

EDITORIAL

Methodological issues in psychiatric case-identification¹

The use of the term 'case' in clinical psychiatric research implies that the investigator wishes to identify the presence or absence of some clinically relevant disorder or disorders in a human population; for example, to distinguish X from non-X, Y from non-Y, and X from Y. The main reasons for making such identifications are to clarify or refine the descriptive features of the disorder, to calculate an incidence or a prevalence rate, and to test hypotheses concerning, for example, causes, treatments or outcomes. Many of the ideas used in the research are derived from clinical practice and it is usually hoped that the results will, in turn, be found clinically useful.

The simplest technique of case-identification, therefore, is for a well-trained psychiatrist to interview all the members of the population under review and to 'make a diagnosis'. A substantial proportion of the research literature has been, and still is, based upon this technique. Its advantages are that it is simple, it has achieved (particularly in Scandinavian epidemiological studies of the functional psychoses) some remarkably replicable results, and there is no question as to its clinical reference. The disadvantages are also considerable. The well-known differences between the way 'schizophrenia' is diagnosed in various parts of the world, even by experienced psychiatrists, illustrate the major problem (Cooper *et al.* 1972; WHO, 1973). Less severe disorders give rise to greater differences, even among the Scandinavian investigators, and it is difficult to say how far these are due to different methods of reaching a diagnosis rather than to variations in true incidence or prevalence.

The central methodological issue in case-identification is how, in the absence of simple physiological or biochemical indices, to construct techniques which will ensure a useful degree of comparability between studies. This is not just a matter of achieving 'reliability' between a few close collaborators, but the much more difficult task of ensuring that teams working independently, perhaps in different countries and using different languages, can replicate each others' studies, thus adding to the stock of knowledge available to investigators all over the world.

The problems involved in achieving this kind of comparability can conveniently be reviewed by considering the information collected by a psychiatrist in order to make a diagnosis. This includes a knowledge of which out of a range of possible symptoms and signs are present, how severe (i.e. intense and continuous) they are, their time relationships, the extent to which they cluster together to form syndromes, and the relative diagnostic weighting of each syndrome. Symptom episodes in the recent, and perhaps the more distant, past, and the presence or absence of pathological indices and possible causative factors, are also considered. Kendell (1973) showed that a decision approximating to the final diagnosis could be reached quite quickly but this does not mean that the process can easily be standardized. It is not, of course, implied that any technique can achieve perfect standardization. Nor is it necessary, in most investigations, to try to standardize everything. Specific scales can be constructed for limited purposes (e.g. to measure improvement following treatment for a phobia of flying), or one of the useful schedules available for measuring change in conditions such as depression (Beck *et al.* 1961; Hamilton, 1960; Zung, 1965) or general anxiety (Taylor, 1953) can be used. It should be recalled, however, that the criteria for entry into the study also need specification and that this always includes an element of differential diagnosis.

Several factors which could be used as indirect measures of the severity of the effects of disorder are best kept separate, as far as possible, from the diagnostic elements mentioned so far. These include decline in social performance or status, degree of personal distress or dissatisfaction, and amount of contact with medical or social services. As we shall see, confusion between such factors

¹ Address for correspondence: Professor J. K. Wing, MRC Social Psychiatry Unit, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF.

and symptoms makes it difficult to test hypotheses about aetiology, treatment or course. The advantage of introducing standardization into the diagnostic process is that each element can be specified separately and various combinations of elements can then be investigated.

For many investigators, standardization begins and ends with symptoms. The simplest technique is to construct a set of questions, each representing a symptom, which can be answered by the respondent filling in a form, or giving his or her answers to an interviewer. The latter situation allows other observations to be made and provides for a limited degree of clarification if the subject asks what a question means. In general, however, the question is left for the respondent to interpret. One of the first such questionnaires was the Neuropsychiatric Screening Adjunct (Stouffer *et al.* 1950). Scores on the NSA differentiated quite well between groups of American soldiers and neurotic in-patients and between individuals who were judged by psychiatrists to be suitable or not suitable for army service. The most useful sections were concerned with so-called 'psychosomatic' items (mostly autonomic symptoms, insomnia and nervousness) and 'personal adjustment' (including depression, self-pity, worries and self-confidence).

Many such scales have been developed since, two early and much-used examples being the Cornell Medical Index (Brodman *et al.* 1952) and the Health Opinion Survey (Macmillan, 1957). Similar (though unfortunately not identical) instruments were used in the Stirling County and Midtown Manhattan projects (Leighton *et al.* 1963; Srole *et al.* 1962). The investigators adopted a uni-dimensional concept of mental disorder – an amalgam of symptoms and disability which was regarded as being quantitatively reactive to the degree of social stress. 'Social stress' was conceptualized in somewhat different ways in the two studies but it included adverse life events, disadvantage, low prestige and difficulty in maintaining 'a place in the social system.' The opportunities for contamination between social and clinical measures are evident.

Many authors (e.g. Dohrenwend & Dohrenwend, 1969) have pointed out that disorders defined through the use of such questionnaires cannot directly be related to the disorders diagnosed by psychiatrists in hospital practice. The fact that the latter tend to have high scores on the questionnaires does not give much information as to the nature of the condition measured. Tyhurst (1957) argued that transient stressful situations could produce transient psychological responses which need not be regarded as abnormal. By the same token, longer-term difficulties might be expected to produce more chronic distress and dissatisfaction without a psychiatric diagnosis being called for.

Recent work with the General Health Questionnaire (Goldberg, 1972) has confirmed the value of a simple self-rating instrument for detecting people in a state of personal distress and dissatisfaction. Such people are often socially disadvantaged or physically impaired. The GHQ can be used to monitor the effects of various kinds of intervention, such as extra counselling by a general practitioner (Johnstone & Goldberg, 1976). A simple cut-off score allows 'cases' to be differentiated, for screening purposes, from 'non-cases', but no suggestion is made that disorders thus defined are equivalent to any particular psychiatric diagnosis.

More comprehensive checklists of symptoms have been devised, with the twin characteristics of ease of administration and lack of control over the responses (Lorr *et al.* 1963; Overall & Gorham, 1962; Wittenborn, 1955). Foulds (1965) pointed out that the Minnesota Multiphasic Personality Inventory (MMPI) (Hathaway & McKinley, 1951) and other schedules contain items representing psychological symptoms mixed with others representing personality traits. He could have added that yet other items represent attitudes. His own checklists have the merits of keeping symptoms and traits separate and of basing the symptom scales firmly on psychiatric experience. The latter is not true of most of the other scales, perhaps because of Lorr's judgement that 'in much of American psychiatry, formal diagnosis is actually ignored as relatively unimportant and outmoded, or disparaged as nondynamic and useless' (Lorr, 1966).

The situation in the USA has changed a good deal since Lorr made this statement and it could not, even at that time, have been applied with the same force to European diagnostic practices. Foulds (1965), who reviewed the evidence concerning the reliability with which psychiatrists classified mental disorders, pointed out that there was very little evidence in the few, mostly inadequate, studies then published, to justify the wholesale condemnation often made, whereas there were studies indicating

a useful degree of reliability (Kreitman, 1961). Further work has shown how psychiatric nosology, particularly in the relevant sections of the International Classification of Diseases, can be improved by providing more precise definitions for use by clinical psychiatrists (Feighner *et al.* 1972; Scharfetter, 1971; Spitzer *et al.* 1978; WHO, 1974). This practical line of development has been paralleled by attempts to standardize more of the elements of the diagnostic process in order to be able to make the diagnoses used in scientific research more comparable.

Again, these efforts had to begin with symptoms. The Present State Examination (PSE), for example, is based on a glossary of differential definitions derived from the clinical practice taught in one particular school of British psychiatry. Apart from a few items, social elements are excluded from the definitions. Rules are laid down for rating the presence or absence of symptoms during the month before interview and, if present, their intensity and stability. The interviewer is free to ask further questions, beyond those laid down in the schedule, in order to elicit sufficient detail from the subject to allow an adequate decision to be made. Ratings on the 140 items allow the calculation of sub-scores and a total score and clustering in order to provide a syndrome profile which can be used as a descriptive summary of the clinical condition (Wing *et al.* 1967, 1974; Wing & Sturt, 1978).

The PSE is intended to cover the whole range of symptoms found in the functional neuroses and psychoses. The fact that it is based on the month before interview and excludes possible aetiological factors means that certain diagnoses (e.g. personality disorders, mental retardation and alcoholic hallucinosis) cannot be derived from the PSE alone. An Aetiology Schedule can be used to help codify the presence or absence of possible causal factors. During acute episodes of disorder sufficient symptoms are usually present to allow a clinical diagnosis to be made, but in a small proportion of cases additional information is required about symptoms present more than a month previously (Simon *et al.* 1971; Sartorius *et al.* 1970). This restriction on clinical diagnosis applies *ipso facto* to any technique designed to simulate it. Extra data are therefore extracted from contemporary case-records on a Syndrome Check List. Reliability is quite reasonable but obviously depends on the quality of the records (Wing *et al.* 1974, p. 106). If good records are not available, it is possible to ask informants, but memory, even for acute and severe episodes, may well be unreliable.

The other major element needed to simulate a diagnosis is a set of rules, laid down sufficiently precisely to obviate the necessity for further subjective interpretation, for differentiating between clinical classes. This is provided by the CATEGO computer program (Wing *et al.* 1974; Wing & Sturt, 1978). The procedures used are clinical throughout. Obviously, no higher claim is made for the validity of the classes produced than for the clinical diagnostic groups they simulate, and the system is explicitly not intended to be used clinically as a substitute for a diagnosis (Wing *et al.* 1974, p. 136). Many clinicians do find it useful to know the outcome of a standard procedure but the use they make of it still depends on their clinical judgement. It is also found useful in teaching. But the chief value is in research, since it allows comparability in the testing of hypotheses. A useful degree of concordance between CATEGO classes and broad clinical diagnoses has been observed, and studies of the small proportion of discrepancies has led to suggestions for improving both types of classification (Scharfetter *et al.* 1976; WHO, 1973, p. 243; Wing & Nixon, 1975; Wing *et al.* 1977a).

This system is firmly grounded in the clinical experience of psychiatrists examining patients with acute disorders. A further element of incomparability needs to be eliminated, as far as possible, before it can be used with samples of people from a general population. This is the threshold at which it can be said that sufficient key symptoms are present to allow the recognition of a disorder. The question rarely arises when examining recently admitted in-patients, since most are suffering from severe disorders. Recalling the earlier discussion about case-identification, it is clear that the procedure governing the determination of a threshold must be clinical, specifiable and replicable.

One such technique is the Index of Definition (Wing *et al.* 1977b, 1978; Wing & Sturt, 1978). The solution adopted is to lay down, on the basis of clinical experience, rules determining a number of levels of certainty that sufficient PSE symptoms are present to allow recognition of a 'disorder'. The threshold is, of course, defined separately for each major type of symptom, since no single criterion could be applied to all the possible combinations of symptoms found in a population survey. The extent of the problem can be understood if it is recalled that a total PSE score of 1 can indicate the

presence of any one of more than 120 symptoms, ranging from a moderate degree of muscular tension to a first-rank symptom of schizophrenia.

The Index of Definition allows 8 different levels of certainty as to the presence of a recognizable functional disorder. Level 5 indicates that certain minimal criteria have been met and levels 6–8 indicate more definite disorders. Disorders at or above the threshold can be classified using the CATEGO program. Thus, the PSE-ID-CATEGO system can be used to classify the conditions found in people admitted to hospital, referred to out-patient clinics, or interviewed in general population surveys. The Syndrome Check List can be used to provide a degree of standardization of the symptoms present in previous episodes of disorder in subjects interviewed in population surveys but detailed case-records are rarely available and the subject's memory is an unreliable guide. The Index of Definition cannot be applied. Estimates of incidence or descriptions of 'disorders' not present at the time of examination must therefore be regarded as very approximate. This is, of course, even more true of any less specified judgement about past episodes.

One test of the distinction between 'cases' and 'non-cases' as defined by the Index of Definition is to compare it with an independent clinical judgement. Two published series allow such a comparison. The first was the 'second Camberwell survey' in which a team of sociologists from Bedford College interviewed a sample of 237 women aged 18–65 (Brown *et al.* 1975). Three psychiatrists (J. P. Leff, S. A. Mann and J. K. Wing) interviewed 95 of these women, using the ninth edition of the PSE and rated tapes of the Bedford interviews with a further 28. They made global clinical judgements of 'caseness' and compared these with the distinction based on the subsequently developed Index of Definition. The concordance (level 5 and above) was 90% (Wing *et al.* 1978). The second survey was carried out by Orley, who used the PSE to interview the adult population of two small Ugandan villages. The concordance with the Index of Definition was 91% (Orley & Wing, 1979).

The global clinical judgement in these studies is therefore likely to have been based mainly on the presence or absence of key symptoms. The non-medical interviewers in the second Camberwell survey were trained by members of the MRC Social Psychiatry Unit to use the short form of the PSE (the first 40 of the 140 items) and were able to do so with fair reliability (Wing *et al.* 1977*b*). The training technique used was to provide experience with rating severe disorders in in-patients (see also Cooper *et al.* 1977). Whether the criteria used by such interviewers are likely to 'drift' after a period of time without further training has yet to be established. The Index of Definition and the CATEGO program can be applied to PSE ratings with absolute reliability. The system can therefore be used wherever teams of trained investigators are available. (Problems of translation are discussed elsewhere: Orley & Wing, 1979; WHO, 1973.) The same cannot be said of a global judgement. Members of a team often come to achieve very close agreement on such judgements but other teams may reasonably adopt different, though equally reliable, criteria. Moreover, the threshold may be affected by the setting in which the interviews are undertaken. Urwin & Gibbons (1979) found that level 5 (the threshold level) of the Index of Definition could be divided into two roughly equal sub-levels, the lower one (5*a*) approximating to what they regarded, clinically, as 'not a case', the upper one (5*b*) approximating to their judgement of a 'case'. In studies of a population in North Uist, conducted by the Bedford College team (Brown *et al.* 1977), we have suggested that the psychiatrist adopted a somewhat similar threshold (i.e. somewhere in the middle of level 5) to that used by the Southampton investigators (R. Prudo, unpublished; Brown & Harris, 1978, p. 580). These differences in private clinical judgement can only be made manifest in comparison with the more detailed and public standards of a more standardized system.

The same must be true, *a fortiori*, when comparing 'cases' detected in a general population with severe depressive or other disorders in in-patients. It is only when this problem of methodological comparability is solved that the clinical nature of 'psychiatric disorders' found in the general population can be investigated. The familiar method of working from the known to the unknown can then be used (Wing *et al.* 1978).

One further common misunderstanding of standardized techniques needs to be considered. This is that standardization is unnecessarily arbitrary or rigid. It is true that *all* standardization requires a degree of rigidity, in the sense that the techniques can only be used if the rules are adhered to.

However, the advantages of the techniques stem from these restrictions. The more precisely and communicably the limits are laid down, the greater is the value of the technique to research workers, who can then criticize and improve it. The criticisms, however, must be specified in just as precise and communicable a manner.

We return now to the starting point of this review. All the techniques discussed are 'clinical', in the sense that they are more or less based on the experience of clinical psychiatrists and are intended, in the longer or shorter run, to be of use to people with psychiatric disorders. There is no point in using a sledgehammer to crack a nut. The simplest technique that will serve the purposes of the study is the most efficient one to use. But many of the problems of concern to contemporary psychiatrists are highly complex; for example (a) the nature of the relationship between the severe affective disorders observed in in-patients and the less severe conditions found in some members of general population samples, and (b) the comparison of disorders found in samples of people living under different social and cultural conditions. The more complex the theories under test, the more important it is to be sure that the techniques used are open to scrutiny by the scientific community.

J. K. WING

REFERENCES

- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry* 4, 561–571.
- Brodman, K., Erdman, A. J. & Lorge, I. (1952). The Cornell Medical Index Health Questionnaire: the evaluation of emotional disturbance. *Journal of Clinical Psychology* 8, 119–124.
- Brown, G. W. & Harris, T. (1978). Social origins of depression: a reply. *Psychological Medicine* 8, 577–588.
- Brown, G. W., Bhrolchain, M. & Harris, T. (1975). Social class and psychiatric disturbance among women in an urban population. *Sociology* 9, 225–254.
- Brown, G. W., Davidson, S., Harris, T., Maclean, U., Pollock, S. & Prudo, R. (1977). Psychiatric disorder in London and North Uist. *Social Science and Medicine* 11, 367–377.
- Cooper, J. E., Kendell, R. E., Gurland, B. J., Sharpe, L., Copeland, J. R. M. & Simon, R. (1972) *Psychiatric Diagnosis in New York and London*. Maudsley Monograph no. 20. Oxford University Press: London.
- Cooper, J. E., Copeland, J. R. M., Brown, G. W., Harris, T. & Gourlay, A. J. (1977). Further studies on interviewer training and inter-rater reliability of the PSE. *Psychological Medicine* 7, 517–523.
- Dohrenwend, B. P. & Dohrenwend, B. S. (1969). *Social Status and Psychological Disorder: A Causal Enquiry*. Wiley: New York.
- Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G. & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* 26, 57–63.
- Foulds, G. A. (1965). *Personality and Personal Illness*. Tavistock: London.
- Goldberg, D. P. (1972). *The Detection of Psychiatric Illness by Questionnaire*. Oxford University Press: London.
- Hamilton, M. (1960). A rating scale for depression. *Journal of Neurosurgery and Psychiatry* 23, 56–62.
- Hathaway, S. R. & McKinley, J. C. (1951). *The Minnesota Multiphasic Personality Inventory Manual (Revised)*. Psychological Corporation: New York.
- Johnstone, A. & Goldberg, D. (1976). Psychiatric screening in general practice: A controlled trial. *Lancet* i, 605–608.
- Kendell, R. E. (1973). Psychiatric diagnoses: a study of how they are made. *British Journal of Psychiatry* 122, 437–445.
- Kreitman, N. (1961). Reliability of psychiatric diagnosis. *Journal of Mental Science* 107, 876.
- Leighton, D. C., Harding, J. S., Macklin, D. B., Macmillan, A. M. & Leighton, A. H. (1963). *The Character of Danger: Psychiatric Symptoms in Selected Communities*. Basic Books: New York.
- Lorr, M. (ed.) (1966). *Explorations in Typing Psychotics*. Pergamon: New York.
- Lorr, M., Klett, C. J., McNair, D. M. & Lasky, J. (1963). *Inpatient Multidimensional Psychiatric Scale (Manual)*. Consulting Psychologists Press: Palo Alto.
- Macmillan, A. M. (1957). The health opinion survey: technique for estimating prevalence of psychoneurotic and related types of disorder in communities. *Psychological Reports* 3, 325–329.
- Orley, J. & Wing, J. K. (1979). Psychiatric disorders in two African villages. *Archives of General Psychiatry* 36, 513–520.
- Overall, J. E. & Gorham, D. R. (1962). The brief psychiatric rating scale. *Psychological Reports* 10, 799–812.
- Sartorius, N., Brooke, E. & Lin, T. Y. (1970). Reliability of psychiatric assessment in international research. *Psychiatric Epidemiology* (ed. E. H. Hare and J. K. Wing), pp. 138–148. Oxford University Press: London.
- Scharfetter, C. (1971). *Das AMP-System: Manual zur Dokumentation psychiatrischer Befunde*. Springer-Verlag: Berlin.
- Scharfetter, C., Moerbt, H. & Wing, J. K. (1976). Diagnosis of functional psychoses: comparison of clinical and computerized classification. *Archiv für Psychiatrie und Nervenkrankheiten* 222, 61–67.
- Simon, R. J., Gurland, B. J., Fleiss, J. L. & Sharpe, L. (1971). Impact of a patient history interview on psychiatric diagnosis. *Archives of General Psychiatry* 24, 437–440.
- Spitzer, R. L., Endicott, J. & Robins, E. (1978). Research diagnostic criteria: rationale and reliability. *Archives of General Psychiatry* 35, 773–782.
- Srole, L., Langer, T. S., Michael, S. T., Opler, M. K. & Rennie, T. A. C. (1962). *Mental Health in the Metropolis: The Midtown Manhattan Study*. McGraw-Hill: New York.
- Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., & Clausen, J. A. (1950). *Measurement and Prediction*. Studies in Social Psychology in World War II, Volume iv. Princeton University Press: Princeton.
- Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology* 48, 285–290.
- Tyhurst, J. S. (1957). The role of transition states – including disasters – in mental illness. In *Symposium on Preventive and Social Psychiatry* pp. 149–169. Government Publishing Office: Washington, D.C.
- Urwin, P. & Gibbons, J. L. (1979). Psychiatric diagnosis in

- self-poisoning patients. *Psychological Medicine* 9, 501–507.
- Wing, J. K. & Nixon, J. (1975). Discriminating symptoms in schizophrenia. *Archives of General Psychiatry* 32, 853–859.
- Wing, J. K. & Sturt, E. (1978). *The PSE-ID-CATEGO System: A Supplementary Manual*. Institute of Psychiatry, Mimeo: London.
- Wing, J. K., Birley, J. L. T., Cooper, J. E., Graham, P. & Isaacs A. (1967). Reliability of a procedure for measuring and classifying 'present psychiatric state'. *British Journal of Psychiatry* 113, 499–515.
- Wing, J. K., Cooper J. E. & Sartorius, N. (1974). *Measurement and Classification of Psychiatric Symptoms*. Cambridge University Press: London.
- Wing, J. K., Nixon, J., von Cranach, M. & Strauss, A. (1977a). Further developments of the PSE and CATEGO system. *Archiv für Psychiatrie und Nervenkrankheiten* 224, 151–160.
- Wing J. K., Nixon, J. M., Mann, S. A. & Leff, J. P. (1977b). Reliability of the PSE (ninth edition) used in a population survey. *Psychological Medicine* 7, 505–516.
- Wing, J. K., Mann, S. A., Leff, J. P. & Nixon, J. N. (1978). The concept of a 'case' in psychiatric population surveys. *Psychological Medicine* 8, 203–217.
- Wittenborn, J. R. (1955). *Wittenborn Psychiatric Rating Scales*. Psychological Corporation: New York.
- World Health Organization (1973). *Schizophrenia: Report of an International Pilot Study*. WHO: Geneva.
- World Health Organization (1974). *Glossary of Mental Disorders and Guide to their Classification*. WHO: Geneva.
- Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry* 12, 63–70.