. Arch Gen Psychiatry 38:408

Henkelman RM, Kay I, Bronskill MJ (1990) Receiver operating characteristic (ROC) analysis without truth. Med Decision Making 10:24–29

Hui SL, Walter SD (1980) Estimating the error rates of diagnostic tests. Biometrics 36:167–171

Kendell RE (1974) The stability of psychiatric diagnoses. Br J Psychiatry 124:352

Kendell RE (1975) The Role of Diagnosis in Psychiatry. Blackwell Scientific Publications, London

Kraemer HC (1979) Ramification of a population model for κ as a coefficient of reliability. Psychometrika 44:461

Masserman JH, Carmichael HT (1938) Diagnosis and prognosis in psychiatry. J Ment Sci 84:893

Murphy GE, Woodruff RA, Herjanic M, Fischer JR (1974) Validity of the diagnosis of primary affective disorder. Arch Gen Psychiatry 30:751

Ødegaard O (1966) An official diagnostic classification in actual hospital practice. Acta Psychiatr Scand 42:329

Prusoff BA, Merikangas KR, Weissman MM (1988) Lifetime prevalence and age-of-onset of psychiatric disorders: Recall four years later. J Psychiatr Res 22:107–117

Reich T, James JW, Morris CA (1972) The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. Ann Hum Genet 36:163–184

Rice J, Reich T (1985) Familial analysis of qualitative traits under multifactorial inheritance. Genet Epidemiol 2:301–315

Rice JP, McDonald-Scott P, Endicott J, Coryell W, Grove WM, Keller MB, Altis D (1986) The stability of diagnosis with an application to Bipolar II disorder. Psychiatry Res 19:285–296

Rice JL, Endicott J, Knesevich MA, Rochberg N (1987a) The estimation of diagnostic sensitivity using stability data: An application to major depressive disorder. J Psychiatr Res 21(4):337–345

Rice J, Reich T, Andreasen NC, Endicott J, Van Eerdewegh M, Fishman R, Hirschfeld RMA, Klerman GL (1987b) The familial transmission of bipolar illness. Arch Gen Psychiatry 44:441–447

Rice J, Neuman J, Moldin SO (1991) Methods for the inheritance of qualitative traits. In: Rao CR, Chakraborty R (Eds) Handbook of Statistics. Elsevier Science Publishers, Amsterdam, pp 1–27

Rice JP, Rochberg N, Endicott J, Lavori PW, Miller C (1992) Stability of Psychiatric diagnoses: An application to the affective disorders. Arch Gen Psychiatry 49:824–830

Rice JP, Todorov AA (1993) The stability of diagnosis: Application to phenotype definition. Schizophr Bull, in press

Robins E, Guze SB (1970) Establishement of diagnostic validity in psychiatric illness: Its application to schizophrenia. Am J Psychiatry 126:107–111

Schrott HG, Becher KA, Clarke WR, Lauer RM (1979) The Muscatine hyperlipidemia family study program. In: Sing CF, Skolnick M (Eds) Genetic Analysis of Common Diseases: Predictive Factors in Coronary Disease. Alan R. Liss, New York, p 619

Schulzer M, Anderson DR, Drance SM (1991) Sensitivity and specificity of a diagnostic test determined by repeated observations in the absence of an external standard. J Clin Epidemiol 44:1167–1179

Tsuang MT, Woolsen RF, Winokur G, Crowe RR (1981) Stability of psychiatric diagnosis. Arch Gen Psychiatry 38:535

Uebersax JS, Grove WM (1989) Latent Structure Agreement Analysis. RAND Corporation, Santa Monica, Calif.

Walter SD, Irwing LM (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. J Clin Epidemiol 41:923–937