Available online at www.sciencedirect.com
SCIENCE DIRECT®
Schizophrenia Research xx (2006) xxx–xxx
SCHIZOPHRENIA RESEARCH
www.elsevier.com/locate/schres

Diagnosing delusions: A review of inter-rater reliability

Vaughan Bell [1], Peter W. Halligan *, Hadyn D. Ellis

School of Psychology, Cardiff University, Park Place, Cardiff, CF10 3YG, UK

Received 3 May 2006; received in revised form 20 June 2006; accepted 20 June 2006

Abstract

Although several studies have examined the reliability of diagnosing delusions there is no comprehensive review of the literature. Therefore, the reliability of diagnosing 'delusions in general' and the subcategory of 'bizarre delusions' was reviewed, including both structured interview and standardized instrument methods. The literature suggests that delusions in general can be diagnosed reliably with both structured interview and standardized instruments. However, bizarre delusions are not reliably diagnosed by either, suggesting that this concept may have little clinical validity. Nevertheless, many of the studies reviewed are poorly designed or subject to significant confounds. Criteria are suggested for adequate future studies.

Keywords: Delusion; Diagnosis; Reliability; Psychosis; Kappa

1. Introduction

The diagnostic criteria for delusions have been widely and deservedly criticised for being conceptually incoherent and subject to significant counter-examples (David, 1999; Leeser and O'Donohue, 1999; Spitzer, 1990), despite the fact that the concept of a core psychopathological feature indicative of a substantial break with reality continues to hold widespread clinical acceptance (Bell et al., 2006). Notwithstanding issues regarding the nature of delusions, the practical use of any clinical diagnostic construct ultimately depends on its reliability and provision for prognosis (the course or consequences of a condition, including recovery). It is notable, however, that a comprehensive review of this topic has yet to be published — an omission that this paper directly addresses.

The attribution of the subcategory of 'bizarre delusions' has particular importance for the clinical diagnosis of schizophrenia and can be used as a DSM exclusion criteria for delusional disorder, although this is another area where no current overview exists (despite some previous studies which have suggested poor reliability; e.g. Mojtabai and Nicholson, 1995). No formal criteria are provided in the DSM-IV (American Psychiatric Association [APA], 1994) for defining the bizarreness of delusions, although previous editions have considered a bizarre delusion to be "patently absurd with no possible basis in fact" (DSM-III; APA, 1980); or that it "involves a phenomenon that the person's culture would regard as totally implausible" (DSM-III-R; APA, 1987).

While much clinical diagnosis still relies on clinical interview, a number of structured methods have also emerged to detect and rate the presence of delusions. It is not clear how these different methods compare. The

* Corresponding author. Tel.: +44 29 208 76911; fax: +44 29 2087 4858.
E-mail address: HalliganPW@cardiff.ac.uk (P.W. Halligan).
[1] Institute of Psychiatry, King's College London, Psychology Dept Box 78, De Crespigny Park, London, SE5 8AF, UK.

0920-9964/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.schres.2006.06.025

aim of this paper is briefly to review the reliability of the diagnosis for both delusions in general, and 'bizarre delusions' in particular, and to determine the current status of reliability using these different classifications and assessment methods.

2. Method

Relevant papers were collated from the PubMed and MEDLINE database searches using the search term "delusion* AND reliability". Owing to the fact that reliability findings are often reported incidentally to a study's main result and so would not necessarily show up in a database search, the original papers reference lists were used as a basis for finding further papers that reported reliability data (akin to a 'snowball sampling' method). All papers reporting an inter-rater reliability result for the diagnosis or detection of delusions, and/or bizarre delusions, in a general adult sample were included. Results that were only from specific populations (e.g. patients with dementia, patients with intellectual impairments) were excluded, as were results from broader diagnostic categories that did not specifically describe delusional phenomena (e.g. Schneiderian first-rank symptoms).

3. Results

3.1. Diagnosing delusions

The studies summarised in Table 1 suggest that the diagnosis of delusions using structured interview can be made with an acceptable level of inter-rater reliability (falling in the 'substantial agreement' range of 0.61–0.80; Landis and Koch, 1977). However, there are some important caveats, particularly with regard to using the clinical interview for diagnosis. Harper (1999) noted that the World Health Organisation (1979) study used a limited statistical method (using simple correlations rather than unweighted estimates), had a small sample size, and could not rule out the possibility of clinicians discussing their findings before deciding on a diagnosis. The Andreasen et al. (1982) study reported a particularly high level of agreement, although the ratings were based on videotaped interviews after the raters were given a "careful and systematic training program", all of which suggest that the results may be unlikely to generalise to typically less structured clinical situations.

Although some studies have relied purely on given diagnosis (rather than requiring a specific method) in an attempt to capture the clinical reality delusion classification (Garety and Hemsley, 1994), research studies are now increasingly favouring the use of standardised scales. It is clear from this review that standardised scales seem to provide the best level of inter-rater reliability among the methods, with the Scale for the Assessment of Positive Symptoms (SAPS; Andreasen, 1984) and the Positive and Negative Syndrome Scale

Table 1
Studies of inter-rater reliability for diagnosis/detection of delusions in general

	Inter-rater reliability	
	Kappa	*r*
Structured interview		
World Health Organisation (1979)		
Inter-centre		0.83–0.98
Intra-centre		0.61
Andreasen et al. (1982)	0.88	
Endicott et al. (1982)		
NHSI[a] delusions	0.81	
Carpenter et al.[b] widespread delusions	0.28	
DSM-III[c] non-persecutory/jealous delusions	0.59	
RDC[d] non-persecutory/jealous delusions	0.77	
Helmes et al. (1983) nihilistic delusions	0.156	
Vignette assessment		
Junginger et al. (1992)	0.46–0.78	
SAPS[e]		
Moscarelli et al. (1987)	0.88	
Norman et al. (1996)	0.86	
CASH[f]		
Andreasen et al. (1992)	0.64	
PANNS[g]		
Bell et al. (1992)	0.93	
Norman et al. (1996)	0.88	
Muller and Wetzel (1998)	0.42–0.85	
PSYRATS (Delusions Subscale)[h]		
Haddock et al. (1999)		0.88–1.0[i]
Mean	0.69	0.76
Standard deviation	0.24	0.21

[a] New Haven Schizophrenia Index (Astrachan et al., 1972).
[b] Carpenter et al. criteria, part of the WHO international Pilot Study of Schizophrenia (Carpenter et al., 1973).
[c] Diagnostic and Statistical Manual of Mental Disorders III (APA, 1980).
[d] Research Diagnostic Criteria (Spitzer et al., 1978).
[e] Scale for the Assessment of Positive Symptoms (Andreasen, 1984).
[f] Comprehensive Assessment of Symptoms and History (Andreasen et al., 1992).
[g] Positive and Negative Syndrome Scale (Kay et al., 1987).
[h] Psychotic Symptoms Rating Scales (Haddock et al., 1999).
[i] Haddock et al. used the unbiased estimate of reliability (intra-class correlation) described by Winer (1971).

(PANNS; Kay et al., 1987) being reported as the most reliable of those with more than one reliability study available.

3.2. Diagnosing bizarre delusions

Table 2 summarises reliability studies for the diagnosis of bizarre delusions, using structured interview, vignette assessment and psychometric scales.

Although the mean reliability lies within the moderate agreement range (0.41–0.6; Landis and Koch, 1977), the studies reporting the most substantial levels of agreement are again subject to significant methodological flaws: diagnoses in the Goldman et al. (1992) study were based on "consensus judgment" and involved raters who, according to Spitzer et al. (1993), "participated in weekly discussions of ratings of psychiatric symptoms, including bizarre delusions, for a year and a half before the study". In turn, the methodology of Spitzer et al. has been criticised, because the raters were either the authors or the authors' associates, which, as pointed out by Mojtabai and Nicholson (1995), is a potentially serious flaw as "it is not surprising that a homogenous group of raters would attain a higher level of agreement". To reinforce this point, Mojtabai and Nicholson used a random sample of psychiatrists and, as expected, obtained far lower levels of inter-rater reliability. Similarly, the high level of diagnostic reliability reported by Nakaya et al. (2002) can be accounted for by the fact that the diagnosis of bizarre delusions was based on the use of both the PANSS, a review of case notes, the fact that the raters both practised joint interviews before the study began and made collective ratings at regular points throughout the testing period. The Tanenberg-Karant et al. (1995) study used a similar technique, where diagnosis was aided by a structured interview, 3–6 months training and consensus decision for each case.

Table 2
Studies of inter-rater reliability for diagnosis/detection of bizarre delusions

	Inter-rater reliability	
	Kappa	*r*
Structured interview		
Endicott et al. (1982)		
Carpenter et al. criteria	0.66	
DSM-III criteria	0.29	
Kendler et al. (1983)	0.27–0.30	
Flaum et al. (1991)	0.28–0.31	
Goldman et al. (1992)	0.78	
Garety and Hemsley (1994)	0.31	
Nakaya et al. (2002)[a]	0.85–0.92	
Tanenberg-Karant et al. (1995)[b]	0.68	
Vignette assessment		
Junginger et al. (1992)	0.45	
Spitzer et al. (1993)	0.64–0.65	
Mojtabai and Nicholson (1995)	0.38–0.43	
Psychometric scale		
Eisen et al. (1998)		0.02
Mean	0.52	0.02
Standard deviation	0.22	0

[a] Also used Positive and Negative Syndrome Scale (Kay et al., 1987).
[b] Structured Clinical Interview for DSM-II-R.

When these problematic studies are removed from the reliability calculations reported in Table 2, the mean kappa drops to 0.39 (SD = 0.14). With these drawbacks in mind, only one method (the Carpenter et al., 1973, criteria from the study by Endicott et al., 1982) reported acceptable reliability, and it is not clear that this study used either a random sample of psychiatrists or prevented pre-rating discussions as criticised previously. These weaknesses suggest that on the basis of diagnostic reliability a reliable concept of bizarre delusion has yet to be achieved.

4. Discussion

Despite the poor reliability for the diagnosis of bizarre delusions, the diagnosis of delusions *per se* seems to be reliable, both with clinical interview and with standardised scales. This discrepancy suggests that the concept of a bizarre delusion may have little clinical validity and should be rejected until a reliable diagnostic method can be demonstrated.

Nevertheless, one notable feature of this review is the relatively small number of studies on which to base conclusions about the reliability of diagnosing delusions, and the fact that many of the studies reviewed lack adequate controls for potentially confounding factors. As recommended by Mojtabai and Nicholson (1995), studies of diagnostic reliability should include randomly selected individuals who make independent ratings of the same phenomena if they are to truly reflect the reliability of the methods under investigation. Furthermore, studies that report reliability, perhaps as a side issue to their main focus, should also report potentially confounding factors — such as any specific training raters may have completed, whether or not they conferred during the process and what proportion of the raters were from the same institutions. This would allow a balanced assessment of how well the findings might generalise.

It is also worth noting that self-report methods of rating delusions or aspects of delusional experience

(such as the Peters et al. Delusions Inventory; Peters et al., 1999, 2004) are becoming increasing popular and typically show good psychometric properties. As opposed to the categorical nature of the diagnostic approach, such scales represent a dimensional view of psychosis (Johns and van Os, 2001). Although such scales are unable to diagnose delusions, they may provide valuable additional information (for example, ratings of related distress and preoccupation) and allow for a comparison between clinical and non-clinical populations in terms of the extent to which delusional themes are present in each.

References

American Psychiatric Association, 1980. Diagnostic and Statistical Manual of Mental Disorders: DSM-III, 3rd edn. Author, New York.

American Psychiatric Association, 1987. Diagnostic and Statistical Manual of Mental Disorders: DSM-III-R, 3rd revised edn. Author, Washington, DC.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV, 4th edn. Author, Washington, DC.

Andreasen, N.C., 1984. Scale for the Assessment of Positive Symptoms (SAPS). Department of Psychiatry, University of Iowa College of Medicine, Iowa City.

Andreasen, N.C., McDonald-Scott, P., Grove, W.M., Keller, M.B., Shapiro, R.W., Hirschfeld, R.M., 1982. Assessment of reliability in multicenter collaborative research with a videotape approach. Am. J. Psychiatry 139 (7), 876–882.

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The Comprehensive Assessment of Symptoms and History (CASH). An instrument for assessing diagnosis and psychopathology. Arch. Gen. Psychiatry 49 (8), 615–623.

Astrachan, B.M., Harrow, M., Adler, D., Brauer, L., Schwartz, A., Schwartz, C., Tucker, G., 1972. A checklist for the diagnosis of schizophrenia. Br. J. Psychiatry 121 (564), 529–539.

Bell, M., Milstein, R., Beam-Goulet, J., Lysaker, P., Cicchetti, D., 1992. The Positive and Negative Syndrome Scale and the Brief Psychiatric Rating Scale. Reliability, comparability, and predictive validity. J. Nerv. Ment. Dis. 180 (11), 723–728.

Bell, V., Halligan, P.W., Ellis, H.D., 2006. Explaining delusions: a cognitive perspective. Trends Cogn. Sci. 10, 219–226.

Carpenter, W.T., Strauss, J.S., Bartko, J.J., 1973. Flexible system for the diagnosis of schizophrenia: report from the WHO International Pilot Study of Schizophrenia. Science 182 (118), 1275–1278.

David, A.S., 1999. On the impossibility of defining delusions. Philos. Psychiatr. Psychol. 6, 17–20.

Eisen, J.L., Phillips, K.A., Baer, L., Beer, D.A., Atala, K.D., Rasmussen, S.A., 1998. The Brown Assessment of Beliefs Scale: reliability and validity. Am. J. Psychiatry 155, 102–108.

Endicott, J., Nee, J., Fleiss, J., Cohen, J., Williams, J.B., Simon, R., 1982. Diagnostic criteria for schizophrenia: reliabilities and agreement between systems. Arch. Gen. Psychiatry 39 (8), 884–889.

Flaum, M., Arndt, S., Andreasen, N.C., 1991. The reliability of "bizarre" delusions. Compr. Psychiatry 32 (1), 59–65.

Garety, P.A., Hemsley, D.R., 1994. Delusions: Investigations into the Psychology of Delusional Reasoning. Oxford University Press, Oxford.

Goldman, D., Hien, D.A., Haas, G.L., Sweeney, J.A., Frances, A.J., 1992. Bizarre delusions and DSM-III-R criteria. Am. J. Psychiatry 149, 494–499.

Haddock, G., McCarron, J., Tarrier, N., Faragher, E.B., 1999. Scales to measure dimensions of hallucinations and delusions: the Psychotic Symptom Rating Scales (PSYRATS). Psychol. Med. 29, 879–889.

Harper, D.J., 1999. Deconstructing paranoia: an analysis of the discourses associated with the concept of paranoid delusion. Doctoral thesis, Manchester Metropolitan University.

Helmes, E., Landmark, J., Kazarian, S.S., 1983. Inter-rater reliability of twelve diagnostic systems of schizophrenia. J. Nerv. Ment. Dis. 171 (5), 307–311.

Johns, L.C., van Os, J., 2001. The continuity of psychotic experiences in the general population. Clin. Psychol. Rev. 21, 1125–1141.

Junginger, J., Barker, S., Coe, D., 1992. Mood theme and bizarreness of delusions in schizophrenia and mood psychosis. J. Abnorm. Psychology 101, 287–292.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276.

Kendler, K.S., Glazer, W.M., Morgenstern, H., 1983. Dimensions of delusional experience. Am. J. Psychiatry 140 (4), 466–469.

Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Leeser, J., O'Donohue, W., 1999. What is a delusion? Epistemological dimensions. J. Abnorm. Psychology 108, 687–694.

Mojtabai, R., Nicholson, R.A., 1995. Interrater reliability of ratings of delusions and bizarre delusions. Am. J. Psychiatry 152, 1804–1806.

Moscarelli, M., Maffei, C., Cesana, B.M., Boato, P., Farma, T., Grilli, A., Lingiardi, V., Cazzullo, C.L., 1987. An international perspective on assessment of negative and positive symptoms in schizophrenia. Am. J. Psychiatry 144 (12), 1595–1598.

Muller, M.J., Wetzel, H., 1998. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. Acta Psychiatr. Scand. 98 (2), 135–139.

Nakaya, M., Kusumoto, K., Okada, T., Ohmori, K., 2002. Bizarre delusions and DSM-IV schizophrenia. Psychiatry Clin. Neurosci. 56, 391–395.

Norman, R.M., Malla, A.K., Cortese, L., Diaz, F., 1996. A study of the interrelationship between and comparative interrater reliability of the SAPS, SANS and PANSS. Schizophr. Res. 19, 73–85.

Peters, E., Day, S., McKenna, J., Orbach, G., 1999. Delusional ideation in religious and psychotic populations. Br. J. Clin. Psychol. 38 (1), 83–96.

Peters, E., Joseph, S., Day, S., Garety, P., 2004. Measuring delusional ideation: the 21-item Peters et al. Delusions Inventory (PDI). Schizophr. Bull. 30, 1005–1022.

Spitzer, M., 1990. On defining delusions. Compr. Psychiatry 31, 377–397.

Spitzer, R.L., Endicott, J., Robins, E., 1978. Research diagnostic criteria: rationale and reliability. Arch. Gen. Psychiatry 35 (6), 773–782.

Spitzer, R.L., First, M.B., Kendler, K.S., Stein, D.J., 1993. The reliability of three definitions of bizarre delusions. Am. J. Psychiatry 150 (6), 880–884.

Tanenberg-Karant, M., Fennig, S., Ram, R., Krishna, J., Jandorf, L., Bromet, E.J., 1995. Bizarre delusions and first-rank symptoms in a first-admission sample: a preliminary analysis of prevalence and correlates. Compr. Psychiatry 36, 428–434.

Winer, B.J., 1971. Statistical Principles in Experimental Design, 2nd ed. McGraw Hill, Tokyo.

World Health Organization, 1979. Schizophrenia: an International Follow-up Study. Wiley, Chichester.