TECHNICAL COMMENTARY

Schizophrenia diagnosis

NeuRA
Discover. Conquer. Cure.
SCHIZOPHRENIA LIBRARY

Introduction

Diagnostic scales are widely used within clinical practice and research settings to ensure consistency of illness ratings. These scales have been extensively validated and provide a set of criteria that is used to define and diagnose an illness. Two key examples include the American Psychiatric Association's *Diagnostic and Statistical Manual* (DSM) and the World Health Organisation's *International Classification of Diseases* (ICD), which both provide a comprehensive definition of schizophrenia. Both the DSM and ICD criteria are regularly updated, and the most recent versions are the DSM-5 and the ICD-10.

For a DSM-5 diagnosis of schizophrenia, at least two symptoms need to have been present for at least six months, and for a significant portion of time over a one-month period. Symptoms include delusions, hallucinations, disorganized speech and behaviour, catatonic behaviour, and negative symptoms such as diminished emotional expression, poverty of speech, and lack of purposeful action. At least one symptom of delusions, hallucinations, or disorganized speech needs to be present, and there also needs to be significant social or occupational dysfunction.

For an ICD-10 diagnosis of schizophrenia, *either* at least one symptom of delusions, hallucinations, or thought symptoms (thought echo, insertion, withdrawal, or broadcasting) needs to be present, *or* at least two symptoms of hallucinations, negative symptoms, catatonic behaviour, or incoherent/irrelevant speech needs to be present for most of the time for at least one month.

Method

We have included only systematic reviews (systematic literature search, detailed methodology with inclusion/exclusion criteria) published in full text, in English, from the year 2000 that report results separately for people with a diagnosis of schizophrenia, schizoaffective disorder, schizophreniform disorder or first episode schizophrenia. Reviews were identified by searching the databases MEDLINE, EMBASE, CINAHL, Current Contents, PsycINFO and the Cochrane library. Hand searching reference lists of identified reviews was also conducted. When multiple copies of reviews were found, only the most recent version was included. Reviews with pooled data are given priority for inclusion.

Review reporting assessment was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist that describes a preferred way to present a meta-analysis[1]. Reviews with less than 50% of items checked have been excluded from the library. The PRISMA flow diagram is a suggested way of providing information about studies included and excluded with reasons for exclusion. Where no flow diagram has been presented by individual reviews, but identified studies have been described in the text, reviews have been checked for this item. Note that early reviews may have been guided by less stringent reporting checklists than the PRISMA, and that some reviews may have been limited by journal guidelines.

Evidence was graded using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group approach where high quality evidence such as that gained from randomised controlled trials (RCT) may be downgraded to moderate or low if review and study quality is limited, if there is inconsistency in results, indirect comparisons, imprecise or sparse data and high probability of reporting bias. It may also be downgraded if risks associated with the intervention or other matter under review are high. Conversely, low quality evidence such as that gained from observational studies may be upgraded if effect sizes are large, there is a dose dependent response or if results are reasonably consistent, precise and direct with low associated risks (see end of table for an

Margarete Ainsworth Building, Barker Street, Randwick NSW 2031. Phone: 02 9399 1000. Email: info@neura.edu.au

explanation of these terms)[2]. The resulting table represents an objective summary of the available evidence, although the conclusions are solely the opinion of staff of NeuRA (Neuroscience Research Australia).

Results

We found 14 systematic reviews that met our inclusion criteria[3-16].

- Moderate to high quality evidence suggests the DSM-III, DSM-III-R or DSM-IV diagnostic criteria assigns significantly more males with psychosis to a schizophrenia diagnosis, rather than any other psychosis diagnosis, compared to the ICD-9, which shows no differences in gender distribution across diagnoses.
- Moderate quality evidence suggests moderate predictive value and good kappa agreement for schizophrenia and schizophrenia spectrum diagnoses, good predictive value and kappa agreement for psychotic disorder, but poor predictive value and kappa agreement for schizoaffective disorder.
- Moderate to high quality evidence suggests better test-retest reliability for a diagnosis of schizophrenia than for a diagnosis of schizoaffective disorder, but lower test-retest reliability than for a diagnosis of bipolar disorder or unipolar depression.
- Moderate to high quality evidence suggests poorer interrater reliability for a diagnosis of schizoaffective disorder than for a diagnosis of schizophrenia, bipolar disorder, or unipolar depression.
- Moderate quality evidence suggests excellent specificity (91%) and good sensitivity (71%) of latent semantic analysis in recognising a diagnosis of schizophrenia. Latent semantic analysis quantifies the degree of incoherence in language use.
- Moderate quality evidence suggests support vector machines combined with other machine learning techniques applied to neuroimaging data may assist with clinical diagnosis of schizophrenia.
- Moderate to high quality evidence suggests a significant association of increased deficit schizophrenia (predominately negative symptoms) vs. non-deficit schizophrenia in males, a finding that was not moderated by sampling method, diagnosis criteria, or duration of illness.
- Moderate quality evidence suggests Black people in the United States are more likely to be diagnosed with schizophrenia than White people in the United States. The effect size was largest in studies with more males, more White patients, more young patients, studies in hospital or military settings, and studies conducted in the Midwest, Southeast, National, or multistate USA. There were no differences in risk according to diagnostic method (structured vs. unstructured), DSM version (DSM-III or DSM-IV), or study year.
- Moderate to high quality evidence suggests the proportion of first-episode psychosis patients retaining a diagnosis of schizophrenia over time is around 90% and 72% for schizoaffective disorder. Diagnostic stability for schizophreniform disorder is 29% over time.
- Moderate quality evidence suggests the rate of a schizophrenia diagnosis following a diagnosis of schizophreniform disorder is around 65% (over 4 years). Following brief, atypical, or not otherwise specified psychoses, the rate of a schizophrenia diagnosis is around 36%. Following a substance-induced psychosis, the rate of a schizophrenia diagnosis is around 25%. The rates were highest for cannabis, hallucinogens, and amphetamines, and lowest for opioids, alcohol, and sedatives.

Margarete Ainsworth Building, Barker Street, Randwick NSW 2031. Phone: 02 9399 1000. Email: info@neura.edu.au

Beauchamp G, Gagnon A

Influence of diagnostic classification on gender ratio in schizophrenia

Social Psychiatry and Psychiatric Epidemiology 2004; 39: 1017–1022

View review abstract online

Comparison	**Assessment of the gender distribution in schizophrenia vs. other psychoses, depending on the diagnostic tool used.**
Summary of evidence	**Moderate to high quality evidence (consistent, imprecise, direct, large sample) suggests the DSM-III, DSM-III-R or DSM-IV diagnostic criteria assigns significantly more males with psychosis to a schizophrenia diagnosis, rather than any other psychosis diagnosis, compared to the ICD-9, which shows no differences in gender distribution across diagnoses.**
Gender distribution	
A male with psychosis is significantly more likely to obtain a diagnosis of schizophrenia than a female with psychosis. 12 studies, N = 827, OR† = 1.70, 95%CI 1.27 to 2.30, $p = 0.0003$, $Q_w = 11.1$, $p = 0.43$ However, this finding depends on the diagnostic tool used as when the ICD-9 and DSM were analysed separately, only the DSM criteria assigned significantly greater proportions of males to a schizophrenia diagnosis. ICD-9: OR = 1.13, 95%CI 0.69 to 1.85 DSM: OR = 2.41, 95%CI 1.60 to 3.61 $Q_B = 5.7$, $p = 0.017$	
Consistency in results‡	Consistent
Precision in results§	Imprecise
Directness of results‖	Direct

Chang WC, Chan SSM, Chung DWS

Diagnostic Stability of Functional Psychosis: a Systematic Review

Hong Kong Journal of Psychiatry 2009; 19: 30-41

View review abstract online	
Comparison	**Assessment of the stability over time of schizophrenia diagnosis (the degree to which the diagnosis remains the same over subsequent clinical evaluations).** **Note: this review also reports on other psychoses, the results here are only reported for schizophrenia.**
Summary of evidence	**Moderate quality evidence (unable to assess consistency or precision, direct, large sample) suggests a schizophrenia diagnosis is associated with reasonably high stability over subsequent evaluations (>70% consistency).**
Diagnostic stability	
12/14 studies reported over 70% consistency of a first-episode schizophrenia diagnosis, with 9/12 reporting over 90% consistency. 11/14 studies reported over 70% consistency of a schizophrenia diagnosis. The remaining 3 studies reported considerable shift between schizophrenia, bipolar and schizoaffective disorder diagnoses.	
Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Unable to assess; no measure of precision is reported.
Directness of results	Direct

Davis KAS, Sudlow CLM, Hotopf M

Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses

BMC Psychiatry 2016; 16: 263

View review abstract online

Comparison	**Diagnostic integrity in administrative databases using ICD-10 vs. reference comparison (e.g. clinical chart or research diagnosis).**
Summary of evidence	**Moderate quality evidence (unable to assess consistency, appears imprecise, direct, large samples) suggests moderate predictive value and kappa agreement for schizophrenia and schizophrenia spectrum diagnoses, good predictive value and kappa agreement for psychotic disorder, but poor predictive value and kappa agreement for schizoaffective disorder.**

Diagnostic integrity	
Moderate predictive value and kappa agreement for schizophrenia and schizophrenia spectrum diagnoses, good predictive value and kappa agreement for psychotic disorder, but poor predictive value and kappa agreement for schizoaffective disorder; Schizophrenia: 18 studies, N = 5,016, median PPV ~75% (range 40-100%), Kappa ~0.45 (range 0-72%). Schizophrenia spectrum (any non-affective psychosis, including schizophrenia): 13 studies, N = 2,662, median PPV ~85% (range 55-95%), Kappa ~0.45 (range 20-65%). Psychotic disorder: 4 studies, N = 1,561, median PPV ~90% (range 85-100%), Kappa ~0.80 (range 35-90%). Schizoaffective disorder: 5 studies, N = 823, median PPV ~50% (range 10-58%), Kappa ~0.38 (range 10-45%).	
Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Appears imprecise
Directness of results	Direct

de Boer JN, Voppel AE, Begemann MJH, Schnack HG, Wijnen F, Sommer IEC

Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis

Neuroscience and Biobehavioral Reviews 2018; 93: 85-92

View review abstract online

Comparison	**Specificity and sensitivity of semantic space calculations for recognising people with a diagnosis of schizophrenia vs. people without a clinical diagnosis.** **Semantic space calculations quantify the degree of incoherence in spoken language using latent semantic analysis.** **Most of the sample included people with schizophrenia, and some studies included people with first-episode psychosis.**
Summary of evidence	**Moderate quality evidence (medium-sized sample, unable to assess consistency, appears precise, direct) suggests excellent specificity (91%) and good sensitivity (71%).**
Specificity and sensitivity	

Excellent specificity and good sensitivity; 4 studies, N = 298, specificity = 91%, 95%CI 0.87% to 0.93% 4 studies, N = 298, sensitivity = 71%, 95%CI 0.63% to 0.78%	
Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Appears precise
Directness of results	Direct

De Filippis R, Carbone EA, Gaetano R, Bruni A, Pugliese V, Segura-Garcia C, De Fazio P

Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: A systematic review

Neuropsychiatric Disease and Treatment 2019; 15: 1605-27

View review abstract online

Comparison	**Assessment of machine learning techniques on functional and structural neuroimaging data as a diagnostic tool for schizophrenia.** **Machine learning techniques are most useful when analysing highly complex datasets. They include support vector machines which is a high-dimensional, pattern recognition, supervised learning algorithm. Other techniques include multivariate pattern analysis and random forest analysis.**
Summary of evidence	**Moderate quality evidence (large samples, unable to assess consistency, appears imprecise, direct) suggests support vector machines combined with other machine learning techniques applied to neuroimaging data may assist the clinical evaluation of schizophrenia.**
Accuracy	
27 functional neuroimaging studies (N = 4,137), accuracy ranged from 41% to 99% 9 structural neuroimaging studies (N = 1,767), accuracy ranged from 70% to 88% Authors reported that accuracy was greatest in studies using support vector machine plus other machine learning techniques, and that the prefrontal and temporal cortices appeared to be the most useful brain regions for diagnosis.	
Consistency in results	Unable to assess; no measure of consistency is reported.

Precision in results	Appears imprecise for the functional studies.
Directness of results	Direct

Fusar-Poli P, Cappucciati M, Rutigliano G, Heslin M, Stahl D, Brittenden Z, Caverzasi E, McGuire P, Carpenter WT

Diagnostic Stability of ICD/DSM First Episode Psychosis Diagnoses: Meta-analysis

Schizophrenia Bulletin 2016; 42(6): 1395-406

View review abstract online

Comparison	**Diagnostic stability in people with first-episode psychosis using ICD or DSM over 4-5 years.**
Summary of evidence	**Moderate to high quality evidence (unable to assess consistency, precise, direct, large sample) suggests the proportion of first-episode psychosis patients retaining the diagnosis of schizophrenia over time is around 90% and 72% for schizoaffective disorder. Diagnositic stability for schizophreniform disorder is 29% over time.**

Diagnostic stability

42 studies N = 14 484, follow-up average 4.5 years

The proportion of first-episode psychosis patients retaining the diagnosis of schizophrenia over time is around 90% and 72% for schizoaffective disorder, but only 29% for schizophreniform disorder;

Schizophrenia: 0.90, 95%CI 0.85 to 0.95

Schizoaffective disorder: 0.72, 95%CI 0.61 to 0.73

Schizophreniform disorder: 0.29, 95%CI 0.22 to 0.38

Other diagnoses;

All affective spectrum psychoses: 0.84, 95%CI 0.79 to 0.89

Substance-induced psychotic disorder: 0.66, 95%CI 0.51 to 0.81

Delusional disorder: 0.59, 95%CI 0.47 to 0.71

Acute and transient psychotic disorder/brief psychotic disorder: 0.56, 95%CI 0.62 to 0.60

Psychosis not otherwise specified: 0.36, 95%CI 0.27 to 0.45

About 10% of patients with diagnosis of affective spectrum psychoses changed to schizophrenia spectrum psychoses over time.

Moderator analyses revealed more recent studies and studies of inpatient (vs. mixed) settings had

higher diagnostic instability over time. There was no effect on diagnostic stability according to patients' age, gender, substance abuse, baseline ICD-10 or DSM-IV diagnostic criteria, baseline functioning level and duration of follow-up.

Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Precise
Directness of results	Direct

Linscott RJ, Allardyce J, van Os J

Seeking verisimilitude in a class: a systematic review of evidence that the criterial clinical symptoms of schizophrenia are taxonic.

Schizophrenia Bulletin 2010; 26(4): 811-829

View review abstract online

Comparison	**Assessment of the latent structure of the schizophrenia diagnosis: if the symptoms of schizophrenia can be organised into discrete (latent) classes vs. a dimensional (continuum) scale.**
Summary of evidence	**Low quality evidence (unclear sample size, appears inconsistent, unable to assess precision, direct) is unclear as to any latent structure underlying the schizophrenia diagnosis.**

Latent classes
There was no consistent evidence from 24 studies to support a two-, three-, four-, five-, or six-class interpretation of a latent class structure of schizophrenia symptoms (e.g. positive, negative, disorganisation, affective). Fourteen analyses identified a two-class structure of schizophrenia; five analyses identified a three-class structure; six analyses identified a four-class structure; three analyses identified five classes; and two studies identified six latent classes.

Consistency in results	Appears imprecise.
Precision in results	Unable to assess; no measure of precision is reported.
Directness of results	Direct

Menezes N, Milovan E

First-episode psychosis: a comparative review of diagnostic evolution and predictive variables in adolescents versus adults

Canadian Journal of Psychiatry 2000; 45: 710-716

View review abstract online

Comparison	**Assessment of the diagnostic stability over time of first-episode psychosis in adolescents and adults (the degree to which the diagnosis remains the same over subsequent clinical evaluations).**
Summary of evidence	**Low quality evidence (unclear sample size, unable to assess consistency or precision, direct) is unable to ascertain any differences in stability of diagnosis of psychosis in adults compared to adolescents.**
Diagnostic stability	
Four studies reported higher rates of misdiagnosis in adolescents compared to adults, and an overall high rate of diagnostic shift for psychosis. Two studies report a comparative stability of the schizophrenia diagnosis.	
Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Unable to assess; no measure of precision is reported.
Directness of results	Direct

Murrie B, Lappin J, Large M, Sara G

Transition of Substance-Induced, Brief, and Atypical Psychoses to Schizophrenia: A Systematic Review and Meta-analysis

Schizophrenia Bulletin 2019; 46(5): 505-516

View review abstract online

Comparison	**Diagnosis of schizophrenia following a diagnosis of schizophreniform, brief, atypical, not otherwise specified (NOS), or substance-induced psychoses.**
Summary of evidence	**Moderate quality evidence (large samples, inconsistent, mostly precise, direct) suggests the rate of a schizophrenia diagnosis**

	following a diagnosis of schizophreniform disorder is around 65%. Following brief, atypical, or NOS psychoses, the rate of a schizophrenia diagnosis is around 36%. Following a substance-induced psychosis, the rate of a schizophrenia diagnosis is around 25%. The rates were highest for cannabis (34%), hallucinogens (26%), and amphetamines (22%), and lowest for opioids (12%), alcohol (10%), and sedatives (9%).
Diagnosis of schizophrenia	
Diagnosis of schizophrenia at follow-up; Schizophreniform: 20 estimates, N = 590, rate = 65%, 95%CI 57% to 72%, I^2 = 54% Brief, atypical and NOS: 34 estimates, N = 5,969, rate = 36%, 95%CI 30% to 43%, I^2 = 92% Substance induced: 25 estimates, N = 34,244, rate = 25%, 95%CI 18% to 35%, I^2 = 99% The rates were highest for cannabis (34%), hallucinogens (26%), and amphetamines (22%), and lowest for opioids (12%), alcohol (10%), and sedatives (9%). Mean follow-up period was 4 years.	
Consistency in results	Inconsistent
Precision in results	Appears mostly precise.
Directness of results	Direct

Olbert CM, Nagendra A, Buck B

Meta-analysis of black vs. white racial disparity in schizophrenia diagnosis in the United States: Do structured assessments attenuate racial disparities?

Journal of Abnormal Psychology 2018; 127: 104-15

View review abstract online

Comparison	**Racial disparity (Black vs. White) in the diagnosis of schizophrenia.** **Studies primarily used the DSM-III or DSM-IV for diagnosis of schizophrenia.**
Summary of evidence	**Moderate quality evidence (large sample, inconsistent, imprecise, direct) suggests Black people in the United States are more likely to be diagnosed with schizophrenia than White people in the United States. The effect size was largest in studies with more males, more White patients, more young patients, studies in hospital or military settings, and studies conducted in**

	the Midwest, Southeast, National, or multistate USA. There were no differences in risk according to diagnostic method (structured vs. unstructured), DSM version (DSM-III or DSM-IV), or study year.
Racial disparity	
A medium-sized, significant effect of greater rates of diagnosis of schizophrenia in Black individuals; 52 studies, N = 2,099,506, OR = 2.42, 95%CI 1.59 to 3.66, p = 0.00003, I^2 = 98%, p < 0.001 The effect size was largest in studies with more males, more White patients, more young patients, studies in hospital or military settings, and studies conducted in the Midwest, Southeast, National, or multistate USA. There were no significant moderating effects of diagnostic method (structured vs. unstructured), DSM version, or study year. Authors report there was little evidence of publication bias.	
Consistency in results	Inconsistent
Precision in results	Imprecise
Directness of results	Direct

Roy MA, Maziade M, Labbe A, Merette C

Male gender is associated with deficit schizophrenia: a meta-analysis

Schizophrenia Research 2001; 47(2-3): 141-147

View review abstract online

Comparison	**Association of male sex in deficit schizophrenia compared to non-deficit schizophrenia.** **Note: deficit schizophrenia is a descriptive rather than diagnostic term, and reflects severe negative symptoms.**
Summary of evidence	**Moderate to high quality evidence (consistent, imprecise, direct, large sample) suggests significant association of increased deficit schizophrenia in males. This effect was not moderated by sampling method, diagnosis criteria, or duration of illness.**
Association with sex	

23 studies, N = 1765 consider the association of sex with deficit schizophrenia. The sample contained 28.7% of people with deficit schizophrenia. A large effect size suggests a significant association of increased number of males with deficit schizophrenia vs. non-deficit schizophrenia. OR = 1.75, 95%CI 1.39 to 2.21, *p* = 0.000002, Q = 15.87, *p* = 0.82
Subgroup analyses No effect of sampling method was reported: no significant difference in effect size between convenience and systematic sampling, Q = 0.04, *p* = 0.86 No effect of deficit assessment method was reported: no significant difference in effect size between three assessment methods: Schedule for the Deficit Syndrome (SDS) vs Proxy Deficit Syndrome (PDS), Q = 0.32, *p* = 0.57 SDS vs detailed review (DR), Q = 0.85, *p* = 0.36 PDS vs. DR, Q = 0.85, *p* = 0.18 Unlike SDS and DR, the pooled OR for PDS alone did not yield a significant association with male gender, 3 studies, OR = 1.33, 95%CI 0.86 to 2.05 No effect of diagnosis inclusion criteria: no significant difference between schizophrenia or schizoaffective disorder, Q = 0.49, *p* = 0.49 Pooled OR for schizophrenia alone was significantly different to pooled OR including schizoaffective: OR = 1.68, 95%CI 1.29 to 2.19 No effect of duration of illness: no significant difference between duration less than or greater than 10 years, Q = 1.17, *p* = 0.28

Consistency in results	Consistent
Precision in results	Imprecise
Directness of results	Direct

Roy MA, Merette C, Maziade M

Subtyping Schizophrenia According to Outcome or Severity: A Search for Homogeneous Subgroups

Schizophrenia Bulletin 2001; 27(1): 115-138

View review abstract online

Comparison	**Validity of three potential schizophrenia diagnostic subdivisions: deficit vs. nondeficit; Kraepelinian vs. non-Kraepelinian and**

	congenital vs. adult-onset.
Summary of evidence	**Low quality evidence (unclear sample size, unable to assess consistency or precision, direct) is unclear as to the reliability of defining schizophrenia in terms of deficit vs. non-deficit, Kraepelinian vs. non-Kraepelinian, or congenital vs. adult-onset.**
Schizophrenia subgroups	
Deficit vs. non-deficit schizophrenia Deficit schizophrenia is defined by the presence of at least two negative symptoms for at least 12 months. 31 studies have compared deficit and non-deficit schizophrenia. People with deficit schizophrenia were reported to have more severe negative symptoms but less severe depressive symptoms. There was no difference between deficit and non-deficit in positive symptom severity (hallucinations or delusions). Deficit schizophrenia has been associated with poorer premorbid adjustment and social functioning, is more common in males, and has a high degree of stability over time. *Kraepelinian vs. non-Kraepelinian* Kraepelinian schizophrenia is defined by either continuous hospitalisation or complete dependence on care; lack of useful employment; and absence of complete remission. Six studies have compared Kraepelinian and non-Kraepelinian schizophrenia. Kraepelinian schizophrenia has been associated with more severe negative symptoms and more severe disorganisation, and poorer social functioning, but no difference in positive symptoms (hallucination and delusions). Kraepelinian schizophrenia has also been associated with poorer treatment response. *Congenital vs. adult-onset* Congenital schizophrenia occurrs before age 25, with poor premorbid adjustment, and a chronic course. Adult-onset schizophrenia is defined by abrupt onset at any age, without significant premorbid impairment, with acute psychotic episodes. 1 study assessed congenital vs. adult-onset schizophrenia. There was preliminary evidence that congenital psychosis was associated with familial psychosis, and adult-onset schizophrenia associated with familial depression.	
Consistency in results	Unable to assess; no measure of consistency is reported.
Precision in results	Unable to assess; no measure of precision is reported.
Directness of results	Direct

Santelmann H, Franklin J, Busshoff J, Baethge C

Interrater reliability of schizoaffective disorder compared with

schizophrenia, bipolar disorder, and unipolar depression - A systematic review and meta-analysis

Schizophrenia Research 2016; 176: 357-63

View review abstract online

Comparison	**Interrater reliability of diagnosis of schizoaffective disorder compared to schizophrenia, bipolar disorder and unipolar depression.**
Summary of evidence	**Moderate to high quality evidence (inconsistent, appears precise, direct, large sample) suggests lower interrater reliability for a diagnosis of schizoaffective disorder than for a diagnosis of schizophrenia, bipolar disorder or unipolar depression.**

Interrater reliability

25 studies, N = 7,912

Interrater reliability for schizoaffective disorder is lower than for schizophrenia, bipolar disorder and unipolar depression;

Schizoaffective disorder = Cohen's kappa = 0.57, 95%CI 0.41 to 0.73, I^2 = 98%

Schizophrenia = Cohen's kappa = 0.80, 95%CI 0.76 to 0.84, I^2 = 70%

Bipolar disorder = Cohen's kappa = 0.82, 95%CI 0.77 to 0.86, I^2 = 38%

Unipolar depression = Cohen's kappa = 0.75, 95%CI 0.70 to 0.81, I^2 = 82%

These results did not change according to diagnostic or kappa method used, sample size, number of differential diagnoses, or year of publication. There was no evidence of publication bias.

Consistency in results	Inconsistent
Precision in results	Appears precise
Directness of results	Direct

Santelmann H, Franklin J, Bushoff J, Baethge C

Test-retest reliability of schizoaffective disorder compared with schizophrenia, bipolar disorder, and unipolar depression-a systematic review and meta-analysis

Bipolar Disorders 2015; 17: 753-68

View review abstract online

Comparison	**Test-retest reliability of a schizophrenia diagnosis compared to a bipolar disorder, schizoaffective disorder or unipolar depression.**
Summary of evidence	**Moderate to high quality evidence (inconsistent, appears precise, direct, large sample) suggests better test-retest reliability for a diagnosis of schizophrenia than for a diagnosis of schizoaffective disorder, but lower than for a diagnosis of bipolar disorder or unipolar depression.**
Test-retest reliability	
49 studies, N = 14,314 *Test-retest reliability is higher for a diagnosis of schizophrenia than for a diagnosis of schizoaffective disorder, with similar test- retest reliability when comparing schizophrenia to bipolar disorder or unipolar depression;* Schizophrenia = 42 studies, Cohen's kappa = 0.69, 95%CI 0.64 to 0.74, I^2 = 90% Schizoaffective disorder = 48 studies, Cohen's kappa = 0.50, 95%CI 0.40 to 0.59, I^2 = 96% Bipolar disorder = 33 studies, Cohen's kappa = 0.77, 95%CI 0.73 to 0.82, I^2 = 92% Unipolar depression = 35 studies, Cohen's kappa = 0.73, 95%CI 0.66 to 0.79, I^2 = 91% In studies of schizophrenia, kappa was significantly higher in; blinded vs. non-blinded studies; studies using consistent vs. inconsistent use of diagnostic interview; studies with a short vs. long follow-up period (< 2 months vs. > 12 months). There were no differences in kappa according to; diagnostic tool (ICD-10 vs. DSM 111, DSM 1V or DSM 5); risk of bias in studies; similar vs. different rater identity; first-episode vs. chronic illness; or inpatient vs. outpatient status.	
Consistency in results	Inconsistent
Precision in results	Appears precise
Directness of results	Direct

Margarete Ainsworth Building, Barker Street, Randwick NSW 2031. Phone: 02 9399 1000. Email: info@neura.edu.au

Explanation of acronyms

CI = confidence interval, DSM = American Psychiatric Association's Diagnostic and Statistical Manual, I^2 = the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance), I^2 = the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance), ICD = World Health Organisation's International Classification of Diseases, N = number of participants, OR = odds ratio, *p* = statistical probability of obtaining that result ($p < 0.05$ generally regarded as significant), PDS = Proxy Deficit Syndrome (measurement tool to assess Deficit schizophrenia), PPV = Positive Predictive Value; the percentage of patients with a positive test who actually have the disease, Q = Q statistic for the test of heterogeneity, Q_w = test for within group differences (heterogeneity in study results within a group of studies – measure of study consistency), Q_B = test for between group differences (heterogeneity between groups of studies for an outcome of interest), SDS = Schedule for the Deficit Syndrome (measurement tool to assess deficit schizophrenia), vs. = versus

Margarete Ainsworth Building, Barker Street, Randwick NSW 2031. Phone: 02 9399 1000. Email: info@neura.edu.au

Explanation of technical terms

* Bias has the potential to affect reviews of both RCT and observational studies. Forms of bias include; reporting bias – selective reporting of results; publication bias - trials that are not formally published tend to show less effect than published trials, further if there are statistically significant differences between groups in a trial, these trial results tend to get published before those of trials without significant differences; language bias – only including English language reports; funding bias - source of funding for the primary research with selective reporting of results within primary studies; outcome variable selection bias; database bias - including reports from some databases and not others; citation bias - preferential citation of authors. Trials can also be subject to bias when evaluators are not blind to treatment condition and selection bias of participants if trial samples are small[17].

† Different effect measures are reported by different reviews.

Prevalence refers to how many existing cases there are at a particular point in time. Incidence refers to how many new cases there are per population in a specified time period. Incidence is usually reported as the number of new cases per 100,000 people per year. Alternatively some studies present the number of new cases that have accumulated over several years against a person-years denominator. This denominator is the sum of individual units of time that the persons in the population are at risk of becoming a case. It takes into account the size of the underlying population sample and its age structure over the duration of observation.

Reliability and validity refers to how accurate the instrument is. Sensitivity is the proportion of actual positives that are correctly identified (100% sensitivity = correct identification of all actual positives) and specificity is the proportion of negatives that are correctly identified (100% specificity = not identifying anyone as positive if they are truly not).

Weighted mean difference scores refer to mean differences between treatment and comparison groups after treatment (or occasionally pre to post treatment) and in a randomised trial there is an assumption that both groups are comparable on this measure prior to treatment. Standardised mean differences are divided by the pooled standard deviation (or the standard deviation of one group when groups are homogenous) that allows results from different scales to be combined and compared. Each study's mean difference is then given a weighting depending on the size of the sample and the variability in the data. 0.2 represents a small effect, 0.5 a moderate effect, and 0.8 and over represents a large effect[17].

Odds ratio (OR) or relative risk (RR) refers to the probability of a reduction (< 1) or an increase (> 1) in a particular outcome in a treatment group, or a group exposed to a risk factor, relative to the comparison group. For example, a RR of 0.75 translates to a reduction in risk of an outcome of 25% relative to those not receiving the treatment or not exposed to the risk factor. Conversely, a RR of 1.25 translates to an increased risk of 25% relative to those not receiving treatment or not having been exposed to a risk factor. A RR or OR of 1.00 means there is no difference between groups. A medium effect is considered if RR > 2 or < 0.5 and a large effect if RR > 5 or < 0.2[18]. lnOR stands for logarithmic OR where a lnOR of 0 shows no difference between groups. Hazard ratios measure the effect of an explanatory variable on the hazard or risk of an event.

Correlation coefficients (eg, r) indicate the strength of association or relationship between variables. They can provide an

indirect indication of prediction, but do not confirm causality due to possible and often unforseen confounding variables. An r of 0.10 represents a weak association, 0.25 a medium association and 0.40 and over represents a strong association. Unstandardised (*b*) regression coefficients indicate the average change in the dependent variable associated with a 1 unit change in the independent variable, statistically controlling for the other independent variables. Standardised regression coefficients represent the change being in units of standard deviations to allow comparison across different scales.

‡ Inconsistency refers to differing estimates of effect across studies (i.e. heterogeneity or variability in results) that is not explained by subgroup analyses and therefore reduces confidence in the effect estimate. I^2 is the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance) - 0% to 40%: heterogeneity might not be important, 30% to 60%: may represent moderate heterogeneity, 50% to 90%: may represent considerable heterogeneity and over this is considerable heterogeneity. I^2 can be calculated from Q (chi-square) for the test of heterogeneity with the following formula[17];

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100\%$$

§ Imprecision refers to wide confidence intervals indicating a lack of confidence in the effect estimate. Based on GRADE recommendations, a result for continuous data (standardised mean differences, not weighted mean differences) is considered imprecise if the upper or lower confidence limit crosses an effect size of 0.5 in either direction, and for binary and correlation data, an effect size of 0.25. GRADE also recommends downgrading the evidence when sample size is smaller than 300 (for binary data) and 400 (for continuous data), although for some topics, these criteria should be relaxed[19].

‖ Indirectness of comparison occurs when a comparison of intervention A versus B is not available but A was compared with C and B was compared with C that allows indirect comparisons of the magnitude of effect of A versus B. Indirectness of population, comparator and/or outcome can also occur when the available evidence regarding a particular population, intervention, comparator, or outcome is not available and is therefore inferred from available evidence. These inferred treatment effect sizes are of lower quality than those gained from head-to-head comparisons of A and B.

Margarete Ainsworth Building, Barker Street, Randwick NSW 2031. Phone: 02 9399 1000. Email: info@neura.edu.au

References

1. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMAGroup (2009): Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *British Medical Journal* 151: 264-9.
2. GRADEWorkingGroup (2004): Grading quality of evidence and strength of recommendations. *British Medical Journal* 328: 1490.
3. Beauchamp G, Gagnon A (2004): Influence of diagnostic classification on gender ratio in schizophrenia - a meta-analysis of youths hospitalized for psychosis. *Social Psychiatry & Psychiatric Epidemiology* 39: 1017-22.
4. Chang WC, Chan SSM, Chung DWS (2009): Diagnostic stability of functional psychosis: A systematic review. *Hong Kong Journal of Psychiatry* 19: 30-41.
5. Linscott RJ, Allardyce J, van Os J (2010): Seeking verisimilitude in a class: a systematic review of evidence that the criterial clinical symptoms of schizophrenia are taxonic. *Schizophrenia Bulletin* 36: 811-29.
6. Menezes NM, Milovan E (2000): First-episode psychosis: a comparative review of diagnostic evolution and predictive variables in adolescents versus adults. *Canadian Journal of Psychiatry - Revue Canadienne de Psychiatrie* 45: 710-6.
7. Roy MA, Maziade M, Labbe A, Merette C (2001): Male gender is associated with deficit schizophrenia: a meta-analysis. *Schizophrenia Research* 47: 141-7.
8. Roy MA, Merette C, Maziade M (2001): Subtyping schizophrenia according to outcome or severity: A search for homogeneous subgroups. *Schizophrenia Bulletin* 27: 115-38.
9. Davis KAS, Sudlow CLM, Hotopf M (2016): Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. *BMC Psychiatry* 16: 263.
10. Fusar-Poli P, Cappucciati M, Rutigliano G, Heslin M, Stahl D, Brittenden Z*, et al.* (2016): Diagnostic Stability of ICD/DSM First Episode Psychosis Diagnoses: Meta-analysis. *Schizophrenia Bulletin* 42: 1395-406.
11. Santelmann H, Franklin J, Busshoff J, Baethge C (2016): Interrater reliability of schizoaffective disorder compared with schizophrenia, bipolar disorder, and unipolar depression - A systematic review and meta-analysis. *Schizophrenia Research* 176: 357-63.
12. Santelmann H, Franklin J, Bushoff J, Baethge C (2015): Test-retest reliability of schizoaffective disorder compared with schizophrenia, bipolar disorder, and unipolar depression--a systematic review and meta-analysis. *Bipolar Disorders* 17: 753-68.
13. de Boer JN, Voppel AE, Begemann MJH, Schnack HG, Wijnen F, Sommer IEC (2018): Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews* 93: 85-92.
14. Olbert CM, Nagendra A, Buck B (2018): Meta-analysis of black vs. white racial disparity in schizophrenia diagnosis in the United States: Do structured assessments attenuate racial disparities? *Journal of Abnormal Psychology* 127: 104-15.
15. De Filippis R, Carbone EA, Gaetano R, Bruni A, Pugliese V, Segura-Garcia C*, et al.* (2019): Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: A systematic review. *Neuropsychiatric Disease and Treatment* 15: 1605-27.
16. Murrie B, Lappin J, Large M, Sara G (2019): Transition of Substance-Induced, Brief, and Atypical Psychoses to Schizophrenia: A Systematic Review and Meta-analysis. *Schizophrenia Bulletin* 46: 505-16.
17. CochraneCollaboration (2008): Cochrane Handbook for Systematic Reviews of Interventions. Accessed 24/06/2011.
18. Rosenthal JA (1996): Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research* 21: 37-59.

Margarete Ainsworth Building, Barker Street, Randwick NSW 2031. Phone: 02 9399 1000. Email: info@neura.edu.au

19. GRADEpro (2008): [Computer program]. Jan Brozek, Andrew Oxman, Holger Schünemann. *Version 32 for Windows*