Targeted scoring criteria reduce variance in global impressions

Steven D. Targum1,2*, Joan Busner3 and Allan H. Young4

1Massachusetts General Hospital, Boston, Massachusetts, USA
2BrainCells Inc, San Diego, California, USA
3United Biosource Corporation, Wayne, Pennsylvania, USA
4University of British Columbia, Vancouver, Canada

Objective This study examined the confounding effect of treatment emergent physical or psychic symptoms on clinical global impression (CGI) ratings in CNS trials and examined the benefit of targeted scoring criteria on clarifying ratings and reducing scoring variance.

Methods Twenty-four raters participating in an investigator meeting training session scored a series of scripted CGI scenarios that included treatment emergent symptoms.

Results The addition of treatment emergent gastrointestinal (GI) symptoms or anxiety symptoms significantly changed the rating of clinical global improvement and caused a broad CGI-improvement (CGI-I) scoring variance reflecting scoring ambiguity amongst these raters. Re-rating after a presentation of well-defined criteria that addressed these scoring issues narrowed the variance and significantly improved inter-rater reliability.

Conclusions It is clear that CNS trials must define scoring criteria for global ratings prior to the initiation of a study to assure ratings consistency. The actual definition of global must be study-specific and may depend upon the targeted symptoms of interest and mechanism of drug action. The targeted criteria that define global must be included in all published reports about the trial. Copyright © 2008 John Wiley & Sons, Ltd.

Key words—scoring criteria; CGI; CNS trials; depressive disorder

INTRODUCTION

The assessment of global improvement in CNS trials includes an evaluation of symptomatic, behavioral, and functional changes during treatment. The clinical global impressions scale for improvement (CGI-I) is often used in clinical trials as a measure of overall clinical change although its clinical validity is not well established (Guy, 1976; Haro et al., 2003; Kadouri et al., 2007; Spielmans and McFall, 2006). As a sequential measure of clinical improvement or worsening from baseline, the properties of the CGI-I scale presume that the clinical changes will be internally consistent in order to make a clinically valid comparison (baseline to specific interval). In reality, patients often experience treatment-emergent physical or psychic symptoms that were not present at baseline and may obfuscate the assessment of change. The etiology of these treatment-emergent symptoms may be related to the specific disease, or the drug treatment, or may be co-morbid symptoms that are entirely unrelated to either the disease or the tested drugs. Scoring guidelines often suggest that raters simply score what they observe and not attribute these symptoms to any specific etiology unless it is absolutely clear (e.g., flu symptoms in the presence of obvious flu) (Beneke and Rasmus, 1992;
Busner and Targum, 2007; Kadouri et al., 2007). Despite these general guidelines, extraneous treatment-emergent symptoms often obscure the CGI-I rating, particularly when the symptoms affect behavior or function. We have previously demonstrated that treatment-emergent physical or psychic symptoms can cause ratings ambiguity and significant scoring variability on the CGI-I scale (Targum et al., 2008). Some raters score the CGI-I as if it were a therapeutic index. In fact, the full form of the CGI instrument includes a therapeutic index as a distinct component from the CGI-I (Guy, 1976). We contend that well-defined criteria that addresses scoring of treatment emergent symptoms and specifically target symptoms of interest can significantly narrow the variance between raters and enhance the utility of the CGI-I in CNS trials. In the present study, we sought to replicate our earlier findings and to ascertain the impact of CGI training with targeted operational scoring criteria on inter-rater reliability.

METHODS

The study was conducted with 24 CGI raters attending an investigator meeting prior to starting a clinical trial for a novel antidepressant medication. The raters were well trained and very experienced: 16 raters (67%) were MD’s or PhD’s and 17 raters (71%) had more than 5 years ratings experience with the CGI. Only one rater had less than 1 year of ratings experience. As part of the rating instrument review and training session, the CGI raters were given a basic introduction to CGI assessments with well-established guidelines for accurate determination of the 1–7 range of global severity and improvement scores. CGI-severity (CGI-S) scores increase from 1 to 7 with increasing levels of illness severity. Alternatively, CGI-I scores measure change from baseline (randomization); a score of 4 reflects no change from baseline, scores of 3, 2, and 1 reflect progressive improvement and scores of 5, 6, and 7 progressive worsening.

Using audience response scoring (ARS), the raters read and scored a scripted baseline clinical scenario for CGI-S for an acutely depressed woman followed by three alternative clinical outcome scenarios for CGI-I. ARS scoring of these examples preceded audience discussion about the exercise.

Raters read the following text to make the CGI-S baseline visit assessment:

A 34-year-old woman presents with melancholic depression confirmed by MINI and DSM-IV classification. She has struggled to work as a clinical trial manager while planning for her upcoming wedding, in which she is no longer interested. She complains of sleep disturbance (early morning awakening), loss of appetite, anhedonia, guilt, and despondency particularly in the morning.

The first follow-up scenario presented the patient as achieving clinical remission at the week 4 visit. Raters read the following text for scenario 1:

Clinically, the patient has improved. She currently has no clinically meaningful symptoms of depression. She has regained her zest and ability to experience pleasure, is sleeping, and eating without difficulty. She is happy about a wedding shower scheduled for the weekend and is looking forward to the wedding.

Alternative CGI-I scenarios 2 and 3 added treatment-emergent gastrointestinal (GI) symptoms or psychic (anxiety) symptoms to the straightforward scenario 1.

- Scenario 2 added treatment-emergent GI symptoms to scenario 1:

  She had intermittent abdominal cramps, nausea, and diarrhea on 3 days in the past week. She plans to attend the wedding shower anyway. She has no previous history of GI problems.

- Scenario 3 added treatment-emergent anxiety symptoms to scenario 1:

  She has experienced waves of anxiety that occurred on 3 days in the past week that disrupted her ability to work. She worries excessively about her ability to achieve a successful marriage and balance in her life. Anxiety symptoms were not present at baseline. The patient has no history of anxiety disorder and tends not to worry.

After the first round of ARS scoring, operational scoring criteria for the CGI-I was presented, discussed, and followed by re-scoring of the treatment-emergent symptom scenarios 2 and 3. The definition of global in this study was targeted to the symptoms of
the identified patient population which was anxious depression. Hence, the operational definition of global included anxious symptoms in the CGI baseline severity (CGI-S) assessment. Raters were told: (a) rate all identified symptoms without attribution and judge the impact of those symptoms on behavior and function, and (b) treatment-emergent symptoms can influence CGI ratings provided they impact behavior or function. It was emphasized that CGI scoring must consider the impact of any emergent or existing symptom within the context of all of the targeted symptoms and the overall clinical changes since baseline.

Statistical evaluation of the data included calculation of means (±SD), medians, and student’s t tests analyzing CGI-I change scores for each scenario before and after audience discussion. In addition, the average absolute deviation from the median was examined as a relevant measure of ratings dispersion which is highly correlated with inter-rater agreement (Burke et al., 1999).

RESULTS

Ninety-one per cent of raters scored the CGI-S baseline scenario as a 4 (moderately ill) or 5 (markedly ill) reflecting high inter-rater agreement. The mean CGI-S was 4.52 ± 0.67 (SD).

The first follow-up scenario (1) indicated that the patient had no clinical symptoms of depression within the past week. CGI-I scores improved significantly with a mean score of 1.17 ± 0.38 (SD) reflecting very much improvement from baseline (t = 36.5; df = 46; p < 0.0001). Twenty raters (83%) scored this scenario as 1 (very much improvement) and four raters (17%) as 2 (much improved). The absolute deviation from the median score of 1 was only 0.11 reflecting high inter-rater consistency (minimal dispersion) for this straightforward example.

The second CGI-I scenario (2) presented the same patient with no clinical symptoms of depression but introduced treatment-emergent GI symptoms (intermittent abdominal cramps, nausea, and diarrhea) for 3 days in the previous week. The mean CGI-I was 2.09 ± 0.92 which was still significantly improved from baseline (t = 10.2; df = 44; p < 0.0001). CGI-I scores ranged from 1 to 5. The absolute deviation from the median score of 2 was 0.55. The CGI-I for scenario 2 was significantly less improved than scenario 1 which had no treatment-emergent physical symptoms (t = 4.52; df = 44; p < 0.0001).

CGI-I scenario 3 presented the same patient with no clinical symptoms of depression but introduced treatment-emergent anxiety symptoms (excessive worry, waves of anxiety) occurring 3 days in the past week that disrupted her ability to work. This scenario indicated that there were no pre-existing co-morbid anxiety symptoms present at baseline. CGI-I scores ranged from 2 to 6 with a mean CGI-I score of 4.21 ± 1.28. The CGI-I score was not significantly different from baseline (t = 0.8; df = 46; p = 0.43). The average absolute deviation from the median score of 4 (no change) was 1.12 reflecting broad scoring variance. In this scenario, 10 raters (42%) scored the patient as improved (CGI-I = 1, 2, or 3) whereas 11 raters (46%) scored the patient as worse (CGI-I scores = 5, 6, or 7) reflecting the scoring ambiguity created by the symptom of treatment-emergent anxiety.

The audience was presented with scoring criteria for the CGI-I that addressed the scoring ambiguity. Emphasis was placed on assessing the impact of treatment emergent symptoms, as well as all targeted symptoms, on the overall clinical presentation, and on changes in behavior and function since baseline.

Re-rating with clearly defined operational criteria narrowed the scoring variance on scenarios 2 and 3. Re-rated mean CGI-I for scenario 2 (GI symptoms) was 2.08 ± 0.51 with a score range from 1 to 3 and an average absolute deviation from the median of 0.25. Hence, the standard deviation narrowed from 0.92 to 0.51 after discussion of targeted criteria.

Re-rated mean CGI-I for scenario 3 (anxiety symptoms) was 2.25 ± 0.44 with a score range of 2–3, a median score of 2, and an average absolute deviation of 0.25. The standard deviation narrowed from 1.28 to 0.44 subsequent to the additional criteria training. The re-rated scenario 3 was significantly different from baseline (t = 19.3; df = 42; p < 0.0001). Further, a comparison of the mean CGI-I scores for scenario 3 between the pre- and post-criteria discussion was also significantly different (t = 6.49; df = 42; p < 0.0001). Figure 1 depicts the narrowing of the scoring range achieved on the CGI-I scores for the treatment-emergent anxiety scenario achieved after the training discussion.

DISCUSSION

This study examined the impact of treatment-emergent physical (GI) or psychic (anxiety) symptoms on CGI-I scoring using brief scripted examples during an investigator meeting. Both physical and psychic symptoms significantly diminished (worsened) the CGI-I scoring of global improvement relative to the straightforward scenario 1 which described no
remaining symptoms of depression with no treatment emergent symptoms.

These results replicate our earlier findings with a different rater population and confirm that treatment emergent physical or psychic symptoms can confound CGI scoring of depressed patients and generate very broad scoring variance (Targum et al., 2008).

Treatment emergent symptoms frequently occur in CNS trials. Most raters will appropriately adjust the score of the CGI-I in the presence of treatment emergent symptoms within the context of their impact on behavior or function. We anticipated an adjusted CGI-I score of 2 or 3 for the examples that added treatment emergent symptoms. However, in this study, 3 days of treatment emergent GI symptoms or anxiety generated a much broader, dispersed range of CGI-I scores. The scoring ambiguity was corrected by providing clear operational criteria. The re-rated CGI-I scores re-established the narrow variance observed in scenario 1 for both the treatment-emergent GI and anxiety symptom scenarios.

The re-rated median CGI-I score for scenario 3 (treatment-emergent anxiety) was 2 (much improved) in contrast to 4 (no change) as observed prior to the Discussion. The average absolute deviation from the median score improved from 1.12 to 0.25 after re-rating reflecting high inter-rater agreement. We believe that a CGI-I of 2 or 3 is reasonable given that the anxiety symptoms slightly disrupted function in the past week, whereas a score of 4 overemphasizes the impact of 3 days of anxiety within the overall context of clinical change achieved by this patient.

Anxiety symptoms are often part of the clinical presentation of MDD. In the NIMH-sponsored STAR*D study, 53% of depressed patients met criteria for anxious depression and responded significantly less well to antidepressant intervention than non-anxious depressed patients (Fava et al., 2004; Fava et al., 2008). The DSM-IV-R criteria for MDD do not include anxiety symptoms as part of the symptom cluster, although these criteria have often been challenged (APA, 2000; Fava et al., 2008; Kessler et al., 1996; Maj, 2005; Targum et al., 2008). Therefore, scoring ambiguity on the CGI-I in the presence of anxious depression may be, in part, a residual of diagnostic uncertainty. Regardless, the current study demonstrates the benefit of introducing well-defined operational criteria for CGI scoring to improve inter-rater reliability by reducing rater scoring variance. These CGI-I scoring criteria must identify the specific symptoms of interest, address the scoring of treatment emergent symptoms, and clearly define the term “global” to be used in the assessments.

It is noteworthy that the CGI as originally conceived included three components related to severity, improvement, and a therapeutic efficacy index that evaluated efficacy relative to drug-related adverse events (Guy, 1976). This component did not include
non-drug-related (extraneous) adverse events. Although rarely used, the therapeutic efficacy measure reflects the awareness of the scale developers of the potential confounding influence of adverse events on clinical assessments.

The meaning of “global” needs to be defined before a CNS trial is initiated. It is clear that the CGI-I is not a therapeutic index weighing the benefit of drug (efficacy) over tolerability and safety. We believe that the definition of “global” must be targeted to the specific objectives of the CNS trial and the anticipated mechanisms of action and clinical effects of the drug candidate. Clearly defined criteria will improve both the reliability and validity of the CGI for the specific disease/symptom construct under study. A study using strictly defined MDD according to DSM-IV-R criteria would apply restrictive operational criteria for CGI-I scoring that specifically excludes the presence of anxious symptoms at baseline. In this instance, treatment emergent anxiety would be recorded as an adverse event but might still influence the CGI-I scoring if it impacted behavior or function. Alternatively, a study of anxious-depression would use a less restrictive operational criterion in which global is more broadly defined to include anxious symptoms at baseline. As mechanisms of action become more specific, it is likely that new drug candidates may target some but not all of the symptoms of a particular disease or syndrome and deviate from the strict DSM-IV-R criteria (Targum et al., 2008). For instance, some novel drugs may target only the anxiety or sleep symptoms of depressed patients and the CGI-I scoring criteria would accordingly target these specific symptoms to be meaningful for that CNS trial. However, while departing from more conventional categorization systems, the CGI-I must sustain both ecological and face validity for the nosological entity it is being used to study (Targum et al., 2008).

Beyond the definition of “global,” CGI raters need to use standardized scoring conventions. Generally, CGI raters are instructed to rate the symptoms they observe (identify) and to judge the impact of these symptoms on behavior and function regardless of attribution. Both exacerbation of pre-existing symptoms and newly identified treatment emergent symptoms can impact behavior or function. Thus, an impactful symptom is included in the CGI scoring unless it is unequivocally due to an identifiable unrelated event (flu symptoms in a case of obvious flu). Scoring must consider the impact within the context of the overall clinical presentation and other clinical changes since baseline.

The pre-emptive step of providing well-defined targeted scoring criteria for the CGI prior to the initiation of a CNS trial will reduce variance and facilitate more accurate and clinically relevant trial outcome data. It is equally important that all published reports about the trial provide the criteria used to define “global” in the study. Hopefully, this paper will generate more effort to standardize the criteria for scoring this potentially useful instrument.

ACKNOWLEDGEMENTS

This study was funded in part by BrainCells Inc. (San Diego, California).

REFERENCES


