DR. JAN SCOTT (Orcid ID : 0000-0002-7203-8601)

DR. BRUNO ETAIN (Orcid ID : 0000-0002-5377-1488)

DR. MIRKO MANCHIA (Orcid ID : 0000-0003-4175-6413)

DR. PIERRE ALEXIS GEOFFROY (Orcid ID : 0000-0001-9121-209X)

DR. MARTIN ALDA (Orcid ID : 0000-0001-9544-3944)

Article type : Original Article

An examination of the quality and performance of the Alda scale for classifying lithium response phenotypes.

[1,2]Scott J.,
[2,3]Etain B.,
[4,5]Manchia M.,
[2,6]Brichant-Petitjean C.,
[2]Geoffroy P.,
[7]Schulze T.,
[8,9]Alda M.,
[2,3]Bellivier F.
and ConLiGen collaborators.

1. Institute of Neuroscience, Newcastle University, Newcastle, UK.
2. Université Paris Diderot and INSERM UMRS1144, Paris, France.

3. Département de Psychiatrie et de Médecine Addictologique, AP-HP, GH Saint-Louis-Lariboisière-F. Widal, Paris, France.
4. Section of Psychiatry, Department of Medical Sciences and Public Health, University of Cagliari, Cagliari, Italy.
5. Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada.
6. EPS Maison Blanche, CMP 18, rue Salneuve, Secteur 75G20-21, Paris, France.
7. Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, LMU Munich, Germany.
8. Department of Psychiatry, Dalhousie University, Halifax, Canada.
9. National Institute of Mental Health, Klecany, Czech Republic.

ConLiGen collaborators include:
A Amare (Adelaide, Australia), R Ardau (Cagliari, Italy), L Backlund (Sweden, Stockholm), B Baune (Melbourne, Australia), A Barboza (San Pedro Garza Garcia, Mexico), A Benabarre (Barcelona, Catalonia, Spain), B Chaumette (Montreal, Canada), H Chen (Taipei, Taiwan), C Chillotti (Cagliari, Italy), S Clark (Adelaide, Australia), F Colom (Barcelona, Catalonia, Spain), M Del Zompo (Cagliari, Italy), N Dalkner (Graz, Austria), C Dantas (Camprias, Brazil), P Ferentinos (London, England), J Garnham (Dalhousie, Canada), S Jamain (Paris, France), E Jimenez (Barcelona, Catalonia, Spain), J-P Khan (Nancy, France), P Kuo (Taipei, Taiwan), C Lavebratt (Stockholm, Sweden), M Maj (Naples, Italy), V Millischer (Stockholm, Sweden), P Monteleone (Salerno, Italy), C Pisanu (Monserrato, Italy), J Potash (Baltimore, USA), A Reif (Frankfurt, Germany), E Reininghaus (Graz, Austria), M Schalling (Stockholm, Sweden), P Schofield (Sydney, Australia), K Schubert (Adelaide, Australia), G Severino (Monserrato, Italy), C Slaney (Dalhousie, Canada), D Smith (Glasgow, Scotland), A Squassina (Monserrato, Italy), L Tondo (Boston, USA), E Vieta (Barcelona, Catalonia, Spain), S Witt (Mannheim, Germany).

Corresponding author: Jan Scott, Institute of Neuroscience, Newcastle University, Newcastle, UK. Email: jan.scott@newcastle.ac.uk.

Abstract word count: 250
Main text: 5314 words; 45 references; 2 tables; 2 figures.
Supplementary materials: Tables 1S-9S.

Abstract (words=250)

Objectives: The Retrospective Assessment of the Lithium Response Phenotype Scale (Alda scale) is the most widely used clinical measure of lithium response phenotypes. We assess its performance against recommended psychometric and clinimetric standards.

Methods: We used data from the Consortium for Lithium Genetics and a French study of lithium response phenotypes (combined sample >2500) to assess reproducibility, responsiveness, validity and interpretability of the A scale (assessing change in illness activity), the B scale and its items (assessing confounders of response) and the previously established response categories derived from the Total Score for the Alda scale.

Results: The key findings are that the B scale is vulnerable to error measurement. For example, some items contribute little to overall performance of the Alda scale (e.g. B2) and that the B scale does not reliably assess a single construct (uncertainty in response). Machine learning models indicate that it may be more useful to employ and algorithm for combining the ratings of individual B items in a sequence that clarifies the noise to signal ratio instead of using a composite score.

Conclusions: This study highlights three important topics. First, empirical approaches can help determine which aspects of the performance of any scale can be improved. Second, the B scale of the Alda is best applied as a multidimensional index (identifying several independent confounders of the assessment of response). Third, an integrated science approach to precision psychiatry is vital, otherwise phenotypic misclassifications will undermine the reliability and validity of findings from genetics and biomarker studies.

Introduction

Clinical practice guidelines identify lithium (Li) as a first line treatment for mood stabilization in bipolar disorders (BD); however, only about 30% of patients show an optimal response in day-to-day settings[1]. Variability in response is poorly understood, making it difficult for clinicians to reliably predict which patients will benefit from Li without a lengthy treatment trial[2]. Many genetics studies aim to identify biomarkers for Li response in the hope that this will enable stratification of BD cases into treatment-relevant subgroups. As such, the method for classifying clinical phenotypes of Li response is a critical element of each study[3].

In research, the ideal assessment of Li response would involve systematic follow-up of Li naive cases to allow prospective evaluation of changes in illness activity after initiation of Li alongside monitoring of treatment exposure[4-6]. This gold standard is difficult to achieve, so most publications

regarding putative markers of Li response have relied on retrospective evaluation. For example, a recent genetics study measured Li response according to case note recordings, although the reliability and validity of these clinical judgements was unreported[7]. Others have employed brief retrospective ratings of the course of illness in those prescribed Li (e.g. Affective Morbidity Index[8]; Illness Severity Index[9]). These methods offer a simple way of evaluating change in illness activity after Li initiation, but their utility is undermined by the failure to address differences between the natural course of illness and treatment effects or to assess potential confounders of treatment (e.g. co-prescription of other mood stabilizers, level of Li adherence, etc.). In recent decades, these approaches have mostly been replaced by the Retrospective Assessment of Response to Lithium Scale (which we will refer to as the Alda scale) to the extent that this is now the most widely used measure of the Li response phenotype[10].

The Alda scale comprises of two subscales; the A scale (which measures response on a 0-10 continuum) and the B scale (which describes five potential confounders of response). The A scale requires assessors to determine change in illness activity whilst receiving Li, with response rated on a 0-10 continuum. The anchor points are: '*no change or worsening'* (score=0) and *'complete response, no recurrences in the course of adequate treatment, no residual symptoms and full functional recovery'* (score=10). The B scale describes five items that may confound the response and/or the interpretation of the magnitude of any response to Li, namely: number (B1) and frequency of episodes before Li (B2); duration of (B3) and adherence with Li treatment (B4) and use of additional medications (B5). Each item is rated 0-2 and a higher B score indicates a lower level of confidence that any observed clinical improvement is a consequence of the introduction of Li. The overall rating of Li response (referred to as the Total Score or TS) is calculated by subtracting the B scale score from the A scale score, with the convention that a negative TS (i.e. B scale > A scale score) is recalibrated as TS=0. Traditionally, the TS is employed to classify response phenotypes categorically with a cut-off score of TS>=7 identifying the good response (GR) group (and some earlier studies further subdivided the group with TS<7 into partial (PR) and non-response (NR) categories (10)).

Increased reliance on the use of the Alda scale to assess Li response has brought to light variations in the interpretation or procedure for combining scale scores (see Scott et al[11]). As no consensus exists on how the A, B or TS ratings may be modified beyond the original proposals (10), decisions to change how the scale is employed lack uniformity or empiricism. For instance, genetics studies that have employed the Alda scale have operationalized Li response according to categorical divisions of the TS score[12], using continuous scores on the A scale either alone[13] or using continuous A scale scores in those individuals with a low B scale score[12]. Other investigators have reported Li response in selected subgroups, such as those demonstrating 'good' adherence (which can be identified by a low score on item B4)[14]. These variations in the use of the B scale score[10-14] suggest that some

investigators are unclear as to whether this subscale is best viewed as a measure of a unidimensional concept (overall uncertainty about Li response) or if it represents a multi-dimensional index (i.e. a method for assessing several independent modifiers of the noise to signal ratio).

Widespread use of the Alda scale has also drawn attention to the limited data on its basic psychometric properties (only two studies exist). A large-scale study highlighted that the inter-rater reliability of the TS was sub-optimal and that it decreased rather than increased after the assessors had participated in a training course[15]. Also, the researchers reported the reliability of the A scale score was especially impaired in cases with high B scale scores[15]. This led Manchia et al[15] to suggest that an additional method of assessing Li response is to report the A scale score (as a continuous rating) in individuals with a B scale score =<3. In a small-scale study, Tighe et al[16] demonstrated that the test-retest reliability of the Alda scale is moderately good if the rating is undertaken in conjunction with a detailed one-hour assessment aimed at standardizing the evaluation of the scales and items. These proposals were based on the (unproven) assumption that all the problems encounter when employing the Alda scale were associated with the accuracy of the ratings and in the absence of any knowledge about other aspects of the quality of the performance of the scale[16]. However, as illustrated by Scott et al[11], the sub-optimal reliability of the Alda scale means that employing any of these different approaches for scoring the Alda scale currently leads to the identification of different patterns of clinical predictors of Li response. As such, a major implication of all of the above issues (i.e. variations in methods of rating the Alda scale and the sub-optimal reliability of the scale) is that they undermine our ability to compare findings across studies.

An expressed goal of the Consortium of Lithium Genetics (ConLiGen) is to optimize the measurement of Li response phenotypes, including examination of the accuracy of the Alda scale[17]. Since the scale was introduced, the assessment of the performance of measurement tools has extended beyond psychometrics (often described as classical test theory (CTT)), to include item response theory (IRT) and clinimetrics (a scientific method focused on the quality and performance of measurements in medical research and clinical practice). Adoption of these newer approaches has been accompanied by several publications on how to undertake a systematic assessment of the quality of performance of rating scales and clinical instruments[18-25]. We decided to evaluate the performance of the Alda scale using a database containing >2500 cases where the A and B scale scores had been used to classify Li response phenotypes in genetics research projects. We primarily followed the approaches described in these key recommendations in the COSMIN (Consensus-based standards for the selection of health measurement instruments[26]) and according to template used to report criteria and findings for the core COSMIN components (for an example see Terwee et al[25]). The specific aims of this study were:

(a) to assess the performance of the Alda scale and its components according to four core clinimetric and psychometric parameters: namely the reproducibility, responsiveness, validity and interpretability;
(b) to identify if any B-scale items consistently fail to meet quality criteria for moderate or good performance;
(c) to use the findings to inform discussions about empirically-based modifications (if appropriate) of either the content of the Alda scale (including sub-scales and items) or clinical application of the scale that could benefit future research on Li response.

A secondary goal of this paper is to provide clinicians and researchers with a template that could be employed to review the performance of other existing rating scales or a procedure that could be followed in future projects that are designed to assess the clinimetric and psychometric properties of any scale.

Methods

The Consortium of Lithium Genetics (ConLiGen) research subcommittee approved a written application (from JS and FB) requesting access to the existing database to allow an examination of the performance of the Alda scale. Ethical approval exists for ConLiGen to undertake a range of studies of Li response (including projects regarding response measures)[17]. The dataset contained de-identified ratings for the Alda scale (A and B subscale scores, individual B item ratings and TS (10)) for 2321 individuals aged >=18 years.

For some sub-studies we used data on the Alda scale and/or symptom rating scales (e.g. Inventory for Depressive Symptoms and Altman Mania Rating Scale) that was extracted from the French ConLiGen dataset[11,27]. The French studies were approved by the French Ethics and Data Protection and Freedom of Information Commissions (CPPRB, RCB:2008-AO14-65-50).

Strategy for Assessing Performance of the Alda Scale

A priori, we identified the core parameters for measuring the quality of performance of the Alda scale. To assist readers, Table 1 provides information on the definition of each component (metric) and the quality criteria employed. The strategy for the assessment and the analyses chosen primarily reflect the COSMIN recommendations (Consensus-based standards for the selection of health measurement instruments[26]), with additional standards derived from GRRAS (Guidelines for reporting reliability and agreement studies[23]) or from similar guidance (e.g. protocols describing approaches to signal detection, such as machine learning or decision tree analysis[28,29]).

Table 1 about here

As shown in table 1, we defined each clinimetric and psychometric property, and then selected a maximum of two criteria to 'bench mark' good or moderate performance according to the published recommendations and standards. For reliability, responsiveness and validity, we identified two key aspects that could be examined using the available data (the exceptions were inter-rater reliability and face validity, where only one criterion was chosen). For interpretability, we examined sample size mean A and B scale scores and proportion of GR from the current ConLiGen database and other studies that recruited samples with similar demographic and clinical characteristics[11,12].

Statistical Analyses

Below, we give an overview of the analyses used to assess the criteria (described in Table 1). Additional details are provided for those approaches that may be less familiar to readers or that require further explanation (further details are available from the authors upon request). It should be noted that, by necessity, some steps in the statistical analytic plan were iterative. For instance, we could not determine in advance whether to use item response theory (IRT) or multi-nomial logistic regression (MNLR) to assess incremental validity (#3.4) as until findings from the factor analysis (#1.2) were available (if this indicated that the B scale was unidimensional, we could proceed with IRT; if multi-dimensional, MNLR is the approach recommended in the literature, etc.) (20). As such, Table 1 identifies the analyses that could be used to explore each criterion. Variables used in the analyses are extracted from the ConLiGen database unless otherwise stated. We used SPSS version 24 for most analyses, but some were undertaken using R software (e.g. the network analysis). The assessment proceeded as follows:

1. REPRODUCIBILITY

1.1 Reliability: We assessed inter-rater reliability by measuring the extent to which three independent raters assigned the same score to the same item, scale or category for the A, B and TS components of the Alda scale for 30 sets of clinical case records from the French site. The three raters (BE, CB, PG) were all psychiatrists with established clinical expertise in BD who had worked in a specialist mood disorders clinic and who had used the Alda scale in research; one is a member of ConLiGen and another participated in the previous study of reliability of the Alda scale (15). We estimated inter-rater reliability using the weighted kappa as this is recommended for analyses that include assessments of ratings of single items[30], and a weighted kappa >= 0.7 as the criterion for good reliability.

1.2 Internal consistency: This was examined by estimating the Cronbach alpha for the B scale. As the B scale has a limited number of items and may measure several constructs, we followed published recommendations and report the standardized Cronbach alphas[31]. For the same reason, we also examined the values of the corrected item-total correlations to determine if any items might warrant

removal (a low value means the item is poorly correlated with the overall scale)[32]. The poorest performing item was removed, and the reliability was re-estimated. The criterion values were 0.7 for the standardised Cronbach alpha and 0.15-0.5 for the mean item-total correlation respectively[32].

We undertook a factor analysis to assess the dimensionality of the B scale. We used the maximum likelihood procedure as it is argued that it does not require the items to have a normal distribution[33]. Eigen values and scree plots were examined, and we used an Eigen value >1 and the point of inflection to determine the number of factors. The minimum factor loading for any item was set at 0.3 (according to recommendations)[33].

1.3. Agreement: We used data from a subsample of 30 cases (the same group used in 1.1) to explore a key aspect of agreement, namely error measurement. First, we calculated the standardised error measurement (SEM) for the A and B scales. The SEM is a useful additional measure of reproducibility as it reflects both the reliability of the scale and its range of scores. Second, we estimated the smallest detectable change (SDC) using an established formula (see Table 1). Then we compared the SDC to the minimal important change (MIC) that is regarded as clinically relevant for each Alda subscale. We defined the MIC according to published recommendations, using a score of seven for the A scale and TS, and three for the B scale[15]. For evaluative purposes, agreement is rated as positive if the SDC is smaller than the MIC[19].

2. RESPONSIVENESS

2.1 Treatment effects: The data available are not ideally suited to this analysis (as the Alda scale relies on a single retrospective assessment of change), but we report the effect sizes (ES) of the A and B scales (compared to the TS) to give an indication of the ES for response or confounding respectively. Also, we examined the area under the curve (AUC) for classifying cases as GR.

2.2 Longitudinal construct validity: A subsample of 50 individuals had symptom severity ratings and Alda scores available. We explored the proportion of individuals whose scores on the 16-item self-rated version of the Inventory for Depressive Symptoms (QIDS-SR) and the 5-item self-rated Altman Mania Rating Scale (ASRM) indicated that they had minimal or no BD symptoms (QIDS-SR <6; ASRM <5) by 18 months follow-up. We used ROC (receiver operating curve) analysis to estimate the AUC for symptomatic improvement compared to Li response as measured by the TS and then for the A score. We then examined the correlation between change in the symptom ratings and Alda scores (criterion value: r >=.5)

3. VALIDITY

3.1 Content validity: Two senior researchers who are ConLiGen investigators (BE, FB) independently rated the content validity (CV) of the A scale and B scale items using the Content Validity Index (CVI)[34]. Each CVI component (relevance, clarity, simplicity, ambiguity) is rated on a 1-4 scale (with a low score indicating inferior CV) and we compared scores with the criteria listed in Table 1. Inter-rater differences in scores were resolved by consensus, and the assessors then completed a qualitative review, providing written feedback on the nature of any perceived problems regarding those components with sub-optimal CV scores.

3.2 Construct validity: The data available allowed an examination of structural validity only. Guidelines vary in recommendations for assessing structural validity, so we selected two options that offered additional insights into the performance of the Alda scale and extended the analyses of reproducibility. We explored associations between items measured by the B scale and their relationship to the A scale using correlational and regression analyses. Also, we used network analysis to explore the connections between each B item and the A scale[35]. We used the partial correlation procedure as this produced the most robust model for the ConLiGen data[36] and provide the network diagram as this summarizes the findings of the other analyses of construct validity (we have not included the centrality plots or tables).

3.3 Face validity: Floor or ceiling effects are said to be present if >=15% of the ratings fall into the lowest or highest scoring band[24]. To assess this, we estimated the proportion of ratings on the A and B scales and the TS that scored zero or 10.

3.4 Incremental validity: As well as reviewing correlational and regression analyses, we undertook MNLR to model how each item on the B scale contributes to the separation of the sample into GR versus the other (PR/NR) categories. Next, we used a machine learning approach, namely classification and regression tree (CART) modelling to produce a decision tree to demonstrate the incremental validity of each B scale item to the correct categorization of Li response phenotypes[37]. It is emphasised that the focus is on determining the contribution of each B items to the overall classification and to generate algorithms that provide insights into which B items were critical to classification (versus those that played a non-significant role). With this goal in mind, we employed the CHAID (chi-squared automatic interaction detection) procedure which predicts categorical classifications from several predictor variables. We cross-validated the models by undertaking training and test analyses (each using 50% of the data)[38].

The classification tree provides a graphical representation of a series of decision rules, with the best predictor appearing at the first step; the tree stops growing when no improvement on the classification is possible. We provide the model derived from the analysis of a two-group classification of GR

versus other response categories, but (following our iterative procedure), we explored other classifications (e.g. NR, PR and GR) to determine if the same items contributed to different models and/or improved the optimal classification.

4. INTERPRETABILITY

We compared mean scores for the A and B scale scores and the number of cases classified as GR (according to a TS score >=7) using the current dataset (N=2321) and data from three published studies[11,12,39]. The studies were selected as they report Alda ratings for Li treated subgroups (i.e. subsamples receiving Li monotherapy and/or other mood stabilizers), the samples had similar sociodemographic and illness characteristics and all centres had experience of using the Alda scale. We compared Alda scale information using standardized mean differences (for A and B scale scores) or differences in the proportion of cases classified as GR (for TS categories).

Table 2 about here

Results

In this section, we provide an overview of the outcomes of each analysis in the text and highlight the key findings in Table 2 and Figures 1 and 2. Further information, including comments on the interpretation of each analysis are provided in Table 1S in the supplementary materials and additional statistical outputs are detailed in Supplementary Tables 2S-9S. To briefly summarise the results:

1. REPRODUCIBILITY- clinical assessors most reliably identify individuals categorised as GR according to their TS classification. Reliability is poor for the B scale (both categorical or continuous approaches).

As the B scale has a limited number of items and scoring ranges, it is unsurprising that the Cronbach alpha for the scale is sub-optimal (see Table 1S). The alpha can be improved if the item measuring duration of Li treatment (B3) is removed. Findings for internal consistency suggest that the B scale is not measuring a single underlying construct, which is confirmed by the factor analysis. The B scale comprises of two interpretable factors: illness activity (number and frequency of episodes) and treatment complexity (polypharmacy and adherence). Item B3 fails to load onto a factor (it shows a sub-threshold loading on the treatment dimension).

It is important to note that the agreement for the B scale is suboptimal (SDC>MIC); the measurement error in the B scale score could lead to phenotypic misclassifications (e.g. sampling frames using a cut-off of B scale score=<3 may be subject to error).

Table 2 about here

2. RESPONSIVENESS- the small sample available for the analysis of the longitudinal construct validity should be noted. Overall, the A scale shows the best performance, with sub-optimal performance for the B scale and TS.

3. VALIDITY- the CVI ratings of the Alda subscales indicate that the core themes identified by each individual B item are clinically relevant and important, but the CV of item B2 was sub-optimal (score=11/16) and items B4 and B5 just met the *a priori* criterion (score=12/16). The qualitative assessment identified potential benefits from clarifying descriptions of each item (the introductory text) and/or modifying text regarding the anchor points (for scoring each component). Feedback highlighted that (i) the content of B items related to illness activity (B1 and B2) were not consistent with the parameters of illness activity assessed by the A scale (e.g. severity is included in the A rating but not the B scale); (ii) the content of item B5 varies between different published versions of the scale (10, 17); and (iii) it may be helpful to simplify and clarify ratings of the treatment-related items (e.g. B4 could be split into two items: plasma Li level and adherence level, etc).

Figure 1 about here

Statistical analyses demonstrate the lower utility of item B2 (e.g. the MNLR identifies that B2 does not contribute to the differentiation of response subgroups; the network analysis highlights that B2 is positively correlated with the A scale score as well as with B1). Also, the network map (see Figure 1), shows that B1 and B3 are negatively correlated with each other.

Figure 2 provides a diagram of the classification tree generated by one of the CART analyses. This tree shows that items B4, B5, B3 and B1 could be used in an algorithm to classify about 90% of individuals into the correct response category (GR versus PR/NR). Item B2 does not contribute to the classification tree for any CART model and the classification rate does not increase even when item B2 is forced into the model.

Figure 2 about here

4. INTERPRETABILITY- The mean scores for the A scale were similar across the four studies (range 5.6-6.35), but the mean score for the B scale differed significantly (range 2.1-3.2; standardized mean difference: 0.32, 95% confidence intervals: 0.21-0.45, $p<0.001$). These differences were not significantly associated with variations in demographic or illness characteristics across samples. The

difference in B scale scores largely explained the difference in the proportion of cases classified as GR across studies (range 17-31%).

Discussion

This paper demonstrates that clinimetrics offer an important extension to traditional, psychometric methods for assessing the quality and performance of clinical measurement tools. Further, it highlights that findings from precision approaches to Li response (e.g. genetic and biomarker studies) will disappoint unless we pay equal attention to precision measurement of clinical phenotypes. Additionally, the study raises many topics for discussion, although various constraints mean that we have focused the discussion on an overview of the study strengths, examination of a selection of key findings, and summarizing the main implications of the findings and offering recommendations for research reporting.

The current study has several strengths, some of the most notable are briefly highlighted. For instance, several aspects of the methodology are innovative. The study uses guidelines (such as COSMIN[26]) that are frequently employed to determine quality and performance of general health outcome measures and applied this assessment template to a rating scale employed for in BD for the first time. The study combined contemporary guidance on how to optimally evaluate the Alda scale with the use of state-of-the-art statistical approaches such as network and CART analyses to gain a full understanding of the data. The sample size and multi-centre, multi-national recruitment of >2500 individuals is a significant strength of the current project. Furthermore, a consistent package of assessments was employed across the consortium and assessors were trained in the use of the rating scales. The study sampling and basic methodology contrasts markedly with evidence from the most recent meta-analysis of Li response which commented on the small, often biased samples recruited to studies and low-quality ratings applied to much of the data pooled for the estimation of Li effects[40]. Importantly, this study demonstrates the benefits of an integrated science approach (i.e. the collaboration between those specialising in clinical phenotypes and psychopathology, genetics, biomarkers and psychopharmacology) to determine empirically which aspects of the performance of the Alda scale could be improved and offers a template that can be applied to the examination of other clinical measurement tools.

An example of the strength of the study is that specific findings were revealed by combining psychometrics and clinimetrics. Traditional psychometric approaches assume that clinical scales are measuring an underlying latent variable. Whilst this assumption is largely true for the A scale of the Alda, it is less applicable to the B scale. In this study, classical psychometric tests confirmed that the B scale has modest reliability and lacks internal consistency[15,16], but also demonstrated for the first time that it is vulnerable to error measurement (the SDC > MIC) and that it is multidimensional.

Further, correlational, regression and network analyses show that the B items and A scale have significant associations in the expected directions, but also some in the opposite direction. Taken together, the findings suggest that the B scale is not reliably measuring a single construct (uncertainty in Li response), but that it appears to perform as an index or inventory (identifying several independent factors that undermine confidence that observed changes in illness activity are attributable to exposure to Li). As such, this study has two key implications. First, it suggests that simply summing together the scores for the individual items on the B scale (and then subtracting this composite score from the A scale score) may not represent the optimal approach to phenotypic classification. Second, current methods for minimising the previously recognized problems associated with the rating of the B scale[15,16], may have less influence than desired because they were focused on improving its basic psychometrics whilst failing to address the weaknesses exposed by clinimetric assessment.

Clinimetrics is a methodological discipline that evaluates consistency, validity and responsiveness and promotes a set of rules to govern the structure of indexes and the choice of component variables[28,41]. This approach generally brings greater flexibility and provides more sophisticated information about the actual performance of measures employed in making clinical judgements than reliance on psychometrics alone[26]. For example, in the current study, the network, MNLR and CART analyses highlight that the weaknesses of the Alda scale are mainly due to the quality of performance of the B scale, but also suggest that some individual items make a minimal, non-significant or unreliable contribution to the differentiation of responders and non-responders (e.g. B2). Critically, the CART model of incremental validity shows that an advantageous approach to utilizing the information gathered by the B scale may be to consider the items in sequence, as part of an algorithm (with 'if-then' rules). For instance, the first step in such an algorithm might be to assess 'noise' by examining the scores for items B4 (Li adherence) and B5 (additional medications prescribed during Li treatment). If the 'noise' level is low, the A scale can be used to rate change in illness activity during Li treatment, followed by an assessment of confounders associated with the strength of the response 'signal' (e.g. prior illness activity as measured by item B1), etc. This strategy has clinical validity as, for instance, it would be counter-intuitive to assess Li response in an individual who is persistently non-adherent (3); and a recent international consensus statement has emphasized the importance of differentiating non-response from non-adherence[42].

The above findings alongside the qualitative and quantitative assessment of content validity highlight some weakness in the B scale, but also indicate that opportunities exist to enhance its performance. This work is now underway, with ConLiGen collaborators examining simple solutions such as ensuring consistency in the wording of B scale items in the different published versions of the Alda scale (e.g. item B5)[10,17], as well as discussing more radical options (such as extending the scoring

range of existing items; modifying the list of B subscale items; proposing revisions based on the dimensions identified by factor analysis, etc). However, future revisions of the Alda scale need careful consideration and any proposed changes would need to be tested against the same standards and criteria described in this paper. In the interim, it is important to emphasize three issues. First, any assessment tool subjected to intense scrutiny is likely to show some deficits and so, whilst noting possible weaknesses in the Alda scale, it is noteworthy that it remains the only measure of Li response that considers change in illness activity in the context of confounders. This speaks to its significance to the research field and its clinical applicability. As such, whilst we are advocating for some modifications to the scale and its modus operandi, we are not recommending its withdrawal. Second, the items in the B scale are highly relevant (e.g. level of adherence; polypharmacy; illness activity prior to Li initiation) and any alterations must not undermine their established utility. Third, whilst this study highlights empirical approaches to assessing the performance of the Alda scale, our analyses do not guarantee that any proposed modifications will enhance the correct classification of Li responders, minimise false positives, etc. This critical step can only be achieved through further study, such as comparing genetic or biomarker findings when using the different approaches to assessing Li response, such as using the Alda scale in an algorithm (as reported by Scott[43]).

There are several limitations to this project. For example, it is known that classical test theory is less useful when examining scales with fewer items and, despite using approaches to minimise the impact on the analyses, the brevity of the B scale may limit the interpretability of some of our findings. Also, assessments such as the analysis of treatment effects (#2.1) and longitudinal construct validity (#2.2) relied on data from smaller subsamples whilst other findings were cross-validated by creating subgroups from within the ConLiGen sample (e.g. CART analyses) rather than testing the new models in independent samples. Importantly, there was no gold standard measure of Li response against which the Alda scale can be compared. These and other shortcomings should be considered when reviewing the findings, and it can be argued that our findings are applicable to the current study populations but need further replication in new samples.

In conclusion, genetics and biomarker studies are resource intensive, and phenotypic misclassifications can be costly, increase sample size requirements and ultimately produce unreliable findings[3,44,45]. The Alda scale is important in helping to characterise the Li response phenotype but, to date, we lacked a complete understanding of its strengths and weaknesses. This innovative study attempts to start the process of enhancing our ability to assess Li response in a reliable and valid way. In the short term, we have three recommendations regarding the use of the Alda scale: researchers should report which version of the Alda scale was employed[10,17]; publications should provide a detailed breakdown of the item by item scores for the B scale; and the methodology should describe which scoring system was used to classify Li response or non-response (e.g. excluding high B scale

scores, only rating the A scale, excluding non-adherers, or using the syntax for the algorithm presented in this manuscript, etc). This will aid researchers when making comparisons of findings across genetics or biomarker studies. Ideally, a flowchart (like a 'CONSORT' diagram for randomized trials) could be provided to denote the proportion of individuals excluded from the study at each step in the selection procedure or to give further insights into how cases were classified into Li responder categories. In the longer term, revisions of the Alda scale may be instituted by ConLiGen as part of its commitment to an integrated science approach to precision psychiatry.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Additional references e.g. for clinimetrics, symptom rating scales, etc. are available from the authors

1. Bellivier F, Marie-Claire C. Lithium Response Variability: New Avenues and Hypotheses. In: Malhi G, Masson M, Bellivier F. (eds) The Science and Practice of Lithium Therapy. 2017. Springer: London. p23-35.

2. Machado-Vieira R, Luckenbaugh DA, Soeiro-de-Souza MG, et al. Early improvement with lithium in classic mania and its association with later response. J Affect Disord. 2013;144(1-2):160-164.

3. Scott, J., Etain, B., & Bellivier, F. Can an integrated science approach to precision medicine research improve lithium treatment in bipolar disorders? Frontiers in Psychiatry, 21st August 2018; 9, 1-10.

4. Perlis R, Smoller J, Ferreira M, et al. A genome-wide association study of response to lithium for prevention of recurrence in bipolar disorder. Am J Psychiatry 2009; 166:718-725.

5. Oedegaard KJ, Alda M, Anand A, Andreassen OA, Balaraman Y, Berrettini WH, et al. The Pharmacogenomics of Bipolar Disorder study (PGBD): identification of genes for lithium response in a prospective sample. BMC Psychiatry 2016, 16:129.

6. Response to Lithium Network: R-LiNK (C.I.: F. Bellivier, Paris). https://rlink.eu.com. Accessed December 1st, 2018.

7. Song J, Bergen S, Di Florio A, Karlsson R, Charney A, Ruderfer D, et al. Genome-wide association study identifies SESTD1 as a novel risk gene for lithium-responsive bipolar disorder. Molecular Psychiatry, 2016; 21(9), 1290-1297.

8. Coppen A, Peet M, Bailey J, Noguera B, Burns B, Swani, M, Maggs, R, Gardner R. Double-blind and open prospective studies of lithium prophylaxis in affective disorders. Journal of Neurology, Neurosurgery and Psychiatry, 1973, 76:501-510.

9. Maj M, Arena F, Lovero N, Pirozzi R, Kemali D. Factors associated with response to lithium prophylaxis in DSM III major depression and bipolar disorder. Pharmacopsychiatry, 1985, 18 (5) 309-313.

10. Grof P, Duffy A, Cavazzoni P, et al. Is response to prophylactic lithium a familial trait? J Clin Psychiatry 2002; 63:942-947.

11. Scott J, Geoffroy P, Sportiche S, Brichant-Petit-Jean C, Gard S, Kahn JP, Azorin J, Henry C, Etain B, Bellivier F. Cross-validation of clinical characteristics and treatment patterns associated with phenotypes for lithium response defined by the Alda scale. J Affect Disord. 2017; 208:62–7.

12. Hou L, Heilbronner U, Degenhardt F, Adli M, Akiyama K, Akula N, et al. Genetic variants associated with response to lithium treatment in bipolar disorder: a genome-wide association study. Lancet 2016; 387:1085–93.

13. Lee MTM, Chen CH, Lee CS, et al. Genome-wide association study of bipolar I disorder in the Han Chinese population. Mol Psychiatry 2011; 16:548-556.

14. Chen C, Lee C, Chen H, Wu L, Chang J, Liu C, et al. GADL1 variant and medication adherence in predicting response to lithium maintenance treatment in bipolar I disorder. BJPsych Open 2016, 2:301–6.

15. Manchia M, Adli M, Akula N, et al. Assessment of response to lithium maintenance treatment in bipolar disorder: a Consortium on Lithium Genetics (ConLiGen) report. PLoS One 2013; 8: e65636.

16. Tighe S, Ritchey M, Schweizer B, Goes F, MacKinnon D, Mondimore F, Raymond DePaulo J, McMahon F, Schulze T, Zandi P, Potash J. Test-retest reliability of a new questionnaire for the retrospective assessment of long-term lithium use in bipolar disorder. J Affect Disord. 2015 Mar 15;174: 589-93.

17. Schulze T, Alda M, Adli M, Akula N, Ardau R, Bui E, et al. The International Consortium on Lithium Genetics (ConLiGen): an initiative by the NIMH and IGSLI to study the genetic basis of response to lithium. Neuropsychobiology 2010, 62:72–8.

18. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol. 2001, Dec;54(12):1204-17.

19. de Vet H, Terwee C, Knol D, Bouter L. When to use agreement versus reliability measures. J Clin Epidemiol. 2006, 59, 1033-1039.

20. Fayers P, Hand D Causal variables, indicator variables and measurement scales: an example of quality of life. J. R. Stat. Soc. 2002, 165, 233-261.

21. Feinstein A. An additional basic science for clinical medicine: IV. The development of clinimetrics. Ann Intern Med. 1983, Dec;99(6):843-8.

22. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000 May;53(5):459-68.

23. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011 Jan;64(1):96-106.

24. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007 Jan; 60:34-42.

25. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012 May;21(4):651-7.

26. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HC. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010 Mar 18; 10:22.

27. Sportiche S, Geoffroy P, Brichant-PetitJean C, Gard S, Khan JP, Azorin J, Henry C, Leboyer M, Etain B, Scott J, Bellivier F. Clinical factors associated with lithium response in bipolar disorders. Aust N Z J Psychiatry. 2017, May;51(5):524-530.

28. Feinstein A. Clinimetrics. 1987. (Chapter 6: The theory and evaluation of sensibility). Murray: Westford, MA. p141-166.

29. Marx R, Bombardier C, Hogg-Johnson S, Wright J. Clinimetric and psychometric strategies for the development of a health measurement scale. J Clin Epidemiol. 1999; 52, 105-111.

30. Kramer M, Feinstein A. The biostatistics of concordance. Clin Pharmacol Ther, 1981; 29, 111-123.

31. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011, Jun 27; 2:53-55.

32. Neuendorf K. The Content Analysis Guidebook. 2nd Edition. 2016. Sage Publications: London. p23-61.

33. Lawley D, Maxwell A. Factor analysis as a statistical method. J. R. Stat. Soc., 1962; 12 (3) pp. 209-229.

34. Yaghmale F. Content validity and its estimation. J Medical Education, 2003; 3 (1), 25-27

35. Epskamp S, Maris G, Waldorp L, Borsboom D. Network psychometrics. In: Irwing, P., Hughes, D., & Booth, T. (Eds.), Handbook of Psychometrics. 2018. New York: Wiley. p1-17.

36. Epskamp S, Fried E. A tutorial on regularized partial correlation networks. Psychological Methods. 2018, 48 (4), 1-18.

37. Loh W, Chen C, Hordle W, Unwin A. Improving the precision of classification trees. Ann Appl Stat 2009, 3:1710–1737.

38. Kass GV. An exploratory technique for investigating large quantities of categorical data. Appl Stat 1980, 29:119–127.

39. Garnham J, Munro A, Slaney C, et al. Prophylactic treatment response in bipolar disorder: results of a naturalistic observation study. J Affect Disord 2007; 104:185-190.

40. Hui TP, Kandola A, Shen L, Lewis G, Osborn DPJ, Geddes JR, Hayes JF. A systematic review and meta-analysis of clinical predictors of lithium response in bipolar disorder. Acta Psychiatr Scand. 2019 Aug;140(2):94-115.

41. Fava G, Tomba E, Sonino N. Clinimetrics: the science of clinical measurements. Int J Clin Pract. 2012 Jan;66(1):11-5.

42. Howes OD, McCutcheon R, Agid O, de Bartolomeis A, van Beveren NJ, Birnbaum ML, et al. Treatment-Resistant Schizophrenia: Treatment Response and Resistance in Psychosis (TRRIP) Working Group Consensus Guidelines on Diagnosis and Terminology. Am J Psychiatry. 2017 Mar 1;174(3):216-229.

43. Scott J. Barriers to optimising benefits of lithium- reliability and utility of response scales and clinical strategies for recognizing the efficacy-effectiveness gap. Session 15, 31st ECNP Conference, Barcelona, Spain, October 6-9, 2018. European Neuropsychopharmacology, Supplement. 2018 November, In press.

44. Buyske S, Yang G, Matise TC, Gordon D When a case is not a case: effects of phenotype misclassification on power and sample size requirements for the transmission disequilibrium test. Hum Hered 2009, 67: 287–292.

45. Fava G, Guidi J, Grandi S, Hasler G. The missing link between clinical states and biomarkers in mental disorders. Psychother Psychosom. 2014;83(3):136-41.

Table 1: Key metrics selected to assess the performance of the Alda scale, operationalization of definitions and quality criteria employed.

Metric	Definition[a]	Quality criteria (derived from consensus recommendations & standards reported in guidelines)[b]
REPRODUCIBILITY		
1.1 Reliability	The extent to which patients can be distinguished from each other, despite measurement errors	Weighted Kappa ≥ 0.70.
1.2 Internal consistency	Extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct	Factor analyses performed on adequate sample size to assess if scale is unidimensional or multidimensional. Cronbach's alpha(s) calculated, with the aim of achieving an alpha of 0.7. Also, as the B scale has a small number of items (defined as <10 items &/or a scale using ordinal ratings), scale composition was determined by examining item-total correlations.
1.3. Agreement	The extent to which the scores on repeated measures are close to each other (absolute measurement error)	Several options are suggested, we note the most commonly reported are: Measurement error estimated as the Standardised error of measurement (SEM) using the formula SEM= Standard Deviation x √1-reliability. Minimal important change (MIC) > Smallest detectable change (SDC). The SDC was calculated using the formula: SDC=1.96 x √2 x SEM.
RESPONSIVENESS		
2.1 Pre- to post-test or 'Treatment' effects	The ability of a questionnaire to detect clinically important changes	Effect Size (ES) >0.3 for moderate effect, or >0.8 for large effect.

VALIDITY[c]	over time. Distribution based concept, where largest change indicates greatest response	Area under the curve (AUC) ≥ 0.70.
		Criteria for content validity (CV) highlight that a clear description is provided of the measurement aim, target population, concepts that are being measured, and item selection.
3.1 Content validity **2.2 Longitudinal construct validity**	Anchor-based measurement that examines the extent to which scores on a questionnaire and change scores relate to other, external measures or standards Extent to which the construct of interest is comprehensively sampled by the items in the questionnaire	Area under the curve (AUC) ≥ 0.70. Correlation of responsiveness rating and change scores on another appropriate instrument is >0.5. We employed the Content Validity Index (CVI) to assess whether all components of the Alda scale are relevant, clearly described & measure all key aspects of the construct. The CVI score for each item or scale assessed

		can range from 4-16. We set two criteria for acceptable CV for the A scale and each B item: CVI >= 12 & no individual component score <3.
3.2 Construct Validity	Construct validity includes three core components (substantive, structural & external) that reflect whether the scale consists of effect indicators	We explored the structural validity by assessing correlational, regression & network analyses.
3.3 Face validity	Closely related to content validity. Judgement regarding the extent to which the instrument adequately reflects clinical observation	Can be assessed using Floor and Ceiling effects. We employed a typically reported criterion for excluding these effects i.e. that ≤15% of the respondents achieve the highest or lowest possible scores.
3.4 Incremental validity	Extent to which each questionnaire item can increase knowledge or prediction of response beyond what is already known or based on the total score of an existing measure	Several analyses can be used to examine this construct- Individual item or subscale contributions can be assessed using regression models. Item response theory (IRT) if the scale is unidimensional otherwise Multinomial logistic regression (MNLR) is recommended. Other approaches include signal detection or machine learning e.g. Classification and Regression Tree (CART) analysis
INTERPRETABILITY		
4. Interpretability	Degree to which one can assign	A recommended approach to assessing this item is to examine mean scores

	qualitative meaning to quantitative scores	and standard deviations (SD) for at least four relevant subgroups of patients.

[a]Several definitions of each term are available, but for consistency we employed those given in the guidelines & recommendations listed in the introduction & methods (i.e. the same publications from which we derived the quality criteria);

[b]It should be noted that findings from the same analyses can be used to explore more than one quality criterion;

[c]We could not assess criterion validity as no appropriate data exist to permit this analysis; so, the concept is excluded from the table.

Table 2: Summary of findings regarding performance of the Alda scale and its components

PERFORMANCE MEASURE:	ALDA SCALE COMPONENTS									
	Continuous scores			Categories[a]		Items				
	TS	A	B	TS	B	B1	B2	B3	B4	B5
1. REPRODUCIBILITY										
1.1 Reliability		+	--	+	--	+	+/-	+	+	+/-
1.2 Internal consistency						+	+	+/-	+	+

1.3. Agreement		+	+/-	+						
2. RESPONSIVENESS										
2.1 Treatment effects		++	+	+/-						
2.2 Longitudinal construct validity		+		+/-						
3. VALIDITY										
3.1 Content validity		+				+	-	+	+	+
3.2 Construct validity		++				+/-	-	+/-	+	+
3.3 Face validity	+	++	+							
3.4 Incremental validity		++		+	+/-	+	--	+/-	++	++
4. INTERPRETABILITY										
4.1 Comparison of four populations		++	-	+						

Shaded boxes indicate that analyses were not undertaken (because it was inappropriate or did not offer additional insights into the performance of the Alda scale)

[a] Categories defined as: TS (>=7 vs <7); B (>=4 vs B<3)

++ Good performance; + Performance>=criterion; +/- Performance moderate or borderline acceptable;

- Suboptimal performance; -- Poor performance

Table 1: Key metrics selected to assess the performance of the Alda scale, operationalization of definitions and quality criteria employed.

Metric	Definition[a]	Quality criteria (derived from consensus recommendations & standards reported in guidelines)[b]
REPRODUCIBILITY		
1.1 Reliability	The extent to which patients can be distinguished from each other, despite measurement errors	Weighted Kappa ≥ 0.70.
1.2 Internal consistency	Extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct	Factor analyses performed on adequate sample size to assess if scale is unidimensional or multidimensional. Cronbach's alpha(s) calculated, with the aim of achieving an alpha of 0.7. Also, as the B scale has a small number of items (defined as <10 items &/or a scale using ordinal ratings), scale composition was determined by examining item-total correlations.
1.3. Agreement	The extent to which the scores on repeated measures are close to each other (absolute measurement error)	Several options are suggested, we note the most commonly reported are: Measurement error estimated as the Standardised error of measurement (SEM) using the formula SEM= Standard Deviation x √1-reliability. Minimal important change (MIC) > Smallest detectable change (SDC). The SDC was calculated using the formula: SDC=1.96 x √2 x SEM.
RESPONSIVENESS		
2.1 Pre- to post-test or 'Treatment' effects	The ability of a questionnaire to detect clinically important changes over time. Distribution based concept, where largest change indicates greatest response	Effect Size (ES) >0.3 for moderate effect, or >0.8 for large effect. Area under the curve (AUC) ≥ 0.70.
2.2 Longitudinal construct validity	Anchor-based measurement that examines the extent to which scores on a questionnaire and change scores relate to other, external measures or standards	Area under the curve (AUC) ≥ 0.70. Correlation of responsiveness rating and change scores on another appropriate instrument is >0.5.

VALIDITY[c]		
3.1 Content validity	Extent to which the construct of interest is comprehensively sampled by the items in the questionnaire	Criteria for content validity (CV) highlight that a clear description is provided of the measurement aim, target population, concepts that are being measured, and item selection. We employed the Content Validity Index (CVI) to assess whether all components of the Alda scale are relevant, clearly described & measure all key aspects of the construct. The CVI score for each item or scale assessed can range from 4-16. We set two criteria for acceptable CV for the A scale and each B item: CVI >= 12 & no individual component score <3.
3.2 Construct Validity	Construct validity includes three core components (substantive, structural & external) that reflect whether the scale consists of effect indicators	We explored the structural validity by assessing correlational, regression & network analyses.
3.3 Face validity	Closely related to content validity. Judgement regarding the extent to which the instrument adequately reflects clinical observation	Can be assessed using Floor and Ceiling effects. We employed a typically reported criterion for excluding these effects i.e. that ≤15% of the respondents achieve the highest or lowest possible scores.
3.4 Incremental validity	Extent to which each questionnaire item can increase knowledge or prediction of response beyond what is already known or based on the total score of an existing measure	Several analyses can be used to examine this construct- Individual item or subscale contributions can be assessed using regression models. Item response theory (IRT) if the scale is unidimensional otherwise Multinomial logistic regression (MNLR) is recommended. Other approaches include signal detection or machine learning e.g. Classification and Regression Tree (CART) analysis
INTERPRETABILITY		
4. Interpretability	Degree to which one can assign qualitative meaning to quantitative scores	A recommended approach to assessing this item is to examine mean scores and standard deviations (SD) for at least four relevant subgroups of patients.

[a]Several definitions of each term are available, but for consistency we employed those given in the guidelines & recommendations listed in the introduction & methods (i.e. the same publications from which we derived the quality criteria);
[b]It should be noted that findings from the same analyses can be used to explore more than one quality criterion;
[c]We could not assess criterion validity as no appropriate data exist to permit this analysis; so, the concept is excluded from the table.

Table 2: Summary of findings regarding performance of the Alda scale and its components

PERFORMANCE MEASURE:	ALDA SCALE COMPONENTS									
	Continuous scores			Categories[a]		Items				
	TS	A	B	TS	B	B1	B2	B3	B4	B5
1. REPRODUCIBILITY										
1.1 Reliability		+	--	+	--	+	+/-	+	+	+/-
1.2 Internal consistency						+	+	+/-	+	+
1.3. Agreement		+	+/-	+						
2. RESPONSIVENESS										
2.1 Treatment effects		++	+	+/-						
2.2 Longitudinal construct validity		+		+/-						
3. VALIDITY										
3.1 Content validity		+				+	-	+	+	+
3.2 Construct validity		++				+/-	-	+/-	+	+
3.3 Face validity	+	++	+							
3.4 Incremental validity		++		+	+/-	+	--	+/-	++	++
4. INTERPRETABILITY										
4.1 Comparison of four populations		++	-	+						

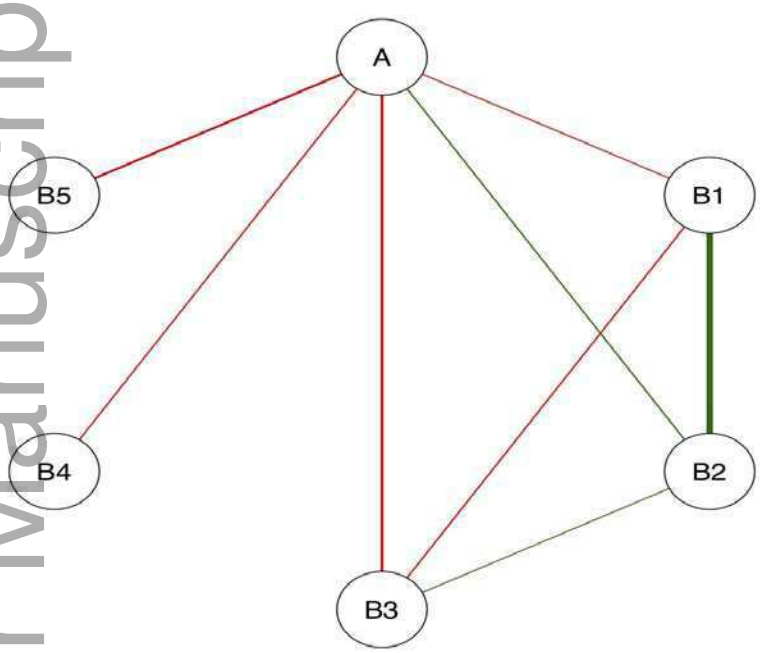
Shaded boxes indicate that analyses were not undertaken (because it was inappropriate or did not offer additional insights into the performance of the Alda scale)

[a] Categories defined as: TS (>=7 vs <7); B (>=4 vs B<3)

++ Good performance; + Performance>=criterion; +/- Performance moderate or borderline acceptable;

- Suboptimal performance; -- Poor performance

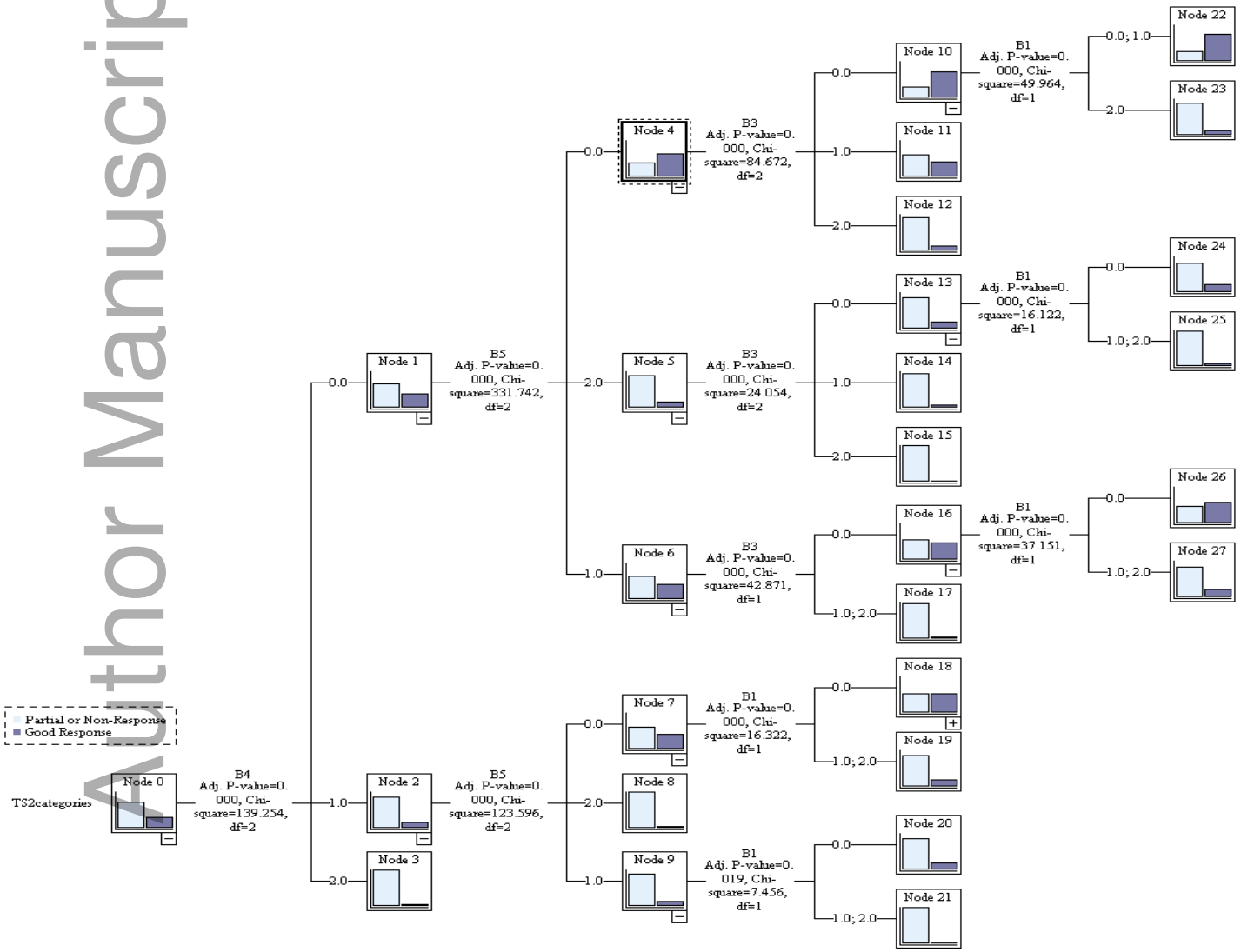
Figure 1: Network analysis summarizing the significant partial correlations between A scale & B items & any significant inter-relationships between B items (further details are given in the main text and supplementary tables)



Legend

The thickness of the line indicates the strength of the association;
Green lines indicate positive correlations; Red lines indicate negative correlations;
As shown- B1,B3,B4 & B5 are negatively correlated with the A scale score; B2 is positively correlated with the A scale score; B1 & B2 are strongly positively correlated; B3 is positively correlated with B2, but negatively correlated with B1.

Figure 2: Example of a decision tree showing the contribution of each B scale item to the classification of Lithium Responders (see main text and supplementary tables for details)



Legend: The classification tree offers a graphical representation of the series of decision rules produced via the CART (classification and regression tree) machine learning model. The analysis shown is the one undertaken for a two-group classification of Good Responders versus other categories, using B scale items only (other classification trees are available on request). The model is used to examine incremental validity and the contribution of each B item to the final classification. The best predictor appears at the first step (B4= adherence); then, depending on the score for this item, the sequence may move to item B5 as the next step (decision), followed by B3 or B5 followed by B1. The tree stopped growing when no improvement on the classification was possible. It is notable that item B2 is not included in this (or any other) tree generated. The classification table is provided in the appendix.

Author/s:
Scott, J;Etain, B;Manchia, M;Brichant-Petitjean, C;Geoffroy, PA;Schulze, T;Alda, M;Bellivier, F;ConLiGen collaborators,

Title:
An examination of the quality and performance of the Alda scale for classifying lithium response phenotypes.

Date:
2020-05

Citation:
Scott, J., Etain, B., Manchia, M., Brichant-Petitjean, C., Geoffroy, P. A., Schulze, T., Alda, M., Bellivier, F. & ConLiGen collaborators, (2020). An examination of the quality and performance of the Alda scale for classifying lithium response phenotypes.. Bipolar Disord, 22 (3), pp.255-265. https://doi.org/10.1111/bdi.12829.