ELSEVIER

Sleep Medicine 4 (2003) 121–132

SLEEP MEDICINE

www.elsevier.com/locate/sleep

Original article

Validation of the International Restless Legs Syndrome Study Group rating scale for restless legs syndrome

The International Restless Legs Syndrome Study Group

Received 11 June 2002; received in revised form 22 October 2002; accepted 7 December 2002

Abstract

Background: There is a need for an easily administered instrument which can be applied to all patients with restless legs syndrome (RLS) to measure disease severity for clinical assessment, research, or therapeutic trials. The pathophysiology of RLS is not clear and no objective measure so far devised can apply to all patients or accurately reflect severity. Moreover, RLS is primarily a subjective disorder. Therefore, a subjective scale is at present the optimal instrument to meet this need.

Methods: Twenty centers from six countries participated in an initial reliability and validation study of a rating scale for the severity of RLS designed by the International RLS study group (IRLSSG). A ten-question scale was developed on the basis of repeated expert evaluation of potential items. This scale, the IRLSSG rating scale (IRLS), was administered to 196 RLS patients, most on some medication, and 209 control subjects.

Results: The IRLS was found to have high levels of internal consistency, inter-examiner reliability, test–retest reliability over a 2–4 week period, and convergent validity. It also demonstrated criterion validity when tested against the current criterion of a clinical global impression and readily discriminated patient from control groups. The scale was dominated by a single severity factor that explained at least 59% of the pooled item variance.

Conclusions: This scale meets performance criteria for a brief, patient completed instrument that can be used to assess RLS severity for purposes of clinical assessment, research, or therapeutic trials. It supports a finding that RLS is a relatively uniform disorder in which the severity of the basic symptoms is strongly related to their impact on the patient's life. In future studies, the IRLS should be tested against objective measures of RLS severity and its sensitivity should be studied as RLS severity is systematically manipulated by therapeutic interventions.

Keywords: Rating scale; Restless legs syndrome; Assessment; Reliability; Validity; Factor analysis; Psycchometrics

1. Introduction

Restless legs syndrome (RLS) is a common condition which may affect as many as 15% of the general adult population, at least in countries whose populations derive from Western Europe [1–3]. In recent years, a number of effective medications have been developed to treat this condition [4,5]. Because this is both a common and treatable disorder, it is necessary to have adequate means of assessing its severity, both for clinical management and to guide the development of further therapies. Measures of RLS severity may also be quite useful in epidemiological and pathophysiological studies.

In the past, a variety of subjective [6–9] and objective [10–12] means have been used to evaluate the severity of RLS and its response to treatment [5]. None of the subjective instruments have been extensively tested for their psychometric properties or their validity when used to assess populations of RLS patients. The most common objective measures – sleep efficiency as calculated from an overnight sleep study or various indices of periodic limb movements (PLM) – are well established in their fields as good measures of specific aspects of sleep. However, they have never been shown to reliably measure the severity of RLS in all individual patients. While some patients have major sleep complaints, others have none. While some patients have numerous periodic limb movements in sleep (PLMS), other have few or none (a significant number of PLMS is regarded as more than five per hour of sleep) [13,14]. In the absence of validated, universally applicable objective measures, the criterion for assessment of RLS remains the expert clinician's judgment or the clinical global impression (CGI). However, the CGI may not always be available or practicable and a means of assessing RLS severity that can be

* Corresponding author. Arthur S. Walters, New Jersey Neuroscience Institute, JFK Medical Center, 65 James Street, Edison, NJ 08818, USA. Tel.: +1-732-321-7000x68177; fax: +1-732-632-1584.

E-mail address: artumdnj@aol.com (A.S. Walters).

1389-9457/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S1389-9457(02)00258-7

used by trained, but not necessarily expert, interviewers would be quite helpful. Such an instrument might also be modified to be useful for self-assessment by patients. In order to provide such an instrument, the International RLS Study Group (IRLSSG) decided to develop a rating scale for measuring severity (International Restless Legs Scale or IRLS). Since RLS is a condition defined largely by its subjective impact, such a subjective rating scale is an appropriate instrument for examining different degrees of severity of the disorder. The resulting ten-question instrument was based, in large part, on the consensus clinical features of RLS as previously delineated by the IRLSSG in 1995 [15] (Table 1).[1] The scale (Appendix A) reflects both subjective assessment of the primary features (diagnostic features 1 through 3 reflected in questions 1 through 3 and 6 of the scale), intensity and frequency of the disorder (questions 7 and 8 of the scale) and associated sleep problems (features 5 and 6 reflected in questions 4 and 5 of the scale). The scale also includes questions which probe the impact of symptoms on the patients' mood and daily functioning (questions 9 and 10 of the scale).

In order to test the psychometric properties of the scale and to begin assessing its validity, the IRLSSG initiated an international, multi-center study of the scale. We now report the results of that study.

Preliminary forms of this rating scale have already been employed in published therapeutic studies [16,17]. The current version of the rating scale was utilized in a large multi-center and multi-national study of pergolide (Permax) in RLS, which has been reported in abstract form [18]. Preliminary results from the current study have also been reported in abstract form [19].

2. Methods

2.1. Development of the rating scale

The rating scale was developed on the basis of questions proposed by members of the IRLSSG, who possess clinical expertise with this condition (see list of contributors in Appendix B). Numerous members of the group then subjected the scale to several rounds of refinement with commentary. An attempt was made to establish content validity by having this large panel of RLS experts ensure that no significant aspect of RLS was omitted from the scale. This was balanced by the need to generate a scale sufficiently brief to permit use in a clinical or interview setting. Preliminary versions of the scale varied between 28 and six questions. The final scale is ten questions in length. The number of questions was reduced by the decision not to use questions in multiple formats to redundantly probe the same aspects of the disorder. It was determined that all questions should have a similar format and a similar polarity. Each question had a set of five response options graded from no RLS or impact (score = 0) to very severe RLS or impact (score = 4). This produced a total scale whose overall score could range from 0 to 40. During the development process, the scale was expanded to include all critical aspects of RLS designated by the expert group. The period of development took 18 months. Besides those who participated in the actual trial, many other centers and individuals contributed to the formation of the scale (see list of contributors in Appendix B). The final scale is reproduced in Appendix A.

Table 1
Features of RLS by IRLSSG consensus

Diagnostic features
(1) A desire to move the extremities usually associated with some definable discomfort
(2) Motor restlessness
(3) Worsening of symptoms at rest with at least temporary relief by activity
(4) Worsening of symptoms later in the day or at night

Associated features
(5) Involuntary movements awake and asleep (PLM)
(6) Sleep disturbance and its consequences
(7) Normal neurological examination in idiopathic cases
(8) Variable age of onset with typical chronic, progressive course
(9) Frequent familial history of cases

These features were recently refined on the basis of a consensus conference on RLS diagnosis held May 1–3, 2002 at the NIH.

2.2. Centers and subjects

A total of 20 centers from six countries (Germany, Ireland, Italy, Spain, Sweden, and the United States) were included in the study. RLS patients were recruited from 17 centers and control subjects from 14 centers. The total number of valid subjects recruited broken down into patient and control subjects, overall and by country, is shown in Table 2 together with their demographic information. Overall, there were 405 subjects, 196 RLS patients and 209 controls.

2.3. Subject recruitment: inclusion and exclusion criteria

RLS patients were required to have a diagnosis of RLS according to IRLSSG criteria [15]. Diagnoses were made by members of the IRLSSG involved in the study. Controls and patients were excluded if they could not complete the questionnaire for any reason, e.g. dementia or aphasia. Controls were excluded if they met the criteria for RLS or

[1] These criteria have recently been refined on the basis of a consensus conference held at the NIH (May 1–3, 2002). See this issue of Sleep Medicine. 'Allen RP, Hening WA, Montplaisir J, Picchietti D, Trenkwalder C, Watters AS. Restless Legs Syndrome: Diagnostic criteria, special considerations and epidemiology'. The new criteria delete criterion 2 on motor restlessness (Table 1), because it has been found to be confusing. In addition, criterion 3 is split into two separate criteria: provocation at rest and relief with activity. These changes should not have any impact on the design of the IRLS.

Table 2
Demographics of patient and control groups

	Patients	All controls	Normal controls	Sleep disorder controls
Total no.	196	209	110	99
US	89	103	48	55
Germany	41	32	22	10
Italy	35	49	27	22
Spain	10	12	0	12
Sweden	5	0	0	0
Ireland	16	13	13	0
Age (years)				
Mean (SD)	61.8 (11.9)	56.1 (15.0)	58.4 (15.0)	53.5 (14.7)
Range	34–90	22–91	39–91	22–88
Sex, *n* (%)				
Male	70 (36)	109 (52)	52 (47)	57 (58)
Female	126 (64)	100 (48)	58 (53)	42 (42)

had a history of neuroleptic exposure, neuroleptic-induced akathisia, peripheral neuropathy, radiculopathy or any other condition that could be confused with RLS. Patients with RLS were excluded if they had a history of neuroleptic exposure or neuroleptic-induced akathisia, but were not excluded if they had ‘secondary’ forms of RLS, i.e. RLS associated with peripheral neuropathy or radiculopathy. The control subjects were broken down into two groups: those with known or clinically suspected sleep disorders ($N = 99$) and those drawn from a normal population without known or suspected sleep disorders ($N = 110$).

2.4. Testing protocol

Prior to the study patients were asked to remain on stable dosages of RLS medications and any other medications known to affect the severity of RLS symptoms for 1 month prior to day 1 (first administrations of the rating scales) and for the 2 week interval between the two administrations of the rating scale. All records were reviewed to ensure that these conditions were met. In some cases, it was necessary for patients to have their medications changed or follow-up at the 2 week interval was not possible. We excluded patients from test–retest evaluations if their medications changed or the interval between the two tests was less than 12 or greater than 30 days.

On each of the testing days, patients were asked to rate themselves twice on the ten-question rating scale (see Appendix A) in the presence of different examiners. This duplicate rating was performed in order to determine whether differences in the responses, help, or instructions of the examiner might influence ratings. Each examiner was available throughout the entire time the patient was filling out the rating scale in order to explain the rating scale and to clarify any misunderstandings the patients might have regarding the questions on the scale. The protocol dictated two examiners at the first administration, but a second examiner was optional at the second administration. If two examiners were used, each remained blind to the answers given by the patient to the other examiner. The patients were also asked to give each examiner an overall rating of the severity of their symptoms over the course of the previous 2 weeks, ranging from 0 (no symptoms) to 8 (most severe) (patient global impression rating, PGI). A third, expert examiner was also asked to conduct a general analysis of patient symptoms and severity and to generate his or her own CGI of the severity of the patient’s symptoms. This was also scored on a scale from 0 (no symptoms) to 8 (most severe). The CGI was required by the protocol on the first day, but was optional for the second administration. This third examiner was also required to be blind to all answers given to the first two examiners by the patient. The coordinating center audited all records to be sure that these conditions were met. In some cases, where the protocol was not followed exactly (e.g. as to rater blinds), those scores were not used in analyses that required independent scores. In other cases, not every rating was completed. In that case, the patient was not included in analyses requiring, for example, two PGI scores. In all cases where patients were used for results related to the ten-question scale, answers to every question were available.

Controls had only a single administration of the ten-question rating scale on day 1 and were not asked to re-do the rating scale on another date. No PGIs or CGIs were generated for the controls.

2.5. Statistical analyses

For construct validity, we performed a factor analysis and examined item convergent validity. Prior to the factor extraction, we examined the dataset for the Kaiser–Meyer–Olkin (KMO) value to see if the dataset supported valid factor extraction. As a general rule, a KMO value greater than 0.6 is considered adequate for extraction and a KMO value greater than 0.9 is considered excellent [20,21]. The factor analysis was first performed on the average ratings obtained from the first administration ($N = 196$). To avoid spurious assignments of variance, we selected a principal factor extraction using the Kaiser criterion of accepting only those factors with an eigenvalue greater than 1 and also evaluating the factors with a scree plot, including use of the objective scree test [22]. In order to confirm the validity of this choice, we also explored other factor solutions stipulating multiple factor solutions using both varimax orthogonal and oblimin oblique solutions.

The factor results obtained from the first administration scores were then compared with those obtained from the second administration ($N = 187$) using the same procedures. First, we correlated the factor loads derived from the two sets of scores. Then, we calculated factor scores for the second administration using the factor score coefficient matrix generated for the first administration. We then correlated those scores with the factor scores we extracted directly from the second administration to determine the similarity of the separate factors extracted from the two administrations. We accepted as related to the factor all those questions which loaded at levels greater than 0.4.

Our reliability analysis consisted of examination of internal consistency, inter-examiner reliability, and test–retest reliability. For internal consistency, we performed a Cronbach alpha analysis [23]. We used a criterion of 0.7 to indicate adequate internal consistency [23,24]. For inter-examiner reliability, we used an intra-class correlation coefficient (ICC) equivalent to a weighted kappa analysis [25]. We have designated this an inter-examiner reliability test since the patients themselves provided the ratings, but did so in the presence of different examiners who might conduct the testing differently, provide different information, or simply influence the patients in different ways. We used the same statistic to compute a test–retest reliability. We used a criterion of an ICC of 0.7 as indicating a satisfactory performance [26]. We also compared scores between the first and second administrations using a paired *t*-test. Our hypothesis was that there would be no significant difference between the scores at the two time points. We set the significance level at 0.05.

Validity analysis consisted of criterion validity, concurrent validity, and discriminant validity. For criterion validity, we regressed the IRLS scores against the CGI. For concurrent validity we regressed the IRLS scores against the PG1. For discriminant validity, we performed a one-way ANOVA with three groups, patients and two types of controls (sleep disorder and normal controls). We then did post-hoc *t*-tests (Scheffé, Bonferroni) to locate any significant differences. Our hypothesis was that the ANOVA would show significant group differences and that the patients would be significantly different from either control group, but that the control groups would not differ from each other.

Because differences between the raters were so low (see Section 3.2.2), we averaged two ratings of a subject (IRLS scores, PGI) for all other analyses of the scores, where two ratings were available.

3. Results

3.1. Construct validity

3.1.1. Factor analysis

Because the KMO value was in the excellent range (0.908), we felt justified to proceed with the factor analysis of the first administration scores. Only one factor had an eigenvalue greater than 1 (6.28) using a principal factor extraction on the dataset ($N = 196$). The eigenvalue for the next factor was 0.88. The scree plot showed a clear break at the second factor. We therefore accepted a one factor solution which accounted for 59.2% of the variance. We therefore call this a general severity factor. All items except question 3 had factor loads in excess of 0.7 (Table 3).

Further exploration of multiple factor solutions indicated that, with rotation, there emerged two separate factors with primary loading on symptom measures (questions 1, 2, 4, 6, 7, and 8) and disease impact measures (questions 5, 9, and 10). However, there was considerable overlap between factors with variables contributing to one factor having significant contributions to the other factors (loadings >0.4). It was therefore concluded that the scale was truly unified around one very strong factor and that the addition of another factor only partially teased apart highly related variables. The exploratory analysis also indicated that the two sleep items (questions 4 and 5) and the two symptom prevalence measures (questions 7 and 8) were highly related to each other. Question 3 did not load well on either of the factors.

For the second administration ($N = 187$; nine subjects lost to follow-up between administrations, KMO = 0.920), the general severity factor was also seen with an eigenvalue of 6.88 and accounting for 65.0% of the variance. As in the first administration, all items except question 3 had factor loads in excess of 0.7 (Table 3). Further analysis also showed that a two factor solution broke down into two factors representing symptom measures and disease impact measures with much overlap. The most distinct items for each factor (symptom measures: questions 1, 2, and 6; impact measures, questions 5,

Table 3
Loadings on general severity factor for two administrations

Question number	Administration	
	First	Second
1	0.872	0.897
2	0.821	0.845
3	0.444	0.572
4	0.799	0.809
5	0.738	0.810
6	0.924	0.931
7	0.723	0.719
8	0.809	0.851
9	0.739	0.831
10	0.726	0.776

Single factor solution – principal factor extraction. Loads for two administrations correlate 0.961 ($P < 0.001$).

9, and 10) were common to the two administrations, but even these showed more than de minimus loads on the other factor (all >0.29). The overall similarity of the weightings for the two administration factor extractions can be seen from the tabulated weightings (Table 3). In fact, the weights correlate 0.961 ($P < 0.001$).

We also further explored the similarity of the factors extracted from the two administrations. We calculated factor scores using data from the second administration with the factor score coefficient matrix from the first administration. The resulting scores correlated 0.982 ($P < 0.001$) with the factor scores directly extracted from the second administration, indicating an almost complete similarity of the factors extracted from the two administrations. This finding indicates that there is very little difference in the factor structure of the scores on the two administrations.

3.1.2. Item convergent validity

Correlations between individual items (questions 1 through 10) and the total score of the questionnaire (minus that item) were always significant and positive. Except for question 3 correlations for the individual items with the total score varied between 0.69 and 0.90 (day 1, $N = 196$; day 2, $N = 187$). Items 1 and 6 had the highest correlations on both days (day 1, 0.83, 0.88; day 2, 0.87, 0.90) ($P < 0.001$), while item 3 (response to movement) had the lowest (day 1, 0.43; day 2, 0.56) ($P < 0.01$). It is usually accepted that item convergent validities above 0.4 are acceptable for rating scales [27].

3.2. Reliability analyses

3.2.1. Internal consistency

Cronbach alpha measures for the two administrations were 0.93 ($N = 196$) and 0.95 ($N = 187$), respectively ($P < 0.001$). There was minimal change in this value when each question was selectively removed. The only question whose exclusion increased the alpha value was question 3, concerning relief with walking.

3.2.2. Inter-examiner reliability

Inter-examiner reliability was measured by ICC, equivalent to a weighted kappa analysis [28]. Subjects were accepted into this analysis if they had two ratings which could be ordered in such a way that the scores for all subjects could be divided into two distinct sets of raters (that is, the same rater was not found in both sets of scores). For the summed rating scale, the ICC was 0.93 for the first administration ($n = 187$ patients) and 0.97 for the second administration ($n = 169$ patients) ($P < 0.001$). Considering the individual questions, ICCs for the individual questions ranged between 0.68 and 0.93 for the first administration and between 0.78 and 0.96 for the second (all $P < 0.01$). For both days, the lowest reliability was for question 3 (relief with walking) and the highest was for question 7 (frequency of symptoms, days per week). For the PGI, the ICCs were 0.95 ($N = 155$) and 0.94 ($N = 130$) respectively ($P < 0.001$).

Because of the extremely high inter-rater reliability, the two scores (either on the rating scale or PGI) were combined for further analyses. In cases where only a single score was available, that single score was used.

3.2.3. Test–retest reliability

A total of 145 patients met criteria for our test–retest evaluation within the 12–30 day window. Of the remaining subjects, 15 returned too soon for retest (2–11 days), five returned too late (34–104 days), nine did not return at all, and 22 were not on constant medications. For those who met the criteria, the mean period between testing was 15.0 days (SD 4.1). The ICC for these patients' scores was 0.87 ($P < 0.001$). Mean scores went from 21.37 (SD 8.35) to 20.88 (SD 8.89) between the two administrations. This difference was not significant ($t = 1.35$, d.f. $= 144$, paired t-test, $P > 0.05$). The mean difference in scores was -0.49 (SD 4.37).

3.3. Validity analyses

3.3.1. Criterion validity

We assessed criterion validity by determining the correlation of the summed questionnaire score to the CGI. This yielded an r value of 0.74 on day 1 and 0.73 on day 2 ($P < 0.001$). The relationship for both days is plotted in Fig. 1. It is evident from these graphs that the intercepts are near zero and that there is a tendency for the most extreme values (low or high) of either measurement to be associated with more moderate values of the other.

To examine this relationship further, we performed an ANOVA after dividing the subjects into groups based on their IRLS score, using a priori designations that mirrored the proposed designations for levels of CGI (mild: scores from 0 to 10; moderate: 11 to 20; severe: 21 to 30; very severe: 31 to 40). An ANOVA using these groups with the CGI as the dependent variable found that the F values for the two administrations were 57.42 and 47.90, respectively ($P < 0.001$), for those subjects with an independently rated CGI. Post-hoc tests all revealed that for the first administration there was a significant difference in CGI scores for all comparisons of the four IRLS levels, while for the second administration the same was true except that the difference between the CGI scores for the severe and very severe IRLS levels was not significant. Scores are given in Table 4 for those subjects with both tests available and independent CGI. As can be seen, the proposed severity levels for the IRLS summed score in general correspond to the same anchored levels of the CGI, except that the mean CGI score for the very severe IRLS level falls into the severe range rather than the very severe range.

The mean CGI for all subjects independently rated ($N = 182$, 153) was 4.03 on day 1 and 3.84 on day 2, also near the middle of the scale of 4. Standard deviations were 1.98 and 2.04, respectively. The distribution of scores on the

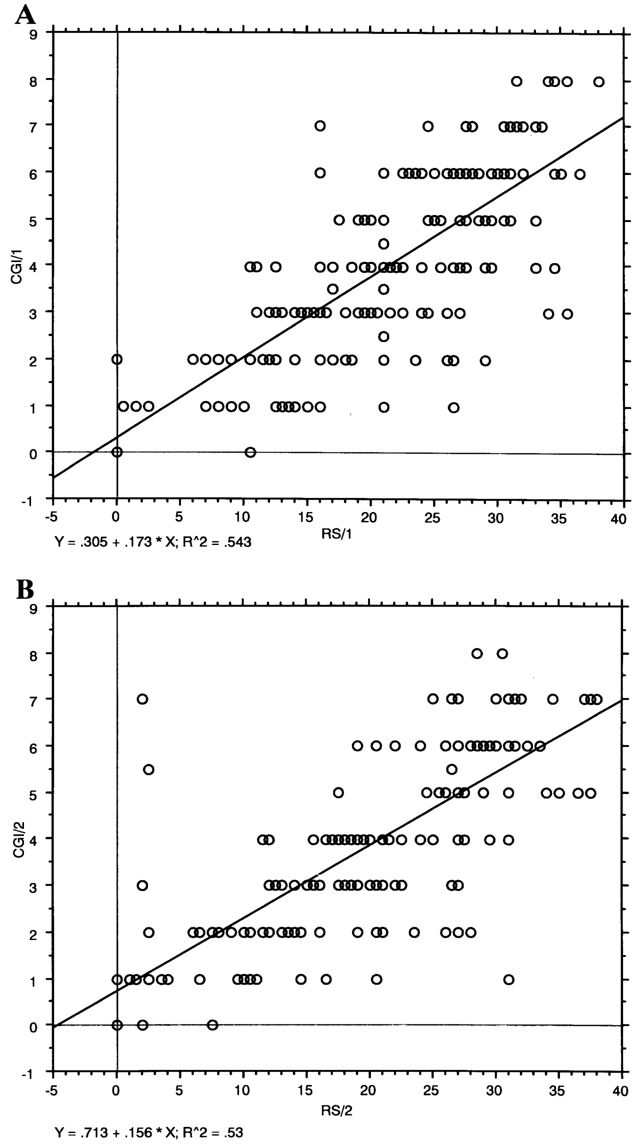

Fig. 1. The ratings assigned by the independent expert raters (CGI) are plotted against the averaged rating scale summed scores for individual subjects for the first (A) and second (B) administrations of the rating scale for all cases with independent ratings. RS/1, averaged rating scale sum, first administration; RS/2, second administration; CGI/1, clinical global rating first administration; CGI/2, second administration.

CGI is given in Table 5. There was a fairly even distribution of scores between 1 and 7 while few patients were scored as 8, most severe. A couple of patients who were without current complaint were scored as 0.

3.3.2. Concurrent validity

Concurrent validity was also examined by comparing the IRLS summed score to the PGI. This correlation was 0.82 on day 1 and 0.78 on day 2 (for both, $P < 0.001$). In an additional analysis, we correlated the PGI on the CGI. This yielded an r value of 0.80 on day 1 and 0.84 on day 2 ($P < 0.001$).

Table 4
Distribution of CGI scores by proposed severity levels of IRLS

IRLS level	N	Mean CGI	SD CGI
First administration			
Mild	22	1.77	1.27
Moderate	58	3.07	1.46
Severe	81	4.81	1.61
Very severe	29	6.07	1.46
Second administration			
Mild	29	1.64	1.47
Moderate	54	3.30	1.38
Severe	56	5.06	1.68
Very severe	22	5.82	1.44

IRLS, IRLS summed score levels: mild, 0–10; moderate, 11–20; severe, 21–30; very severe, 31–40. CGI – clinical global impression: 0, asymptomatic; 1–2, mild; 3–4, moderate; 5–6, severe; 7–8, very severe.

3.3.3. Discriminant validity

The majority of the controls had scores of zero on the IRLS. For those controls recruited in centers also recruiting patients, there were four non-zero scores in controls drawn from the normal population (normal controls, 4/105 non-zero or 4%) and 13 non-zero scores in controls with a sleep disorder (sleep disorder controls, 13/71 or 18%). Overall, 17 of 176 controls from these centers had non-zero scores (9.7%). Five additional normal controls and 28 sleep disorder controls were recruited at centers that did not recruit patients: all of these controls had zero scores. For either control group or all controls, the median and mode were both zero.

A one-way ANOVA found a highly significant F ratio for the main factor of group ($F = 577.0$, d.f. $= 2$, $P < 0.01$). Post-hoc tests between the patients and both the normal control subjects and the sleep disordered control subjects showed that the patients had significantly higher scores than either group of controls (both $P < 0.001$), but that there was no significant difference between the two control groups.

3.4. Distributions of scores

3.4.1. Total rating scale score

The mean averaged summed scores for the two

Table 5
Distribution of CGI scores

CGI level	Administration	
	First	Second
0	2	3
1	18	19
2	28	25
3	28	25
4	32	25
5	22	16
6	30	20
7	17	18
8	5	2
Total	182	153

administrations for all subjects were 21.91 and 20.27, near the center of the possible range of scores. Standard deviations were 8.39 and 9.24, respectively, and the full range for both administrations was from 0 to 38. Distributions of scores for the RLS patients are plotted in Fig. 2. It is apparent that the scores are rather evenly distributed from 12 to 32 with a longer tail towards the lower scores. Median values for the two administrations were 23.25 and 20.5, respectively, while modal values for scores grouped into intervals of 4 were in the intervals 24–27 and 28–31, respectively (Fig. 2).

3.4.2. Global impressions

The mean PGIs for all subjects with at least one rating ($N = 188$, 179) were 4.27 and 4.12, near the middle score of 4. Standard deviations were 2.01 and 2.08, respectively.

4. Discussion

4.1. Summary of results

All the reliability and validity analyses revealed highly significant results that met or exceeded minimum quality standards for an instrument of this kind. Internal consistency revealed that, with the possible exception of question 3, this scale was very highly unified, a conclusion supported by the emergence of a single strong factor with highly significant loadings from each question except 3. This factor can be termed a severity factor, and notably draws strong support not only from primary measures of symptom severity (questions 1, 2, and 6) and intensity/frequency (questions 7 and 8) but also from those which related to impact on sleep (questions 4 and 5) and impact on mood and daily functions (questions 9 and 10). Within each group of questions there were tighter relations than across groups, but the overlap among groups was so high that the one factor solution was the optimal one. Similar trends were manifest in the high degree of convergent validity found. Such a result argues that RLS may be a relatively unified condition in which the severity of diagnostic symptoms largely determines the impact on the patients. This conclusion is also supported by the indication in circadian studies that all features of the condition tend to co-vary with the circadian cycle [29,30] and that, in therapeutic trials, subjective and objective measures usually indicate a similar result [5]. This conclusion will need to be explored in other studies. The degree of consistency observed in the rating scale, however, may be greater than optimal [31]. As a result, it may be feasible either to reduce the number of questions or to add items that deal with other aspects of RLS symptoms and impact.

Inter-examiner reliability was very high (0.93 and 0.97) and suggested that there should be no difficulty in having this scale administered by diverse raters. Test–retest reliability measured on patients with a consistent medication profile and at intervals up to 30 days revealed that the rating scale scores

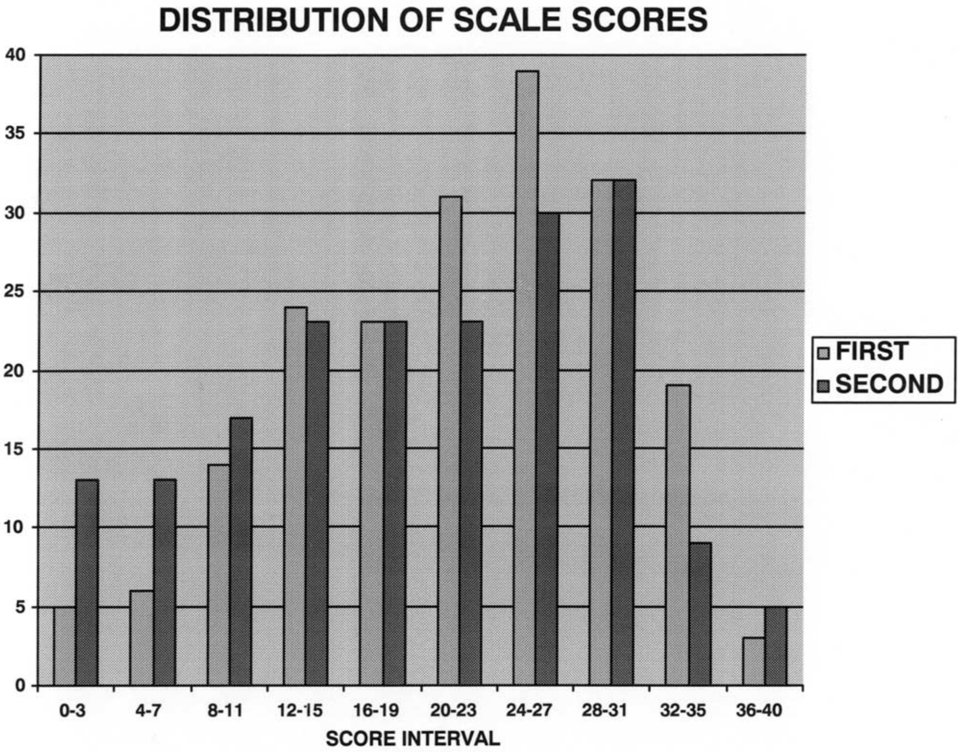

Fig. 2. The averaged rating scale summed scores are displayed in 4 point bins for the first and second administrations. The data include scores from all RLS patients with valid scores ($N = 196$, first administration; $N = 187$, second administration).

were quite stable over this time period. This was also true of the CGI, currently the most generally accepted means of assessing RLS severity.

In criterion validity, the rating scale score performed well in comparison to the criterion of the CGI, with correlations of 0.74 and 0.73 ($P < 0.001$) (Fig. 1). In discriminant validity, it was clear that control groups, even those with sleep disorders that might cause overlapping symptoms, had mostly zero scores, dramatically different from the RLS patients. This result is probably explained in large part by the fact that all the questions were anchored by a reference to RLS.

The analysis of the rating scale's distribution showed that it had most scores in a range that corresponded to moderate to severe CGI ratings (Fig. 1). The scale scores showed a good spread of values from zero through the highest values of the scale (>30). This should allow for adequate discrimination of patients with a wide range of severities.

4.2. Utility of scale

Because the scale is brief and apparently posed few problems to any of the patients, it offers the possibility of ready use in clinical practice, in epidemiological and pathophysiological research, and in clinical trials. It is notable that, under the conditions of administration, all but one questionnaire was completed with all questions answered. However, since the current study had the scale completed in the presence of a knowledgeable professional, it is not clear from these results that the scale, if administered by a telephone canvasser or in mailed questionnaires, would perform equally well without such a professional being available. However, the very high inter-examiner reliability and strong test–retest stability suggest that it will be useful in situations where scores need to be obtained by diverse individuals. The results also suggest that changes in the scale are likely to reflect true changes in the underlying condition. Because it was well correlated to CGI, the scale is validated as a satisfactory instrument for use without contribution from a sustained, expert clinical interview.

Since the patients examined had by and large been treated and were on a variety of medications, the scale should be useful for assessment not only of untreated patients, but of those who are on different medication regimens. This study was conducted on the typical patient populations regularly seen in RLS centers.

4.3. Comparison to other measures

This rating scale has an advantage over other measures since it has been subjected to intensive evaluation of its reliability and aspects of validity. Unlike objective measures of RLS, it can be easily and effectively applied to all patients. However, it does not examine all aspects of RLS. The Johns Hopkins RLS severity scale (JHRLSS) takes a different approach, examining severity by time of day of onset of symptoms. That scale has been validated against objective measures of RLS such as sleep efficiency and PLM index [32]. It has also proven useful in correlating severity to biological measures such as serum ferritin [33] or brain iron in the substantia nigra [34]. However, that scale, while complementing the IRLS, does not cover as many aspects of the RLS condition. Other areas not covered by either scale include the number of involved limbs or the rapidity with which symptoms develop when a patient first sits or lies down.

4.4. Future requirements and prospects

In further work, it will be necessary to establish the relationship of the IRLS to such objective measures of RLS as sleep efficiency and PLM indices. Some of this work is currently under way: in a large parallel double blind placebo/drug trial, changes in the scale were found to be significantly related to changes in PLM indices, sleep efficiency, and CGI (Trenkwalder, personal communication). This suggests that the scale is sensitive to changes in or manipulations of the severity of RLS as is expected in clinical trials. The scale should be able to discriminate between different levels of RLS severity at different time points within the same individual.

We have considered whether elimination of question 3 (relief with walking) from the scale would benefit its psychometric properties. While the scale more than meets all performance standards with this question included, it is the one question that repeatedly stands out as less related to the overall scale or the remainder of the scale items. It does not contribute at a high level to the main factor, or even at lower levels to any multiple factor rotated solution. This may be due to its answers having a somewhat different format (Appendix A) or to the fact that almost all patients experience significant relief with walking, a possibility supported by the low mean scores for this question and the minimal standard deviation. However, several factors mitigate against removing the item from the scale. First, it measures relief with walking, a key diagnostic feature of RLS. Second, it meets threshold standards for good performance. Third, the scale, as presently constructed, is more coherent than recommended [31]. If question 3 were eliminated, this coherence would only increase. The authors are aware of ongoing studies using the scale in other contexts, and if question 3 is repeatedly found to suffer from these deficits, particularly an insensitivity to changed clinical status within therapeutic trials, future editions may decide that it should be eliminated.

Another question for future study is whether either a shorter or a longer scale would be equally useful. By eliminating some questions, a shorter scale with comparable psychometric properties but a lesser degree of coherence might be achieved. A longer scale could incorporate additional aspects of the condition (such as time of day of symptom onset or rapidity of symptom development at rest) and include additional quality of life measures. Such an extended scale might better capture disease impact for purposes of clinical evaluation or measurement of therapeutic response. It is also possible that in the future a more complex scale or different scales aimed at different aspects of RLS (e.g. symptom severity versus disease impact or quality of life) will prove more useful for different evaluative contexts.

The scale is also flexible and can be used, with minor modification, for either shorter (e.g. 1 week) or longer (e.g. 1 month) periods of assessment. While the IRLS was administered with an examiner present in this study, in the future it may prove possible to have patients fill out the scales by themselves or be queried over the telephone to expand the contexts in which the scale may be used.

Acknowledgements

We wish to thank Mark Atkinson, Linda Hirsch, David Streiner, and Barbara Tabachnick for providing statistical consultations during the analysis of data and development of the manuscript.

Appendix A. IRLSSG restless legs syndrome rating scale for severity (IRLSSGRS)

This scale is copyrighted by the Interntional Restless Legs Syndrome Study Group 2002 and this version IS NOT TO BE USED OR DISTRIBUTED. A slightly modified version of the scale that is re-worded for better clarity is presented in an accompanying editorial in this issue of Sleep Medicine. The English version of the modified scale and translation into other languages can be obtained through "Caroline Anfray, Information Resources Centre, MAPI Research Institute, 27 rue de la Villette, 69003 Lyon, France. Phone + 33(0) 472 13 66 67. FAX + 33 (0) 472 13 66 82. E-mail canfray@mapi.fr or instdoc@mapi.fr".

Rate your symptoms for the following ten questions. Unless otherwise instructed, you should rate the average symptoms that you have experienced for the most recent two week period.

(1) Overall, how would you rate the RLS discomfort in your legs or arms?

(4) Very severe
(3) Severe
(2) Moderate
(1) Mild
(0) None

(2) Overall, how would you rate the need to move around because of your RLS symptoms?

(4) Very severe
(3) Severe
(2) Moderate
(1) Mild
(0) None

(3) Overall, how much relief of your RLS arm or leg discomfort do you get from moving around?

(4) No relief
(3) Slight relief
(2) Moderate relief
(1) Either complete or almost complete relief
(0) No RLS symptoms and therefore question does not apply

(4) Overall, how severe is your sleep disturbance from your RLS symptoms?

(4) Very severe
(3) Severe
(2) Moderate
(1) Mild
(0) None

(5) How severe is your tiredness or sleepiness from your RLS symptoms?

(4) Very severe
(3) Severe
(2) Moderate
(1) Mild
(0) None

(6) Overall, how severe is your RLS as a whole?

(4) Very severe
(3) Severe
(2) Moderate
(1) Mild
(0) None

(7) How often do you get RLS symptoms?

(4) Very severe (This means 6 to 7 days a week)
(3) Severe (This means 4 to 5 days a week)
(2) Moderate (This means 2 to 3 days a week)
(1) Mild (This means 1 day a week or less)
(0) None

(8) When you have RLS symptoms how severe are they on an average day?

(4) Very severe (This means 8 hours per 24 hour day or more)
(3) Severe (This means 3 to 8 hours per 24 hour day)
(2) Moderate (This means 1 to 3 hours per 24 hour day)

(1) Mild (This means less than 1 hour per 24 hour day)
(0) None

(9) Overall, how severe is the impact of your RLS symptoms on your ability to carry out your daily affairs, for example carrying out a satisfactory family, home, social, school or work life?

(4) Very severe
(3) Severe
(2) Moderate
(1) Mild
(0) None

(10) How severe is your mood disturbance from your RLS symptoms – for example angry, depressed, sad, anxious or irritable?

(4) Very severe
(3) Severe
(2) Moderate
(1) Mild
(0) None

Appendix B

B.1. Writing, central data collection and data analysis committee

New Jersey Neuroscience Institute at JFK Medical Center, Edison, NJ: Arthur S. Walters, MD, Cheryl LeBrocq, Anjana Dhar, MD; UMDNJ-Robert Wood Johnson Medical School, New Brunswick, NJ: Arthur S. Walters, MD, Wayne Hening, MD, PhD, Ray Rosen, PhD; Seton Hall University School of Graduate Medical Education, South Orange, NJ: Arthur S. Walters, MD; Department of Neurology, Johns Hopkins University, Baltimore, MD: Wayne Hening, MD, PhD, Richard P. Allen, PhD; Center for Molecular and Behavioral Neuroscience, Rutgers University, Piscataway, NJ: Wayne Hening, MD, PhD; Max Planck Institute of Psychiatry, Munich, Germany: Claudia Trenkwalder, MD; Department of Clinical Neurophysiology, University of Goettingen, Goettingen, Germany: Claudia Trenkwalder, MD.

B.2. Members of the group and other contributors

Parkinson's Disease and Movement Disorders Center, Mayo Clinic, Scottsdale, AZ: Charles Adler*, Stephanie Newman, Cynthia Reiners.

Department of Neurology, Erciyes University Medical Faculty, Kayseri, Turkey: Murat Aksu.

Department of Neurology, Johns Hopkins University, Baltimore, MD: Richard P. Allen*, David Buchholz*, Wayne A. Hening*.

Sleep Disorders Center, St. Joseph Hospital, Orange, CA: Melanie Anderson, Sarah Mosko*.

Department of Psychiatry, University of California, San Diego, CA: Sonia Ancoli-Israel*.

National Institute of Neurological Disease and Stroke, Bethesda, MD: William Bara Jimenez*, Mark Hallett*.

Neurologische Poliklinik, Universitatsspital Zurich, Zurich, Switzerland: Claudio Bassetti*, Sandra Clavadetscher.

Department of Neurology, Emory University, Atlanta, GA: Donald L. Bliwise*, Paul Gurecki, David B. Rye*.

Sleep Wake Disorders Center of the New York-Presbyterian Hospital, New York, NY: Lauren L. Broch, Rochelle Zak*.

St. Vincent's Medical Center, New York Medical College, New York, NY: Sudhansu Chokroverty*.

Department of Neurological Sciences, University of Bologna, Bologna, Italy: Giorgio Coccagna*, Elio Lugaresi*, Filomena Miele, Pasquale Montagna*, Giuseppe Plazzi*, Federica Provini*.

Psicobiologia Universidade Federal de Sao Paulo, Sao Paulo, Brazil: Marco Tulio de Mello*, Sergio Tufik.

Centre for Sleep and Wake Disorders, Westeinde Hospital, The Hague, The Netherlands: Al W. de Weerd*, Roselyne M. Rijsman*.

New Jersey Neuroscience Institute at JFK Medical Center, Edison, NJ: Anjana Dhar, Cheryl LeBrocq, Arthur S. Walters*.

Department of Neurology, Tufts New England Medical Center, Boston, MA: Bruce Ehrenberg*.

Department of Neurology, Ludwig Maximilians Universitat, Munich, Germany: Ilonka Eisensehr*.

Department of Neurology, Huddinge University Hospital, Huddinge, Sweden: Karl Ekbom Jr.*, Ake Ljungdahl.

Department of Neurology, Fundacion Jimenez Diaz, Madrid, Spain: Diego Garcia-Borreguero*, Oscar Larrosa.

UMDNJ-Robert Wood Johnson Medical School, New Brunswick, NJ: Wayne A. Hening*, Ray Rosen*, Arthur S. Walters*.

Center for Molecular and Behavioral Neuroscience, Rutgers University, Piscataway, NJ: Wayne A. Hening*, Linda Hirsch.

Universitatsklinik fur Neurologie, Innsbruck, Austria: Birgit Hogl*.

Department of Psychiatry, Shimane Medical University, Izumo City, Japan: Jun Horiguchi*.

Department of Psychiatry and Psychotherapy, Albert-Ludwigs-University, Freiburg, Germany: Magdolna Hornyak*, Ulrich Voderholzer*.

Sleep Disorders Center, St. Boniface Hospital-Research Center, Winnipeg, Manitoba, Canada: Meir Kryger*, Robert Skomrow.

Clinical Neuroscience Research Foundation, Concord, MA: Joseph F. Lipinski*.

Department of Pulmonary and Critical Care Medicine, University of Kentucky Medical Center, Lexington, KY: Ahmed Masood, Barbara Phillips*.

Department of Neurology, Philipps University, Marburg, Germany: Wolfgang H. Oertel*, Karin Stiasny*.

St. Michael's Hospital, Dublin, Ireland: Shaun O'Keeffe*.

Sleep Disorders Center, San Raffaele Scientific Institute, Milan, Italy: Alessandro Oldani, Marco Zucconi*.

Department of Neurology, Baylor College of Medicine, Houston, TX: William G. Ondo*.

Carle Clinic, University of Illinois, Champaign-Urbana, IL: Daniel Picchietti*.

Division of Neurology, Scripps Clinic, La Jolla, CA: J. Steven Poceta*.

Department of Neurology, Pacific Sleep Program, Portland, OR: Gerald B. Rich*.

Sleep Alertness Center, Aurora, CO: Larry Scrima*.

San Diego Sleep Disorders Center, San Diego, CA: Renata Shafor*.

Tulane University Hospital and Clinic, New Orleans, LA: Denise Sharon*.

Mayo Clinic, Rochester, MN: Michael Silber*.

Department of Medicine, Michigan State University, East Lansing, MI: Robert Smith*.

Max Planck Institute of Psychiatry, Munich, Germany: Claudia Trenkwalder*, Thomas C. Wetter*, Juliane Winkelmann*.

Department of Clinical Neurophysiology, University of Goettingen, Goettingen, Germany: Claudia Trenkwalder*.

Department of Neurology, UCLA School of Medicine, Los Angeles, CA: Zeba Vanek.

Department of Pharmacy Practice, Rutgers University, Piscataway, NJ: Mary Wagner.

Seton Hall University School of Graduate Medical Education, South Orange, NJ: Arthur S. Walters.

*Indicates member of the International Restless Legs Syndrome Study Group.

References

[1] Lavigne GJ, Montplaisir JY. Restless legs syndrome and sleep bruxism: prevalence and association among Canadians. Sleep 1994; 17(8):739–43.

[2] Phillips B, Young T, Finn L, et al. Epidemiology of restless legs symptoms in adults. Arch Intern Med 2000;160(14):2137–41.

[3] Rothdach AJ, Trenkwalder C, Haberstock J, et al. Prevalence and risk factors of RLS in an elderly population: the MEMO study. Memory and Morbidity in Augsburg Elderly. Neurology 2000;54(5):1064–8.

[4] Chesson Jr. AL, Wise M, Davila D, et al. Practice parameters for the treatment of restless legs syndrome and periodic limb movement disorder. An American Academy of Sleep Medicine Report. Standards of Practice Committee of the American Academy of Sleep Medicine. Sleep 1999;22(7):961–8.

[5] Hening W, Allen R, Earley C, et al. The treatment of restless legs syndrome and periodic limb movement disorder. An American Academy of Sleep Medicine Review. Sleep 1999;22(7):970–99.

[6] O'Keeffe ST, Gavin K, Lavan JN. Iron status and restless legs syndrome in the elderly. Age Ageing 1994;23(3):200–3.

[7] Wagner ML, Walters AS, Coleman RG, et al. Randomized, double-blind, placebo-controlled study of clonidine in restless legs syndrome. Sleep 1996;19(1):52–8.

[8] Earley CJ, Yaffee JB, Allen RP. Randomized, double-blind, placebo-controlled trial of pergolide in restless legs syndrome. Neurology 1998;51(6):1599–602.

[9] Wetter TC, Stiasny K, Winkelmann J, et al. A randomized controlled study of pergolide in patients with restless legs syndrome (see comments). Neurology 1999;52(5):944–50.

[10] Walters AS, Wagner ML, Hening WA, et al. Successful treatment of the idiopathic restless legs syndrome in a randomized double-blind trial of oxycodone versus placebo. Sleep 1993;16(4):327–32.

[11] Montplaisir J, Nicolas A, Denesle R, Gomez-Mancilla B. Restless legs syndrome improved by pramipexole: a double-blind randomized trial. Neurology 1999;52(5):938–43.

[12] Stiasny K, Robbecke J, Schuler P, Oertel WH. Treatment of idiopathic restless legs syndrome (RLS) with the D2-agonist cabergoline – an open clinical trial. Sleep 2000;23(3):349–54.

[13] Montplaisir J, Boucher S, Poirier G, et al. Clinical, polysomnographic, and genetic characteristics of restless legs syndrome: a study of 133 patients diagnosed with new standard criteria. Mov Disord 1997; 12(1):61–5.

[14] Nicolas A, Michaud M, Lavigne G, Montplaisir J. The influence of sex, age and sleep/wake state on characteristics of periodic leg movements in restless legs syndrome patients. Clin Neurophysiol 1999;110(7):1168–74.

[15] The International Restless Legs Syndrome Study Group (Arthur S. Walters MD – Group Organizer and Correspondent), Towards a better definition of the restless legs syndrome. Mov Disord 1995;10: 634–42.

[16] Becker PM, Ondo W, Sharon D. Encouraging initial response of restless legs syndrome to pramipexole (see comments). Neurology 1998;51(4):1221–3.

[17] Ondo W. Ropinirole for restless legs syndrome (see comments). Mov Disord 1999;14(1):138–40.

[18] Trenkwalder C, Brandenburg U, Hundemer H-P, et al. A randomized long-term placebo-controlled multicenter trial of pergolide in the treatment of RLS – the PEARLS-Study (abstract). Neurology 2001; 56(Suppl 3):A5.

[19] Hening WA, Walters AS, Rosen R, et al. The international RLS study group rating scale: a reliable and valid instrument for assessing severity of the restless legs syndrome. Neurology 2001;56(Suppl 3): A4.

[20] Kaiser HF, Rice J. Little Jiffy Mark IV. Educ Psychol Measure 1974; 34:111–7.

[21] Sharma S. Applied multivariate techniques. New York: Wiley; 1996.

[22] Gorsuch RL. Factor analysis. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.

[23] Cronbach LJP. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

[24] Nunnally JCJ. Psychometric theory, 2nd ed. New York: McGraw-Hill; 1978.

[25] Guyatt GH, Walter S, Norman GJCD. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40(2):171–8.

[26] Hays R, Anderson R, Reviki DA. Assessing reliability and validity of measurement in clinical trials. In: Staquet MJ, Hays RD, Fayers PM, editors. Quality of life assessment in clinical trials. New York: Oxford University Press; 1998.

[27] Campbell DT, Fiske JL. Convergent and discriminant validation by the multitrait multimethod matrix. Psychol Bull 1959;56:85–105.

[28] Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Measure 1973;33:613–9.

[29] Trenkwalder C, Hening WA, Walters AS, et al. Circadian rhythm of periodic limb movements and sensory symptoms of restless legs syndrome. Mov Disord 1999;14(1):102–10.

[30] Hening WA, Walters AS, Wagner M, et al. Circadian rhythm of motor restlessness and sensory symptoms in the idiopathic restless legs syndrome. Sleep 1999;22(7):901–12.

[31] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use, 2nd ed. New York: Oxford University Press; 1995.

[32] Allen RP, Earley CJ. Validation of the Johns Hopkins restless legs severity scale. Sleep Med 2001;2(3):239–42.

[33] Sun ER, Chen CA, Ho G, et al. Iron and the restless legs syndrome. Sleep 1998;21(4):371–7.

[34] Allen RP, Barker PB, Wehrl F, et al. MRI measurement of brain iron in patients with restless legs syndrome. Neurology 2001;56(2):263–5.