A review of the reliability and validity of OPCRIT in relation to its use for the routine clinical assessment of mental health patients

PHILIP J. BRITTAIN,¹ DANIEL STAHL,² JAMES RUCKER,¹ JAMIE KAWADLER¹ & GUNTER SCHUMANN¹

1 MRC-SGDP Centre, Institute of Psychiatry, King's College London, London, UK

2 Department of Biostatistics, Institute of Psychiatry, King's College London, London, UK

Key words

OPCRIT, reliability, validity

Correspondence

Philip Brittain, MRC-SGDP
Centre, Institute of Psychiatry,
King's College London, London,
SE5 8AF, UK.
Telephone (+44) (0) 207 848 0798
Email: Philip.Brittain@kcl.ac.uk

Abstract

The OPCRIT program is a symptom checklist with accompanying algorithms producing operationally defined diagnoses. We undertook a review of studies which had used OPCRIT and had reported statistics concerning its reliability and validity, producing summary measures from 44 studies. The first main measure of interest was inter-rater reliability where mean kappa values indicated that agreement between raters was “substantial” with a marginal improvement at the diagnostic (0.76) versus individual item (0.69) level. The second main measure of interest was convergent validity – the agreement between OPCRIT and clinical diagnoses. Most studies reported these figures as concordance rates suggesting mean agreement, unadjusted for chance, of 69%. Very few studies used the chance-adjusted kappa statistic but where this was used agreement was “fair” (0.39). Agreement between OPCRIT and other research diagnoses was “moderate” (0.60). We also considered differences between the way OPCRIT has traditionally been used in research settings and the naturalistic manner in which it will be employed in the hospital ward. This review provides a summary of the reliability and validity of OPCRIT, which will be considered during the preparation for its use in the routine characterization of mental health patients in clinical settings. Copyright © 2013 John Wiley & Sons, Ltd.

Introduction

Whilst the collection and use in medical research of routine clinical data has an established history, the emergence in recent years of electronic medical records has fundamentally altered the potential of this practice. The flexibility afforded by this advance opens up new opportunities to better integrate clinical, research and even training

activities, allowing the possibility of greater productivity in these areas via the cross-fertilization of efforts and ideas.

An example would be the large scale re-use of anonymized patient symptom data, routinely collected during clinical assessments, as phenotypic information in genetic research.

Since 2006, all patient records of the South London and Maudsley National Health Service trust (SLaM) have been maintained and stored electronically on an interface known

as the “Electronic Patient Journey System” (ePJS). The Institute of Psychiatry’s “Biomedical Research Centre” is in the process of implementing a project to routinely assess patients under the care of this trust using a standardized electronic assessment tool called “OPCRIT+” (Rucker *et al.*, 2011), which is an expansion of the original “OPCRIT” program (McGuffin *et al.*, 1991).

In its original form, OPCRIT is a checklist of symptoms constructed from operational criteria for the major psychiatric classificatory systems and accompanying algorithms allowing the generation of diagnoses. Its research use in the last two decades has been extensive, not only for its diagnostic capacities but also as a platform for systematic and comprehensive symptom assessment. It has been widely used in genetic research (Fanous *et al.*, 2009) but also in the exploration of dimensional approaches to illness (Allardyce *et al.*, 2007).

OPCRIT+ offers, amongst other additional features, a comprehensive hyperlink training system, the ability to produce standardized patient summaries and, compared with the original OPCRIT which focuses on psychotic and affective disorders only, a considerably increased diagnostic repertoire. The successful uptake of such a tool could potentially bring many benefits to clinicians.

We undertook a review of previous studies which had used OPCRIT and had reported reliability and validity measures in relation to its use. This paper reports on the findings of that review and records some general observations concerning the methods of its prior use and the differences from the way in which it will be used in this project.

Methods

Literature review

The original article describing the development of OPCRIT had been cited 608 times as of June 2011. We reviewed those articles which appeared in an OVID Medline search (from 1950 to 2010) and a PsychINFO search (from 1806 to 2010) using “OPCRIT” as the search term. This returned 97 non-overlapping articles with “OPCRIT” in the title, abstract or listed as a keyword. This subset putatively contained those, of all citing OPCRIT, where the tool received greater focus. After review, several of these pointed towards relevant studies that had been overlooked in the searches and several more were included from other sources. In all, 44 studies were found which contained reliability and/or validity statistics.

Categorization and reporting

The statistics reported were separable into five categories (see Tables 1–5), ordered here in the descending relevance we considered they had to our project:

1. Studies reporting on the agreement between raters when using OPCRIT, both at an individual item (e.g. elevated mood) and diagnostic level – a measure of “inter-rater reliability”
2. Studies reporting on the agreement between OPCRIT and clinical diagnoses – a measure of “convergent validity”
3. Studies reporting on the agreement between OPCRIT diagnoses and other diagnostic methods, including other research and consensus methods – a measure of “inter-method reliability”
4. Studies reporting on the agreement between repeated OPCRIT diagnoses made on the same patients but when based on different information sources (e.g. case records versus clinical abstracts) – a further measure of “inter-method reliability”
5. Studies reporting on the agreement of repeated OPCRIT ratings on the same patients, both at an individual item and diagnostic level – a measure of “test-retest reliability”

We only reported diagnostic statistics based on the popular classificatory systems of International Classification of Disease (ICD), Diagnostic and Statistical Manual of Mental Disorders (DSM) and Research Diagnostic Criteria (RDC), although OPCRIT produces diagnoses from a number of lesser-used classificatory systems (e.g. Feighner). Furthermore, we did not report statistics relating specifically to agreement between systems (e.g. ICD-10 versus DSM-IV).

The statistics in Tables 1–5 are generally reported verbatim, with the exception of several cases where, for the sake of brevity, summary variables were produced. When the same study reported statistics for more than one of the five categories, they were reported separately in each table.

Analysis

To summarize agreement, we calculated mean values from all studies within each of categories one to five. Agreement was summarized initially by the most commonly reported statistic type within each category. With 32 out of 44 studies reporting Cohen’s kappa (Cohen, 1960), only category two had the majority of articles reporting another method, in this case concordance rates, which were transformed into percentage agreement.

Due to uncertainty in several studies concerning the number of subjects on which the kappa value was based,

Table 1. Studies reporting on the agreement between raters when using OPCRIT

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject N & patient type of entire cohort	Statistic types reported in article	Description of OPCRIT variables reported	Values
Allan <i>et al.</i> , 2009	NK/NK	All available clinical information	224/proband-wise twin pairs where proband had suffered psychotic symptoms or an episode of RDC mania or hypomania 464/first episode	k, p, icc	Nine illness history items on a subset of 30 cases. Values were reportedly means from "pairs of raters". The k and p values reported here are means of those reported.	2 = 0.48 (k), 1 = 0.9 (icc), 3 = 0.64 (p), 3 < 0.41 (statistic type not reported)
Allardyce <i>et al.</i> , 2000	2/Specialist Registrars	Case records	ICD-9 & -10 psychotic disorders NK/first episode	k	ICD-10 diagnosis of schizophrenia	0.79
Allardyce <i>et al.</i> , 2001	3/psychiatrists	Case records	ICD-9 & 10 psychotic disorders	k	RDC diagnosis of schizophrenia 28 symptom items reported here as a mean and one described verbally. 10 demographic and risk items which were reported as all being over a defined threshold apart from one.	0.79
Allardyce <i>et al.</i> , 2007	2/experienced psychiatrists	Case records	464/first episode psychosis 335/schizophrenia, schizotypal and delusional disorders	k	Values were for a subset of 45 cases. ICD-10 diagnoses on a subset of 50 cases	28 = 0.77, 1 = "poor"
Balestrieri <i>et al.</i> , 1997	2/NK	Case records	1055/first episode ICD-9 & -10 psychotic disorders	k	RDC diagnosis of schizophrenia on a subset of unknown size	0.87
Boydell <i>et al.</i> , 2003	1(rating against a previously scored database)/NK	Case records		k		0.79

Author	Year	Study Design	Measures	Results	Notes
Cardno <i>et al.</i> , 1998	4/NK	All available clinical information	191/sibling pairs with DSM-IV schizophrenia or schizoaffective disorder 191/sibling pairs with DSM-IV schizophrenia or schizoaffective disorder	k or icc	27 symptom and demographic variables on a subset of 30 cases. 27 ≥ 0.61 (not specified as k or icc)
Cardno <i>et al.</i> , 1999	4/NK	All available clinical information	448/twin pairs with proband effected by psychotic symptomatology or an episode of RDC mania or hypomania	k	22 psychotic symptom items on a subset of 30 cases 19 = 0.64–0.92, 3 = 0.52–0.54
Cardno <i>et al.</i> , 2001	NK/NK	All available clinical information	448/twin pairs with proband effected by psychotic symptomatology or an episode of RDC mania or hypomania	k	30 symptom items on a subset of 30 cases 28 > 0.41, 2 < 0.41
Cardno <i>et al.</i> , 2002	NK/NK	All available clinical information	448/ twin pairs with proband effected by psychotic symptomatology or an episode of RDC mania or hypomania	k	Eight symptom items on a subset of 30 cases reported as a mean 0.66
Cardno <i>et al.</i> , 2008	NK/NK	All available clinical information	486/first-contact ICD-9 defined "schizophrenic psychosis", "paraphrenia" or "other nonorganic psychosis"	k	15 symptom items 15 > 0.41
Castle <i>et al.</i> , 1991	2/NK	Case records	DSM-III and RDC diagnoses on a subset of 50 cases	k	0.76 (DSM-III), 0.82 (RDC)

(Continues)

Table 1. (Continued)

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject N & patient type of entire cohort	Statistic types reported in article	Description of OPCRIT variables reported	Values
Castle <i>et al.</i> , 2006	24/NK	Observed or participatory interviews with information collated on the "Diagnostic interview for psychoses"	108/NK 180/patients admitted to a Mother and Baby Unit at a psychiatric hospital 397/first-episode ICD-8 or –9 schizophrenia, paranoia, paranoid disorders, paraphrenia or psychosis not elsewhere specified	k, par	65 items and ICD-10 and DSM-III-R diagnoses for "narrow" and "broad" (three-digit) classifications on a subset of 20 cases	Item level: 18 > 0.8, 27 = 0.6–0.80, 13 = 0.01–0.39, 6 = 0 (k) 65 = 0.8–1.0 (par) Narrow diagnoses: ICD-10 0.73 (k) 0.81 (par) DSM-III-R 0.49 (k) 0.67 (par) Broad diagnoses: ICD-10 0.74 (k) 0.82 (par) DSM-III-R 0.62 (k) 0.78 (par)
Davies <i>et al.</i> , 1995	2/NK	Case records	Unit at a psychiatric hospital 397/first-episode ICD-8 or –9 schizophrenia, paranoia, paranoid disorders, paraphrenia or psychosis not elsewhere specified	pa	Schizophrenia diagnosis on a subset of 10+	90–95%
Farmer <i>et al.</i> , 1992	2/experienced clinicians	Case records	Case records, clinical interview and collateral information	k	DSM-III and RDC diagnoses on a subset of 50 cases 79 items and ICD-10, DSM-IV, DSM-III-R, DSM-III and RDC diagnoses	0.76 (DSM-III), 0.82 (RDC) Item level: 30 > 0.9, 26 = 0.8–0.9, 18 = 0.7–0.8, 5 = 0.6–0.7
Matsuura <i>et al.</i> , 2004	6/neuropsychiatrists		116/epileptic psychosis and schizophrenia patients	k		

	from family members			Diagnostic level: 0.8–0.9 (ICD-10, DSM-III-R, DSM-III, RDC) 0.7–0.8 (DSM-IV)
McConville and Walker, 2000	2/experienced clinicians	449/patients from a psychiatric case register with a range of ICD-7 to ICD-9 diagnoses	cr	5/6 Item level: 0.6 Diagnostic level: 0.74 (DSM-III-R), 0.85 (DSM-III), 0.71 (RDC) (not specified as k or wk)
McGuffin <i>et al.</i> , 1991	3/NK	Case records and patient interviews in a proportion	k, wk	ICD-10 diagnosis of schizophrenia on a subset of 45 cases 68 items and DSM-III-R, DSM-III and RDC diagnoses. Item level value reported here is a mean. Individual items on a subset of 20 cases; k figure was reported as a median.
Mihalopoulos <i>et al.</i> , 2000	3/experienced psychopathology raters	50/first episode psychosis patients	k, s-iqr	0.84 (k), 0.19 (s-iqr) Item level: $75 \geq 0.6$, $64 > 0.8$, $12 < 0.4$, 5 items = 0 (k), 88 in range of 0.8 to 1.0 (par), $87/90 \geq 0.80$ (ppos) $71/90 \geq 0.80$ (pneg) Narrow diagnoses: ICD-10 0.84 (k) 0.73–0.96 (ci) 0.87 (par) DSM-IV 0.74 (k) 0.59–0.84 (ci) 0.78 (par)
Rossi <i>et al.</i> , 2010	7/psychiatrists and psychologists	60/schizophrenia, affective disorder, personality disorder and other diagnostic groups	k and associated ci, par, ppos, pneg	90 items and ICD-10 and DSM-IV diagnoses reported as specific subtypes (narrow) or aggregated into five broad groupings (broad).

(Continues)

Table 1. (Continued)

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject N & patient type of entire cohort	Statistic types reported in article	Description of OPCRIT variables reported	Values
Rucker <i>et al.</i> , 2011	20/junior doctors	Case summaries	10/patients with multiple, co-morbid diagnoses	wk and pa	158 items and diagnoses categorized into four mixed ICD-10/DSM-IV diagnostic spectra (psychotic/affective, substance misuse, anxiety and personality disorder spectra). Item level value was reported as a mean. At the diagnostic level, a combined k value was reported and a mean pa rate is reported here. 90 items and ICD-10 and DSM-IV diagnoses reported separately for five diagnoses and also split into three broad categories	Broad diagnoses: ICD-10 0.85 (k) 0.71–0.97 (ci) 0.89 (par) DSM-IV 0.74 (k) 0.57–0.88 (ci) 0.80 (par) Item level: 0.80 (wk)
		Interview with the “Diagnostic Interview for Psychosis” and case notes in a proportion				
Skorven and Mikalsen, 2010	7/psychology students, psychologists, psychiatrists and resident Doctors	27/psychiatric inpatients		k, ci, icc, par, ppos, pneg		Diagnostic level: Affective/psychotic: 0.63 (wk) 68% (pa). Anxiety disorders: 0.71 (wk) 80% (pa), Substance misuse disorders: 0.76 (wk) 84% (pa), Personality disorder: 0.83 (k), 92% (pa). 0.70 (combined k), 81% (mean pa) Item level: 24 > 0.80, 39 = 0.60–0.79, 11 = 0.40–0.59, 1 = 0.39, 11 = “not significant” 4 = “not calculated” (mix of k and icc). The

Van Os <i>et al.</i> , 1996	3/NK	Case records Abstracts produced from semi- structured interviews and case records together with full records in a proportion	110/First admission "non-affective psychosis" patients	k	RDC diagnosis of schizophrenia	0.74	(bipolar/manic, depressive, non-affective disorders). Diagnostic values reported here are means.	following only reported for 24 items where kappa was calculated: 15/24 ≥ 0.80 (ppos), 22/24 ≥ 0.80 (pneg). 23/24 > 0.80, 1/24 = 0.76 (par). Narrow diagnoses: ICD-10 0.70 (k) 0.41–0.99 (ci) 0.87 (par) 0.89 (ppos) 0.80 (pneg) DSM-IV 0.55 (k) 0.15–0.96 (ci) 0.88 (par) 0.92 (ppos) 0.63 (pneg) Broad diagnoses: ICD-10 0.76 (k) 0.49–1.04 (ci) 0.90 (par) 0.85 (ppos) 0.92 (pneg) DSM-IV 0.74 (k) 0.46–1.02 (ci) 0.89 (par) 0.83 (ppos) 0.90 (pneg)
Vares <i>et al.</i> , 2006 Williams <i>et al.</i> , 1996	2/experienced psychiatrists 30/mainly psychiatrists	Case summaries	73/schizophrenia or related psychoses	k, ua wk and associated sd	DSM-III-R diagnoses on subsets of 20 to 22. Figures reported here are means of two values reported for each stat type. ICD-10, DSM-III-R, DSM-III and	0.73 (k), 0.85 (ua) 0.70/0.08 (ICD-10),		(Continues)

(Continues)

Table 1. (Continued)

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject N & patient type of entire cohort	Statistic types reported in article	Description of OPCRIT variables reported	Values
	but a few psychologists and nurses		30/range of diagnoses with an emphasis on major psychoses		RDC diagnoses clustered into seven semi-hierarchical categories.	0.73/0.08 (DSM-III-R), 0.75/0.10 (DSM-III), 0.75/0.09 (RDC)

Note: Statistic types reported in table: concordance rates (cr), 95% confidence intervals (ci), diagnostic frequency as a percentage (dfp), intra-class correlation (icc), kappa (k), negative predictive power (npp), overall pairwise agreement (par), percent agreement (pa), positive predictive power (ppp), proportion of cases for which agreement occurred (pc), proportions of negative agreement (pneg), proportions of positive agreement (ppos), semi-interquartile range (s-iqr), sensitivity (se), Spearman's correlation (p), specificity (sp), standard error (ste), standard deviation (sd), standardized incidence ratio (sir), unadjusted agreement (ua), weighted kappa (wk).

Table 2. Studies reporting on the agreement between OPCRIT and clinical diagnoses

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject N & patient type of entire cohort	Statistic types reported in article	Description of variables reported	Values
Allardyce <i>et al.</i> , 2000	2/Specialist Registrars	Case records	464/first episode ICD-9 & -10 psychotic disorders		OPCRIT ICD-10 and DSM-IV schizophrenia diagnosis versus rates of ICD-9 & -10 clinically diagnosed schizophrenia over a 20-year period	Significant decline in the rate of clinically diagnosed schizophrenia (summary SIR = 0.77) no significant decline when same cases rated with OPCRIT (ICD-10 summary SIR = 0.98, DSM-IV summary SIR = 1.10)

Balestrieri <i>et al.</i> , 1997	2/NK	Case records	381/schizophrenia, schizotypal and delusional disorders	cr	OPCRIT ICD-10 diagnosis of schizophrenia or delusional disorder versus ICD-10 clinically diagnosed grouping called "Schizophrenia, schizotypal and delusional disorders" on a subset of 335 cases OPCRIT ICD-10, DSM-III-R, DSM-III and RDC schizophrenia diagnoses versus subset of 45 patients with a clinical diagnosis of schizophrenia OPCRIT DSM-IV diagnoses versus ICD-8, -9 or -10 clinical diagnoses on a subset of 134. Values are from three sets of comparisons reflecting different diagnostic clusterings (a comparison consisting of eight different diagnoses, a schizophrenic psychoses versus all other diagnoses comparison and a schizophrenia versus all other diagnoses comparison). Each figure reported here is a mean from two values reflecting OPCRIT diagnoses produced from different amounts of clinical data.	284/335 82% (ICD-10) 62% (DSM-III-R) 53% (DSM-III)
Davies <i>et al.</i> , 1995	2/NK	Case records	180/patients admitted to a Mother and Baby Unit at a psychiatric hospital	pa		60% (RDC) Eight diagnoses: 0.22 (k) 0.60 (ua) Schizophrenic psychoses versus all other: 0.19 (k) 0.74 (ua)
Ekholm <i>et al.</i> , 2005	3/trained clinicians	Case records and information from semi-structured interviews	143/schizophrenia or related psychoses	k, ua		Schizophrenia versus all other: 0.27 (k) 0.68 (ua)

(Continues)

Table 2. (Continued)

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject N & patient type of entire cohort	Statistic types reported in article	Description of variables reported	Values
Forrester <i>et al.</i> , 2001	1/investigator	Case records	204/ICD-9 index admission of schizophrenic psychoses, affective psychoses, paranoid states or other non-organic psychoses	pa	OPCRIT ICD-10 and DSM-III-R diagnoses categorised into "Affective disorders" (Ad), "Schizophrenia" (Sc), "Other psychotic disorders" (Op) and "Other conditions" (Oc) versus ICD-9 clinical diagnoses categorised the same. Figure reported here is the mean agreement across five admissions.	Ad: 66% (ICD-10) 72% (DSM-III-R) Sc: 75% (ICD-10) 75% (DSM-III-R) Op: 40% (ICD-10) 38% (DSM-III-R)
Humphreys & Ogilvie, 1996	NK/NK	Case records	10/Schizophrenia or "... various degrees of disturbance of personality"	cr	OPCRIT ICD-10 and DSM-III-R diagnoses versus clinical diagnoses for index admission (I) and "clinical outcome" (CO)	Oc: 29% (ICD-10) 36% (DSM-III-R) I: 1/10 (ICD-10) I: 2/10 (DSM-III-R) CO: 6/9 (ICD-10)
Jakobsen <i>et al.</i> , 2005	2/experienced psychiatrists	Case records and information from semi-structured interviews and collateral informants in a proportion	353/Inpatients and outpatients with a history of psychosis	k	OPCRIT ICD-10 schizophrenia diagnosis versus ICD-10 clinical diagnosis of schizophrenia on a subset of 100.	0.60 Schizophrenia: 74 OPCRIT/79 clinical
Jakobsen <i>et al.</i> , 2007	2/psychiatrists, 1 described as "experienced"	Structured interviews	359/patients participating in a psychiatric biobank	cr	OPCRIT ICD-10 diagnoses versus clinical ICD-10	Schizophrenia spectrum: 96 OPCRIT/89 clinical BPD versus non-BPD: 0.28
Jakobsen <i>et al.</i> , 2008	155/patients with ICD-10 affective			k		

		disorder diagnoses and history of ECT		diagnoses. Results reported as bipolar disorder (BPD) versus non-BPD and major depressive disorder (MDD) versus non-MDD. OPCRIT ICD-10 aggregate diagnostic category of "affective disorder" versus first admission	MDD versus non-MDD: 0.35
	1/research consultant psychiatrist	Case records and patient interviews		ICD-8 clinical diagnosis of manic-depressive psychosis in a subset of 100.	
Kessing, 1998	NK/NK	Case records	cr	OPCRIT RDC diagnosis of schizophrenia versus clinical diagnosis of schizophrenia. OPCRIT ICD-10, DSM-IV, DSM-III-R, DSM-III and RDC diagnosis of schizophrenia versus clinical diagnosis of schizophrenia on subset of 58 schizophrenia spectrum patients. Figures are only agreement on frequency of schizophrenia diagnosis NOT case-by-case agreement.	95/100
MacCabe <i>et al.</i> , 2002	1/NK	Case records	cr	OPCRIT ICD-10 diagnoses categorized as "Schizophrenia & related conditions" and "Affective disorders" versus ICD-10 coded "incident register" (IR) and "register lifetime" (sp)	71/98
Matsuura <i>et al.</i> , 2004	6/neuropsychiatrists	Case records, clinical interview and collateral information from family members	dfp	OPCRIT ICD-10 diagnoses categorized as "Schizophrenia & related conditions" and "Affective disorders" versus ICD-10 coded "incident register" (IR) and "register lifetime" (sp)	97% (ICD-10), 80% (DSM-IV), 89% (DSM-III-R), 61% (DSM-III), 89% (RDC) Schizophrenia & related conditions: IR = 0.50 (k), 0.52 (se), 0.97 (sp) RL = 0.69 (k), 0.83 (se), 0.95 (sp)
McConville & Walker, 2000	2/experienced clinicians	Case records	k, se, sp		

(Continues)

Table 2. (Continued)

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject N & patient type of entire cohort	Statistic types reported in article	Description of variables reported	Values
Pihlajamaa <i>et al.</i> , 2008	5/experienced psychiatrists	Case records	806/ICD-8 and -9 defined schizophrenia spectrum disorder patients (schizophrenia, schizoaffective disorder and schizophreniform disorder)	pa, ci	(RL) clinical diagnoses categorized the same. Figures are based on subsets of between 159–183. OPCRIT ICD-10, DSM-IV and DSM-III-R schizophrenia or schizophrenia spectrum (schizophrenia, schizophreniform & schizoaffective disorders) diagnoses versus “lifetime ever” or most recent ICD-8 and -9 clinical diagnoses categorized the same on subsets of between 448 and 626; ci figures range data only is reported here due to space considerations. OPCRIT DSM-IV schizophrenia, schizophreniform and schizoaffective disorder diagnoses versus ICD-8, -9 or -10 clinical diagnoses of the same OPCRIT RDC diagnosis of schizophrenia or “other psychosis” versus aggregate clinical diagnosis category “ICD non-affective psychosis”.	Affective disorders: IR = 0.53 (k), 0.69 (se), 0.84 (sp) RL = 0.63 (k), 0.87 (se), 0.76 (sp) Schizophrenia: 78% (ICD-10) 74% (DSM-IV) 76% (DSM-III-R) 71–81% (ci) Schizophrenia spectrum: Lifetime ever = 78% (ICD-10) 76% (DSM-IV) 75% (DSM-III-R) Most recent = 82% (ICD-10) 80% (DSM-IV) 80% (DSM-III-R) 72–85% (ci) 77/126 (Schizophrenia) 2/16 (Schizophreniform) 15/26 (Schizoaffective)
Reutfors <i>et al.</i> , 2010	2/NK	Case records	168/ICD-8, -9 or -10 defined schizophrenia, schizophreniform or schizoaffective disorder patients	cr		
Van Os <i>et al.</i> , 1996	3/NK	Case records	110/first admission “non-affective psychosis” patients	cr		110/110

Vares <i>et al.</i> , 2006	2/experienced psychiatrists	Semi-structured interview and case records Interview, case records, clinical observation and information from informants	73/schizophrenia or related psychoses	k, ua, se, sp, ppp, npp	OPCRIT DSM-III-R diagnoses versus ICD-6 to – 10 defined hierarchical clinical diagnoses. Values for three methods of clustering diagnoses reported (a seven diagnosis, a schizophrenia versus all other diagnoses and a schizophrenic psychoses versus all other diagnoses method).	7 diagnoses: 0.32 (k) 0.73 (p) Schizophrenia versus all other: 0.31 (k) 0.76 (p) 0.78 (se) 0.62 (sp) 0.91 (ppp) 0.36 (npp)
					Values for two methods of producing the OPCRIT diagnoses based on differing amounts of information were reported and values reported here are a mean of those. Based on a subset of 66.	Schizophrenic psychoses versus all other: 0.08 (k) 0.80 (p) 0.81 (se) 0.50 (sp) 0.98 (ppp) 0.08 (npp) 5/6 (ICD-10)
Vogels <i>et al.</i> , 2004	2/psychiatrists		6/Prader–Willi Syndrome patients with a psychotic disorder	cr	diagnoses versus clinical diagnoses described only as “psychotic disorder”, therefore values reported here are OPCRIT diagnoses amalgamated as such. OPCRIT ICD-10 schizoaffective diagnosis versus clinically diagnosed ICD-10 defined schizoaffective disorder patients	6/6 (DSM-IV) <

See footnote to Table 1 for statistic types reported in table.

Table 3. Studies reporting on the agreement between OPCRIT diagnoses and other diagnostic methods

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject N & patient type of entire cohort	Statistic types reported in article	Description of variables reported	Values
Castle <i>et al.</i> , 2006	NK/NK	Observed or participatory interviews with information collated on the "Diagnostic interview for psychoses"	108/NK	cr	OPCRIT ICD-10 diagnoses versus SCAN-generated ICD-10 diagnoses at a "broad" (three-digit level) classification on a subset of 10 cases OPCRIT DSM-III-R and RDC diagnoses versus individual and consensus best-estimate DSM-III-R and RDC diagnoses. Figures reported here are means. OPCRIT DSM-IV diagnoses versus DSM-IV "standard research diagnosis". Values for three methods of clustering the diagnoses were reported (an eight diagnoses method, a schizophrenia versus all other diagnoses method and a schizophrenic psychosis versus all other diagnoses method). Each figure reported	9/10 DSM-III-R: 0.84 (k) 0.04 (se) 0.88 (p) 0.03 (se) RDC: 0.81 (k) 0.04 (se) 0.84 (p) Eight diagnoses: 0.75 Schizophrenia versus all other: 0.87
Craddock <i>et al.</i> , 1996	2/Clinical research fellows	All available data	100/psychiatrically ill individuals from families multiply affected by bipolar or schizophrenia	k, pc, se		
Ekholm <i>et al.</i> , 2005	3/trained clinicians	Case records and information from semi-structured interviews	143/schizophrenia or related psychoses	k	Schizophrenic psychosis versus all other: 0.60	

Gil <i>et al.</i> , 2009	4/psychiatrists and psychologists	Case records and collateral information from informants	102/outpatients enrolled in a schizophrenia program	cr	here is a mean from two values reflecting OPCRIT diagnoses produced from different amounts of clinical data. Based on subsets of between 138 and 139.	OPCRIT diagnosis versus clinical assessment following DSM-IV and ICD-10 criteria for schizophrenia OPCRIT DSM-III-R diagnoses (not for PD and substance use) versus DSM-III-R consensus diagnoses from two senior researchers (which included rechecking the researchers OPCRIT). Values were reported as a grouping of “schizophrenia versus non- schizophrenia” and a four-class grouping of “schizophrenia, schizophrenia spectrum, other psychoses and non-psychotic disorders” based on a subset of 140. OPCRIT ICD-10 and DSM-III-R diagnoses versus ICD-10 and DSM-III-R diagnoses from a computerized diagnostic system called “CATEGO5” OPCRIT DSM-III-R diagnoses versus a “Gold Standard”	≥ 99/102 Four-class: 0.55 (k)
Isohanni <i>et al.</i> , 1997	1/junior researcher	Case records	249/DSM-III-R defined psychosis, personality disorder and psychoactive substance use disorders	k, se, sp	Schizophrenia versus non-schizophrenia: 0.64 (k), 0.76 (se), 0.96 (sp) 0.43 (ICD-10)		
Lazartigues <i>et al.</i> , 1997	NK/NK	SCAN interview and case records on a proportion	40/adolescent inpatient psychotics 462/ICD-8 or –9 defined schizophrenia	k	0.11 (DSM-III-R) 0.89 (wk), 0.96 (se), 0.94 (sp),		
Makikyro <i>et al.</i> , 1998	1/junior researcher	Case records		sp, ppp, npp			

(Continues)

(Continues)

Table 3. (Continued)

Study	Rater numbers & description	Description of data source on which OPCRIT completed	Subject <i>N</i> & patient type of entire cohort	Statistic types reported in article	Description of variables reported	Values
Meszaros <i>et al.</i> , 1996	NK/NK	NK	or schizophrenia spectrum diagnoses		DSM-III-R consensus diagnosis from two experienced researchers (which also included rechecking the researchers OPCRIT). Values reflect a diagnostic grouping of "schizophrenia versus non-schizophrenia" based on a subset of 73 cases. OPCRIT ICD-10, DSM-III-R and RDC diagnoses versus DSM-III-R consensus diagnoses on a subset of 42 seemingly within the "schizophrenia spectrum" OPCRIT DSM-III-R diagnoses based on either case notes or clinical abstracts versus four alternative methods of assigning a DSM-III-R diagnosis; the DSM-IV field trial instrument, the RPMIP, the Munich Diagnostic Checklists and a consensus procedure on a subset of between 45 and 46. Figures reported here are means of all comparisons.	0.98 (ppp), 0.89 (npp)
Mihalopoulos <i>et al.</i> , 2000	3/experienced 12psychopathology raters	Case records or clinical abstracts	50/first episode psychosis patients 60/schizophrenia, affective disorder, personality disorder and other diagnostic groups	k, ste, ua		
Rossi <i>et al.</i> , 2010	7/psychiatrists and psychologists	Interview with the "Diagnostic Interview for Psychosis"		cr	OPCRIT diagnoses versus SCAN-generated diagnoses on a subset of 18 patients	0.38 (k), 0.08 (ste), 51% (ua) 12/18

Vares <i>et al.</i> , 2006	2/experienced psychiatrists	Semi-structured interview and case records	73/schizophrenia or related psychoses	k, ua, se, sp, ppp, npp	OPCRIT DSM-III-R diagnoses versus DSM-III-R research diagnoses. Values for three methods of clustering the diagnoses were reported (a seven diagnoses, a schizophrenia versus all other diagnoses and a schizophrenic psychoses versus all other diagnoses method). Values for two methods of producing the OPCRIT diagnoses and two different research diagnoses methods (SCID-I and "traditional research method") were used and are reported here as means. Based on subset of 69. OPCRIT ICD-10 diagnoses versus ICD-10 consensus diagnoses. Six diagnostic categories are reported (schizophrenia, affective illness, misuse/organic, schizoaffective, other non-organic psychosis, no diagnosis)	Seven diagnoses: 0.49 (k) 0.79 (ua) Schizophrenia versus all other: 0.64 (k), 0.87 (ua), 0.88 (se), 0.85 (sp), 0.94 (ppp), 0.70 (npp) Schizophrenic psychoses versus all other: 0.33 (k), 0.86 (ua), 0.88 (se), 0.62 (sp), 0.97 (ppp), 0.48 (npp)
Vollmer-Larsen <i>et al.</i> , 2006	2/psychiatrists	Clinical summaries and entire case records in a proportion	59/schizoaffective disorder	cr		39/59

See footnote to Table 1 for statistic types reported in table.

Table 4. Studies reporting on the agreement between repeated OPCRIT diagnoses made on the same patients but when based on different information sources

[illegible]

Vares <i>et al.</i> , 2006	2/experienced psychiatrists	Semi-structured interview and case records	73/schizophrenia or related psychoses	k, ua, se, sp, ppp, npp	DSM-III-R diagnoses based on case records versus based on case records and interview. Both were made by the same raters. Values for three methods of clustering the diagnoses were reported (a seven diagnoses method, a schizophrenia versus all other diagnoses method and a schizophrenic psychoses versus all other diagnoses method). Based on a subset of 69.	Seven diagnoses: 0.90 (k) 0.97 (ua) Schizophrenia versus all other: 0.89 (k), 0.97 (ua), 0.97 (se), 1.00 (sp), 1.00 (ppp), 0.83 (npp)
					Schizophrenic psychoses versus all other: 0.79 (k), 0.99 (ua), 0.99 (se), 1.00 (sp), 1.00 (ppp), 0.67 (npp)	

See footnote to Table 1 for statistic types reported in table.

Table 5. Studies reporting on the agreement of repeated OPCRIT ratings on the same patients

Study	Rater numbers & description	Description of data source	Subject N & patient type of entire cohort	Statistic types reported in article	Description of OPCRIT variables reported	Values
Castle <i>et al.</i> , 2006	10/NK	Observed or participatory interviews with information collated on the "Diagnostic interview for psychoses"	108/NK	k, par	Individual items and ICD-10 and DSM-III-R diagnoses for "narrow" and "broad" (3-digit) classifications on a subset of 7 cases rated twice with intervals of 2–11 weeks	Item level: 42% ≥ 0.6 , +33% < 0.4 (k) "Majority of items" 0.80–1.00 (par) Narrow diagnoses: ICD-10 0.51 (k) 0.57 (par) DSM-III-R 0.48 (k) 0.57 (par) Broad diagnoses: ICD-10 0.65 (k) 0.71 (par) DSM-III-R 0.81 (k) 0.86 (par) Item level: With symptom durations; 4 > 0.8 , 14 = 0.60–0.79, 14 = 0.40–0.59, 58 = –0.02–0.39 (k). Without symptom durations; 20 ≥ 0.6 , 21 = 0.4–0.60 (k). 45 > 0.80 –1.0 (PAR) 68/90 ≥ 0.80 (Ppos).
Rossi <i>et al.</i> , 2010	7/psychiatrists and psychologists	Interview with the "Diagnostic Interview for Psychosis"	60/Schizophrenia, affective disorder, personality disorder and other diagnostic groups	k, par, ppos, ci	Narrow diagnoses: ICD-10; 0.21 (k) 0.08–0.36 (ci) 0.34 (par) DSM-IV; 0.24 (k) 0.09–0.41 (ci) 0.36 (par) Broad diagnoses: ICD-10; 0.22 (k) 0.01–0.45 (ci) 0.44 (par) DSM-IV; 0.40 (k) 0.18–0.64 (ci) 0.54 (par)	

Skorven & Mikalsen, 2010	7/psychology students, psychologists, psychiatrists and resident Doctors	Interview with the "Diagnostic Interview for Psychosis" and case notes in a proportion	14/psychiatric inpatients	k, icc, par, ppos, npes	Item level: $3 > 0.80$, 11 = $0.60-0.79$, 15 = $0.40-0.59$, 39 = "not significant", 22 = "not calculated" (mix of k and icc). The following reported for the 10 items where kappa was "obtainable and significant": $5/10 > 0.80$ (ppos), $9/10 > 0.80$ (pneg). $7/10 > 0.80$ (par). Narrow diagnoses: ICD-10 0.50 (k) 0-1 (ci) 0.78 (par) 0.83 (ppos) 0.67 (pneg) DSM-IV 0.69 (k) $-0.13-0.88$ (ci) 0.71 (par) 0.78 (ppos) 0.60 (pneg). Broad diagnoses: ICD-10 0.65 (k) 0.86 (par) $0.20-1.09$ (ci) 0.76 (ppos) 0.89 (pneg) DSM-IV 0.82 (k) 0.93 (par) $0.56-1.09$ (ci) 0.88 (ppos) 0.95 (pneg)
					90 items and ICD-10 and DSM-IV diagnoses reported separately for four diagnoses and also split into three broad categories (bipolar/manic, depressive, non-affective disorders). Rated twice separated by a mean of 19.5 days. Most diagnostic values reported here are means.

See footnote to Table 1 for statistic types reported in table.

we choose in the first instance not to weight the kappa values by subject number. However, to investigate the difference this would make, we also calculated the weighted kappa values, based on the data which was available (although only for diagnostic level agreement and not for item-level agreement, where a further complication was the varied number of individual OPCRIT items the studies reported values for). Several guides are available as to the interpretation of kappa but we use that proposed by Landis and Koch (1977): < 0, poor agreement; 0.0–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; 0.81–1.00, almost perfect agreement. A number of studies provided supporting values (e.g. sensitivity) and these are summarized where available.

Results

Results for the five categories of reliability and validity, in terms of average kappa values, are summarized in Figure 1 (individual studies are listed in Tables 1–5).

Category 1. Studies reporting on the agreement between raters when using OPCRIT, both at an individual item and diagnostic level

Item level: 14 studies were found reporting agreement statistics at the level of the individual OPCRIT item, all of which used kappa (Allan *et al.*, 2009; Allardyce *et al.*, 2007; Cardno *et al.*, 1998, 1999, 2001, 2002, 2008; Castle *et al.*, 2006; Matsuura *et al.*, 2004; McGuffin *et al.*, 1991; Mihalopoulos *et al.*, 2000; Rossi *et al.*, 2010; Rucker *et al.*, 2011; Skorven and Mikalsen, 2010). However, only six of these reported values in a way which allowed the calculation of a mean (Allan *et al.*, 2009; Allardyce *et al.*, 2007; Cardno *et al.*, 2002; McGuffin *et al.*, 1991; Mihalopoulos *et al.*, 2000; Rucker *et al.*, 2011); typically, in the other studies, values were reported as all being below a certain threshold or between a certain range. The mean kappa of these six studies was 0.69 (range: 0.48–0.84, number of subjects = 189) indicating “substantial agreement”. For the remainder, we conservatively categorized each study’s values into one of Landis and Koch’s bandings based on the values which were reported. Following this approach, two studies reported “moderate agreement” (Cardno *et al.*, 2001, 2008), four reported “substantial agreement” (Cardno *et al.*, 1998, 1999; Castle *et al.*, 2006; Skorven and Mikalsen, 2010) and two reported “almost perfect agreement” (Matsuura *et al.*, 2004; Rossi *et al.*, 2010). Two studies reported values for the proportions of positive and negative agreement (ppos/pneg), the means of which were, respectively, 80%

Figure 1. The reliability and validity of OPCRIT as represented by average kappa bandings from 32 studies. X axis categorized into: Category 1, Agreement between raters; Category 2, Agreement between OPCRIT and clinical diagnoses; Category 3, Agreement between OPCRIT diagnoses and other diagnostic (research) methods; Category 4, Agreement between OPCRIT diagnoses on the same patients but based on different information sources; Category 5, Agreement between repeated OPCRIT ratings on the same patients. Y axis categorized into Landis and Koch’s (1977) proposed “agreement” bandings. Where applicable, ratings are split into diagnostic (solid line) and item level (dashed line) agreement. All values are means except item-level agreement for Category 5 which is an approximation. Values represent a mix of classificatory systems (i.e. DSM, ICD and RDC).

and 85% of items at or above 0.80 (Rossi *et al.*, 2010; Skorven and Mikalsen, 2010).

Diagnostic level: 17 studies were found reporting agreement statistics at the diagnostic level (Allardyce *et al.*, 2000, 2001; Balestrieri *et al.*, 1997; Boydell *et al.*, 2003; Castle *et al.*, 1991, 2006; Davies *et al.*, 1995; Farmer *et al.*, 1992; Matsuura *et al.*, 2004; McConville and Walker, 2000; McGuffin *et al.*, 1991; Rossi *et al.*, 2010; Rucker *et al.*, 2011; Skorven and Mikalsen, 2010; Van Os *et al.*, 1996; Vares *et al.*, 2006; Williams *et al.*, 1996) only two of which did not report kappa (Davies *et al.*, 1995; McConville and Walker, 2000). Fourteen of these studies produced a mean kappa of

0.76 (range: 0.64–0.87, $n \geq 945$) indicating “substantial agreement”, however, this included all classificatory systems. When considered separately the mean kappa values were: ICD-10, 0.78 (Allardyce *et al.*, 2000; Balestrieri *et al.*, 1997; Castle *et al.*, 2006; Rossi *et al.*, 2010; Skorven and Mikalsen, 2010; Williams *et al.*, 1996); DSM-IV, 0.69 (Rossi *et al.*, 2010; Skorven and Mikalsen, 2010); DSM-III-R, 0.69 (Castle *et al.*, 2006; McGuffin *et al.*, 1991; Vares *et al.*, 2006; Williams *et al.*, 1996); DSM-III, 0.78 (Castle *et al.*, 1991; Farmer *et al.*, 1992; McGuffin *et al.*, 1991; Williams *et al.*, 1996); RDC, 0.77 (Allardyce *et al.*, 2001; Boydell *et al.*, 2003; Castle *et al.*, 1991; Farmer *et al.*, 1992; McGuffin *et al.*, 1991; Van Os *et al.*, 1996; Williams *et al.*, 1996), values all also falling within the “substantial agreement” banding. The three remaining studies reported values supporting at least this level of agreement (Davies *et al.*, 1995; Matsuura *et al.*, 2004; McConville and Walker, 2000). One study reported several ppos and pneg values, the means of which were 0.87 and 0.81, respectively (Skorven and Mikalsen, 2010). Finally, four studies suggested a marginal improvement in kappa when broad (mean k : 0.74) versus narrow (mean k : 0.68) diagnostic categories were used (Castle *et al.*, 2006; Rossi *et al.*, 2010; Rucker *et al.*, 2011; Skorven and Mikalsen, 2010).

Category 2. Studies reporting on the agreement between OPCRIT and clinical diagnoses

The majority of studies reported values either in terms of concordance rates or percentage agreement. We converted the former into the latter and across 12 studies (Balestrieri *et al.*, 1997; Davies *et al.*, 1995; Forrester *et al.*, 2001; Humphreys and Ogilvie, 1996; Jakobsen *et al.*, 2007; Kessing, 1998; MacCabe *et al.*, 2002; Pihlajamaa *et al.*, 2008; Reutfors *et al.*, 2010; Van Os *et al.*, 1996; Vogels *et al.*, 2004; Vollmer-Larsen *et al.*, 2006) mean agreement was 68.9% (range 6.8–100%). Considering separate classificatory systems, mean agreement rates were: ICD-10, 68.3% (Balestrieri *et al.*, 1997; Davies *et al.*, 1995; Forrester *et al.*, 2001; Humphreys and Ogilvie, 1996; Jakobsen *et al.*, 2007; Kessing, 1998; MacCabe *et al.*, 2002; Pihlajamaa *et al.*, 2008; Reutfors *et al.*, 2010; Vogels *et al.*, 2004; Vollmer-Larsen *et al.*, 2006); DSM-IV, 73.5% (Pihlajamaa *et al.*, 2008; Reutfors *et al.*, 2010; Vogels *et al.*, 2004); DSM-III-R, 60.8% (Davies *et al.*, 1995; Forrester *et al.*, 2001; Humphreys and Ogilvie, 1996; Pihlajamaa *et al.*, 2008); DSM-III, 53% (Davies *et al.*, 1995); RDC, 77.5% (Davies *et al.*, 1995; MacCabe *et al.*, 2002; Van Os *et al.*, 1996).

It is important to consider the classificatory systems of the clinical diagnoses against which the OPCRIT diagnoses were compared. Of the 12 studies, eight reported ICD but in four, no mention was made.

Five studies were found reporting kappa (Ekholm *et al.*, 2005; Jakobsen *et al.*, 2005, 2008; McConville and Walker, 2000; Vares *et al.*, 2006) producing a mean of 0.39 (range 0.24–0.6, $n = 614$) indicating “fair agreement”. Two of these studies provided sensitivity and specificity values the means of which were 0.76 and 0.72, respectively (McConville and Walker, 2000; Vares *et al.*, 2006). One study reported several values for positive and negative predictive power (ppp/npp), the means of which were 0.94 and 0.22, respectively (Vares *et al.*, 2006).

Two other studies reported statistics other than kappa or percentage agreement. One of these indicated low-levels of agreement (Allardyce *et al.*, 2000) whilst the other suggested high-levels (Matsuura *et al.*, 2004).

None of the earlier studies used the explicit terms of “narrow” and “broad” diagnostic groupings but by defining “narrow” as those studies reporting values for single disorders (e.g. schizophrenia), against those reporting values for aggregate groupings (e.g. “schizophrenia spectrum disorders”), we see a considerably better agreement rate between OPCRIT and “broad” clinical diagnostic groupings (84.2%; Balestrieri *et al.*, 1997; Forrester *et al.*, 2001; Jakobsen *et al.*, 2007; Kessing, 1998; Pihlajamaa *et al.*, 2008; Van Os *et al.*, 1996; Vogels *et al.*, 2004) than when considering “narrow” diagnoses (59.4%; Davies *et al.*, 1995; Humphreys and Ogilvie, 1996; Jakobsen *et al.*, 2007; MacCabe *et al.*, 2002; Pihlajamaa *et al.*, 2008; Reutfors *et al.*, 2010; Vollmer-Larsen *et al.*, 2006).

Finally, two studies indicated improved agreement when OPCRIT diagnoses were compared with “lifetime” versus “first contact” clinical diagnoses (Humphreys and Ogilvie, 1996; Isohanni *et al.*, 1997).

Category 3. Studies reporting on the agreement between OPCRIT diagnoses and other diagnostic methods, including other research and consensus methods

The mean kappa across seven studies (Craddock *et al.*, 1996; Ekholm *et al.*, 2005; Isohanni *et al.*, 1997; Lazartigues *et al.*, 1997; Makikyro *et al.*, 1998; Mihalopoulos *et al.*, 2000; Vares *et al.*, 2006) was 0.60 (range 0.27–0.89, $n = 605$), indicating agreement at the top-end of the “moderate” banding. The most commonly used classificatory system in these studies was DSM.

Three of these studies reported sensitivity and specificity values, producing means of 0.87 and 0.88, respectively (Isohanni *et al.*, 1997; Makikyro *et al.*, 1998; Vares *et al.*, 2006), whilst two reported ppp and npp values producing means of 0.97 and 0.74, respectively (Makikyro *et al.*, 1998; Vares *et al.*, 2006).

Four other studies (Castle *et al.*, 2006; Gil *et al.*, 2009; Rossi *et al.*, 2010; Vollmer-Larsen *et al.*, 2006) reported concordance rates producing a mean percentage agreement rate of 80%. One other study merely gave a verbal report suggesting good agreement (Meszaros *et al.*, 1996).

Category 4. Studies reporting on the agreement between repeated OPCRIT diagnoses made on the same patients but when based on different information sources

The mean kappa value from four studies (Azevedo *et al.*, 1999; Ekholm *et al.*, 2005; Mihalopoulos *et al.*, 2000; Vares *et al.*, 2006) was 0.72 (range 0.32–0.88, $n = 392$) indicating “substantial” agreement. One study reported several values for sensitivity and specificity the means of which were 0.98 and 1.00, respectively (Vares *et al.*, 2006). The same study reported several values for ppp and npp, the means of which were 1.00 and 0.75, respectively.

Category 5. Studies reporting on the agreement of repeated OPCRIT ratings on the same patients, both at an individual item and diagnostic level

Three studies were found reporting such statistics (Castle *et al.*, 2006; Rossi *et al.*, 2010; Skorven and Mikalsen, 2010). At the level of the individual item, kappa values were not reported in a way allowing the calculation of a mean, thus, we categorized them as before, leading one study (Skorven and Mikalsen, 2010) to report “fair” and one to report “substantial” agreement (Castle *et al.*, 2006). The Rossi *et al.* (2010) study reported item values twice, once considering the associated “durations” of the symptom items and once not. The former resulted in “fair” and the latter in “moderate” agreement. Two of the studies (Rossi *et al.*, 2010; Skorven and Mikalsen, 2010) together indicated that ppos and pneg values for the majority of items were equal or greater than 0.80.

At the diagnostic level, the mean kappa was 0.51 (range 0.26–0.66, $n = 65$), suggesting “moderate agreement”, with one study reporting several values for ppos and pneg the means of which were 0.81 and 0.78, respectively (Skorven and Mikalsen, 2010).

Weighted kappa values

For all categories, except Category 5, mean kappa values weighted by subject number were marginally higher. Category 1: 0.78, Category 2: 0.40, Category 3: 0.65, Category 4: 0.79, Category 5: 0.39.

Discussion

In preparation for an introduction within clinical services of an expanded version of the OPCRIT program, we reviewed 44 studies which had used the tool and had reported measurements concerning its reliability or validity. We separated the values into five different types of agreement and produced summary measures, primarily with mean kappa values. We interpreted these values according to the levels of agreement proposed by Landis and Koch (1977) which, in ascending order are; “poor”, “slight”, “fair”, “moderate”, “substantial” and “almost perfect”. We will now summarize the levels of agreement found for each category (see also Figure 1).

For Category 1 (inter-rater reliability using the OPCRIT), agreement was best described as “substantial”, with a marginally improved value for diagnostic (0.76) versus item-level agreement (0.69). For Category 2 (agreement between OPCRIT and clinical diagnoses) the mean percentage agreement figure was 68.9%, however, this is unadjusted for agreement by chance. A “maximum possible” kappa can be calculated from this figure and this was 0.43, very close to the actual mean kappa value of 0.39 from the five studies which did report this measure. This reflects agreement at the top-end of the “fair” banding. For Category 3 (agreement between OPCRIT and other diagnostic methods) agreement was at the top-end of the “moderate” banding (0.60). For Category 4 (agreement between repeated OPCRIT diagnoses on the same patients but based on different information sources), although agreement was “substantial” (0.72), this was based on only four studies therefore caution should be applied to this result. Finally, for Category 5 (agreement between repeated OPCRIT ratings on the same patients), item-level agreement was between “fair” and “moderate” whilst diagnostic level agreement was “moderate” (0.51). However, with only three studies found, as for Category 4, caution should be applied.

We observed in studies which reported item- and diagnostic-level agreement (Categories 1 and 5), a modest improvement in agreement at the diagnostic level. When diagnostic level agreement could be compared using “narrow” versus “broad” definitions of illness, we found an improvement in agreement for the latter (Categories 1 and 2). We also noted a high level of agreement between the three different operational criteria reported: ICD, DSM and RDC.

Agreement between OPCRIT and clinical diagnoses (Category 2), at a mean kappa of 0.39, was “fair” whereas agreement between OPCRIT and other diagnostic (research) methods, at a mean kappa of 0.60, was higher. Furthermore, several studies reviewed here reference reports of poor or moderate agreement between clinical and other research diagnoses (Makikyro *et al.*, 1998; Vares *et al.*, 2006) which

suggests that clinical diagnoses are discordant from OPCRIT as well as other research diagnoses. This might reflect a discrepancy between clinical diagnostic practices and the operationalized use of research criteria (Isohanni *et al.*, 1997; Reutfors *et al.*, 2010; Vollmer-Larsen *et al.*, 2006) and indicates that the routine use of OPCRIT+ could improve the validity of diagnoses being given in the clinic; although we will have to be aware of this potential source of disagreement during the implementation of OPCRIT+ into the clinical environment.

We also noted a number of methodological differences between the way OPCRIT has been traditionally used in research settings and the way in which it will be employed on the hospital ward. For example, most of the studies reviewed used a retrospective method of rating OPCRIT, often based on case records. We will be using a consecutive completion method whereby the OPCRIT+ is completed shortly after, and based upon, the “clerking” of a patient when they enter a ward, a process which includes a mental state assessment. A further difference is that OPCRIT has typically been rated on the “life-time ever” existence of symptoms, whereas in this project they will be completed on the “present episode” of illness only (enabling us to build a standardized, longitudinal record of illness history).

There were a number of methodological challenges we encountered during this review. Firstly, the varied manner in which the studies calculated and reported their agreement statistics made the process of producing summary variables challenging and prevented us from undertaking a full meta-analysis. Secondly, using kappa as a measure of agreement is also considered problematic. Criticisms of it have included the fact that it is not informative concerning the source of disagreement (Makikyro *et al.*, 1998) and that it can be unstable when data distributions are skewed (e.g. a non-prevalent symptom or disorder; Azevedo *et al.*, 1999; McGuffin *et al.*, 1991). Therefore, various authors have suggested that it should be accompanied by supporting statistics, including: sensitivity, specificity, ppp and npp and the ppos and pneg agreement. However, only six of the studies reviewed here presented any of these, four of which reported values equal to or greater than their accompanying kappa's (Isohanni *et al.*, 1997; Makikyro *et al.*, 1998; McConville and Walker, 2000; Rossi *et al.*, 2010).

In conclusion, the systematic and comprehensive assessment of symptoms in routine clinical care, via OPCRIT+, has the potential to improve clinical outcomes, may contribute to the training of junior doctors and nurses and can facilitate research applications. The demonstration of “substantial” inter-rater reliability with OPCRIT use indicates its suitability for use in a routine clinical setting, where symptoms will be assessed by different (but suitably trained) medical personnel. The “fair” agreement between OPCRIT and clinical diagnoses can be attributed to differences in the application of diagnoses in clinical versus research settings. This is supported by an improved (“moderate”) agreement between OPCRIT and other structured diagnostic methods and suggests that routine use of the tool in the clinic may improve the validity of the clinical diagnoses being given.

Whilst the results of our review are encouraging, it will be crucial to investigate the reliability and validity of the tool's use in a routine clinical setting. Furthermore, it will be necessary, over time, to demonstrate to clinicians and management the clinical and research benefits of OPCRIT+ in order to fully embed its use.

Acknowledgements

The first, third, fourth and last authors receive salary support from the National Institute for Health Research's (NIHR) Mental Health Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health. The last author is also supported by the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behaviour in normal brain function and psychopathology) (LSHM-CT-2007-037286), the FP7 projects ADAMS (Genomic variations underlying common neuropsychiatric diseases and disease related cognitive traits in different human populations) (242257) and the Innovative Medicine Initiative Project EU-AIMS (115300-2), as well as the Medical Research Council Programme Grant “Developmental pathways into adolescent substance abuse” (93558) and the Swedish Research Council (FORMAS).

Declaration of interest statement

The authors have no competing interests.

References

Allan C.L., Cardno A.G., Rijsdijk F.V., Kalidindi S., Farmer A., Murray R.M., McGuffin P. (2009) Twin study of illness history variables in psychosis. <i>Schizophrenia Research</i>, 115(2), 237–244, DOI: 10.1016/j.schres.2009.09.002	Allardyce J., Boydell J., Van Os J., Morrison G., Castle D., Murray R.M., McCreadie R.G. (2001) Comparison of the incidence of schizophrenia in rural Dumfries and Gallo-way and urban Camberwell. <i>The British Journal of Psychiatry</i>, 179, 335–339, DOI: 10.1192/bjp.179.4.335	Allardyce J., McCreadie R.G., Morrison G., van Os J. (2007) Do symptom dimensions or categorical diagnoses best discriminate

- between known risk factors for psychosis? *Social Psychiatry and Psychiatric Epidemiology*, **42**(6), 429–437, DOI: 10.1007/s00127-007-0179-y
- Allardyce J., Morrison G., Van Os J., Kelly J., Murray R.M., McCreadie R.G. (2000) Schizophrenia is not disappearing in south-west Scotland. *The British Journal of Psychiatry*, **177**, 38–41, DOI: 10.1192/bjp.177.1.38
- Azevedo M.H., Soares M.J., Coelho I., Dourado A., Valente J., Macedo A., Pato M., Pato C. (1999) Using consensus OPCRIT diagnoses. *The British Journal of Psychiatry*, **175**, 154–157, DOI: 10.1192/bjp.175.2.154
- Balestrieri M., Rucci P., Nicolaou S. (1997) Gender-specific decline and seasonality of births in operationally defined schizophrenics in Italy. *Schizophrenia Research*, **27**(1), 73–81, DOI: 10.1016/S0920-9964(97)00071-6
- Boydell J., van Os J., Lambri M., Castle D., Allardyce J., McCreadie R.G., Murray R.M. (2003) Incidence of schizophrenia in south-east London between 1965 and 1997. *British Journal of Psychiatry*, **182**, 45–49, DOI: 10.1192/bjp.182.1.45
- Cardno A.G., Jones L.A., Murphy K.C., Sanders R.D., Asherson P., Owen M.J., McGuffin P. (1998) Sibling pairs with schizophrenia or schizoaffective disorder: associations of subtypes, symptoms and demographic variables. *Psychological Medicine*, **28**(4), 815–823.
- Cardno A.G., Jones L.A., Murphy K.C., Sanders R.D., Asherson P., Owen M.J., McGuffin P. (1999) Dimensions of psychosis in affected sibling pairs. *Schizophrenia Bulletin*, **25**(4), 841–850.
- Cardno A.G., Rijdsdijk F.V., Murray R.M., McGuffin P. (2008) Twin study refining psychotic symptom dimensions as phenotypes for genetic research. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: the Official Publication of the International Society of Psychiatric Genetics*, **147B**(7), 1213–1221, DOI: 10.1002/ajmg.b.30756
- Cardno A.G., Sham P.C., Farmer A.E., Murray R.M., McGuffin P. (2002) Heritability of Schneider's first-rank symptoms. *The British Journal of Psychiatry*, **180**, 35–38, DOI: 10.1192/bjp.180.1.35
- Cardno A.G., Sham P.C., Murray R.M., McGuffin P. (2001) Twin study of symptom dimensions in psychoses. *The British Journal of Psychiatry*, **179**, 39–45, DOI: 10.1192/bjp.179.1.39
- Castle D., Wessely S., Der G., Murray R.M. (1991) The incidence of operationally defined schizophrenia in Camberwell, 1965–84. *The British Journal of Psychiatry*, **159**, 790–794, DOI: 10.1192/bjp.159.6.790
- Castle D.J., Jablensky A., McGrath J.J., Carr V., Morgan V., Waterreus A., Valuri G., Stain H., McGuffin P., Farmer A. (2006) The diagnostic interview for psychoses (DIP): development, reliability and applications. *Psychological Medicine*, **36**(1), 69–80, DOI: 10.1017/S0033291705005969
- Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46, DOI: 10.1177/001316446002000104
- Craddock N., Asherson P., Owen M.J., Williams J., McGuffin P., Farmer A.E. (1996) Concurrent Validity of the OPCRIT Diagnostic System. *The British Journal of Psychiatry*, **169**, 58–63, DOI: 10.1192/bjp.169.1.58
- Davies A., McLvor R.J., Kumar R.C. (1995) Impact of childbirth on a series of schizophrenic mothers: a comment on the possible influence of oestrogen on schizophrenia. *Schizophrenia Research*, **16**(1), 25–31.
- Ekholm B., Ekholm A., Adolfsson R., Vares M., Osby U., Sedvall G.C., Jonsson E.G. (2005) Evaluation of diagnostic procedures in Swedish patients with schizophrenia and related psychoses. *Nordic Journal of Psychiatry*, **59**(6), 457–464, DOI: 10.1080/08039480500360906
- Fanous A.H., Chen X., Wang X., Amdur R., O'Neill F.A., Walsh D., Kendler K.S. (2009) Genetic variation in the serotonin 2A receptor and suicidal ideation in a sample of 270 Irish high-density schizophrenia families. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: the official publication of the International Society of Psychiatric Genetics*, **150B**(3), 411–417, DOI: 10.1002/ajmg.b.30833
- Farmer A.E., Wessely S., Castle D., McGuffin P. (1992) Methodological issues in using a polydiagnostic approach to define psychotic illness. *The British Journal of Psychiatry*, **161**, 824–830, DOI: 10.1192/bjp.161.6.824
- Forrester A., Owens D.G.C., Johnstone E.C. (2001) Diagnostic stability in subjects with multiple admissions for psychotic illness. *Psychological Medicine*, **31**(1), 151–158.
- Gil A., Gama C.S., Rocha de Jesus D., Lobato M.I., Zimmer M., Belmonte-de-Abreu P. (2009) The association of child abuse and neglect with adult disability in schizophrenia and the prominent role of physical neglect. *Child Abuse & Neglect*, **33**(9), 618–624, DOI: 10.1016/j.chiabu.2009.02.006
- Humphreys M., Ogilvie A. (1996) Feigned psychosis revisited – a 20 year follow up of 10 patients. *Psychiatric Bulletin*, **20**, 666–669.
- Isohanni M., Makikyro T., Moring J., Rasanen P., Hakko H., Partanen U., Koironen M., Jones P. (1997) A comparison of clinical and research DSM-III-R diagnoses of schizophrenia in a Finnish national birth cohort. *Social Psychiatry and Psychiatric Epidemiology*, **32**(5), 303–308, DOI: 10.1007/BF00789044
- Jakobsen K.D., Frederiksen J.N., Hansen T., Jansson L.B., Parnas J., Werge T. (2005) Reliability of clinical ICD-10 schizophrenia diagnoses. *Nordic Journal of Psychiatry*, **59**(3), 209–212, DOI: 10.1080/08039480510027698
- Jakobsen K.D., Hansen T., Dam H., Larsen E.B., Gether U., Werge T. (2008) Reliability of clinical ICD-10 diagnoses among electroconvulsive therapy patients with chronic affective disorders. *The European Journal of Psychiatry*, **22**(3), 161–172, DOI: 10.1080/08039480510027698
- Jakobsen K.D., Hansen T., Werge T. (2007) Diagnostic stability among chronic patients with functional psychoses: an epidemiological and clinical study. *BMC Psychiatry*, **7**, 41, DOI: 10.1186/1471-244X-7-41
- Kessing L.V. (1998) Validity of diagnoses and other clinical register data in patients with affective disorder. *European Psychiatry*, **13**(8), 392–398.
- Landis J.R., Koch G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.
- Lazartigues A., Barua U., Garcia-Orad I., Genest Ph., Fermanian J. (1997) A comparative study of PSE-10/CATEGO-5, opcrit-checklist/quick-basic: two psychiatric computer assisted diagnostic systems. *European Psychiatry*, **12**, 168s.
- MacCabe J.H., Aldouri E., Fahy T.A., Sham P.C., Murray R.M. (2002) Do schizophrenic patients who managed to get to university have a non-developmental form of illness? *Psychological Medicine*, **32**, 535–544, DOI: 10.1017/S0033291702005391
- Makikyro T., Isohanni M., Moring J., Hakko H., Hovatta I., Lonnqvist J. (1998) Accuracy of register-based schizophrenia diagnoses in a genetic study. *European Psychiatry*, **13**(2), 57–62, DOI: 10.1016/S0924-9338(98)80019-9
- Matsuura M., Adachi N., Oana Y., Okubo Y., Kato M., Nakano T., Takei N. (2004) A polydiagnostic and dimensional comparison of epileptic psychoses and schizophrenia spectrum disorders. *Schizophrenia Research*, **69**(2–3), 189–201, DOI: 10.1016/S0920-9964(02)00492-9
- McConville P., Walker N.P. (2000) The reliability of case register diagnoses: a birth cohort analysis. *Social Psychiatry and Psychiatric Epidemiology*, **35**(3), 121–127, DOI: 10.1007/s001270050194

- McGuffin P., Farmer A., Harvey I. (1991) A polydiagnostic application of operational criteria in studies of psychotic illness. *Archives of General Psychiatry*, **48**(8), 764–770.
- Meszaros K., Lenzinger E., Furedi T., Hornik K., Willinger U., Isenberg K.E., Todd R.D., Cloninger C.R., Reich T., Fuchs K., Sieghart W., Aschauer H.N. (1996) The influence of phenotype on the outcome of linkage analysis of schizophrenia. *Schizophrenia Research*, **22**(1), 89–90.
- Mihalopoulos C., McGorry P., Roberts S., McFarlane C. (2000) The procedural validity of retrospective case note diagnosis. *The Australian and New Zealand Journal of Psychiatry*, **34**(1), 154–159, DOI: 10.1046/j.1440-1614.2000.00649.x
- Pihlajamaa J., Suvisaari J., Henriksson M., Heila H., Karjalainen E., Koskela J., Cannon M., Lonnqvist J. (2008) The validity of schizophrenia diagnosis in the Finnish Hospital Discharge Register: findings from a 10-year birth cohort sample. *Nordic Journal of Psychiatry*, **62**(3), 198–203, DOI: 10.1080/08039480801983596
- Reutfors J., Bahmanyar S., Jonsson E.G., Ekblom A., Nordstrom P., Brandt L., Osby U. (2010) Diagnostic profile and suicide risk in schizophrenia spectrum disorder. *Schizophrenia Research*, **123** (2–3), 251–256, DOI: 10.1016/j.schres.2010.07.014
- Rossi A., Morgan V., Amadeo F., Sandri M., Grigoletti L., Maggioni F., Ferro A., Rigon E., Donisi V., Venturi V.V., Gorio F., Skre I., Tansella M., Jablensky A. (2010) Diagnosing psychotic disorders: validity, reliability and applications of the Diagnostic Interview for Psychosis (DIP). Italian version. *Epidemiologia e Psichiatria Sociale*, **19**(1), 33–43.
- Rucker J., Newman S., Gray J., Gunasinghe C., Broadbent M., Brittain P., Baggaley M., Denis M., Turp J., Stewart R., Lovestone S., Schumann G., Farmer A., McGuffin P. (2011) OPCRIT+: an electronic system for psychiatric diagnosis and data collection in clinical and research settings. *The British Journal of Psychiatry*, **199**(2), 151–155, DOI: 10.1192/bjp.bp.110.082925
- Skorven C.S., Mikalsen G. (2010) Diagnosing Bipolar Disorder with the Diagnostic Interview for Psychoses (DIP), Masters Thesis, University of Troms.
- Van Os J., Castle D.J., Takei N., Der G., Murray R. M. (1996) Psychotic illness in ethnic minorities: clarification from the 1991 census. *Psychological Medicine*, **26**(1), 203–208, DOI: <http://dx.doi.org/10.1017/S0033291700033845>
- Vares M., Ekholm A., Sedvall G.C., Hall H., Jonsson E.G. (2006) Characterization of patients with schizophrenia and related psychoses: evaluation of different diagnostic procedures. *Psychopathology*, **39**(6), 286–295, DOI: 10.1159/000095733
- Vogels A., De Hert M., Descheemaeker M.J., Govers V., Devriendt K., Legius E., Prinzie P., Fryns J.P. (2004) Psychotic disorders in Prader–Willi Syndrome. *American Journal of Medical Genetics. Part A*, **127A**(3), 238–243, DOI: 10.1002/ajmg.a.30004
- Vollmer-Larsen A., Jacobsen T.B., Hemmingsen R., Parnas J. (2006) Schizoaffective disorder – the reliability of its clinical diagnostic use. *Acta Psychiatrica Scandinavica*, **113**(5), 402–407, DOI: 10.1111/j.1600-0447.2005.00744.x
- Williams J., Farmer A.E., Ackenheil M., Kaufman C.A., McGuffin P., The OPCRIT Reliability Research Group (1996) A multicentre inter-rater reliability study using the OPCRIT computerized diagnostic system. *Psychological Medicine*, **26**(4), 775–783, DOI: <http://dx.doi.org/10.1017/S003329170003779X>