A video clinical global impression scale (CGI) in obsessive compulsive disorder

Abderrahmane Bourredjem, Antoine Pelissolo, Jean-Yves Rotge, Nematollah Jaafari, Sebastien Machefaux, Solene Quentin, Eric Bui, Nicolas Bruno, Jean-Baptiste Pochon, Mircea Polosan, Nicolas Baup, François Papetti, Isabelle Chéreau, Christophe Arbus, Luc Mallet, Sophie Tezenas du Montcel, and for the French “Stimulation dans le Trouble Obsessionnel Compulsif (STOC)” Study Group

Institute of the Santé et de la Recherche Médicale (INSERM), Avenir Team, Behavior, Emotion, and Basal Ganglia, IFR70, France

Clinical Investigation Center Inserm U 802, Poitiers University Hospital, Poitiers, France

Biostatistics and Medical Informatics Unit, Pitié-Salpêtrière University Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

Department of Psychiatry, Saint-Anne, Hospital University Hospital Department, Paris, France

Department of Psychiatry, University Hospital, Toulouse and Traumatic Stress Laboratory, (JE 2511), University of Toulouse, France

Clinical Investigation Center, Pitié-Salpêtrière University Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

Department of Psychiatry, Grenoble University Hospital, Grenoble, France

Paris Descartes University, Paris Descartes Medical Faculty; Laboratory of Pathophysiology of Psychiatric Diseases, Center of Psychiatry and Neurosciences, INSERM U894, Paris, France

Department of Psychiatry, University Hospital, Clermont-Ferrand, France

Institut National de la Santé et de la Recherche Médicale (INSERM), Avenir Team, Behavior, Emotion, and Basal Ganglia, IFR70, France

Clinical Investigation Center Inserm U 802, Poitiers University Hospital, Poitiers, France

A video clinical global impression scale (CGI) in obsessive compulsive disorder

1. Introduction

Reliable metrics are an integral component of outcome-based intervention evaluations, which are the mainstay of evidence-based clinical practice. In clinical trials, one typically wants to evaluate treatment effects on disease status. If measurement reliability is low, the ability to identify the effects of treatment in the different treatment arms decreases. The Clinical Global Impression scale (CGI) is a classic instrument for global assessment in psychiatric disorders (Guy, 1976). The CGI scale requires the clinician to rate the overall severity of the patient's illness at the time of assessment relative to the clinician's experience with patients having the same diagnosis. This scale yields three different measures: severity of illness, global improvement, and the efficacy index.

The CGI is widely used in psychiatric clinical trials and has been shown to correlate with standard drug-efficacy scales in diseases such as major depressive disorder, panic disorder, social anxiety disorder (Bandelow et al., 2006), generalized anxiety disorder (Zaider et al., 2003) and schizophrenia (Leucht et al., 2005; Leucht and Engel, 2006a; Leucht et al., 2006b; Levine et al., 2008). The standard CGI is most commonly used, although revisions have been proposed for some diseases such as dementia (Dahlke et al., 1992), bipolar illness (CGI-BP) (Spearing et al., 1997) and schizophrenia (CGI-S) (Haro et al., 2003).

Studies have tested the reliability of the CGI scale in both its standard (Kadouri et al., 2007) and revised (Dahlke et al., 1992;
Spearing et al., 1997; Haro et al., 2003) versions and have demonstrated good inter-rater reliability in the above diseases. Kadouri et al. showed that the standard CGI scale was as good as an improved CGI scale in depression and used a video protocol to improve CGI reliability in depressive disorders (Kadouri et al., 2007). A video protocol also allows the implementation of quality control procedures. The CGI has not been tested in obsessive compulsive disorder (OCD).

The purpose of this article was to investigate the reliability of the CGI scale and the usefulness of a standardized video protocol in evaluating severe OCD patients, drawing on a clinical trial of the effect of subthalamic nucleus (STN) stimulation in OCD.

2. Methods

2.1. Patients and study design

The “Stimulation dans le Trouble Obsessionnel Compulsif” (STOC) trial was conducted in 10 academic centers in France in accordance with the Declaration of Helsinki and was approved by the ethics committee of the Pitié-Salpêtrière University Hospital. All patients provided written informed consent.

Sixteen patients with highly refractory OCD were included in a 10-month crossover, double-blind, multicenter study assessing the efficacy and safety of STN stimulation (Mallet et al., 2008). Patients were randomly assigned in a 1:1 ratio to either 3 months of stimulation followed by a 3-month sham stimulation period (the On/Off group) or the reverse (the Off/On group). The two phases were separated by a 1-month wash-out period. A psychiatrist and a neurologist blinded to patients' stimulation status evaluated the patients at five time-points: in the 2 months prior to surgery for inclusion (visit 1); at 3 months after surgery with stimulation off (visit 2); at 6 months post-surgery corresponding to the end of the first crossover period (visit 3); at 7 months post-surgery, corresponding to the end of the wash-out period (visit 4); and at 10 months after surgery, corresponding to the end of the second crossover period (visit 5). Drug treatments and dosages were kept unchanged throughout the study and psychotherapy was not allowed (Mallet et al., 2008).

Study psychiatrists confirmed the diagnosis of OCD for all patients, assessed OCD severity using the Yale-Brown Obsessive Compulsive Scale (YBOCS), (Goodman et al., 1989) and evaluated patients with the CGI (Guy, 1976) and scales assessing global health and functioning (Global Assessment of Functioning (GAF)) (American Psychiatric Association, 2000). Psychiatrists completed these assessments at the bedside and were blinded to the patients' stimulation status (face-to-face rating). They used a standardised clinical interview to assess the CGI scale evaluating the severity of illness (CGI-S). The interview structure, adapted from the interview in depressed patients (Kadouri et al., 2007), was similar to day-to-day clinical practice (see Appendix B). According to the response to the first question of “Hello, how are you?”, follow-up questions assessed the intensity and the impact of OCD symptoms in the previous four weeks.

A video of at least 5-min duration of each interview provided the material for raters to form an “impression” of the patient's disease severity, ranging from “Normal” (1) to “Extremely Ill” (7). Patients wore head coverings to maintain blinding to the visit number. We kept the first 5 min of all videos and, for videos lasting more than 10 min, the 5–10 min section for a specific subanalysis. We deleted any information which might reveal the visit number from the final videos.

Six psychologists or psychiatrists rated each video. All had formal training in the clinical assessment of psychiatric disorders. When reviewing the interviews, the order of the patients was the same for all raters but the order of the interviews for each patient was randomized. Thus, for a given patient, videos were presented in an unpredictably different order from one rater to another. For interviews lasting 10 min or more, the first 5 min and the next 5 min were left next to each other. The rater scored the first 5 min before viewing the second 5 min and providing a second rating on the full 10 min.

2.2. Statistical analysis

We used the Spearman correlation test to study the correlation between the YBOCS and GAF scores and the mean of the six video CGI scores and the intraclass correlation coefficient (ICC) to assess both the agreement between face-to-face and video CGI scores and the inter-rater agreement between two video ratings. We calculated the ICC as the ratio of the inter-rater variance to the sum of the inter-rater, inter-rater and residual variance (Shrout and Fleiss, 1979).

For the inter-rater agreement between two videos, we used a linear mixed-effects model taking into account the five visits with a fixed effects structure respecting the trial's crossover design, based on Jones and Kenward (1989), and a random effect for each group of the six raters. The ICC was estimated based on Molenaers et al. (2007) and the influence of different factors was tested by likelihood ratio tests. This model allowed testing of the treatment, carry-over and period effects.

We estimated the inter-rater agreement for the mean of ratings from m independent psychologists or psychiatrists. The ICC was the ratio of the inter-patient variance to the sum of the inter-patient variance, the inter-rater variance divided by m and the residual variance divided by m (Steiner and Norman, 1994). This approach estimates the reliability between the mean of m ratings and the mean of m’ other ratings.

To calculate confidence intervals of the ICC, we used a bootstrap procedure of the residuals of the linear mixed-effects models. Statistical analyses were performed with SAS 9.1 software (SAS Institute, Cary, NC). We considered test results significant at the 0.05 level and reported two-tailed P-values.

3. Results

3.1. Descriptive results

At least one video was available for 11 of the 16 patients randomized in the STOC study, representing 6 of the 10 study centers. A total of 50 videos were available: at visit 1, 11 videos of 5 min and 9 videos of 10 min; at visits 2, 4, and 5, 10 videos of 5 min; and at visit 3, 9 videos of 5 min. Mean patient age was 43 years (S.D. = 9 min; 29, max: 56). Eight patients were male and three female; six patients belonged to the “on/off” group of the STOC study and five to the “off/on” group.

The sample represented a wide range of OCD severity over the five visits, with video CGI scores ranging from 2 to 7 and YBOCS scores ranging from 9 to 38. Overall, the level of OCD was severe (mean video CGI score: 5.3, S.D. = 1.3 and YBOCS: 27.7, S.D. = 7.3). However, severity was less variable at visits 1, 2 and 4 than at visits 3 and 5 (Table 1).

The mean of the six video CGI scores was highly correlated with the GAF scores at visits 3 and 4 (Fig. 1A) and with the YBOCS scores at visits 2, 3, and 4 (Fig. 1B).

3.2. Video CGI score during the crossover study

The association of the video CGI score with stimulation was similar to that of the face-to-face CGI score (Mallet et al., 2008). The CGI score was significantly lower at the end of the “on” period (P = 0.0009) with no period effect (P = 0.8) and no carry-over effect (P = 0.2).

3.3. Reliability of the video CGI

The ICC of the video CGI score at visit 1 (0.304 [0.12–0.495]) was significantly lower than the ICC at visits 2 to 5 (0.678 [0.609–0.797], likelihood ratio test, P = 0.017) (Fig. 2). A post-hoc combined reliability score was therefore calculated for visits 2 to 5. The reliability between the mean of the raters increased with the number of raters (Fig. 3). The ICC was greater than 0.80 when 2 or more evaluations of this combined score were averaged. This was also achieved with two ratings averaged at visits 4 and 5 and three ratings averaged at visits 2 and 3. The ICC remained lower than 0.75 for visit 1 even with six ratings averaged.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Visit1</th>
<th>Visit2</th>
<th>Visit3</th>
<th>Visit4</th>
<th>Visit5</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Video CGI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mean of the 6 raters)</td>
<td>6.1 ± 0.8</td>
<td>5.2 ± 1.0</td>
<td>4.3 ± 1.4</td>
<td>5.3 ± 1.3</td>
<td>5.2 ± 1.3</td>
</tr>
<tr>
<td>Face-to-face CGI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YBOCS</td>
<td>6.2 ± 0.6</td>
<td>5.5 ± 0.7</td>
<td>4.3 ± 1.4</td>
<td>5.4 ± 1.0</td>
<td>5.4 ± 1.3</td>
</tr>
<tr>
<td>GAF</td>
<td>32.9 ± 3.9</td>
<td>29.2 ± 6.3</td>
<td>22.6 ± 8.8</td>
<td>28.3 ± 5.8</td>
<td>25.5 ± 7.6</td>
</tr>
<tr>
<td>Agreement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>between video CGI and face-to-face CGI</td>
<td>0.331</td>
<td>0.649</td>
<td>0.736</td>
<td>0.685</td>
<td>0.753</td>
</tr>
</tbody>
</table>

Data are expressed as mean ± SD ($), intraclass correlation coefficient ($$).
3.4. Factors influencing the reliability of the video CGI

To examine the best video length for scoring the CGI, we tested whether the 10-min duration gave a better reliability than the 5-min duration. Nine videos, at visit 1, were available for this. The ICC was higher for the 10-min video than the 5-min video (0.625 [0.527–0.774] vs 0.404 [0.187–0.605]) but the difference was not statistically significant (Likelihood ratio test: \( P = 0.40 \)).

The investigator performing the video interview was blinded to the patient’s stimulation status and the video order was also randomized. As patients could be on stimulation only at visits 3 and 5, we tested for the influence of stimulation status on reliability using the videos for these visits. The ICC for the videos of patients in the stimulation period was not statistically different from those in the non-stimulation period (0.621 vs 0.616, Likelihood ratio test \( P = 0.23 \)).

3.5. Relation between face-to-face and video CGI

The reliability between the face-to-face and video CGI scoring was similar but slightly lower than the reliability between the video CGI assessments (Table 1). The ICC between the face-to-face and the video CGI scores was low at visit 1 (0.331) and good at the other visits (from 0.649 at visit 2 to 0.753 at visit 5) (Fig. 1C).

4. Discussion

Properties of outcome measures suitable for clinical trials include established utility in research settings, reliability, sensitivity to change, simplicity, brevity, and absence of copyright. The CGI meets most of these criteria. We have shown a high correlation between the CGI, a simple, single rating ranging from 1 to 7, and more complex disease scales such as the YBOCS when the CGI is assessed using video. The video CGI was highly correlated with the YBOCS at visits 2, 3, and 4. At visit 1, both the patient and the psychiatrist conducting the interview knew that the patient was off stimulation. However, all patients were also off stimulation at visits 2 and 4 without an influence on the reliability of the scale or on the correlation between the CGI and the YBOCS. Further, patient severity was very homogenous owing to stringent inclusion criteria, with little comorbidity (Mallet et al., 2008) but this pattern was mostly observed at visit 1 (inclusion). Homogeneity was less marked at the other visits because...
of the effects of STN stimulation and the crossover design. In addition, at visit 5, video CGI was correlated with face-to-face CGI but not with the GAF or YBOCS. We noted that the two outlying patients identified on the scatter plot were in the “off/on” intervention group but did not find any explanation for this.

We have also shown that video rating of the CGI provides a reliable evaluation of the severity of the patient, even in the context of a crossover trial (ICC = 68%). Our ICCs are comparable to those of other studies (Spearing et al., 1997; Haro et al., 2003; Kadouri et al., 2007), even if an adapted form of the CGI was tested in some of the previous studies. The ICC can exceed 80% when using the mean of two raters. Using the mean of ratings from several raters decreases the variance and thus increases the precision of the estimate without introducing bias.

Rather than using the CGI-improvement scale, we used the CGI-S scale and evaluated patient improvement by the difference of the CGI-S before and after each period because of the trial’s crossover design. Beneke and Rasmus do not recommend this approach (Beneke and Rasmus, 1992). However, Dahlke et al. have shown that the reliability scores for CGI-S did not vary greatly, whereas the reliability scores for CGI-improvement showed wide confidence intervals (Dahlke et al., 1992). Further, Kadouri et al. have shown that reliability does not differ between these two approaches (Kadouri et al., 2007).

We have used a more complex model than usual to evaluate reliability so as to take into account the complexity of the crossover data. As shown by Molenberghs et al. (2007), it is essential to take into account both the fixed effects and the variance model underlying the data to estimate reliability. The measurements of the fixed effects in clinical trials are often ‘unstable’, in a psychometric sense, owing to the presence of treatment and time effects as well as the effects of other covariates (Molenberghs et al., 2007). Thus we used the approach proposed by Jones and Kenward (1989) to model the fixed effects of the crossover design. This model takes into account the treatment, period and carry-over effects involved in this kind of design. Further, we used a model based on generalizability theory so as to incorporate all of the sources of variability in the same analysis of variance (Molenberghs et al., 2007). Stimulation status was the only factor that could be tested but we found no influence of it on reliability.

When using longitudinal data to estimate reliability, variances must be equal at each measurement time (Molenberghs et al., 2007). This was the case in our study for visits 2 to 5 but not for visit 1, leading to different ICCs. Using a video to rate the patients can provide raters with only the information needed for rating the CGI and so ensure they are blinded to the patient’s treatment status and evolution of the patient’s condition during the trial. This avoids influencing the CGI through knowledge of when the patient was filmed. However, reliability was lowest at visit 1. This may be because patients were very homogenous at visit 1 because of the stringent inclusion criteria since good reliability is harder to achieve in such a context of low variability. As the order of each patient’s videos for the different raters was random, a training effect is unlikely.

Regarding the patient effect, all the patients included in the STOC protocol had several medical interviews before inclusion. However, for some of the patients, visit 1 was their first interview with the STOC-study psychiatrist. Thus the psychiatrist was less focused in asking the questions and 5 min seemed insufficient to detect all of the patients’ symptoms. The reliability was higher at visit 1 using the 10-min rather than the 5-min videos (even if not statistically significant, possibly owing to a lack of power related to the small number of videos available at that visit). The ICC at visit 1 based on the 10-min videos almost equaled the ICC at the other visits based on the 5-min videos. This first interview is probably a training interview for both the patient and the doctor. Interestingly, we had 10-min videos for almost all interviews at visit 1, but few videos for the other visits. It seems that fewer than 10 min are enough to assess the intensity and impact of OCD symptoms for patients with intensive follow-up (evaluation OCD symptoms every 3 months, visits to evaluate side effects every month). When the patient is not known to the doctor, we therefore suggest an initial training interview before starting the protocol with the video interview. Another protocol could be to allow the patient and the investigator to define the optimal length of the video. However this protocol presents some disadvantages as longer videos take longer to be rated. Videos of 10- rather than 5-min duration could be problematic in the context of a clinical trial with a large number of patients. In addition, we believe that the length of the videos will decrease as the trial progresses, making blinding more difficult to achieve.

The number of patients in the study was relatively small. However, the effect of the deep brain stimulation was large, meaning that we were able to show a significant effect of the stimulation versus sham stimulation (Mallet et al., 2008). For the present reliability study, we had two to five videos at different times of the protocol for each...
patient, giving a total of 50 videos. In addition, each video was rated 6 times, providing 300 different ratings. This allowed us to use linear mixed-effects models.

One can argue that STN stimulation brings about a dramatic and significant clinical improvement on patients’ symptoms and that any rater could have judged the CGI-S. However, we have shown that we cover the full range of OCD severity and that scoring was done blinded to the visit number and stimulation status. Thus, we believe that video CGI may be used even with milder effect in CNS trials or in drug trials.

We have already investigated the role of video and blind rating on reliability in a trial of deep brain stimulation in dystonia (Krystkowiak et al., 2007) and have shown that blinding was more important than video in providing good reliability. However, this study was not double-blind. In the present study, face-to-face CGI was scored blind to stimulation status and the face-to-face CGI and the video CGI were highly reliable, except at visit 1. Thus face-to-face CGI is a simple tool to evaluate patients with OCD disorders, for instance, in clinical practice. The clinical semi-structured interview is close to day-to-day clinical practice. The first question is “Hello, how are you?” According to the nature of the response, other questions are asked to assess the intensity and impact of OCD symptoms in the preceding four weeks: not all questions have to be asked but rather depend on the patient’s responses.

Part I: general questions
- Hello, how are you?
- How are you in your daily life?
- How do you feel?
- Are you currently anxious?

Part II: Specific questions
If not given spontaneously by the patients, questions are asked to address the issues of obsessions and compulsions
- What about your obsessions (or intrusive thoughts, depending on patient’s terminology) and your compulsions (or rituals)?
- Do they take up a lot of your time?
- Does the disorder currently impair you a lot?
- Can you refrain from compulsions or rituals?
- Do you have to fight a lot against these obsessions?
- What repercussions do these disorders have on your daily life?

Appendix B. Structured interview guide for the video CGI in Obsessive Compulsive Disorder Interview

The clinical semi-structured interview is close to day-to-day clinical practice. The first question is “Hello, how are you?” According to the nature of the response, other questions are asked to assess the intensity and impact of OCD symptoms in the preceding four weeks: not all questions have to be asked but rather depend on the patient’s responses.

Acknowledgements
We thank Alexandre Israël for the video digitization, Flore Mathias for her technical assistance and David Evans for reviewing the manuscript. Funding for this study was provided by a national grant (PHRC 2004) from the Direction Régionale de la Recherche Clinique Assistance Publique-Hôpitaux de Paris and by the ANR program for Young Researchers.

Appendix A


References