ELSEVIER

Psychiatry Research 73 (1997) 159–171

PSYCHIATRY RESEARCH

Modification of the Clinical Global Impressions (CGI) scale for use in bipolar illness (BP): the CGI-BP

Melissa K. Spearing[a], Robert M. Post[a,*], Gabriele S. Leverich[a], Diane Brandt[b], Willem Nolen[c]

[a] *NIMH / BPB, NIH, Building 10, Room 3N212, Bethesda, MD 20892-1272, USA*
[b] *Stanley Foundation Bipolar Network, 5430 Grosvenor Lane, Suite 200, Bethesda, MD 20814, USA*
[c] *HC Rumke Group, Willem Arntz Huis, Vrouwjnuttenhof 18, P.O. Box 61, 3500AB, Utrecht, The Netherlands*

Received 2 December 1996; revised 15 September 1997; accepted 26 September 1997

Abstract

The Clinical Global Impressions Scale (CGI) was modified specifically for use in assessing global illness severity and change in patients with bipolar disorder. Criticisms of the original CGI were addressed by correcting inconsistencies in scaling, identifying time frames for comparison, clarifying definitions of illness severity and change, and separating out assessment of treatment side effects from illness improvement during treatment. A Detailed User's Guide was developed to train clinicians in the use of the new CGI-Bipolar Version (CGI-BP) for rating severity of manic and depressive episodes and the degree of change from the immediately preceding phase and from the worst phase of illness. The revised scale and manual provide a focused set of instructions to facilitate the reliability of these ratings of mania, depression, and overall bipolar illness during treatment of an acute episode or in longer-term illness prophylaxis. Interrater reliability of the scale was demonstrated in preliminary analyses. Thus, the modified CGI-BP is anticipated to be more useful than the original CGI in studies of bipolar disorder. © 1997 Elsevier Science Ireland Ltd.

Keywords: Methodology; Global ratings; Mental disorder; Manic-depressive illness

* Corresponding author. Tel.: +1 301 4964805; fax: +1 301 4020052.

0165-1781/97/$17.00
PII S0165-1781(97)00123-6

1. Introduction

Clinicians have access to a wide variety of rating scales designed to help identify, assess, and quantify symptoms of diverse illnesses. Although many of these scales measure the presence and severity of discrete illness variables, rating scales of global illness conditions can contribute valuable information to the overall clinical picture of each patient and can be helpful in assessing the efficacy of a given intervention. Hence, global judgments of illness severity and change during the course of treatment may usefully inform clinicians of whether to continue or alter treatment strategies (i.e. decisions regarding whether or not a treatment is effective) (Rush and Kupfer, 1995).

The utility of physicians' global ratings in differentiating drug response from non-response was demonstrated by Lehmann (1984) in his review of 17 placebo-controlled studies of psychotropic drugs. He found that global ratings of therapeutic effect distinguished drug from placebo in 14 of 17 studies, suggesting that global ratings are valid assessments and can be highly instrumental in determining the efficacy of a drug. Lehmann (1984) also asserted that global ratings can provide more comprehensive information about patients' clinical status (or change during treatment) than scores on itemized symptom rating scales.

Global ratings have been used in studies of bipolar disorder to gauge acute phase treatment response or overall prophylactic benefit. For example, Okuma (1993) assessed the prophylactic effects of lithium or carbamazepine on patients with bipolar illness, classifying their treatment response as 'markedly effective', 'moderately effective', or 'unchanged or aggravated'. These categories were defined by the degree of symptom remission and episode duration during treatment as compared to the illness prior to treatment. Similarly, Calabrese et al. (1993) classified acute or prophylactic treatment response into categories of 'marked', 'moderate', or 'no response' based on global ratings of change in episode severity, duration, and frequency. In addition, many literature reviews employ these global categories of marked or moderate improvement when grouping responders versus non-responders to a particular treatment for bipolar illness (e.g. Post et al., 1990, 1996a,b). Thus, global ratings are a commonly used method with which to assess clinically meaningful degrees of treatment response, hence guiding the clinical management of bipolar disorder.

The Clinical Global Impressions Scale (CGI) (Guy, 1976) is a standard measure for making global assessments of illness. The complete CGI scale consists of three different global measures designed to rate the effectiveness of a particular treatment: I, Severity of Illness — assessment of the patient's current severity of symptoms; II, Global Improvement — comparison of the patient's baseline condition to his/her current condition; and III, Efficacy Index — comparison of the patient's baseline condition to a ratio of current therapeutic benefit and severity of side effects (Guy, 1976). The CGI has been widely used as a primary outcome measure in clinical trials of medications for depression (e.g. Salzmann and Robin, 1995), schizophrenia (e.g. Honer et al., 1995), and many other illness conditions (e.g. Rabkin et al., 1995; Ravizza et al., 1995). However, most studies using the CGI scale have utilized only one or two of the global measures to assess drug response (e.g. Remick et al., 1994 (used both Items I and II, Severity of Illness and Global Improvement); DeWilde et al., 1993 (used Item I, Severity of Illness); Kanowski et al., 1990 (used Item III, Efficacy Index); Fieve et al., 1986 (used Item II, Global Improvement)).

The CGI Global Improvement measure, a commonly used item, offers a range of rating choices between 'Very Much Improved' and 'Very Much Worse' based on the degree of change during treatment. A general convention among studies utilizing this measure has been to classify treatment responders as patients rated 'Very Much Improved' or 'Much Improved' and non-responders as those rated either 'Minimally Improved', 'No Change', or 'Worse' (e.g. Quitkin et al., 1984; Ontiveros et al., 1991; Colle et al., 1994). In studies of antidepressant medications, this convention has been supported by validating Global Improvement ratings against percent changes in scores on the Hamilton Rating Scale

for Depression (HRSD) (Sato et al., 1984; Fieve et al., 1986). Both Sato et al. (1984) and Fieve et al. (1986) found CGI Global Improvement ratings to correlate significantly with changes in HRSD scores, indicating that Global Improvement validly measures overall change in depressive symptoms during treatment.

Despite its frequent use in efficacy studies, the CGI has been criticized in the literature for being inconsistent, unreliable, and too general to provide meaningful information about patient clinical status or treatment response (Beneke and Rasmus, 1992; Dahlke et al., 1992). Inconsistency and asymmetry in scaling, lack of standard definitions of illness severity and change, lack of clearly identified time frames for evaluating change, redundancy and different scaling in the two change measures, and an inability to differentiate improvement ascribable only to the treatment effect have been described as primary flaws in the CGI leading to its questionable validity and reliability (Beneke and Rasmus, 1992; Dahlke et al., 1992). While the CGI continues to be used as an integrated measure of clinical improvement and treatment efficacy, there is a clear need for a standardized global rating scale that is more consistently applied and interpreted to improve validity, reliability, and utility of the ratings.

There was a consensus among experienced clinicians at five sites of the Stanley Foundation Bipolar Network that while the CGI had the many liabilities noted above, no other global instrument was available as a better alternative. Bipolar disorder is often difficult to assess due to its pleomorphic and variable components (i.e. severity, duration, and frequency of both manic and depressive phases). To address the criticisms of the CGI as well as to better adapt it to patients with bipolar illness, we attempted to revise the scale by correcting its scaling and definitional problems and allowing for individual assessments of mania, depression, and overall bipolar illness. In addition, we sought to make the revised CGI scale amenable to evaluations of both acute treatment efficacy (i.e. for treatment of a single affective episode) and prophylaxis of episodes (i.e. for treatment and prevention of episode recurrence or cycling). The new CGI-Bipolar Version (CGI-BP), therefore, is expected to improve the accuracy and reliability of global assessments of illness severity and degree of improvement occurring in mania, depression, and overall illness on a given treatment.

The original CGI scale was apparently not tested for reliability at the outset, and the scale was used in studies for many years before its reliability was formally assessed. We believe the first study to demonstrate and document inter-rater and test-retest reliability of the CGI was conducted by Dahlke et al. (1992). This study revealed relatively good reliability scores for CGI severity ratings but not for the change (i.e. Global Improvement) ratings. A later study by Weitkunat et al. (1993) also found the CGI change measure to have unimpressive test-retest reliability. Both studies clearly indicated a need for improved CGI guidelines and assessment criteria in order to strengthen the reliability of the ratings. Since the scaling flaws and conceptual problems of the original CGI were corrected in the CGI-BP, and since precise definitions and guidelines for rating bipolar illness were provided in the CGI-BP User's Guides, we anticipated that the reliability and clinical utility of the new scale would be very much improved.

2. Methods

Each stated criticism (Table 1) of the CGI (Fig. 1) was specifically addressed in the development of the CGI-BP scale (Fig. 2). A group of clinicians (consisting of one psychiatrist, two clinical social workers, and three research clinicians experienced with bipolar illness) reviewed, piloted, and reached consensus on the re-scaling, conceptual modifications in the change measures (Items II and III), designation of the relevant time domains for assessment, and other definitions and instructions for the scale. The new CGI-BP form was pilot tested by clinicians on numerous research patients with refractory bipolar disorder and revised several times for clarity. Each measure of severity or change (Items I–III) was expanded to accommodate separate ratings of the manic and depressive components of the disorder, as well as the overall bipolar illness.

CLINICAL GLOBAL IMPRESSIONS*

Instructions: Complete Item 1 - *Severity of Illness* at the initial and subsequent assessments.
Items 2 and 3 may be omitted at the initial assessment by marking 0 - "Not Assessed."

CLINICAL GLOBAL IMPRESSIONS

1. SEVERITY OF ILLNESS
Considering your total clinical experience with this particular population, how mentally ill is the patient at this time?

0 = Not Assessed
1 = Normal, not at all ill
2 = Borderline mentally ill
3 = Mildly ill
4 = Moderately ill
5 = Markedly ill
6 = Severely ill
7 = Among the most extremely ill patients

THE NEXT TWO ITEMS MAY BE OMITTED AT THE INITIAL ASSESSMENT BY MARKING "NOT ASSESSED" FOR BOTH ITEMS

2. GLOBAL IMPROVEMENT - Rate total improvement whether or not, in your judgment, it is due entirely to drug treatment.

Compared to his condition at admission to the project, how much has he changed?

0 = Not assessed
1 = Very much improved
2 = Much improved
3 = Minimally improved
4 = No change
5 = Minimally worse
6 = Much worse
7 = Very much worse

3. EFFICACY INDEX - Rate this item on the basis of DRUG EFFECT ONLY.

Select the terms which best describe the degrees of therapeutic effect and side effects and record the number in the box where the two items intersect.

THERAPEUTIC EFFECT	SIDE EFFECTS			
	None	Do not significantly interfere with functioning	Significantly interferes with patient's functioning	Outweighs therapeutic effect
MARKED - Vast improvement. Complete or nearly complete remission of all symptoms.	01	02	03	04
MODERATE - Decided improvement. Partial remission of symptoms.	05	06	07	08
MINIMAL - Slight improvement which doesn't alter status of care of patient.	09	10	11	12
UNCHANGED OR WORSE	13	14	15	16
Not Assessed = 00				

* Reproduced from Guy, 1976.

Fig. 1. The Clinical Global Impressions Scale (Guy, 1976).

A Detailed User's Guide (available upon request) was developed to provide a self-contained instruction manual to teach clinicians how to use the revised CGI-BP. The detailed guide includes definitions of illness severity and change and graphical descriptions and examples of how to rate both the acute and prophylactic effects of treatments (Fig. 3 and Fig. 4). A set of practice ratings is provided for clinicians to complete based on sample patient life chart information (Leverich and Post, 1995, 1996). An answer key was provided so that new raters could assess their performance against a consensus standard of experienced clinicians (RMP, GSL). A Brief User's Guide also was created based on a condensation of the information provided in the more detailed version. A condensed list of rating principles for assessing the prophylactic effectiveness of a treatment was included in both versions of the guide. These instructions were intended to assist raters in the selection of the appropriate time domains when considering the ability of a treatment to prevent further recurrence of affective episodes.

A test packet, composed of additional sample patient life charts, was developed (without answers) for clinicians to make CGI-BP ratings of hypothetical treatment response for 10 individual episodes of acute mania or depression and another 10 patients whose illness did or did not respond to prophylactic treatment. This test packet would allow field testing of the CGI-BP and the principles taught in both the Brief and Detailed Users' Guides.

3. Results

3.1. Development of the CGI-BP

A number of important scaling and conceptual changes were implemented in converting the original CGI scale to the new CGI-BP. Firstly, in order for the CGI-BP to address bipolar illness in particular, each item of the original CGI was changed to include separate ratings of manic, depressive, and overall components of the illness. In addition, to allow individual assessments of acute treatment efficacy on a single affective episode as well as prophylactic treatment for recurring episodes (i.e. prevention of cycling), different time domains for reference were established for each type of assessment. These refinements permit the CGI-BP to reflect global illness severity and change in a disease with much inherent variability.

Next, the definitions and corresponding time domains of the two CGI change measures were modified to allow clinicians to make global assessments of (II) improvement during treatment compared to the phase immediately preceding treatment and (III) improvement compared to the prior worst phase of the illness. These changes were implemented so that the efficacy of augmentation or combination regimens could be assessed more readily using the CGI-BP. Moreover, the confound of improvement and side effects, and whether improvement was attributable only to the drug in Item III, was removed. Finally, all three measures were modified to address the scaling criticisms of the original CGI scale. These changes (Table 1) are discussed in more detail below.

Scaling inconsistencies of the original CGI Item I, Severity of Illness, as discussed by Beneke and Rasmus (1992), were corrected in the CGI-BP.

Table 1
Improvements to the Clinical Global Impressions scale

Old CGI Faults	New CGI-BP Improvements
I. Severity of Illness	I. Severity of Illness
• One rating	• Mania
• Last interval (quality different)	• Depression
	• Overall illness
	• Last interval consistent
II. Global Improvement	II. Change From Preceding Phase
• Time domains uncertain	• Time domains clarified
• No worsening available	• Worsening available
• II and III use different scales	• II and III use same scale
• Inaccurate name and instructions	• Meaningful name and instructions
III. Efficacy Index	III. Change From Worst Phase
• Drug effect only	• Time domains clarified
• Side effects combined	• Side effects separated
• Choices not symmetrical	• Choices symmetrical

The wording of category 7 was changed from 'among the most extremely ill patients' to 'Very Severely Ill', so that the scaling would be consistent between categories (e.g. 'Mildly Ill', 'Moderately Ill', 'Severely Ill', and 'Very Severely Ill'). In addition, the original CGI measure includes a category 0, 'not assessed', which would be marked if no information about the patient's current illness condition was available. However, none of the CGI assessments could be made in this circumstance (Beneke and Rasmus, 1992), and therefore, this Severity of Illness category was not included in the CGI-BP.

The scaling of the CGI Item II, Global Improvement, has been considered symmetrical and consistent: positive and negative ratings are distributed equally around a neutral point ('no change') and are verbally constructed in the same manner (Beneke and Rasmus, 1992). Therefore, we left Item II unchanged in the CGI-BP. However, CGI Item III, Efficacy Index, has been criticized for its poor scaling construction, side effects confound, and difficulty in assessing improvement attributable only to the new treatment (see below) (Beneke and Rasmus, 1992). Its categories of improvement, labeled 'Marked', 'Moderate', 'Minimal', and 'Unchanged or Worse', are clearly skewed to the side of improvement (Beneke and Rasmus, 1992). No discrimination is made between unchanged and worse, and therefore only truncated information can be obtained from a rating of this category. To correct this scaling problem, the same categories for evaluating CGI (and CGI-BP) Item II (Global Improvement) were used for rating CGI-BP Item III (Change from Worst Phase of Illness). Hence, differing degrees of improvement or worsening could be assessed along the same consistent, symmetrical scale (Fig. 2).

The underlying concept of the original CGI Item II (Global Improvement, i.e. total improvement since starting a treatment) was left essentially intact in the new CGI-BP. However, due to the number of variable factors which influence bipolar illness presentation (i.e. severity, duration, and frequency of episodes), it was necessary to identify appropriate time frames and domains to be rated for both acute (treatment of a single episode) and prophylactic (treatment and prevention of episode cycling or recurrence) assessments. An emphasis was placed on allowing assessment of the impact of an augmentation strategy as compared to what had been achieved on the antidepressant regimen alone immediately prior to starting the augmentation trial. Global Improvement was thus renamed in the CGI-BP 'Change from Preceding Phase' (where phase refers to treatment phase) in accordance with its new emphasis on these specific time frames. For acute assessments of the treatment of single affective episodes, the patient's condition during the week of the assessment is compared to that during the week immediately preceding the start of the current treatment (Fig. 3). Similarly, for prophylactic ratings on the longer-term prevention of episode recurrence or cycling, the patient's pattern of illness (severity, duration, and frequency of manic and depressive episodes) during the entire current treatment phase (with emphasis on the most recent symptoms or episodes) is compared to the treatment phase (or medication-free period of significant duration) immediately preceding the current intervention (Fig. 4). Thus, the revised Item II of the CGI-BP permits assessment of either the acute efficacy or the prophylactic usefulness of the current treatment in relation to the medication regimen immediately prior.

The original CGI Item III, Efficacy Index, has been criticized not only for scaling inconsistencies, but also for its conceptual basis (Beneke and Rasmus, 1992). The measure was originally conceptualized to reflect the interaction between the therapeutic effectiveness and the side effects of a treatment (Guy, 1976). Even more problematic was that the measure also was intended to gauge improvement specifically due to the treatment and not to other potential factors (Guy, 1976). However, Beneke and Rasmus (1992) argued that there is no reason to differentiate treatment-related improvement in CGI Item III from total (global) improvement in CGI Item II, and that Item III only produces redundant information. In addition, Beneke and Rasmus (1992) criticized the scaling used to assess side effects within

Patient ID ___ ___ ___ - ___ ___ ___ ___ Patient Initials ___ ___ Clinician Initials ___ ___ Date ___ ___ / ___ ___ / ___ ___

Protocol Code _____ Level _____ Visit Number _____ Blinded Rating? ☐ NO ☐ YES ☐ N/A

CLINICAL GLOBAL IMPRESSIONS - BP Version (CGI-BP)

A. Which medication is being evaluated? *(if known):*
_____

C. What was the patient's mood state at trial entry? *(Choose one)*
1. ☐ Depressed 3. ☐ Dysphoric Mania 5. ☐ Cycling
2. ☐ Euthymic 4. ☐ Euphoric Mania

B. Assessment Type:
1. ☐ Admission 2. ☐ Acute 3. ☐ Prophylactic

D. How long had the patient been in this mood state?
___ ___ . ___ ☐Days ☐Weeks ☐Months ☐Years

I. SEVERITY of Illness

Considering your total clinical experience with bipolar patients, how severely ill has the patient been during the assessment period? **ACUTE ASSESSMENTS** reflect the **past week. PROPHYLACTIC ASSESSMENTS** reflect the **time on current prophylactic medication(s) with emphasis on the most recent episodes.**

	Normal, not ill	Minimally ill	Mildly ill	Moderately ill	Markedly ill	Severely ill	Very severely ill
	1	2	3	4	5	6	7
a. Mania	☐	☐	☐	☐	☐	☐	☐
b. Depression	☐	☐	☐	☐	☐	☐	☐
c. Overall Bipolar Illness	☐	☐	☐	☐	☐	☐	☐

II. Change from PRECEDING Phase

Compared to the phase immediately preceding this trial, how much has the patient changed? For **ACUTE ASSESSMENTS**, compare the **current week** to the **week preceding** initiation of treatment for the present episode. For **PROPHYLACTIC ASSESSMENTS**, compare the manic and depressive **episodes on current prophylactic medication(s)** to **episodes immediately preceding this treatment**.

	Very much improved	Much improved	Minimally improved	No change	Minimally worse	Much worse	Very much worse	Not applicable
	1	2	3	4	5	6	7	8
a. Mania	☐	☐	☐	☐	☐	☐	☐	☐
b. Depression	☐	☐	☐	☐	☐	☐	☐	☐
c. Overall Bipolar Illness	☐	☐	☐	☐	☐	☐	☐	☐

III. Change from WORST Phase of Illness

Compared to the patient's worst phase of illness prior to the current medication trial or during the early titration phase, how much has the patient changed? For **ACUTE ASSESSMENTS**, compare the **current week** to the **worst week** within this episode. For **PROPHYLACTIC ASSESSMENTS**, compare the manic and depressive **episodes on current prophylactic medication(s)** to **episodes during the patient's worst phase of illness** when untreated or on ineffective medication(s).

	Very much improved	Much improved	Minimally improved	No change	Minimally worse	Much worse	Very much worse	Not applicable
	1	2	3	4	5	6	7	8
a. Mania	☐	☐	☐	☐	☐	☐	☐	☐
b. Depression	☐	☐	☐	☐	☐	☐	☐	☐
c. Overall Bipolar Illness	☐	☐	☐	☐	☐	☐	☐	☐

E. Is the patient experiencing significant side effects? ☐ 1. Yes ☐ 2. No

If yes, specify and rate:	1. MILD Does not interfere with patient's functioning or comfort	2. MODERATE Significantly interferes with patient's functioning or comfort	3. SEVERE Markedly interferes with patient's functioning or comfort
a. _____	☐	☐	☐
b. _____	☐	☐	☐

F. NOTE: If you feel that this assessment is unreliable, mark the appropriate reason below:
1.☐Not Compliant with Study Meds 2.☐Non-Study Meds Added 3.☐Substance Abuse 4.☐Other_____

Fig. 2. The Clinical Global Impressions Scale-Bipolar Version (CGI-BP).

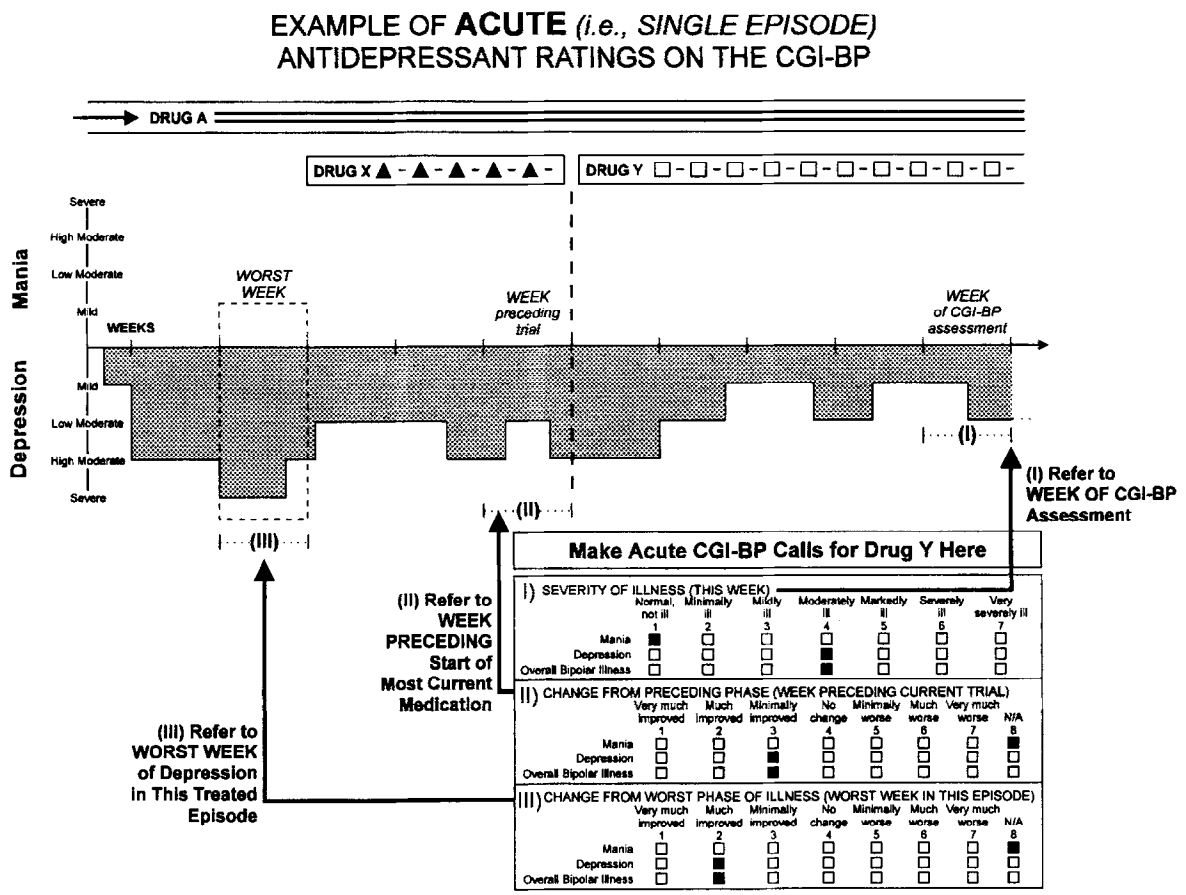

Fig. 3. Example of acute (i.e. single episode) antidepressant ratings on the CGI-BP.

Item III. The side-effect scale items were not constructed consistently, and the scale did not permit the individual assessment of the severity of each side effect.

To address the conceptual and scaling problems of the CGI Item III (Efficacy Index), several changes were made to the original concept of this item and incorporated into the CGI-BP. Item III was changed to 'Change from Worst Phase of Illness' to allow a focused assessment of improvement or worsening during a treatment as compared to when the patient was recently most ill. When assessing the efficacy of a treatment for an acute episode of mania or depression, the worst phase would be considered the worst week within the current episode being treated (Fig. 3). However, when rating the prophylactic effectiveness of a treatment, the worst phase would refer to a period of time in the past when manic and depressive episodes were at their worst, i.e. preferably when the illness was recently untreated or when preventative treatments had been initiated but the illness continued or progressed (Fig. 4).

The new CGI-BP Item III assessment eliminates the redundancy between the old CGI Item II (Global Improvement) and Item III (Efficacy Index), since it asks the rater to compare the current treatment phase to a time period that often may be different from the phase immediately preceding the current treatment rated in Item II. In addition, Change from Worst Phase of Illness attempts to address the question of how

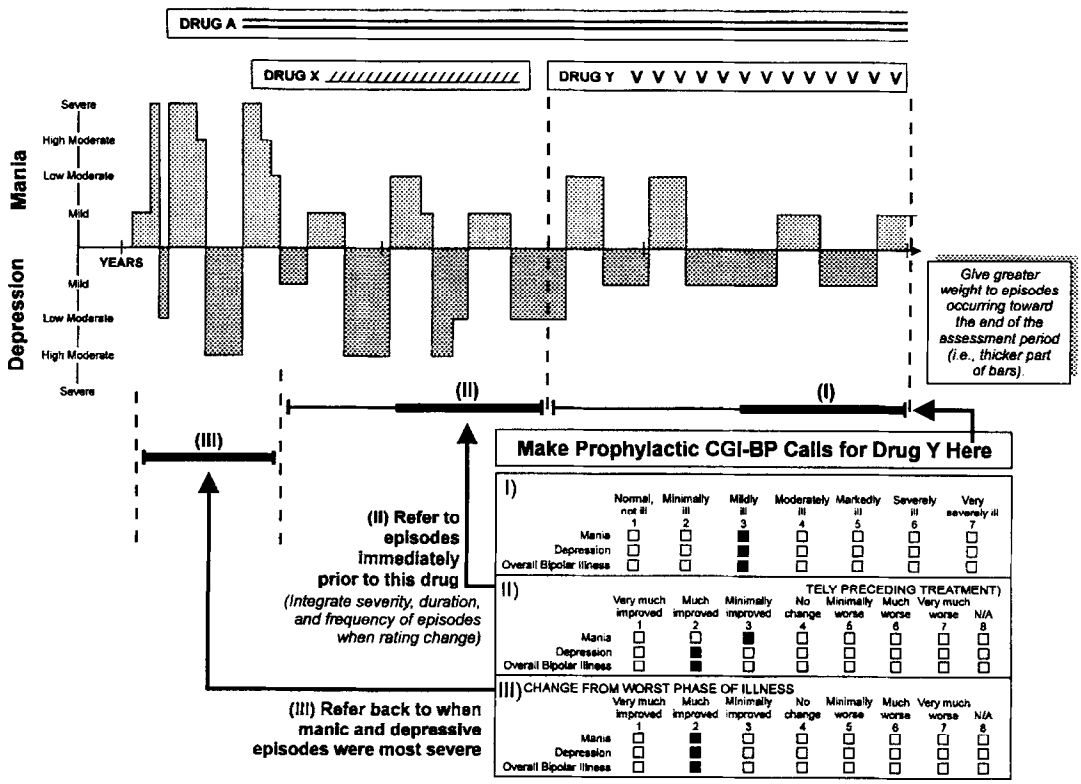

Fig. 4. Example of prophylactic (i.e. cycling prevention) ratings on the CGI-BP.

well a treatment regimen has altered the illness from when the acute episode or cycling course was most fulminant. Although the original CGI Item III was intended to assess in part whether or not the change was attributable to the intervention used, Beneke and Rasmus (1992) argued against this concept and asserted that a global rating of change should reflect simply the degree of clinically observed change. The CGI-BP Item II, Change from Preceding Phase, easily accommodates this idea. However, in addition, the Change from Worst Phase of Illness measure readily allows assessment of the overall degree of improvement that either a single or combined treatment regimen provides. Other approaches (in clinical trial designs or comparisons with the prior or subsequent course of illness when untreated) would have to be utilized in order to further verify if the observed change was or was not likely attributable to the intervention being studied (Pazzaglia et al., 1993; McDermut et al., 1995; Post et al., 1996a;).

A final, important distinction between the CGI Item III (Efficacy Index) and the CGI-BP Item III (Change from Worst Phase of Illness) is that the latter does not integrate a side effects rating into the assessment of improvement. Instead, a separate item was created for the more conventional assessment of the severity of individual side effects based on their degree of interference with patients' comfort or functioning (Fig. 2, bottom).

3.2. User's Guides

Given the multiple changes in emphasis and the clarification of new time domains for each of

the CGI-BP measures, the Detailed and Brief User's Guides were intended to explicate and teach clinicians how to use the revised scale in a consistent fashion. Definitions, graphic illustrations, and 'rating hints' were included in an attempt to secure the accuracy and reliability of ratings among clinicians. The following definitions were developed for rating Item I, Severity of Illness: (1) Normal, Not Ill (no symptoms, not at all ill); (2) Minimally Ill (minimal symptoms, continued effective functioning); (3) Mildly Ill (low level symptoms, subjective distress, little to no functional impairment); (4) Moderately Ill (some prominent symptoms, moderate functional impairment); (5) Markedly Ill (significant symptoms, very substantial functional impairment); (6) Severely Ill (very notable symptoms, unable to function in most areas); and (7) Very Severely Ill (extreme symptoms, completely incapacitated, requiring extra care).

In addition, the following definitions of change were specified for Items II and III, Change From Preceding Phase and Change From Worst Phase of Illness: (1) Very Much Improved (all better or nearly all better; very good level of functioning; minimal residual symptoms; represents a very substantial change); (2) Much Improved (notably better with significant reduction of symptoms; increase in the level of functioning, but some symptoms remain); (3) Minimally Improved (slightly better with little or no clinically meaningful reduction of symptoms; represents very little change in basic clinical status, level of care, or functional capacity); (4) No Change (symptoms remain essentially unchanged); (5) Minimally Worse (slightly worse but not clinically meaningful and represents very little change in basic clinical status or functional capacity); (6) Much Worse (notably worse with significant increase in symptoms and loss of functioning in several areas of usual social or occupational roles); (7) Very Much Worse (distinctly worse with severe exacerbation of symptoms and loss of functioning); and (8) Not Applicable (a particular mood state, i.e. mania or depression, cannot be rated at this time because it has not occurred during this rating period).

A page of additional hints for prophylactic ratings of bipolar disorder (i.e. treatment and prevention of episode recurrence or cycling) was included in both Detailed and Brief guides to further clarify a number of principles helpful for rating global severity and change in this variable illness over time. For example, (a) it is important to integrate episode severity, duration, and frequency in order to capture change in the illness characteristics that may be clinically meaningful along these different dimensions. Additionally, (b) at least one or two cycles of illness (i.e. several manic and depressive episodes) should be considered when assessing illness severity and improvement on a prophylactic treatment in order to accurately reflect bipolar illness presentation and change. Moreover, (c) episodes occurring toward the end of the treatment phase being assessed should be given more weight in the ratings of severity and improvement so as to account for how well the treatment is currently working, i.e. usually after the longest duration at the optimal or current dose.

3.3. Interrater reliability

Interrater reliability of the CGI-BP was examined in two preliminary stages. The first stage assessed the reliability of the Detailed User's Guide in teaching the concepts of the scale, while the second stage demonstrated the reliability of the scale itself (Severity of Illness measure only) with ratings done from videotapes of interviews with bipolar patients.

In the first stage, nine clinicians at field centers of the Stanley Foundation Bipolar Network read the CGI-BP Detailed User's Guide, rated the practice life charts (with answer keys), and then completed the test packet composed of 10 ratings of single episode, acute treatment response and 10 ratings of multiple episode, prophylactic treatment response. Interrater reliability was demonstrated for each of the three CGI-BP measures through intraclass correlation analyses (Shrout and Fleiss, 1979) on acute and prophylactic ratings of overall bipolar illness. Analysis of Item I (Severity of Illness) indicated excellent agreement on ratings of overall bipolar illness for both acute (0.91) and prophylactic (0.75) assessments. Evaluation of Item II (Change from Preceding Phase)

revealed strong interrater reliability for acute ratings (0.86) and moderate agreement for prophylactic ratings (0.63). Item III (Change from Worst Phase of Illness) analysis demonstrated good reliability for acute assessments (0.76) and moderate reliability for prophylactic assessments (0.64). These results provide preliminary support for the reliability of the CGI-BP as taught and assessed by the 20 test cases illustrated in the Detailed User's Guide.

In the second stage of reliability testing, 11 clinicians previously trained on the CGI-BP scale and Detailed User's Guide viewed videotaped clinical interviews of five different patients with bipolar disorder and rated each patient's acute Severity of Illness (Item I) for overall bipolar illness. A 'Gold Standard' rating of each interview was provided (GSL) as an additional measure of the reliability of individual raters. Intraclass correlation analyses (Shrout and Fleiss, 1979) revealed excellent agreement among all raters across ratings of mania (no variability), depression (0.92), and overall bipolar illness (0.93). In addition, *t*-tests (two-tailed) and correlations were performed comparing each rater with the Gold Standard across all five videotapes. There were no significant differences between any rater and the Gold Standard ($P < 0.05$), and correlations across all ratings were very high (0.87–0.99). The mean difference across all tapes for any one rater compared to the Gold Standard was never greater than one severity unit.

4. Discussion

Not only does the CGI-BP correct many of the flaws of the original CGI scale, but also it preserves the fundamental assets of the original global rating instrument while focusing specifically on the different and variable components of bipolar illness. Basically, the CGI-BP scale Items II and III systematize what clinicians inherently do when making decisions regarding the clinical utility of a treatment — that is, judging whether a patient on a given treatment is essentially all better (Very Much Improved; i.e. excellent, a grade of an 'A'), much better (Much Improved; i.e. moderately improved, or a 'B'), somewhat better (Minimally Improved; i.e. mildly improved, or a 'C'), no different (No Change; i.e. a 'D'), somewhat worse (Minimally Worse; i.e. mildly worse, or an 'E'), and so on.

In addition, more specifically, the CGI-BP has been carefully designed and revised for use in bipolar illness in a variety of ways. The scale allows for separate assessments of each phase of the illness (i.e. mania, depression, and overall illness). Also, the scale can be used to assess either the acute (single episode) or prophylactic (long-term prevention of episodes) effects of treatments. For prophylactic assessments, the scale allows clinicians to integrate severity, duration, and frequency of episodes in evaluating a treatment. Because time domains for assessments were clarified for each measure of the scale, clinicians now have a more standardized framework within which they can make global ratings of illness severity and change of both single and adjunctive treatments.

While the CGI-BP improves upon the original CGI scale in the several ways noted above, liabilities of the scale remain due primarily to its nature as a global rating instrument. The concept of a global rating scale inherently involves a degree of integration of information and subjectivity which is not as prevalent in more differentiated, symptom-based rating scales. However, CGI-BP ratings are not intended to replace ratings from more explicit, symptom-driven scales or detailed course of illness information (i.e. severity, duration, or frequency of episodes) available from a continuous longitudinal measure such as the NIMH Life Chart Method® (Leverich and Post, 1995, 1996). It still may be difficult, at times, to integrate all of the illness variables into global ratings of illness severity (I) and change (II and III). However, the flexibility and requirement of the CGI-BP to accommodate the differing illness characteristics into a single rating or set of ratings are unique and add to the overall value of the scale in evaluating degree of response to treatment in bipolar illness.

Although the CGI-BP now has operationalized guidelines and time domains for assessment, there may be situations in which the appropriate time frame to refer to is unclear. For instance, while

the specific weeks to be evaluated in an acute treatment rating are clearly indicated, it may be more difficult when making prophylactic CGI-BP ratings to identify the length of time referred to as the immediately 'preceding phase' or the appropriate most recent 'worst phase of illness'. In addition, in some instances, both change ratings (II and III) may even refer to the same time period (i.e. when the preceding treatment phase was also the worst phase of illness). Choosing these intervals for comparison and giving the greatest weight to the most recent episodes in a prophylactic evaluation can involve a degree of clinical judgment, although the better defined guidelines and specifications of the CGI-BP should assist in the process.

Nevertheless, the lack of an arbitrary definition of a set time interval for the prophylactic ratings has the advantage of dealing with one of the key characteristics of bipolar illness — its inherent variability within and between individuals. For example, to include at least two cycles of illness in the prophylactic assessment of a treatment in one individual might require years, while in another patient with ultra-rapid cycling, only days to weeks. These vastly different time domains can readily be incorporated into the specified rating intervals for the change items II and III of the CGI-BP. Thus, what some critics might consider a liability of the revised CGI-BP can also be seen as an asset in dealing more flexibly with the range of illness variables presented by patients with bipolar disorder for evaluation and treatment.

A final consideration regarding the reliability of the CGI-BP is that the User's Guide teaches clinicians to make ratings based on sample patient life charts and not on direct clinical observation or interaction with patients (as the scale would be used in clinical practice). Clear information on episode severity, duration, and frequency is readily available from the life charts in the User's Guide and test packet. Although these illness variables may be more difficult to assess and quantify from a patient interview, we acquired excellent interrater reliability on CGI-BP acute Severity of Illness ratings using videotaped interviews. Assessments of the prophylactic efficacy of a treatment using the CGI-BP may be particularly challenging if detailed information such as that in a life chart has not been obtained; moreover, videotape-based reliability assessments of the change measures (Items II and III) may be especially difficult.

In addition to the preliminary reliability evaluation presented here, an initial validity assessment of the CGI-BP has been conducted by Denicoff et al. (unpublished results) in the context of a randomized clinical trial study of bipolar patients on 1 year of carbamazepine or lithium prophylaxis, cross-over to the opposite drug in the second year, and a third year on the combination (Denicoff et al., 1997). CGI-BP criteria for rating Item III, Change from Worst Phase of Illness on overall bipolar illness was used to assess treatment response in each phase of the study. Validity of this CGI-BP change measure was examined by comparing it to the degrees of affective episode severity, duration, and frequency as assessed on daily prospective Life Chart Methodology® (NIMH-LCM-p®) measures. Ratings of 'Marked' improvement (i.e. 'Very Much Improved') were associated with negligible morbidity on NIMH-LCM-p® ratings (e.g. only a few days in the year with mild symptoms). Progressively greater degrees of illness were evident in severity, duration, or numbers of episodes rated on the NIMH-LCM-p® as CGI-BP treatment response was classified successively as 'Moderate' (i.e. 'Much Improved'), 'Minimal' (i.e. 'Minimally Improved'), 'Unchanged' (i.e. 'No Change'), or 'Worse' (i.e. 'Minimally', 'Much', or 'Very Much Worse'). While further systematic efforts are required to document the reliability and validity of the CGI-BP measures, these preliminary results support the utility of the scale in prospective assessments of treatment efficacy in bipolar disorder.

Acknowledgements

This work was supported by the Ted and Vada Stanley Foundation. We also thank Amy Thibault, B.A. and Rachel Hayden, M.S.W. of the Stanley Foundation Bipolar Network for their input and advice in the preparation of this scale.

References

Beneke, M., Rasmus, W., 1992. 'Clinical global impressions' (ECDEU): some critical comments. Pharmacopsychiatrie 25, 171–176.

Calabrese, J.R., Woyshville, M.J., Kimmel, S.E., Rapport, D.J., 1993. Predictors of valproate response in bipolar rapid cycling. Journal of Clinical Psychopharmacology 13, 280–283.

Colle, L.M., Belair, J.F., DiFeo, M., Weiss, J., LaRoche, C., 1994. Extended open-label fluoxetine treatment of adolescents with major depression. Journal of Child and Adolescent Psychopharmacology 4, 225–232.

Dahlke, F., Lohaus, A., Gutzmann, H., 1992. Reliability and clinical concepts underlying global judgments in dementia: implications for clinical research. Psychopharmacology Bulletin 28, 425–432.

Denicoff, K.D., Smith-Jackson, E.E., Disney, E.R., Ali, S.O., Leverich, G.S., Post, R.M., 1997. Comparative prophylactic efficacy of lithium, carbamazepine and the combination in bipolar disorder. Journal of Clinical Psychiatry, in press.

DeWilde, J., Spiers, R., Mertens, C., Bartholome, F., Schotte, G., Leyman, S., 1993. A double-blind, comparative, multicentre study comparing paroxetine with fluoxetine in depressed patients. Acta Psychiatrica Scandinavica 87, 141–145.

Fieve, R.R., Goodnick, P.J., Peselow, E.D., Barouche, F., Schlegel, A., 1986. Pattern analysis of antidepressant response to fluoxetine. Journal of Clinical Psychiatry 47, 560–562.

Guy, W. (Ed.), 1976. Clinical Global Impressions. In: ECDEU Assessment Manual for Psychopharmacology, revised. National Institute of Mental Health, Rockville, MD.

Honer, W., MacEwan, G., Kopala, L., Altman, S., Chisholm-Hay, S., Singh, K., Smith, G., Ehmann, T., Ganesan, S., Lang, M., 1995. A clinical study of clozapine treatment and predictors of response in a Canadian sample. Canadian Journal of Psychiatry 40, 208–211.

Kanowski, S., Fischhof, P.K., Grobe-Einsler, R., Wagner, G., Litschauer, G., 1990. Efficacy of xantinolnicotinate in patients with dementia. Pharmacopsychiatrie 23, 118–124.

Lehmann, E., 1984. Practicable and valid approach to evaluate the efficacy of nootropic drugs by means of rating scales. Pharmacopsychiatrie 17, 71–75.

Leverich, G.S., Post, R.M., 1996. Life charting the course of bipolar disorder. In: Rush, A.J. (Issue Ed.), Current Review of Mood and Anxiety Disorders, Issue 1. Current Medicine, Philadelphia, pp. 48–61.

Leverich, G.S., Post, R.M., 1995, revised version. The NIMH Life Chart Manual for Recurrent Affective Illness: The LCM®. Biological Psychiatry Branch Monograph, Bethesda, MD.

McDermut, W., Pazzaglia, P.J., Huggins, T., Mikalauskas, K., Leverich, G.S., Ketter, T.A., Bartko, J., Post, R.M., 1995. Use of single case analyses in on-off-on trials in affective illness: a demonstration of the efficacy of nimodipine. Depression 2, 259–271.

Okuma, T., 1993. Effects of carbamazepine and lithium on affective disorders. Neuropsychobiology 27, 138–145.

Ontiveros, A., Fontaine, R., Elie, R., 1991. Refractory depression: the addition of lithium to fluoxetine or desipramine. Acta Psychiatrica Scandinavica 83, 188–192.

Pazzaglia, P.J., Post, R.M., Ketter, T.A., George, M.S., Marangell, L.B., 1993. Preliminary controlled trial of nimodipine in ultra-rapid cycling affective dysregulation. Psychiatry Research 49, 257–272.

Post, R.M., Ketter, T.A., Denicoff, K., Pazzaglia, P.J., Leverich, G.S., Marangell, L.B., Callahan, A., George, M.S., Frye, M.A., 1996a. The place of anticonvulsant therapy in bipolar illness. Psychopharmacology 128, 115–129.

Post, R.M., Ketter, T.A., Pazzaglia, P.J., Denicoff, K., George, M.S., Callahan, A., Leverich, G.S., Frye, M.A., 1996b. Rational polypharmacy in the bipolar affective disorders. Epilepsy Research 115, 153–180.

Post, R.M., Leverich, G.S., Rosoff, A.S., Altshuler, L.L., 1990. Carbamazepine prophylaxis in refractory affective disorders: a focus on long-term follow-up. Journal of Clinical Psychopharmacology 10, 318–327.

Quitkin, F.M., Rabkin, J.G., Ross, D., Stewart, J.W., 1984. Identification of true drug response to antidepressants: use of pattern analysis. Archives of General Psychiatry 41, 782–786.

Rabkin, J.G., Rabkin, R., Wagner, G., 1995. Testosterone replacement therapy in HIV illness. General Hospital Psychiatry 17, 37–42.

Ravizza, L., Barzega, G., Bellino, S., Bogetto, F., Maina, F., 1995. Predictors of drug response in obsessive-compulsive disorder. Journal of Clinical Psychiatry 56, 368–373.

Remick, R.A., Reesal, R., Oakander, M., Allen, J., Claman, J., Ramirez, C.E., Perry, K., Keller, F.D., 1994. Comparison of fluvoxamine and amitriptyline in depressed outpatients. Current Therapeutic Research 55, 243–250.

Rush, A.J., Kupfer, D.J., 1995. Strategies and tactics in the treatment of depression. In: Gabbard, G.O. (Ed.), Treatments of Psychiatric Disorders, 2nd ed., vol. 1. American Psychiatric Press, Inc., Washington, DC, pp. 1349–1368.

Salzmann, E., Robin, J.L., 1995. Multicentric double-blind study comparing efficacy and safety of minaprine and imipramine in dysthymic disorders. Pharmacopsychiatrie 31, 68–75.

Sato, T.L., Turnbull, C.D., Davidson, J.R.T., Madakasira, S., 1984. Depressive illness and placebo response. International Journal of Psychiatry Medicine 14, 171–179.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86, 420–428.

Weitkunat, R., Letzel, H., Kanowski, S., Grobe-Einsler, S., 1993. Clinical and psychometric evaluation of the efficacy of nootropic drugs: characteristics of several procedures. Zeitschrift für Gerontopsychologie undpsychiatrie 6(1), 51–60.