Diagnosis and screening

Introduction

Bipolar disorders are a group of disorders characterised by episodes of mania or hypomania and depression. In between episodes, mild symptoms of mania and/or depression may, or may not, be present. Bipolar disorders characterised in the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, version 5) include bipolar I disorder, bipolar II disorder, and cyclothymic disorder.

A major depressive episode is a period of at least two weeks in which a person has at least five of the following symptoms (including one of the first two): intense sadness or despair; feelings of helplessness, hopelessness or worthlessness; loss of interest in activities once enjoyed; feelings of guilt, restlessness or agitation; sleeping too little or too much; slowed speech or movements; changes in appetite; loss of energy; difficulty concentrating, remembering or making decisions; and/or thoughts of death or suicide.

A manic episode is a period of at least one week when a person is high spirited or irritable in an extreme way most of the day for most days. A manic episode involves changes in normal behaviour such as showing exaggerated self-esteem or grandiosity, less need for sleep, talking more than usual, talking more loudly and quickly, being easily distracted, doing many activities at once, scheduling more events in a day than can be accomplished, embarking on risky behaviour, uncontrollable racing thoughts, and/or quickly changing ideas or topics. These changes in behaviour are significant and clear to friends and family and are severe enough to cause major dysfunction.

A hypomanic episode is similar to a manic episode, but the symptoms are less severe and need only last four days in a row. Hypomanic symptoms do not lead to the major problems that mania often causes, and the person is still able to function.

The difference between bipolar I disorder and bipolar II disorder is determined by the existence of mania in bipolar I disorder or hypomania in bipolar II disorder.

Cyclothymic disorder is a milder form of bipolar disorder involving many mood swings, with hypomania and depressive symptoms that occur often and fairly constantly. People with cyclothymia experience emotional ups and downs, but with less severe extremes than people with bipolar I or II disorder. Cyclothymic symptoms include at least two years of many periods of hypomanic and depressive symptoms that have lasted for at least half the time and have never stopped for more than two months.

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 2010 that report results for people with a diagnosis of bipolar or related disorders. Due to the high volume of systematic reviews we have now limited inclusion to systematic meta-analyses. Where no systematic meta-analysis exists for a topic, systematic reviews without meta-analysis are included for that topic. Reviews were identified by searching the databases MEDLINE, EMBASE, and PsycINFO. Hand searching reference lists of identified reviews was also conducted. When multiple reviews assessing the same topic were found, only the most recent and/or comprehensive reviews were included.

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 (RCTs) 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 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 15 systematic reviews that met our inclusion criteria[3-17].

Children and adolescents

- Moderate quality evidence suggests the clinical features associated more often in children or youth with bipolar depression than in children or youth with unipolar depression include more psychiatric comorbidities and behavioural problems (oppositional disorder, conduct disorder, anxiety disorders, irritability, suicidal/self-harm, social impairment, and substance use); earlier onset of mood symptoms; more severe depression; and having a family history of psychiatric illness.
- Moderate quality evidence suggests better test-retest reliability for bipolar disorder than for schizophrenia and schizoaffective disorder, but it is lower than for unipolar depression in children and adolescents ≤ 18 years.
- Moderate to high quality evidence suggests good reliability of checklists for identifying bipolar disorder in children and youth. Checklists included the Achenbach System of Empirically Based Assessment, the General Behaviour Inventory, the Mood Disorders Questionnaire, the Young Mania Rating Scale, the Child Mania Rating Scale, the Child and Adolescent Symptom Inventory, and the Child Bipolar Questionnaire. Checklists that focus on manic symptoms, parent-only assessments, and distilled samples (those that included healthy controls or excluded youth with diagnoses similar to bipolar disorder) were most accurate at identifying bipolar disorder.
- Caregiver report was more accurate than youth self-report or teacher report.
- Moderate to low quality evidence found higher bipolar disorder polygenic risk scores were associated with a diagnosis of ADHD, impaired executive functioning, lower IQ, and higher hypomania scores in children. Polygenic risk scores are indirect measures of genetic risk that are calculated by using weighted counts of risk variants, where the risk variants and their weights have been identified in genome-wide association studies.

Adults

- Moderate to high quality evidence finds reasonable diagnostic stability of bipolar disorder over time.
- Moderate to high quality evidence suggests the screening tools Hypomania Checklist, Bipolar Spectrum Diagnostic Scale, and Mood Disorder Questionnaire have good accuracy for detecting bipolar disorders in mental healthcare settings. The Hypomania

Checklist was better at detecting bipolar disorder II than the Mood Disorder Questionnaire.

- Moderate to high quality evidence suggests better inter-rater reliability for a diagnosis of bipolar disorder than for schizoaffective disorder, schizophrenia, or unipolar depression. There is also better test-retest reliability for bipolar disorder.
- Moderate quality evidence suggests reasonable predictive value and moderate kappa agreement for bipolar disorder diagnoses between administrative databases using the ICD-10 and clinical or research diagnoses. However, an estimated 17% of people in primary care settings that were previously diagnosed with depression have undiagnosed bipolar disorder.
- Moderate to low quality evidence suggests machine learning techniques of results from structural and functional neuroimaging studies show similar levels of moderate specificity and sensitivity for determining bipolar disorder diagnosis from other psychiatric diagnoses or healthy controls.
- Compared to people with schizoaffective disorder, moderate quality evidence suggests people with bipolar disorder may be older, with a later age of onset, more years of education, more likelihood of being Caucasian and less likelihood of being African American. People with bipolar disorder are also more likely to be married, have shorter duration of illness with less psychotic and negative symptoms (e.g., social withdrawal), and less depression.

Biederman J, Green A, DiSalvo M, Faraone SV

Can polygenic risk scores help identify pediatric bipolar spectrum and related disorders?: A systematic review

Psychiatry Research 2021; 299: 113843

View review abstract online

Comparison	**Accuracy of polygenic risk scores (PRSs) for identifying bipolar disorder in children.** **PRSs are calculated as weighted counts of thousands of risk variants, where the risk variants and their weights have been identified in genome-wide association studies.**
Summary of evidence	**Moderate to low quality evidence (large samples, unable to assess consistency or precision, indirect) found higher bipolar disorder polygenic risk scores were associated with a diagnosis of ADHD, impaired executive functioning, lower IQ, and higher hypomania scores in children.**

Bipolar disorder polygenic risk scores (BP-PRS)

BP-PRSs are associated with a diagnosis of ADHD, impaired executive functioning, lower IQ, and higher hypomania scores in children;

1 study (N = 5,936, ages 7 to 8) found high BP-PRSs were associated with impaired executive functioning, lower performance IQ, and poorer processing speed.

1 study (N = 3,448, ages 7 to 11) found BP-PRSs were associated with the diagnosis of ADHD and increased hypomania scores.

1 study (N = 495, ages 6 to 18) children with ADHD have a higher probability of being a BP disorder risk allele carrier than controls.

Consistency in results‡	Unable to assess; no measure of heterogeneity is reported.
Precision in results§	Unable to assess; no measure of precision is reported
Directness of results‖	Indirect; polygenic risk scores are considered an indirect measurement technique using imputation.

Carvalho AF, Takwoingi Y, Sales PM, Soczynska JK, Kohler CA, Freitas TH, Quevedo J, Hyphantis TN, McIntyre RS, Vieta E

Screening for bipolar spectrum disorders: A comprehensive meta-analysis

of accuracy studies

Journal of Affective Disorders 2015; 172: 337-46

View review abstract online

Comparison	**Accuracy of screening instruments for bipolar disorder.**
Summary of evidence	**Moderate to high quality evidence (large samples, unable to assess consistency, appears precise, direct) suggests the screening tools HCL-32, BSDS and MDQ have reasonable accuracy for detecting any bipolar disorder in mental healthcare settings. The HCL-32 was better at detecting bipolar disorder II than the MDQ.**
The Hypomania Checklist (HCL-32), Bipolar Spectrum Diagnostic Scale (BSDS), and Mood Disorder Questionnaire (MDQ)	
All three screening tools showed reasonable accuracy for detecting any bipolar disorder in mental healthcare settings; HCL-32 (cut-off score = 14): 9 studies, N = 6,652, sensitivity = 81%, 95%CI 77% to 85% HCL-32 (cut-off score = 14): 9 studies, N = 6,652, specificity = 67%, 95%CI 47% to 82% BSDS (cut-off score = 13): 3 studies, N = 672, sensitivity = 69%, 95%CI 63% to 74% BSDS (cut-off score = 13): 3 studies, N = 672, specificity = 86%, 95%CI 74% to 93% MDQ (cut-off score = 6): 3 studies, N = 612, sensitivity: 66%, 95%CI 57% to 73% MDQ (cut-off score = 6): 3 studies, N = 612, specificity: 79%, 95%CI 72% to 84% The HCL-32 was significantly more accurate than the MDQ for the detection of bipolar disorder II. Authors suggest that a positive screen should be confirmed by a clinical diagnostic evaluation.	
Consistency in results	Unable to assess; no measure of overall heterogeneity is reported.
Precision in results	Unable to assess; appears precise
Directness of results	Direct

Cegla-Schvartzman FB, Ovejero S, Lopez-Castroman J, Baca-Garcia E

Diagnostic Stability in Bipolar Disorder: A Narrative Review

Harvard Review of Psychiatry 2019; 27: 3-14.

View review abstract online	
Comparison	**Diagnostic stability of bipolar disorders over time.** **Prospective consistency is the proportion of subjects in a diagnostic category at first evaluation who received the same diagnosis at last evaluation. Retrospective consistency is the proportion of subjects in a diagnostic category at the last evaluation who were in that same category at first evaluation.**
Summary of evidence	**Moderate to high quality evidence (large samples, unable to assess consistency, appears precise, direct) finds reasonable diagnostic stability of bipolar disorder over time.**
Diagnostic stability	
Reasonable prospective consistency and lower retrospective consistency; Prospective consistency (1-12 years): 4 studies, N = 2,336, weighted mean = 65.7%, 95%CI 64.5% to 66.9% Retrospective consistency (7-9 years): 2 studies, N = 7,178, weighted mean = 32.9%, 95%CI 31.6% to 34.2%	
Consistency	Unable to assess; no measure of consistency is reported
Precision	Appears precise
Directness	Direct

Daveney J, Panagioti M, Waheed W, Esmail A

Unrecognized bipolar disorder in patients with depression managed in primary care: A systematic review and meta-analysis

General Hospital Psychiatry 2019; 58: 71-6

View review abstract online

Comparison	**Prevalence of unrecognised bipolar disorders in people diagnosed with depression.**

Summary of evidence	**Moderate to high quality evidence (large sample, inconsistent, appears precise, direct) finds around 17% of people previously diagnosed with depression who are in primary care have unrecognised bipolar disorder.**
Prevalence of bipolar disorder	
10 studies, N = 3,803, prevalence = 17%, 95%CI 12% to 22%, I^2 = 95% This prevalence was non-significantly higher in studies that used questionnaires as assessment tools for bipolar disorder compared to studies that used clinical interviews, however this difference was not significant (14% vs. 22%).	
Consistency	Inconsistent
Precision	Appears precise
Directness	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 of bipolar disorders 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 a bipolar disorder diagnosis gained from administrative databases.**
Diagnostic integrity	
Reasonable predictive value and moderate kappa agreement for bipolar disorder; 12 studies, N = 2,455, median PPV ~75% (range 22-100%), Kappa ~0.50 (range 18-65%).	
Consistency	Unable to assess; no measure of consistency is reported
Precision	Appears imprecise

Directness	Direct

Librenza-Garcia D, Kotzian BJ, Yang J, Mwangi B, Cao B, Pereira Lima LN, Bermudez MB, Boeira MV, Kapczinski F, Passos IC

The impact of machine learning techniques in the study of bipolar disorder: A systematic review

Neuroscience and Biobehavioral Reviews 2017; 80: 538-54

View review abstract online

Comparison	**Assessment of machine learning techniques for diagnosis of bipolar disorder vs. other psychiatric diagnoses or no psychiatric diagnosis.**
Summary of evidence	**Moderate to low quality evidence (large samples, indirect, unable to assess consistency or precision) suggests machine learning techniques of results from structural and functional neuroimaging studies show similar levels of moderate specificity and sensitivity for determining bipolar disorder diagnosis from other psychiatric diagnoses or healthy controls.**

Bipolar disorder diagnosis

Machine learning of results from structural and functional neuroimaging studies show similar levels of moderate specificity and sensitivity for determining bipolar disorder diagnosis from other psychiatric diagnoses or healthy controls;

7 structural MRI studies, N = 1,031, sensitivity = 0.61, specificity = 0.68

5 functional MRI studies, N = 801, sensitivity = 0.63, specificity = 0.67

Consistency	Unable to assess; no measure of consistency is reported.
Precision	Unable to assess; no measure of precision is reported.
Directness	Indirect

Pagel T, Baldessarini RJ, Franklin J, Baethge C

Characteristics of patients diagnosed with schizoaffective disorder compared with schizophrenia and bipolar disorder

Bipolar Disorders 2013; 15: 229-239

View review abstract online

Comparison	**Assessment of patient characteristics in bipolar vs. schizoaffective disorders.**
Summary of evidence	**Moderate quality evidence (direct, large samples, some inconsistency and imprecision) suggests bipolar disorder patients may be older, with a later age of onset, more years of education and more are Caucasian and less are African American. Bipolar patients are more likely to be married, have shorter duration of illness, less psychotic and negative symptoms (e.g. social withdrawal, speech reduction, loss of interest, blunted emotional response), less depression, and lower IQ.**

Demographic characteristics, hospitalisations and symptoms

15 studies used DSM-IIIR, 14 used DSM-IV, 4 used DSV-III, 11 used RDC, 1 used ICD-9, 1 used ICD-10, and 4 used mixed diagnostic tools.

Bipolar disorder N = 4814, schizoaffective disorder N = 2684

Studies of bipolar disorder vs. schizoaffective disorder report:

Later age at onset: 26.1 vs. 23.3yrs, MD -2.91, CI -4.52 to -1.29, $p < 0.0004$, I^2 77%, $p < 0.00001$

Older sample: 46.7 vs. 42.7yrs, MD -3.03, CI -4.22 to -1.89, $p < 0.0001$, I^2 59%, $p < 0.00001$

More education: 13.3 vs. 12.3yrs, MD -0.92, CI -1.44 to -0.40, $p = 0.0006$, I^2 22%, $p = 0.25$

More Caucasians: 60 vs. 52%, OR 0.52, CI 0.40 to 0.69, $p < 0.0001$, I^2 0%, $p = 0.50$

Less African Americans: 13 vs. 25%, OR 1.50, CI 1.02 to 2.21, $p < 0.04$, I^2 59%, $p = 0.02$

More ever married: 41 vs. 34%, OR 0.63, CI 0.43 to 0.93, $p = 0.02$, I^2 10%, $p = 0.35$

Shorter duration of illness: 11.5 vs. 13.3yrs, MD 2.10, CI 0.10 to 4.09, $p = 0.04$, I^2 56%, $p = 0.03$

Lower BPRS scores (mainly psychotic symptoms): 37.8 vs. 46.6, MD 3.85, CI 1.94 to 5.87, $p < 0.0001$, I^2 0%, $p = 0.48$

Lower HDRS scores (depression symptoms): 10.8 vs. 20.3, MD 7.01, CI 1.67 to 12.36, $p = 0.01$, I^2 80%, $p = 0.002$

Lower SANS scores (negative symptoms): 3.3 vs. 0.9, MD 0.85, CI 0.14 to 1.55, $p = 0.02$, I^2 76%, $p = 0.02$

Lower WAIS-IQ: 103.7 vs. 105.5, MD -7.31, CI -10.22 to -4.08, p = 0.001 , I^2 0%, p = 0.64 There were no significant differences were reported for gender, currently married, number of hospitalisations, age at first hospitalisation, CGI, GAS, GAF, or SAPS. Overall, SDs tended to be larger in bipolar disorder than in schizoaffective studies, indicating higher heterogeneity in bipolar disorder results, although this finding was not significant.	
Consistency	Consistent for education, Caucasians, ever married, BPRS and WAIS-IQ
Precision	Precise for Caucasians only
Directness	Direct

Salamon S, Santelmann H, Franklin J, Baethge C

Test-retest reliability of the diagnosis of schizoaffective disorder in childhood and adolescence - A systematic review and meta-analysis

Journal of Affective Disorders 2018; 230: 28-33

View review abstract online

Comparison	**Test-retest reliability of a diagnosis of a bipolar disorder compared to a diagnosis of schizoaffective disorder, schizophrenia, or unipolar depression in children and adolescents ≤ 18 years.**
Summary of evidence	**Moderate quality evidence (some inconsistency and imprecision, direct, large sample) suggests better test-retest reliability for bipolar disorder than for schizophrenia and schizoaffective disorder, but it is lower than for unipolar depression.**
Test-retest reliability	
7 studies, N = 403 *Test-retest reliability is better for bipolar disorder than for schizophrenia and schizoaffective disorder, but lower than for unipolar depression;* Bipolar disorder = 5 studies, Cohen's kappa = 0.64, 95%CI 0.55 to 0.74, I^2 = 0.2% Unipolar depression = 3 studies, Cohen's kappa = 0.66, 95%CI 0.52 to 0.81, I^2 = 0% Schizophrenia = 7 studies, Cohen's kappa = 0.56, 95%CI 0.29 to 0.83, I^2 = 94% Schizoaffective disorder = 7 studies, Cohen's kappa = 0.27, 95%CI 0.07 to 0.47, I^2 = 91%	

Consistency in results	Consistent for bipolar disorder and unipolar depression.
Precision in results	Appears precise for bipolar disorder.
Directness of results	Direct

Santelmann H, Franklin J, Busshoff J, Baethge C

Inter-rater 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	**Inter-rater reliability of a diagnosis of a bipolar disorder compared to a diagnosis of schizoaffective disorder, schizophrenia, or unipolar depression.**
Summary of evidence	**Moderate to high quality evidence (inconsistent, appears precise, direct, large sample) suggests better inter-rater reliability for a diagnosis of bipolar disorder than for a diagnosis of schizoaffective disorder, schizophrenia, or unipolar depression.**

Inter-rater reliability

25 studies, N = 7,912

Inter-rater reliability kappa is higher for bipolar disorder than for schizoaffective disorder, schizophrenia, or unipolar depression;

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%

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

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

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	Mostly 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 bipolar disorder diagnosis compared to a diagnosis of schizoaffective disorder, schizophrenia, or unipolar depression.**
Summary of evidence	**Moderate to high quality evidence (inconsistent, appears precise, direct, large sample) suggests better test-retest reliability for bipolar disorder than for schizoaffective disorder, schizophrenia or unipolar depression.**
Test-retest reliability	
49 studies, N = 14,314 *Test-retest reliability is higher for bipolar disorder than for schizoaffective disorder, schizophrenia or unipolar depression;* 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% 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% In studies of bipolar disorder, kappa was significantly higher in; low vs. high risk of bias studies (including blinded vs. non-blinded studies); studies using ICD-10 diagnostic tool vs. DSM 111, DSM 1V or DSM 5 diagnostic tools; studies with a short vs. long follow-up period (< 2 months vs. > 12 months). There were no differences in kappa in studies using consistent vs. inconsistent use of diagnostic interview; similar vs. different rater identity; first-episode vs. chronic illness; inpatient vs. outpatient status.	
Consistency in results	Inconsistent
Precision in results	Appears precise
Directness of results	Direct

Wang YY, Xu DD, Liu R, Yang Y, Grover S, Ungvari GS, Hall BJ, Wang G, Xiang YT

Comparison of the screening ability between the 32-item Hypomania Checklist (HCL-32) and the Mood Disorder Questionnaire (MDQ) for bipolar disorder: A meta-analysis and systematic review

Psychiatry Research 2019; 273: 461-6

View review abstract online

Comparison	**Psychometric properties of the 32-item Hypomania Checklist (HCL-32) vs. the Mood Disorder Questionnaire (MDQ).**
Summary of evidence	**Moderate quality evidence (large sample, unable to assess consistency and precision, direct) finds both the HCL-32 and the MDQ have acceptable psychometric properties.**
Psychometric properties	
Both the HCL-32 and the MDQ have acceptable psychometric properties: 9 studies, N = 1,615 HCL-32: sensitivity = 82%, 95%CI 72% to 89%, specificity = 57%, 95%CI 48% to 66% MDQ: sensitivity = 80%, 95%CI 71% to 86%, specificity = 70%, 95%CI 59% to 71%	
Consistency in results	Unable to assess
Precision in results	Unable to assess
Directness of results	Direct

Youngstrom EA, Genzlinger JE, Egerton GA, Van Meter AR

Multivariate meta-analysis of the discriminative validity of caregiver, youth, and teacher rating scales for pediatric bipolar disorder: Mother knows best about mania

Archives of Scientific Psychology 2015; 3: 112-37

View review abstract online

Comparison	**Reliability of bipolar disorder symptom checklists for children and youth < 18 years.** **Checklists included the Achenbach System of Empirically Based**

	Assessment, the General Behavior Inventory, the Mood Disorders Questionnaire, the Young Mania Rating Scale, the Child Mania Rating Scale, the Child and Adolescent Symptom Inventory, and the Child Bipolar Questionnaire.
Summary of evidence	**Moderate to high quality evidence (inconsistent, precise, direct, large sample) suggests good reliability of checklists for identifying bipolar disorder in children and youth. Caregiver report was more accurate than youth self-report or teacher report. Scales that focus on manic symptoms, parent-only assessments, and distilled samples (those that included healthy controls or excluded youth with diagnoses similar to bipolar disorder) were most accurate at identifying bipolar disorder.**
Checklist reliability	
A large effect showed the checklists were reliable at detecting bipolar disorder in youth; 25 studies, N = 11,941, *g* = 1.05, 95%CI 0.83 to 1.27, *p* < 0.05, Q = 738.25, *p* < 0.00005 *Caregiver report was more accurate in detecting bipolar disorder than youth or teacher report;* Caregiver report: N = 10,232, *g* = 1.11, 95%CI 0.93 to 1.28, *p* < 0.05 Youth report: N = 3,018, *g* = 0.49, 95%CI 0.38 to 0.61, *p* < 0.05 Teacher report: N = 1,290, *g* = 0.32, 95%CI 0.15 to 0.49, *p* < 0.05 Authors report that studies using scales that focus on manic symptoms, parent-only assessments, and that use distilled samples (included healthy controls or excluded youth with diagnoses similar to bipolar disorder) were better at detecting bipolar disorder than other studies. There were significant differences in the effect size according to; study design, reporting quality, number of scale items, year of publication, percentage of cases with ADHD, or whether the study had sponsorship from a pharmaceutical company. There was no evidence of publication bias.	
Consistency in results	Inconsistent
Precision in results	Precise
Directness of results	Direct

Youngstrom EA, Egerton GA, Genzlinger J, Freeman LK, Rizvi SH, Van Meter A

Improving the global identification of bipolar spectrum disorders: Meta-analysis of the diagnostic accuracy of checklists

<table>
<tr><td colspan="2"><b>Psycholgical Bulletin 2018; 144(3): 315-342</b>
View review abstract online</td></tr>
<tr><td><b>Comparison</b></td><td><b>Discriminative validity of checklists and rating scales assessing hypomanic and manic symptoms in adults with bipolar disorder.</b>
<b>Checklists included the Altman Self Rating Mania Scale, Behavioral Activation Scale, Bipolar Spectrum Diagnostic Scale, General Behavior Inventory, Hypomania Checklist, Hypomanic Attitudes and Positive Predictions Inventory, Hypomanic Personality Scale, Internal State Scale, Multidimensional Assessment of Thymic States, Mood Disorder Questionnaire, Mood Spectrum Self Reports, Self-Report Mania Inventory , Symptom Checklist-90, Temperament Evaluation of Memphis, Pisa, Paris and San Diego-Auto questionnaire, short version.</b></td></tr>
<tr><td><b>Summary of evidence</b></td><td><b>Moderate to high quality evidence (large sample, inconsistent, mostly precise, direct) suggests good discriminative validity for scales assessing hypomanic and manic symptoms, particularly the Mood Disorder Questionnaire scale. There were larger effect sizes in less recent publications, more distilled samples (not just clinical), and in hospital settings. There were no differences in the effect size according to study quality, study region, and use of translated scales.</b></td></tr>
<tr><td colspan="2"><b>Scales assessing mania or hypomania</b></td></tr>
<tr><td colspan="2"><i>A large, significant effect of good discriminative validity for hypomanic and manic symptoms;</i>
103 studies, N = 50,310, <i>g</i> = 1.10
After controlling for other variables, multiple meta-regression found significant differences in the effect size according to the scale used; the Mood Disorder Questionnaire performed significantly better than the Altman Self Rating Mania Scale and the “other” set of scales (various scales with few data). There was a trend effect of better discriminative validity with the Mood Disorder Questionnaire than the Hypomanic Personality Scale.
After controlling for other variables, there were also larger effect sizes in less recent publications, more distilled samples (not just clinical), and in hospital settings.
There were no effects of study quality, study region, and use of translated scales.</td></tr>
<tr><td><b>Consistency in results</b></td><td>Authors report that data are inconsistent.</td></tr>
<tr><td><b>Precision in results</b></td><td>Precise for the scale subgroup analyses of; Altman Self Rating Mania Scale, Bipolar Spectrum Diagnostic Scale, Hypomanic Personality Scale, Hypomania Checklist, Mood Disorder Questionnaire, Temperament Evaluation of Memphis, Pisa, Paris and San Diego-Auto questionnaire, and the “other” set of scales.</td></tr>
<tr><td><b>Directness of results</b></td><td>Direct</td></tr>
</table>

Uchida M, Serra G, Zayas L, Kenworthy T, Faraone SV, Biederman J

Can unipolar and bipolar pediatric major depression be differentiated from each other? A systematic review of cross-sectional studies examining differences in unipolar and bipolar depression

Journal of Affective Disorders 2015; 176: 1-7

View review abstract online

Comparison	**Clinical differences in children and adolescents with unipolar vs. bipolar disorder.**
Summary of evidence	**Moderate quality evidence (large sample, inconsistent, unable to assess precision, direct) suggests the clinical features associated more often in children or youth with bipolar depression than in children or youth with unipolar depression include; more psychiatric comorbidities and behavioural problems (oppositional disorder, conduct disorder, anxiety disorders, irritability, suicidal/self-harm, social impairment, and substance use); earlier onset of mood symptoms; more severe depression; and having a family history of psychiatric illness.**
Clinical features	
4 studies, N = 1,476 3/4 studies found significantly higher rates of psychiatric comorbidities in children/youth with bipolar disorder, including oppositional defiant disorder, conduct disorder, anxiety disorders, and substance use (in adolescents only). 3/4 studies found significantly higher rates of first-degree relatives with any psychiatric illness in children/youth with bipolar disorder. 3/4 studies found significantly earlier onset of mood symptoms in children/youth with bipolar disorder. 2/4 studies found significantly greater severity, and more frequent episodes, of depression in children/youth with bipolar disorder. 2/4 studies found significantly more sadness, aggression, irritability, hopelessness, and suicidal or self-injurious behaviors in children/youth with bipolar disorder. 2/4 studies found significantly higher level of impairment, including difficulties with peers and family members, and severe behavioral problems in school in children/youth with bipolar disorder.	
Consistency in results	Results appear inconsistent.
Precision in results	Unable to assess; no CIs are reported

Directness of results	Direct

Explanation of acronyms

BPRS = Brief Psychiatric Rating Scale, CGI = Clinical Global Impression, CI = confidence interval, DSM = American Psychiatric Association's Diagnostic and Statistical Manual, *g* = Hedges *g* standardised mean difference, GAF = Global Assessment of Functioning, GAS = Global Assessment Scale, HDRS = Hamilton Depression Rating Scale, 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, MD = mood disorder, N = number of participants, OR = odds ratio, *p* = probability of rejecting a null hypothesis of no differences between groups, Q = test for heterogeneity, RDC = Research Diagnostic Criteria, SANS = Scale for Assessment of Negative Symptoms, SAPS = Scale for Assessment of Positive Symptoms, vs. = versus, WAIS-IQ = Wechsler Adult Intelligence Scale-Intelligence Quotient

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[18].

† 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. Standardsed 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[18].

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[19]. 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. I² 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. I² can be calculated from Q (chi-square) for the test of heterogeneity with the following formula[18];

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

§ 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[20].

‖ 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.

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. Youngstrom EA, Genzlinger JE, Egerton GA, Van Meter AR (2015): Multivariate meta-analysis of the discriminative validity of caregiver, youth, and teacher rating scales for pediatric bipolar disorder: Mother knows best about mania. *Archives of Scientific Psychology* 3: 112-37.
4. 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.
5. Pagel T, Baldessarini RJ, Franklin J, Baethge C (2013): Characteristics of patients diagnosed with schizoaffective disorder compared with schizophrenia and bipolar disorder. *Bipolar Disorders* 15: 229-39.
6. 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.
7. 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.
8. Carvalho AF, Takwoingi Y, Sales PM, Soczynska JK, Kohler CA, Freitas TH*, et al.* (2015): Screening for bipolar spectrum disorders: A comprehensive meta-analysis of accuracy studies. *Journal of Affective Disorders* 172: 337-46.
9. Librenza-Garcia D, Kotzian BJ, Yang J, Mwangi B, Cao B, Pereira Lima LN*, et al.* (2017): The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neuroscience and Biobehavioral Reviews* 80: 538-54.
10. Salamon S, Santelmann H, Franklin J, Baethge C (2018): Test-retest reliability of the diagnosis of schizoaffective disorder in childhood and adolescence - A systematic review and meta-analysis. *Journal of Affective Disorders* 230: 28-33.
11. Youngstrom EA, Egerton GA, Genzlinger J, Freeman LK, Rizvi SH, Van Meter A (2018): Improving the global identification of bipolar spectrum disorders: Meta-analysis of the diagnostic accuracy of checklists. *Psychological Bulletin* 144: 315-42.
12. Uchida M, Serra G, Zayas L, Kenworthy T, Faraone SV, Biederman J (2015): Can unipolar and bipolar pediatric major depression be differentiated from each other? A systematic review of cross-sectional studies examining differences in unipolar and bipolar depression. *Journal of Affective Disorders* 176: 1-7.
13. Cegla-Schvartzman FB, Ovejero S, Lopez-Castroman J, Baca-Garcia E (2019): Diagnostic Stability in Bipolar Disorder: A Narrative Review. *Harvard Review of Psychiatry* 27: 3-14.
14. Daveney J, Panagioti M, Waheed W, Esmail A (2019): Unrecognized bipolar disorder in patients with depression managed in primary care: A systematic review and meta-analysis. *General Hospital Psychiatry* 58: 71-6.
15. Wang YY, Xu DD, Liu R, Yang Y, Grover S, Ungvari GS*, et al.* (2019): Comparison of the screening ability between the 32-item Hypomania Checklist (HCL-32) and the Mood Disorder Questionnaire (MDQ) for bipolar disorder: A meta-analysis and systematic review. *Psychiatry Research* 273: 461-6.
16. Biederman J, Green A, DiSalvo M, Faraone SV (2021): Can polygenic risk scores help identify pediatric bipolar spectrum and related disorders?: A systematic review. *Psychiatry Research* 299: 113843.
17. Ribeiro JS, Pereira D, Salagre E, Coroa M, Oliveira PS, Santos V*, et al.* (2020): Risk calculators in bipolar disorder: A systematic review. *Brain Sciences* 10: 1-13.
18. CochraneCollaboration (2008): Cochrane Handbook for Systematic Reviews of Interventions. Accessed 24/06/2011.

19. Rosenthal JA (1996): Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research* 21: 37-59.

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