The added value of the CGI-I scale in assessing global severity: a cost/benefit analysis using data from four Phase III MDD trials

Nations KR[1], Gandy-Don Sing Z[1], Spiridonescu L[1], Reinhold C[1], Brewer C[2], Baker RA[2]

[1]INC Research/inVentiv Health, [2]Otsuka Pharmaceutical Development and Commercialization

Methodological Question

Does the Clinical Global Impressions-Improvement scale (CGI-I) convey added benefit over the Clinical Global Impressions-Severity scale (CGI-S) in clinical trials, considering relative psychometric performance as well as resource costs associated with scale administration and data cleaning?

Introduction

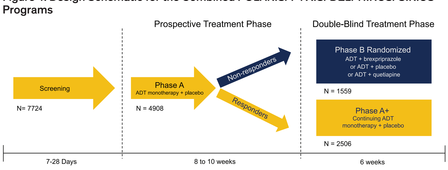
Considered a gold standard of global disease evaluation, the CGI is ubiquitously selected as a key secondary measure in CNS trials. Both the CGI-S and CGI-I are well-validated, highly sensitive to change, and frequently deployed in the same trial. However, the literature is mixed as to whether both scales are warranted, and trial designers may fail to appreciate the significant time and cost associated with administering, monitoring, and data cleaning a measure even as simple as the CGI. In major depressive disorder (MDD) trials, response is typically defined as a CGI-I score of 1 or 2 (much or very much improved), while a CGI-S rating of 1 or 2 is used to define remission (normal or borderline ill). We undertook to understand whether the CGI-S could reasonably address response, in addition to global severity and remission, obviating any need for the CGI-I.

Methods

The current analysis examined validity and reliability of the CGI-S and CGI-I, as well as cost/resource burden, using data from four Phase III adjunctive brexpiprazole trials in major depressive disorder (MDD; POLARIS-NCT01360632, PYXIS-NCT01360645, DELPHINUS-NCT01727726, SIRIUS-NCT02196506).

All four trials evaluated adjunctive brexpiprazole vs placebo in subjects with MDD who had demonstrated inadequate response to 8-10 weeks of monotherapy ADT during a prospective treatment phase. Inadequate responders entered the 6-week randomized brexpiprazole vs placebo phase. Subjects who responded to antidepressant (ADT) monotherapy in the prospective treatment phase continued ADT for an additional 6 weeks. This analysis will focus on those subjects, pooled across all four trials, who received only ADT monotherapy through endpoint.

Figure 1: Design Schematic for the Combined POLARIS/PYXIS/DELPHINUS/SIRIUS Programs



ADT = Anti-depressant treatment

Our goal was to determine whether CGI-S could be used to accurately assess both response and remission. Given the uncontested common definition of remission as CGI-S of 1 or 2 (normal or borderline ill), we focused on examining whether CGI-S could also be used to define response, and if so, what cost savings might be conveyed if CGI-I were to be eliminated from the assessment schedule entirely. The following analytical steps were undertaken to better understand these questions:

1. Convergent validity: correlation between CGI-S and CGI-I with Montgomery-Asberg Depression Rating Scale (MADRS) total score.
2. Reliability: test-retest reliability of the CGI-S and CGI-I at two proximal visits, Week 6 and Week 8, during the lead-in ADT treatment period.
3. Diagnostic accuracy: sensitivity and specificity of the CGI-S and CGI-I, at various cut point definitions, in predicting MADRS response. We evaluated receiver (ROC) operating characteristics curves and report several diagnostic accuracy indices including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), diagnostic odds ratio, and Youden index.
4. Discriminative validity: using the best threshold scores resulting from the ROC analysis, we evaluated the ability for CGI-S and CGI-I to significantly differentiate between MADRS responders and non-responders.
5. Cost analysis: we calculated total costs per scale based on investigator administration grants, rater vendor costs, database and edit check programming costs, monitoring time, and data cleaning time.

Results

Sample Characteristics

A total of 2374 subjects across the POLARIS/PYXIS/DELPHINUS/SIRIUS trials completed at least 14 weeks of monotherapy antidepressant treatment. The combined sample was predominantly female and white. Demographic and clinical characteristics are shown in Table 1.

Table 1: Demographic and Clinical Characteristics

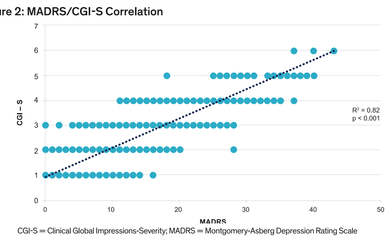
	Enrolled in Prospective ADT Treatment (Phase A) and completed week 14 (Phase A+) N = 2374
Demographics (at screening)	
Age, mean (SD)	44.49 (12.01)
Female sex, n (%)	1637 (68.9%)
Race, n (%)	
White	2066 (87.0%)
Black or African American	252 (10.6%)
Other	56 (2.4%)
BMI, mean (SD)	29.04 (6.96)
Clinical characteristic (at phase A baseline)	
Recurrent episodes, yes, n (%)	2021 (85.1%)
No. of lifetime episodes, mean (SD)	3.33 (2.56)
MADRS total score, mean (SD)	30.04 (4.39)
CGI-S rating, mean (SD)	4.46 (0.58)
Assigned ADT (at phase A baseline)	
Cymbalta, n (%)	493 (21%)
Effexor XR, n (%)	429 (18%)
Lexapro, n (%)	455 (19%)
Paxil, n (%)	291 (12%)
Prozac, n (%)	300 (13%)
Zoloft, n (%)	406 (17%)

ADT = Anti-depressant treatment; CGI-S = Clinical Global Impressions-Severity; MADRS = Montgomery-Asberg Depression Rating Scale; SD = standard deviation

Convergent Validity

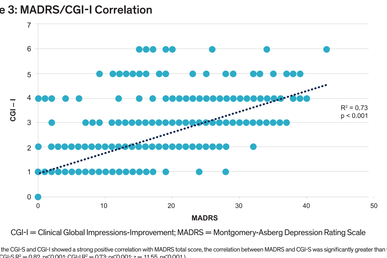
Convergent validity for the CGI-S vs CGI-I was examined by calculating correlations between each scale and the MADRS total score, then testing the difference between these correlations. Scatterplots, correlations, and test statistics are shown in Figures 2 and Figure 3.

Figure 2: MADRS/CGI-S Correlation



CGI-S = Clinical Global Impressions-Severity; MADRS = Montgomery-Asberg Depression Rating Scale

Figure 3: MADRS/CGI-I Correlation



CGI-I = Clinical Global Impressions-Improvement; MADRS = Montgomery-Asberg Depression Rating Scale

While both the CGI-S and CGI-I showed a strong positive correlation with MADRS total score, the correlation between MADRS and CGI-S was significantly greater than the MADRS and CGI-I (CGI-S R^2 = 0.82, p<0.001; CGI-I R^2 = 073; p<0.001; z = 11.56, p<0.001).

Reliability

Test-retest reliability was calculated for the CGI-S and CGI-I at two proximal visits, Week 6 of the prospective treatment phase, the point at which approximately 50% of patients had already responded, and Week 8. [Note: test-retest was the best approximation of reliability in the current program; inter-rater reliability was not assessed, and internal consistency reliability was not a viable statistical option for a single-item measure.]

Table 2: CGI-S and CGI-I Test-Retest Reliability

	CGI-S (Week 6/Week 8)	CGI-I (Week 6/Week 8)	p value
Correlation, R (95% CI)	0.71 (0.69 to 0.73)	0.57 (0.55 to 0.60)	p < 0.001

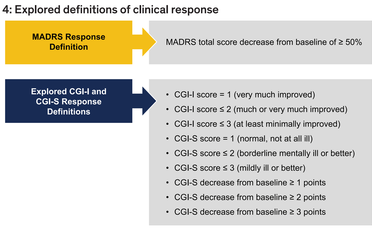
CGI-I = Clinical Global Impressions of Improvement; CGI-S = Clinical Global Impressions of Severity. The p value represents the difference between measures on the reliability coefficient

Test-retest reliability on the CGI-I was significant but fell into the questionable range (CGI-I R = 0.57; p<0.001). Test-retest reliability on the CGI-S was significant and in the acceptable range (CGI-S R = 0.71; p<0.001) and was significantly greater than CGI-I reliability (z = 778, p<0.001).

Discriminant Validity

Figure 4 below outlines CGI-I and CGI-S response definitions explored in the current analysis, all of which were examined using MADRS (≥ 50% decrease from baseline in total score) as the benchmark definition of a "true" response.

Figure 4: Explored definitions of clinical response

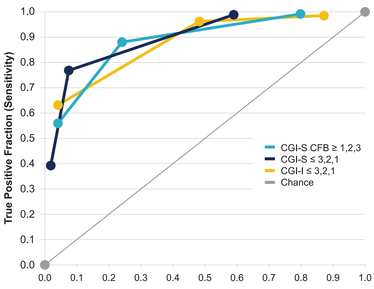


MADRS = Montgomery-Asberg Depression Rating Scale; CGI-I = Clinical Global Impression of Improvement; CGI-S = Clinical Global Impression of Severity

Receiver Operating Characteristics (ROC)

ROC curves were created and area under the curve (AUC) calculated to compare CGI-S vs CGI-I as adequate measures to detect subjects who were MADRS responders. Figure 5 shows the ROC curves representing CGI-S change (improvement) from baseline, CGI-S absolute score, and CGI-I absolute score. All ROC curves are visually comparable, and AUC did not significantly differ between scales (χ^2 = 2.57, p = 0.28).

Figure 5: ROC curves representing the discriminative ability of CGI-S and CGI-I to detect clinical response



CGI-S = Clinical Global Impressions-Severity; CGI-I = Clinical Global Impressions-Improvement ; MADRS = Montgomery-Asberg Depression Rating Scale

Diagnostic accuracy indices

The relative ability of CGI-S and CGI-I, at different cut points, to accurately detect MADRS response in the current dataset was assessed by deriving nine different indices of diagnostic accuracy, where indices were calculated as follows:

- **Sensitivity (TPF; True Positive Fraction)** = Proportion of MADRS responders for whom CGI score also indicated response. Calculation: number of true positives divided by the sum of true positives and false negatives
- **Specificity** = Proportion of MADRS non-responders for whom CGI score did not indicate response. Calculation: number of true negatives divided by the sum of true negatives and false positives
- **False Positive Fraction (FPF)** = 1 – specificity.
- **Positive Predictive Value (PPV)** = Proportion of CGI responders who were true responders as defined by MADRS. Calculation: number of true positives divided by the sum of true positives and false positives.
- **Negative Predictive Value (NPV)** = Proportion of CGI non-responders who were true non-responders as defined by MADRS. Number of true negatives divided by the sum of true negatives and false negatives.
- **Positive Likelihood Ratio (LR+)** = Ratio of the probability that the CGI will show response for those who are MADRS responders to the probability of CGI response in those who are not MADRS responders. Calculation: sensitivity divided by (1-specificity).
- **Negative Likelihood Ratio (LR-)** = Ratio of the probability that the CGI will not show response for those who are MADRS responders to the probability that CGI will show response in those who are not MADRS responders. Calculation: (1-sensitivity) divided by specificity.
- **Youden Index** = Maximum potential of effectiveness in detecting response with CGI. Calculation: (sensitivity + specificity) – 1.
- **Diagnostic Odds Ratio (DOR)** = Global indicator of diagnostic accuracy, taking prevalence into account. Ratio of the probability that the CGI will show response if subject is a MADRS responder relative to the odds that the CGI will show response if subject is not a MADRS responder. Calculation: (true positives/false negatives) ÷ (false positive/true negatives)

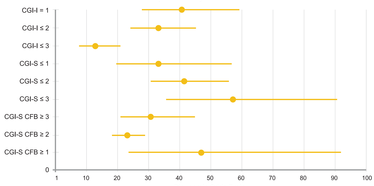
Table 3: Indices of diagnostic accuracy as defined by CGI-S and CGI-I at multiple threshold levels

	Sensitivity/ TPF	Specificity	FPF	PPV	NPV	LR+	LR-	Youden Index	DOR
CGI-I = 1	0.64	0.96	0.04	0.97	0.54	15.36	0.38	0.6	40.57
CGI-I ≤ 2	0.97	0.52	0.48	0.82	0.88	2.01	0.06	0.49	32.95
CGI-I ≤ 3	0.99	0.13	0.87	0.72	0.83	1.13	0.09	0.12	12.68
CGI-S ≤ 1	0.4	0.98	0.02	0.98	0.42	20.43	0.62	0.38	33.14
CGI-S ≤ 2	0.77	0.92	0.08	0.96	0.64	10.16	0.25	0.7	41.33
CGI-S ≤ 3	**0.99**	**0.41**	**0.59**	**0.79**	**0.94**	**1.68**	**0.03**	**0.4**	**56.86**
CGI-S CFB ≥ 1	0.99	0.2	0.80	0.74	0.94	1.25	0.03	0.2	46.56
CGI-S CFB ≥ 2	0.88	0.76	0.24	0.89	0.73	3.64	0.16	0.64	22.98
CGI-S CFB ≥ 3	0.56	0.96	0.04	0.97	0.49	13.98	0.46	0.52	30.57

CGI-I = Clinical Global Impressions of Improvement; CGI-S = Clinical Global Impressions of Severity; CFB = Change from Baseline; TPF = True Positive Fraction; FPF = False Positive Fraction; PPV = Positive Predictive Value; NPV = Negative Predictive Value; LR+ = Positive Likelihood Ratio; LR- = Negative Likelihood Ratio; DOR = Diagnostic Odds Ratio

Based on calculated diagnostic odds ratios (DOR), the CGI-S absolute score of ≤3 was a more accurate measure of response than any other CGI-S definition (χ^2 = 368.68, p <0.001, Diagnostic Odds Ratio (DOR) = 56.86). Furthermore, the CGI-S absolute score of ≤3 was more accurate than the commonly used CGI-I score of 1 or 2 (χ^2 = 219.97, p <0.001, DOR: 32.95). Odds ratios with confidence intervals are presented in Figure 6.

Figure 6: Discrimination of clinical response by CGI-S ≤ 3 and CGI-I ≤ 2

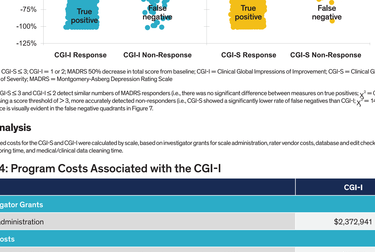


Odds ratios are presented with 95% confidence intervals.

The DOR for CGI-S ≤ 3 is 56.86 ± 0.24, which is greater than any other measure of response. Although the confidence interval is wide, driven by low false negative rates yielding a larger SE, the lower bound of the confidence interval is also greater than any other measure evaluated.

Figure 7 below further illustrates the apparent diagnostic advantage of response as defined by the CGI-S absolute score of ≤ 3 over the CGI-I ≤ 2, which is commonly used to define global response in clinical trials.

Figure 7: Discrimination of MADRS-based clinical response using CGI-S ≤ 3 vs CGI-I ≤ 2



Response = CGI-S ≤ 3; CGI-I = 1 or 2; MADRS 50% decrease in total score from baseline; CGI-I = Clinical Global Impressions of Improvement; CGI-S = Clinical Global Impressions of Severity; MADRS = Montgomery-Asberg Depression Rating Scale

While both CGI-S ≤ 3 and CGI-I ≤ 2 detect similar numbers of MADRS responders (i.e., there was no significant difference between measures on true positives; χ^2 = 0.19, p = 0.66), the CGI-S was using a score threshold of > 3, more accurately detected non-responders (i.e., CGI-S showed a significantly lower rate of false negatives than CGI-I; χ^2 = 14.01, p < 0.01). This difference is visually evident in the false negative quadrants in Figure 7.

Cost Analysis

Total estimated costs for the CGI-S and CGI-I were calculated by scale, based on investigator grants for scale administration, rater vendor costs, database and edit check programming costs, monitoring time, and medical/clinical data cleaning time.

Table 4: Program Costs Associated with the CGI-I

	CGI-I
Investigator Grants	
Scale administration	$2,372,941
CRO Costs	
Database build: CRF design and programming, edit check programming and QC	$3,339
CRA monitoring: On-site eCRF and source data review	$189,343
Query management: Medical/clinical and data manager review	$11,505
Vendor Costs	
Scale Management, rater training, project management	$69,641
Total Cost	**$2,646,769**

The nearly 500 data and medical queries for CGI-I in this program were 21% higher than for CGI-S, with the majority of errors related to the lookback period (rater incorrectly comparing severity to last visit rather than Baseline). These errors translate to $11,500 in medical and data personnel time for data cleaning alone. Importantly, the financial burden of this brief scale also includes much more significant costs associated with scale production, investigator administration, data entry, and database/edit check programming.

By eliminating CGI-I entirely, the overall program budget could have been reduced by approximately $2,646,769. These cost savings translate to approximately $661,692 per study, a savings which in this program could have otherwise been applied to one of the following (for example):

- 2 investigator meetings per study,
- 165 monitoring visits per study, or
- 14 additional subjects in each study, translating to an estimated 1.5% increase in study power, or
- 56 additional subjects in a single study, translating to an estimated 5% increase in study power

Conclusions

The results from the current analysis suggest that the CGI-I scale does not necessarily offer added value beyond what can be measured using the CGI-S alone. The CGI-S's superior validity, reliability, and discriminative accuracy for gauging response (defined as CGI-S ≤3), taken together with the cost savings conveyed by omission of the CGI-I, provide strong evidence in favor of choosing CGI-S as a single global severity measure. The CGI-S as a sole measure of global severity, response and remission, is a viable, valid, and cost effective option in trial design for MDD, with potential for broader applicability in other therapeutic areas.

References

Guy W. Clinical global impressions. In ECDEU Assessment Manual for Psychopharmacology Rockville, MD, USA: US Department of Health, Education, and Welfare; 1976: 217-222.

Hajian-Tilaki KO et al. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med.* 2013; 4(2): 627-635.

Šimundić A.M. Measures of diagnostic accuracy: basic definitions. *EJIFCC.* 2009; 19(4): 203–211.

Thase M.E., Youakim J.M., Skuban A et al. Efficacy and safety of adjunctive brexpiprazole 2mg a Phase 3, randomized, double-blind, placebo-controlled study in patients with inadequate response to antidepressants. *J Clin Psychiatry.* 2015; 76(9): 1224-31.

Thase M.E., Youakim J.M., Skuban A et al. Adjunctive brexpiprazole 1 and 3 mg for patients with major depressive disorder following inadequate response to antidepressants: a phase 3, randomized, double-blind study. *J Clin Psychiatry.* 2015 Sep; 76(9): 1232-40.

Disclosures

Funding for the POLARIS, PYXIS DEPHINUS and SIRIUS trials was provided by Otsuka Pharmaceutical Development and Commercialization (OPDC) and H. Lundbeck A/S. RB and CB are employees of OPDC. INC Research/inVentiv Health was the Contract Research Organization responsible for execution of all four trials. KN, ZS, LS and CR are employees of INC Research/inVentiv Health.

Presented at International Society for CNS Clinical Trials Methodology, Paris, France, August 31-September 2, 2017