Psychological Medicine, 1996, **26**, 1105–1110. Copyright © 1996 Cambridge University Press

EDITORIAL

Diagnostic complexity and depression: time to allow for uncertainty[1]

The last 25 years have seen the evolution of an increasingly ordered and structured approach to psychiatric diagnosis, recognized by some (e.g. Andreason, 1995) as dating from publication of the work of Eli Robins and colleagues in St. Louis (e.g. Robins & Guze, 1970). Their call for a more systematic scientific method to be applied to the diagnostic process, through the development of explicit criteria, has been heeded and has provided the foundation to many of the advances in psychiatry that have been subsequently achieved. This objective has been satisfied at intervals by the publication of modified criteria intended to be ever more precisely crafted and to embody consensus views on the formal requirements to meet diagnostic states. In tandem with these changes in criteria, has been the linked development of diagnostic instruments, each designed to represent the diagnostic rules specified within individual or competing schemes.

Inevitably, while many regard this period during which diagnostic rules have become more formally expressed as a necessary scientific step for psychiatry, others (e.g. Snaith, 1987; Blacker & Tsuang, 1992; Van Praag, 1993; Charlton, 1995) have expressed their concerns over aspects of the resulting classifications; for example, Charlton has recently argued for the complete abandonment of the major syndromal categories in favour of a nosology informed through a new discipline of cognitive neuropsychiatry. These current criticisms of the outcome of diagnostic rule development represent the contemporary face of what has been a continuing concern with aspects of this work over the last 25 years. They have naturally arisen from the formal steps needed to develop and refine diagnostic procedures (concerning, for example, their consensus, comprehensiveness, the standardization of items, together with reliability, validity and cross-cultural considerations), the increasing recognition of the importance of taking the 'longitudinal view' for informing opinions on psychiatric status and from the development of 'expert systems' designed to enhance diagnostic accuracy.

The papers published by Kendell during the early part of this period serve as a reminder of some of the central and continuing issues (e.g. Kendell, 1976, 1982; Kendell & Brockington, 1980). His concerns were with the developing complexity of the classificatory schemes, the need to establish a consensus on how the depressions (in particular) should be classified and, given the existence of rival definitions, the need to employ multiple classificatory schemes simultaneously for research purposes. Jablensky *et al.* (1983) and Jablensky (1987) have set out similar views to those of Kendell, also formulating a critique of aspects of the competing diagnostic systems that were perceived, in particular, as having only limited value for the prognostic evaluation of the affective disorders. The opinion again being expressed was that the problems identified by Kendell had not been resolved but had probably been exacerbated; evidence being marshalled to show that the different operational criteria embodied in the various schemes had low concordance. A further issue raised by Kendell, concerned what he called the 'boundary problem', meaning by this the apparent absence of points of discontinuity or rarity between psychiatric syndromes and their appearance of merging into one another in a seamless fashion. The possibility being considered was that if syndromal neighbours could be mapped out, then this would improve the likelihood of specifying their potentially different aetiologies and perhaps responsivity to treatment. The continuing evolution of diagnostic schemes has depended, in part at least, upon attempts to specify more

[1] Address for correspondence: Dr P. G. Surtees, MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Cambridge CB2 2SR.

https://doi.org/10.1017/S0033291700035820 Published online by Cambridge University Press

precisely those 'boundary disputes' arising from the binary approach to diagnosis (see Blacker & Tsuang, 1992).

Application of a diagnostic scheme depends not only upon the rules specified to govern the relationship between diagnostic classes, but on the way in which the clinical phenomena of symptoms and behaviours are assessed and incorporated within the diagnostic decision making process. Failure to ensure that clinical phenomena are recorded in systematic and standardized ways further serves to compound the difficulty of achieving satisfactory levels of agreement between competing schemes. However, the specification of increasingly explicit rules for distinguishing diagnostic categories and clinical phenomena in psychiatry has enabled their inclusion within computer-based reasoning systems. Such 'expert systems' vary considerably in terms of their structural foundations, some operating entirely within a deterministic rule-based environment while others have employed a variety of formalisms to model uncertainty.

Prior to the wide availability of computers, some efforts were made to support the needs of clinicians through the invention of simple technological devices (such as wooden frame or card arrangements) to aid differential diagnosis, for example of blood diseases (e.g. see Nash, 1954; Lipkin & Hardy, 1958). Such approaches have now been replaced by very large medical expert systems designed, for example, to employ descriptive representations, in the case of CADUCEUS (Pople, 1985), of some 750 disorders. Expert systems applications increasingly depend upon probabilistic theory (see Spiegelhalter *et al.* 1993 for a review), but also upon other methods (e.g. fuzzy set theory).

Expert systems have, therefore, been used in many circumstances to assist clinicians in diagnosis. Some of these act both as a memory store and a guide through specific diagnostic rules, and are designed to yield binary classifications. Others, through statistical inference, provide probabilities that represent the likelihood of diagnoses conditional on the profile of presenting symptoms. In these instances, each presenting symptom acts with a certain pre-assigned weight to help either confirm or deny possible diagnoses. Many of the larger diagnostic systems incorporate both knowledge-based and inferential components. The outcomes from these expert systems have been used in a wide variety of diagnostic and treatment contexts (including teaching aids for those not yet familiar with a diagnostic area, by acting as a reference to the knowledge and reasoning of a more experienced clinician). Such systems can also assist those with a large amount of specialist knowledge by suggesting new (perhaps relatively rare) diagnoses that the clinician may not have considered.

In psychiatry, early attempts to employ such systems included DIAGNO (Spitzer & Endicott, 1968), designed to represent all disorders described in the American Psychiatric Association's *Diagnostic and Statistical Manual* (DSM-II; APA, 1968); Pathfinder, designed to assist the distinction of depression into endogenous and non-endogenous subtypes (Feinberg & Carroll, 1983) and Adinfer, designed to incorporate DSM-III-R rules (APA, 1987) based upon an inferential system (Ohayon, 1993). More recently, an advanced diagnostic support system has been developed for clinical psychiatry dependent both upon a rule-based probabilistic reasoning process together with the use of a deterministic approach to aid differential diagnosis (Do Amaral *et al.* 1995). This system has been elaborated to operate both within the DSM-III-R and the International Classification of Diseases (ICD-9) schemes. The system, designed to aid the decision making of clinicians and to support educational endeavours, is quite distinct from computerized versions of interview protocols that have been developed (e.g. Lewis *et al.* 1988) in the form in which knowledge is represented and used to distinguish between the most likely diagnoses. A further recent development of a computerized version of a fully structured diagnostic assessment is that of the Composite International Diagnostic Interview (CIDI-Auto; Peters & Andrews, 1995) based upon the schedule developed by Lee Robins and colleagues in St. Louis (Robins *et al.* 1988). The interview is designed to be self or interviewer administered and to provide (lifetime) and current psychiatric diagnoses informed through the rules included in both the tenth edition of the ICD (WHO, 1993) and DSM-III-R rules. Despite the rigorous methods applied, the overall level of agreement between the CIDI-Auto and clinician diagnoses have been found to be only fair to modest. Details of the

https://doi.org/10.1017/S0033291700035820 Published online by Cambridge University Press

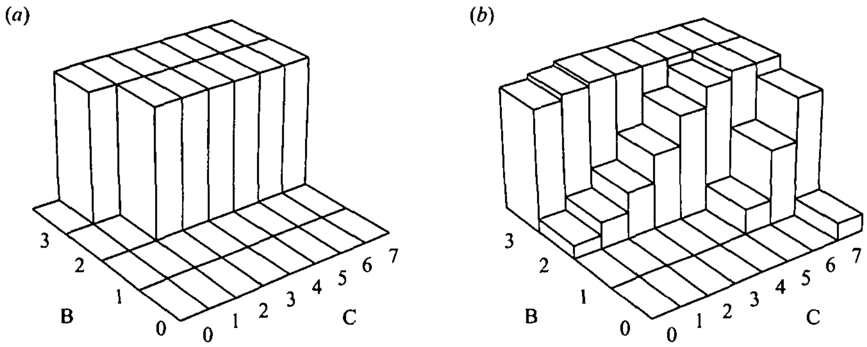

FIG. 1. Illustration (*a*) of the boundaries between the presence and absence of a depressive disorder according to the number of symptoms rated present in groups 'B' and 'C', and according to a probabilistic procedure that provides for the relaxation of the diagnostic boundaries (*b*).

reliability and validity of the CIDI have been reported (Wittchen, 1994; Andrews *et al.* 1995) with recommendations that further procedural validity studies need to be undertaken in association with instruments more widely used by clinicians, particularly the Schedules for Clinical Assessment in Neuropsychiatry (SCAN; Wing *et al.* 1990). Problem areas identified by Wittchen included aspects of how research criteria were operationalized, the use of multiple or long questions and of terms for symptoms that had pivotal diagnostic importance (e.g. depressed mood).

Allied to these objectives to improve reliability and validity have been recent attempts to advance the process of identifying those most likely to meet current diagnostic criteria through the creation of new screening scales. Kessler and colleagues have developed tentative short-form measures of a number of DSM-III-R conditions. These were based upon data obtained from the University of Michigan version of the CIDI as applied to the National Comorbidity Survey (Kessler *et al.* 1994). These scales have been designed to provide, through predicted probabilities, an optimal designation of meeting specific case criteria, and should allow more efficient case detection strategies to be employed than those based upon procedures less well linked to the morbidity criteria.

These issues collectively underscore the very real difficulties associated with aspects of the diagnostic decision making process in psychiatry. While diagnostic uncertainty may not be well tolerated in clinical settings, for some research endeavours, particularly those that are population based, allowance for uncertainty may better mimic reality. The practice of allocating diagnoses in a binary fashion suggests the complete absence of doubt, however much such doubt either pervaded the rating process or contributed to the setting of arbitrary syndromal boundaries.

Modelling uncertainty has been used as a basis upon which to try to clarify some of the problems in diagnostic decision making. We now ask whether similar ideas can be incorporated into prevalence and risk estimation. Typically, a given diagnostic scheme will group those symptoms that have special relevance for particular disorders. The diagnostic process then depends (at least in part) upon the number and pattern of symptoms considered present within these pre-defined groups. Where certain thresholds are exceeded, and other conditions are met, then the requirements for case-status to be achieved are fulfilled; and conversely, case status is not satisfied if (at least) one of these thresholds is not reached. These thresholds can be seen as defining clear diagnostic boundaries, with diagnosis depending upon whether the presenting symptoms do, or do not, exceed them.

Fig. 1*a* illustrates this process, with diagnosis being determined according to the number of symptoms in just two groups 'B' and 'C'; one of these ('B') includes only three, while a second group ('C') includes a possible total of seven symptoms. Each combination of symptoms is represented by a vertical bar with height (0 or 1) corresponding to case status. In this instance, the minimum requirement for disorder is fulfilled with a total of just four symptoms; if at least two symptoms are rated from within each group, or alternatively if three symptoms are present in group

https://doi.org/10.1017/S0033291700035820 Published online by Cambridge University Press

'B' and only one in group 'C', effectively setting the boundary as shown. An individual outside these boundaries does not meet the case criteria, while individuals meeting the criteria are shown as columns in the figure. The placement of these 'boundaries' enables a population to be precisely classified, permitting for example, prevalence and odds ratio estimates to be easily obtained.

However, an inevitable consequence of such arbitrary but precise distinctions is to classify those individuals whose presenting symptoms fall just either side of such a boundary into very different groups. This consideration is further highlighted in those circumstances where multiple diagnostic schemes are applied to the same body of clinical data, to establish levels of agreement between the schemes. Clearly, it would be a desirable goal if the outcome measure (fulfilling diagnostic criteria) and hence inference from it, could be shown to be stable according to alternate diagnostic schemes. To consider this issue, we are investigating a natural extension to this type of diagnostic system that includes provision for the relaxation of diagnostic boundaries through allowing for a degree of uncertainty in diagnosis. The system can be represented by a simple probabilistic structure along with a set of constraints intended to mimic those specified within a particular diagnostic scheme; in this instance based upon a symptom group structure. If a model is specified in terms of the probabilities of displaying individual symptoms, conditional on disorder status, then these can be constrained by the data in such a way so as to retain the symptom group structure of the scheme. The application of Bayes' theorem then permits outcome to be expressed in terms of the probability of disorder, given any individual presenting symptom profile.

As an example, consider the ICD-10, Diagnostic Criteria for Research (DCR) classification of a depressive episode. Other than the need to satisfy the general criteria for a depressive episode (F32.0), this requires at least two of the three symptoms from within one group (Group 'B') and one or more symptoms from among a second group (of seven, group 'C') to give a total symptom count of at least four symptoms. The formal structure imposed by these criteria is represented in Fig. 1*a*. However, these principal rules are further elaborated through setting other constraints (boundaries) depending upon whether the depressive episode is 'mild', 'moderate' or 'severe'.

The procedure that we are investigating allocates the probability of a diagnosis (of depression) within the range zero (representing the certain absence of depression) to one (representing the certain presence of depression), as an alternative to a categorical outcome. Probabilities between zero and one represent uncertainty; for example a probability of 0·5 would mean that presence and absence of depression are considered equally probable. This provides for a natural gradation of probabilities over the complete symptom ranges, ensuring that those satisfying the extremes of the diagnostic rules will be classified with near certainty (as outcome near zero or one), whereas for those nearer the boundaries, there will be some element of doubt and hence they will be assigned probabilities nearer the mid-range.

These methods have now been investigated through application to survey data obtained from the revised version of the Clinical Interview Schedule (CIS-R; Lewis *et al.* 1992). This was used in the Office of Population Censuses and Surveys (OPCS) national household survey of psychiatric morbidity in Great Britain (Meltzer *et al.* 1995), based upon a probability sample of almost 10000 individuals aged 16–64 years of age. These data were used to establish ICD-10 categories of disorder and then to provide the basis for probabilistic estimates of ICD-10 depressive disorders. Fig. 1*b* reveals the consequence of allowing for uncertainty in the diagnosis of depression in this sample, and displays how the probability of satisfying the requirements of ICD-10 for a depressive episode vary according to the extent to which the criteria are formally met. The vertical bar heights, shown in the Figure, represent the probability on a scale of zero to one of meeting depression case criteria. It should be noted how the patterns shown in the Figures differ. In particular, under a probabilistic representation certain combinations of symptoms that would not place individuals within the formal diagnostic boundaries are assigned a relatively high probability of meeting case criteria, while other combinations (that meet the formal diagnostic criteria) are accorded relatively low probabilities. As an example, a minimum requirement for an ICD-10 depressive episode of just two symptoms in each group, has been assigned a relatively low probability of about 0·35, (see Fig. 1*b*). This is not to say that this is an unlikely combination in an individual with disorder, more that there

https://doi.org/10.1017/S0033291700035820 Published online by Cambridge University Press

are sufficient symptoms occurring by chance in the healthy population for this combination to be considered quite plausible in a healthy individual. The reverse is true when symptom combinations considered not to meet the formal requirements for depressive disorder are assigned relatively large probabilities. It is an intriguing question as to whether individuals so classified may be found to differ in terms of risk and particularly whether they differ in terms of their recognition and receipt of care (e.g. by general practitioners).

Our initial results are encouraging and suggest that by using these probabilities as an outcome measure it will be possible to enhance the information gained from the system both in terms of risk and of prevalence estimation. For large scale studies, such properties should be broadly coincident as the system is designed to match the dimensions inherent in reaching the categorical decision. In small scale studies, the somewhat arbitrary delineation between presence and absence of a disorder could distort prevalence comparisons between risk groups. A probabilistic approach should smooth out unwanted large fluctuations in estimates, caused solely by the chance distribution of symptom profiles around diagnostic boundaries, and therefore, provide more reliable comparisons between risk groups. Odds ratios and prevalence estimates could be obtained directly from the probabilities. In addition, since a strict zero/one boundary would no longer be imposed, slight changes in the system would cause less disruption to the outcome measure. The probabilistic procedure outlined is not an expert system, but a routine procedure that could be applied to diagnostic criteria to relax boundaries and enrich risk and morbidity estimation. It is possible that the strategy could be used to reduce some of the shortcomings inherent in employing a categorical diagnostic system as a basis for gaining insights into public health psychiatry.

P. G. SURTEES, N. W. J. WAINWRIGHT AND W. R. GILKS

REFERENCES

American Psychiatric Association (1968). *Diagnostic and Statistical Manual of Mental Disorders* (*2nd edn*). APA: Washington, DC.

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (*3rd edn – Revised*). APA: Washington, DC.

Andreasen, N. C. (1995). The validation of psychiatric diagnosis – new models and approaches. *American Journal of Psychiatry* **152**, 161–162.

Andrews, G., Peters, L., Guzman, A. M. & Bird, K. (1995). A comparison of 2 structured diagnostic interviews – CIDI and SCAN. *Australian and New Zealand Journal of Psychiatry* **29**, 124–132.

Blacker, D. & Tsuang, M. T. (1992). Contested boundaries of bipolar disorder and the limits of categorical diagnosis in psychiatry. *American Journal of Psychiatry* **149**, 1473–1483.

Charlton, B. G. (1995). Cognitive neuropsychiatry and the future of diagnosis – a 'PC' model of the mind. *British Journal of Psychiatry* **167**, 149–153.

Do Amaral, M. B., Satomura, Y., Honda, M. & Sato, T. A. (1995). A psychiatric diagnostic system integrating probabilistic and categorical reasoning. *Methods of Information in Medicine* **34**, 232–243.

Feinberg, M. & Carroll, B. J. (1983). Separation of subtypes of depression using discriminant analysis II: separation of bipolar endogenous depression from nonendogenous 'neurotic' depression. *Journal of Affective Disorders* **5**, 129–139.

Jablensky, A. (1987). Prediction of the course and outcome of depression. *Psychological Medicine* **17**, 1–9.

Jablensky, A., Sartorius, N., Hirschfeld, R. & Pardes, H. (1983). Diagnosis and classification of mental-disorders and alcohol-related and drug-related problems – a research agenda for the 1980s. *Psychological Medicine* **13**, 907–921.

Kendell, R. E. (1976). The classification of depressions: a review of contemporary confusion. *British Journal of Psychiatry* **129**, 15–28.

Kendell, R. E. (1982). The choice of diagnostic criteria for biological research. *Archives of General Psychiatry* **39**, 1334–1339.

Kendell, R. E. & Brockington, I. F. (1980). The identification of disease entities and the relationship between schizophrenia and the affective psychoses. *British Journal of Psychiatry* **137**, 324–331.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H.-U. & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**, 8–19.

Lewis, G., Pelosi, A. J., Glover, E., Wilkinson, G., Stansfeld, S. A., Williams, P. & Shepherd, M. (1988). The development of a computerized assessment for minor psychiatric disorder. *Psychological Medicine* **18**, 737–745.

Lewis, G., Pelosi, A. J., Araya, R. & Dunn, G. (1992). Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychological Medicine* **22**, 465–486.

Lipkin, M. L. & Hardy, J. D. (1958). Mechanical correlation of data in differential diagnosis of hematological diseases. *Journal of the American Medical Association* **166**, 113–125.

Meltzer, H., Gill, B., Petticrew, M. & Hinds, K. (1995). *OPCS Surveys of Psychiatric Morbidity in Great Britain. Report 1. The Prevalence of Psychiatric Morbidity among Adults living in Private Households.* HMSO: London.

Nash, F. A. (1954). Differential diagnosis: an apparatus to assist the logical faculties. *Lancet* i, 874–875.

Ohayon, M. M. (1993). Utilisation des systèmes experts en psychiatrie. *Revue Canadienne de Psychiatrie* **38**, 203–211.

Peters, L. & Andrews, G. (1995). Procedural validity of the computerized version of the Composite International Diagnostic Interview (CIDI-Auto) in the anxiety disorders. *Psychological Medicine* **25**, 1269–1280.

Pople, H. E. (1985). Evolution of an expert system: from INTERNIST to CADUCEUS. In *Artificial Intelligence in Medicine*, (ed. I. de Lotto and M. Stefanelli), pp. 179–208. North Holland: Amsterdam.

Robins, E. & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. *American Journal of Psychiatry* **126**, 983–987.

Robins, L. N., Wing, J., Wittchen, H.-U., Helzer, J. E., Babor, T. F., Burke, J., Farmer, A., Jablenski, A., Pickens, R., Regier, D. A.,

https://doi.org/10.1017/S0033291700035820 Published online by Cambridge University Press

Sartorius, N. & Towle, L. H. (1988). The Composite International Diagnostic Interview: an epidemiological instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* **45**, 1069–1077.

Snaith, R. P. (1987). The concepts of mild depression. *British Journal of Psychiatry* **150**, 387–393.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. & Cowell, R. G. (1993). Bayesian-analysis in expert systems. *Statistical Science* **8**, 219–247.

Spitzer, R. L. & Endicott, J. (1968). DIAGNO: a computer program for psychiatric diagnosis utilizing the differential diagnosis procedure. *Archives of General Psychiatry* **18**, 746–756.

Van Praag, H. M. (1993). Diagnosis, the rate-limiting factor of biological depression research. *Neuropsychobiology* **28**, 197–206.

Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R., Jablenski, A., Regier, D. & Sartorius, N. (1990). SCAN – Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry* **47**, 589–593.

Wittchen, H.-U. (1994). Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research* **28**, 57–84.

World Health Organization (1993). *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. WHO: Geneva.

https://doi.org/10.1017/S0033291700035820 Published online by Cambridge University Press