TECHNICAL COMMENTARY

Outcome assessment tools

Introduction

Standardised assessment tools are vital for assessing a range of variables including symptoms, functioning and quality of life. They are often used within a controlled research environment, but high-quality assessment tools are also useful in practice for both clinical management and outcome prediction.

The quality of assessment tools can be measured in various ways. 'Reliability' refers to the reproducibility of an instrument's results across different assessors, settings and times. 'Construct validity' is the extent to which an instrument measures the theoretical construct it was designed to measure. This involves 'convergent validity', which is the degree of correlation between different scales measuring the same construct, confirming they are measuring the same thing; and 'divergent validity', which is the lack of correlation between scales measuring different constructs, confirming that they are measuring different things. Similarly, 'known groups' validity' is the extent to which an instrument can demonstrate different scores for groups known to vary on the variables being measured. 'Content validity' is the extent to which each individual item on a scale represents the construct being measured, and 'internal consistency' is the degree of correlation between individual items within a scale.

'Predictive validity' refers to sensitivity, which is the proportion of correctly identified positives, and specificity, which is the proportion of correctly identified negatives. Sensitivity and specificity are measured by comparing an instrument's results with known 'gold standard' results. 'Responsiveness' is the extent to which an instrument can detect clinically significant or practically important changes over time, and 'area under the curve' (AUC) is a global measure of test performance.

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 separately for people with a diagnosis of bipolar or related disorders. 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, low or very 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 three systematic reviews that met our inclusion criteria[3-5].

- Moderate to low quality evidence suggests electronic self-monitoring of depression symptoms is reliable, being similar to clinically rated instruments (Montgomery Asberg Depression Rating Scale, the Hamilton Depression Rating Scale or the Inventory of Depressive Symptomatology). Low quality evidence is unsure of the validity of electronic self-monitoring of mania symptoms.
- Moderate to low quality evidence suggests patient-reported measures with the highest clinical utility for assessing symptoms were the Altman Self-Rating Mania Scale, the Quick Inventory of Depressive Symptomatology–Self Report and the Internal State Scale. Clinician-rated measures with the highest clinical utility for assessing symptoms were the Bech-Rafaelsen Mania Rating Scale, the Quick Inventory of Depressive Symptomatology, and the Bipolar Inventory of Symptoms Scale.
- Moderate to low quality evidence suggests the most commonly used scales for assessing functioning were the Global Assessment of Functioning and the Functional Assessment Short Test.

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

Cerimele JM, Goldberg SB, Miller CJ, Gabrielson SW, Fortney JC

Systematic review of symptom assessment measures for use in measurement-based care of bipolar disorders

Psychiatric Services 2019; 70: 396-408

View review abstract online

Comparison	**Utility of bipolar disorder symptom measures.**
Summary of evidence	**Moderate to low quality evidence (direct, unclear sample size, unable to assess precision or consistency) suggests the patient-reported measures with the highest clinical utility include the Altman Self-Rating Mania Scale, the Quick Inventory of Depressive Symptomatology–Self Report and the Internal State Scale. The clinician-rated measures with the highest clinical utility include the Bech-Rafaelsen Mania Rating Scale, the Quick Inventory of Depressive Symptomatology, and the Bipolar Inventory of Symptoms Scale.**
Symptom measures	
Authors report that the patient-reported measures with the highest clinical utility were; The Altman Self-Rating Mania Scale The Quick Inventory of Depressive Symptomatology–Self Report (QIDS-SR) The Internal State Scale (mania and depression) *Authors report that the clinician-observed measures with the highest clinical utility were;* The Bech-Rafaelsen Mania Rating Scale The Quick Inventory of Depressive Symptomatology The Bipolar Inventory of Symptoms Scale (mania and 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

Chen M, Fitzgerald HM, Madera JJ, Tohen M

Functional outcome assessment in bipolar disorder: A systematic

literature review

Bipolar Disorders 2019; 21: 194-214

View review abstract online

Comparison	**Assessment of functioning in people with bipolar disorder.**
Summary of evidence	**Moderate to low quality evidence (direct, unclear sample size, unable to assess precision or consistency) suggests the most commonly used scales were the Global Assessment of Functioning and the Functional Assessment Short Test.**
Functioning scales	
Authors report that the most commonly used scales were; The Global Assessment of Functioning The Functional Assessment Short Test	
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

Faurholt-Jepsen M, Munkholm K, Frost M, Bardram JE, Kessing LV

Electronic self-monitoring of mood using IT platforms in adult patients with bipolar disorder: A systematic review of the validity and evidence

BMC Psychiatry 2016; 16: 7

View review abstract online

Comparison	**Reliability of electronic self-monitoring of depression and mania symptoms compared to clinically rated scales.**
Summary of evidence	**Moderate to low quality evidence (small samples, consistent, direct, unable to assess precision) suggests electronic self-monitoring of depression symptoms is reliable, being similar to clinically rated instruments (Montgomery Asberg Depression Rating Scale, the Hamilton Depression Rating Scale or the Inventory of Depressive Symptomatology).** **Low quality evidence (inconsistent) is unsure of the validity of electronic self-monitoring of mania symptoms.**

Outcome assessment tools

Correlation between scales	
6 of 6 studies (N = 179) found electronic self-monitoring of depression scores significantly correlated with scores on clinical rating scales for depression; the Montgomery Asberg Depression Rating Scale (MADRS), the Hamilton Depression Rating Scale (HDRS), or the Inventory of Depressive Symptomatology, clinician rated (IDS-C). 2 of 7 studies (N = 64 of 206) found electronic self-monitoring of depression scores significantly correlated with scores on clinical rating scales for mania at baseline; the Young Mania Rating Scale (YMRS). 1 additional study (N = 18) found a significant correlation between self-monitored mood and the Young Mania Rating Scale after 6 weeks of mood self-monitoring, but not at baseline.	
Consistency in results	Consistent for depression, inconsistent for mania.
Precision in results	Unable to assess; no confidence intervals are reported.
Directness of results	Direct

Explanation of acronyms

HDRS = Hamilton Rating Scale for Depression, IDS-C = Inventory of Depressive Symptomatology, clinician rated, MADRS = Montgomery Asberg Depression Rating Scale, N = number of participants, YMRS = Young Mania Rating Scale

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[6].

† 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). A receiver operating characteristic (ROC) curve represents sensitivity/specificity pairs corresponding to different cut-off values. A guide for interpreting the area under the curve (AUC) statistic is; 0.90 to 1.00 = excellent, 0.80 to 0.90 = good, 0.70 to 0.80 = fair, 0.60 to 0.70 = poor, and 0.50 to 0.60 = fail.

Weighted mean difference scores refer to mean differences between treatment and comparison groups after treatment (or occasionally pre to post treatment) and in a randomized trial there is an assumption that both groups are comparable on this measure prior to treatment. Standardized 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[6].

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[7]. lnOR stands for

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

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. Unstandardized (*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. Standardized 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[6];

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

§ 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[8].

‖ 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. Cerimele JM, Goldberg SB, Miller CJ, Gabrielson SW, Fortney JC (2019): Systematic review of symptom assessment measures for use in measurement-based care of bipolar disorders. *Psychiatric Services* 70: 396-408.
4. Chen M, Fitzgerald HM, Madera JJ, Tohen M (2019): Functional outcome assessment in bipolar disorder: A systematic literature review. *Bipolar Disorders* 21: 194-214.
5. Faurholt-Jepsen M, Munkholm K, Frost M, Bardram JE, Kessing LV (2016): Electronic self-monitoring of mood using IT platforms in adult patients with bipolar disorder: A systematic review of the validity and evidence. *BMC Psychiatry* 16: 7.
6. CochraneCollaboration (2008): Cochrane Handbook for Systematic Reviews of Interventions. Accessed 24/06/2011.
7. Rosenthal JA (1996): Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research* 21: 37-59.
8. GRADEpro (2008): [Computer program]. Jan Brozek, Andrew Oxman, Holger Schünemann. *Version 32 for Windows*