pISSN 1976-8753 · eISSN 2383-9910
Brain & NeuroRehabilitation Vol. 9, No. 1, March, 2016
http://dx.doi.org/10.12786/bn.2016.9.1.31

Original Article

백서의 신경학적 결손에 대한 평가도구의 검사자간 및 검사자내 신뢰도 분석

[1]서울대학교 의과대학, 서울대학교병원 재활의학과, [2]경북대학교병원 재활의학과

박지홍[1] · 오병모[1] · 민유선[2] · 방문석[1] · 한태륜[1]

The Intra- and Inter-rater Reliability and the Learning Curve for a Simple Neurological Score for Rats

Ji Hong Park, M.D.[1], Byung-Mo Oh, M.D., Ph.D.[1], Yusun Min, M.D.[2], Moon Suk Bang, M.D., Ph.D.[1], Tai Ryoon Han, M.D., Ph.D.[1]

[1]Department of Rehabilitation Medicine, Seoul National University College of Medicine, Seoul National University Hospital, [2]Department of Rehabilitation Medicine, Kyungpook National University Medical Center

Objective: To measure the intra- and inter-rater reliability of a simple sensorimotor performance test for rats, and to evaluate the learning efficiency of a novice rater for the test.

Method: Middle cerebral arteries were occluded by intraluminal sutures in 25 male Sprague-Dawley rats (10～12 weeks old). The sensorimotor performance test was performed by a novice and an experienced rater, with each rater performing the test twice each week for 3 consecutive weeks. A ten-minute standardized video about the rating method was shown to the novice rater after the second test each week.

Results: The intra- and inter-rater agreement was determined using Cohen's weighted kappa coefficient. The intra-rater reliability was initially poor for the novice (κ [95% confidence interval], 0.31[－0.02, 0.64]), but it improved significantly after 3-week self education using the standardized video (0.81 [0.69, 0.93], showing almost perfect agreement. The reliability of the experienced researcher was good at all times (κ =0.64, 0.76, 0.71, for week 1, 2, 3, respectively), indicating substantial agreement. The inter-rater reliability showed clear improvement after self-education (κ = 0.44, 0.69, 0.69, for week 1, 2, 3, respectively). Although the total sum score was highly reliable, some of the individual items showed lower intra-and inter-rater agreement. However, each rater showed greater within-rater variability for different subtests.

Conclusion: The simple sensorimotor performance test showed high degree of intra- and inter-rater agreement when performed by experienced or properly educated raters. The inaccuracy of the novice was rectified by 3-week self-education using a video. **(Brain & NeuroRehabilitation 2016; 9: 31-36)**

Key Words: Learning curve, Motor activity, Behavioral research, Reproducibility of results

Introduction

Both practicality and reliability are essential prerequisites of outcome measures in neuroscience research using laboratory animals. Although the volume of infarct tissue or histological changes such as the number of specific cells and optical density of tissue markers have been frequently used, tracking of these changes longitudinally within the same brain is limited. Therefore, neurological functional tests still serve as the valuable adjuncts even in this golden era of molecular biology.

Received: October 12, 2015, Revised: December 5, 2015,
Accepted: March 27, 2016
Correspondence to: Byung-Mo Oh, Department of Rehabilitation Medicine Seoul National University Hospital, 101 Daehang-ro, Jongno-gu, Seoul 03080, Korea
Tel: 02-2072-2619, Fax: 02-743-7473
E-mail: keepwiz@gmail.com
This work was supported by a grant of the Korea Healthcare technology R&D Project, Ministry for Health, Welfare & Family Affairs, Republic of Korea (A091281).
The authors thank Ms. Jin-Joo Li, a senior researcher working in the Department of Rehabilitation Medicine in Clinical Research Institute, Seoul National University Hospital, for joining this study as the expert, sharing her experience, and helping us make a standardized video.

Garcia et al. introduced a simple test to estimate the extent of brain damage by scoring the sensory and motor function.[1] The sum of the 6 test scores in a Wistar rat with middle cerebral artery occlusion (MCAo) was correlated with the degree of the histological deficit, which enabled easier longitudinal assessment of intra-vital changes. It is indisputably evident that the value of a method that assess neurological deficit can be greatly affected by the variability among raters.[2] Fortunately, the total scores of the Garcia neurological test (GNT) showed high inter-rater agreement in rats with MCAo.[3] In that study, the value of the weighted kappa was 0.79, which means 'substantial' agreement.[4]

However, there is a paucity of literature on the intra-rater variability of GNT and there are no guidelines about how much training is needed for novice researchers to conduct GNT reliably. In this study, we assessed the intra- and inter-rater agreement for a novice and an expert, with the novice self-learning using a standardized video. We also conducted statistical analysis on six individual subtests to compare inter-and intra-rater reliability for each subtest. In addition, we investigated how much time was needed for the trainees to obtain professional competence by drawing the learning curve of the novice.

Materials and Methods

1) Animal preparation

Twenty five male Sprague-Dawley rats (10～12 weeks old) that had undergone middle cerebral artery occlusion (MCAo) surgery were prepared. A permanent right MCAo was made using a modified protocol initially described by Longa.[5] Twenty five rats were randomly assigned into 3 groups for each week's test (N=12, 10, and 13 for week 1, 2, and 3, respectively). Surgery was done 1～4 weeks prior to the time of assessment, which enabled the authors to evaluate the reliability at a wide range of test scores. The experimental protocol was approved by institutional animal care and use committee of Seoul National University Hospital.

2) Raters

Two raters performed the 6-item GNT for each rat according to method proposed by Garcia et al.[1] The first rater had been performing animal experiments for more than 2 years and had extensive experience of neurological tests such as the GNT ('expert' rater). The second rater was a medical doctor without prior experience in laboratory procedures involving animals ('novice' rater). The novice rater underwent introductory 1-day hands-on workshop for animal experiment before participation in the present study. The workshop encompassed biomedical ethics, small animal care and handling, feeding, intervention, and anesthesia. The raters were both blinded to the operative procedure that each rat had undergone and the purpose of the study.

3) Neurological evaluation

The GNTs were done consecutively by each rater independently in one session. Two sessions were performed on the same day. The interval between sessions on the same day was about 2 hours. Evaluations were repeated weekly for 3 weeks, making a total of 6 sessions of GNT for each rater. The rats were randomly rearranged for the second test so that the raters were 'blinded' to the first scores. In addition, the raters carried out GNT without knowing the results scored by the other rater.

4) Standardized video education

We made a video that shows the actual processes of performing GNT with an explanation about each step of the individual evaluations. In the video, each step of the test was explained by using descriptions from the original article[1] and also showed an expert carrying out the test using the method suggested in the original article. Thus, the video had both an explanation and a demonstration of each step of GNT. This 10-minute standardized video was shown to the novice rater once a week after he performed the second test session on each day. This self-education was done a total of 2 times (after the session 2 and 4). The novice had 20 more minutes to freely review the video clip in each session.

5) Learning curve

We drew a learning curve of the novice using a weighted average method in the assumption that the scores of the expert are correct. We focused on the gaps in the scores between the novice and the expert. If the score of the novice was equal to that of the expert, we scored 3 for the item. If the scores of the novice were higher or lower by 1, 2, or 3 degrees, then we scored 2, 1, or 0, respectively, for the item. Values from the items were averaged to generate learning curve.

6) Statistical analysis

Intra-rater agreement was assessed by comparing the 1st and 2nd sessions' scores on the same rats for each rater. For inter-rater agreement, scores of two raters on the same rats were compared for each session. We used the weighted kappa coefficient, which is based on a formula suggested by Cohen.[6] Kappa statistics were reported with a 95% confidence interval in parentheses.[7] We classified the kappa values as ranging from 'less than chance agreement' to 'almost perfect agreement'.[4] The statistical analyses were done with SPSS 17.0 for Windows (SPSS Inc., Chicago, USA) and MedCalc v11.4.2.

Results

The GNT comprises six subtests as the following: 'spontaneous activity', 'symmetry of movements', 'symmetry of forelimbs', 'climbing wall of wire cage', 'reaction to touch on either side of trunk', and 'response to vibrissae touch'. We assigned these subtests number 1～6, as shown in Table 1. We performed statistical analysis on the total sum score and each score using Cohen's weighted kappa coefficient. The kappa values were presented with 95% confidence interval in parentheses. Agreement, as measured by kappa, was interpreted by the criteria of Landis and Koch[4] as follows: ＜0, less than chance agreement; 0.01～0.20, slight agreement; 0.21～0.40, fair agreement; 0.41～0.60, moderate agreement; 0.61～0.80, substantial agreement; and 0.81～0.99, almost perfect agreement.

Table 1. Six Subtests of Garcia's Sensorimotor Test

Subtest		Values	Scale
Subtest 1	Spontaneous activity	0～3	ordinal
Subtest 2	Symmetry of movements	0～3	ordinal
Subtest 3	Symmetry of forelimbs	0～3	ordinal
Subtest 4	Climbing wall of wire cage	1～3	ordinal
Subtest 5	Reaction to touch on either side of trunk	1～3	ordinal
Subtest 6	Response to vibrissae touch	1～3	ordinal

1) Intra-rater reliability

For the total GNT score, the 1st week's kappa value for the novice was 0.31 (−0.02, 0.64) which indicated 'fair agreement', but the value reached that of the expert after 3-week self-education using a standardized video. The novice's kappa value was 0.68 (0.49, 0.87) for the 2nd week and 0.81 (0.69, 0.93) for the 3rd week. The kappa value for the expert indicated 'substantial agreement' at all times: 0.64 (0.48, 0.81), 0.76 (0.63, 0.88) and 0.71 (0.53, 0.89) for the 1st, 2nd and 3rd weeks, respectively (Table 2).

The intra-rater reliability for the novice with regard to the six individual subtests was also evaluated. The novice showed relatively low agreement in subtest 5 and 6 on the 1st week: 0.22 (−0.40, 0.84) and 0.38 (−0.18, 0.93) for test 5 and 6, respectively. These values increased in the 2nd week, but again decreased in the 3rd week (Table 3). The overall intra-rater reliability for the expert during the 3 weeks showed relatively low agreement for categories 3 and 4 (Table 3).

2) Inter-rater reliability

The kappa value for the total GNT score in the 1st week was 0.44 (0.26, 0.62). The value rose to 0.69 (0.57, 0.81) and 0.69 (0.56, 0.82) during the 2nd and 3rd weeks after self-education, respectively. Specifically, the inter-rater reliability for the scores from subtests 5 and 6 was low in the 1st week (Table 4).

Table 2. Rater-specific Intra-rater Reliability across Week

Rater	Week 1 (n=12)	Week 2 (n=10)	Week 3 (n=13)
Expert	0.64 (0.48, 0.81)	0.76 (0.63, 0.88)	0.71 (0.53, 0.89)
Novice	0.31 (−0.02, 0.64)	0.68 (0.49, 0.87)	0.81 (0.69, 0.93)

Values are weighted kappa for the total score of Garcia's test (95% CI).

Table 3. Intra-rater Reliability of the Novice and the Expert

Subtest	Novice			Expert
	Week 1 (n=12)	Week 2 (n=10)	Week 3 (n=13)	Overall (n=35)
Subtest 1	0.57 (0.28, 0.87)	0.90 (0.71, 1.00)	0.58 (0.05, 0.85)	0.67 (0.47, 0.87)
Subtest 2	0.56 (0.33, 0.80)	0.78 (0.48, 1.00)	0.68 (0.41, 0.94)	0.67 (0.48, 0.85)
Subtest 3	0.57 (0.33, 0.82)	0.85 (0.57, 1.00)	0.60 (0.21, 0.99)	0.46 (0.17, 0.74)
Subtest 4	0.57 (0.21, 0.93)	0.63 (0.26, 0.99)	0.66 (0.37, 0.95)	0.40 (0.12, 0.68)
Subtest 5	0.22 (−0.40, 0.84)	0.79 (0.53, 1.00)	0.35 (−0.05, 0.75)	0.70 (0.47, 0.92)
Subtest 6	0.38 (−0.18, 0.93)	0.78 (0.51, 1.00)	0.45 (0.10, 0.81)	0.85 (0.71, 1.00)

Values are weighted kappa for the six subtests of the Garcia's test (95% CI).

3) Learning curve

In the assumption that the scores from the expert are correct, the learning curve of the novice showed great improvement by self-education (Fig. 1A). The shapes of the learning curves in six individual evaluation items were diverse, but tests 5 and 6 showed a clear positive tendency (Fig. 1B).

Discussion

The purpose of this study was to assess the intra- and inter-rater reliability of GNT, and to estimate the learning curve of the novice researcher. The intra-rater agreement was substantial for an experienced rater. After 3-week learning using a video, a novice reached an 'almost perfect' level of intra-rater agreement. Although inter-rater reliability of the total sum was also substantial, which is in line with the previous reports,[3] some individual items showed low level of intra- and inter-rater agreement.

Table 4. Measurement Occasion-specific Inter-rater Reliability

	Week 1 (n=24*)	Week 2 (n=20*)	Week 3 (n=26*)
Subtest 1	0.68 (0.44, 0.91)	0.76 (0.60, 0.92)	0.50 (0.22, 0.78)
Subtest 2	0.66 (0.41, 0.91)	0.52 (0.32, 0.73)	0.79 (0.59, 0.99)
Subtest 3	0.34 (0.06, 0.62)	0.50 (0.20, 0.80)	0.39 (−0.01, 0.78)
Subtest 4	0.23 (0.02, 0.44)	0.19 (−0.09, 0.47)	0.71 (0.51, 0.90)
Subtest 5	0.14 (−0.18, 0.46)	0.56 (0.30, 0.83)	0.66 (0.36, 0.96)
Subtest 6	0.07 (0.19, 0.33)	0.60 (0.35, 0.85)	0.86 (0.72, 1.00)
Total score	0.44 (0.26, 0.62)	0.69 (0.57, 0.81)	0.69 (0.56, 0.82)

Values are weighted kappa (95% CI).
*The test was performed twice on each week (12, 10, and 13 rats for 1st, 2nd, and 3rd week, respectively).

Intra-rater agreement for assessment tools of animal neurological function has seldom been investigated. When Garcia et al. proposed a neurological functional scale, they validated the test by assessing the histological change, primarily by counting cells.[1] However, the qualification of the raters was not clearly described at that time. Three years later, Pantoni et al. reported that the inter-rater reliability between two experienced observers for the total sum score of GNT was high.[3] However, they did not mention the intra-rater agreement. Furthermore, there has been no study on how much time or effort would be needed to educate and train the raters to perform reliable assessment. The same is true to other neurological functional tests frequently used in animal research. For the cylinder test,[8] which is widely used to evaluate the function of a rodent's impaired forelimb, the inter-rater reliability was also shown to be fair even with relatively inexperienced raters.[9] Few details were known, however, on the intra-rater agreement or the training of novices.

In contrast with neurological functional tests for experimental animals, the intra-rater reliability has been emphasized along with the inter-rater reliability for assessing human subjects. In addition, questions such as how researchers trained the raters and what was needed to educate them have been explored in great detail. For example, the intra- and inter-rater reliability were estimated to verify the Copenhagen stroke scale.[10] And also there are studies that mention

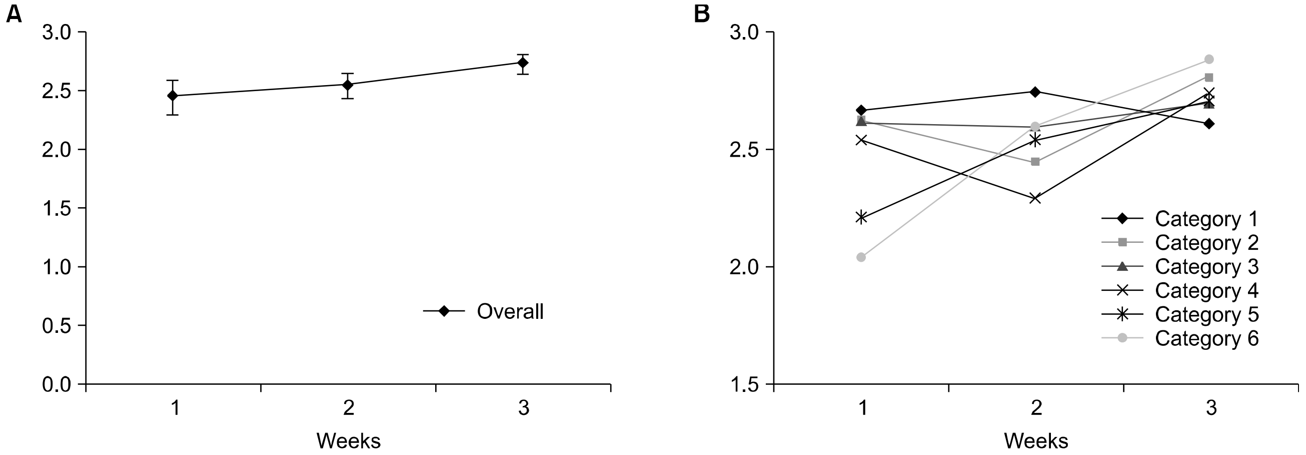

Fig. 1. The learning curves using the weighted average method. (A) The learning curve of the novice for all scores of Garcia's test using the weighted average method. A score of 3.0 represents a perfect match between the expert and the novice. The graph shows a clear increase after self-education. (B) The learning curve of the novice was calculated from the scores of six subtests using the weighted average method. Each graph of the six subtests shows a different shape and slope, but there is a clear upward slope for categories 5 and 6.

intra-rater reliability or the learning process of the raters by the education which corrected the inter-rater variability.[11] Experiments using laboratory animals not only can provide important biological knowledge, but offer invaluable opportunities to explore possibilities of translating results from wet benches to clinical medicine. In this regard, growing interests in translational research is calling researchers for a higher standard of technical refinement as well as research ethics.[12] Further studies will be required to test intra-individual agreement of common assessment tools.

The present study showed that the novice rater is potentially error-prone and has difficulty obtaining reliable results even with a simple neurological functional test. This highlights the importance of quality control for raters. The authors used a standardized video in order to train the novice. Standardized videos are generally accepted as practical and useful in self-education.[13] Training using a standardized video was an effective method to improve the reliability of the National Institutes of Health stroke scale.[14] In addition, another study showed that self-education using a 30-minute standardized video for adult cardiopulmonary resuscitation (CPR) yielded comparable or better CPR performance than the traditional training method.[15] The increase of kappa value for intra-rater reliability and the improvement in the learning curve suggest that video clips are also effective in training beginners of animal experiment.

Our results showed that the intra-rater reliability improved clearly for the novice after self-education. The value of kappa increased from 0.31 (−0.02, 0.64) to 0.68 (0.49, 0.87) after one session of self-education using the standardized video, indicating that the intra-rater variability was effectively reduced by the video education. The kappa values in the 2nd and 3rd weeks reached ‘substantial agreement’ and ‘almost perfect agreement’, respectively. Furthermore, the fact that the lower limit of 95% CI of the 3rd week’s kappa (0.81 [0.69, 0.93]) was higher than the upper limit of 1st week’s value (0.31 [−0.02, 0.64]) solidly proves the efficiency of learning (Table 2). In our study, the novice became competent enough for research after 3-week self-education using video.

Although the total sum score was highly reliable, some of the individual items showed lower inter-and intra-rater agreement. Furthermore, each rater showed greater within-rater variability for different subtests (Table 3, 4). This underlines the necessity of monitoring competence and consistency of the raters, even experienced ones. The reason that the novice showed low agreement in subtests 5 and 6 (‘reaction to touch on either side of the trunk’ and ‘response to vibrissae touch’) at 1st weak could be because judgment in these subtests was more difficult for the novice before the standardized video education (Table 4). For the expert, all values of kappa were high enough to indicate that the test had internal consistency (Table 2). The relatively low values in subtests 3 and 4 (‘symmetry of forelimbs’ and ‘climbing wall of wire cage’) could be due to the fact that repetitive tests on rats led to stress or caused fatigue, causing them to be less active during the second test (Table 3). This could also indicate that the subtest 3 and 4 are vulnerable to physiologic conditions such as fatigue. There have been reports on the correlation between behavior and fatigue or stress.[16,17] The protocol on those studies was not exactly same as GNT, but the change in rats’ environment can cause stress or fatigue, and that can decrease the activity of the rats. On the second test session, the expert’s scores in subtests 3 and 4 decreased in 11 rats, but increased only in 6 rats. Thus there is a possibility that the repetitive examination might have caused fatigue and decreased the rats’ activity in many rats. And the vulnerability to the fatigue of rats could be a flaw of GNT, considering that the intra-rater reliability of the expert was relatively low in certain subtests. Interestingly, in the learning curve for the novice, individual items differed in the velocity of improvement (Fig. 1B). The different shape of the learning curve of each subtest leads us consider item-based education and more focused feedback for researchers who are newly engaged in animal laboratory. And the variability in the shape of each subtest could also be due to the vulnerability of the subtests to fatigue. For subtest 5 and 6, in which the expert had high intra-rater reliability, the curve showed a clear upward slope. However, the fact that only the total sum score is used as a primary outcome in most cases lessens the influence of this item-based variability.

One limitation of this study stemmed from its small sample size. Greater sample size would have yielded narrower confidence interval (CI) for each kappa value. However, the kappa of the intra-rater reliability for the novice increased in the 3rd week to such an extent as to have higher CI without an overlap with that in the 1st week. In addition, the tendency of the change in kappa or the scores of the tests in our study was clear enough to analyze the issues that we set out to address. One other problem was the fact that the novice in our study was a medical doctor who had some background knowledge about physiology and anatomy, which make it difficult to generalize the learning curve described here to all novices.

In animal research, the doctrine of replacing, reducing and refinement (also known as the 3Rs) is now widely accepted and practiced. The reliable outcome measurement is essential to refine the research and experiment process to reduce the need for live subjects

This study showed the learning curve for a novice which is important to undertake reliable research through training. And also it suggested that self-learning using video is likely to facilitate the learning process.

Conclusion

In conclusion, the accuracy of the test scores was higher for an expert than for a novice, although this was rectified by a short period of self-education. The intra- and inter-rater reliability also showed enough agreement. A standardized video for the Garcia score might be effective to educate novice raters in a short period of time. Future study may be necessary to assess the intra-rater reliability for the other frequently used functional tests.

References

1) Garcia JH, Wagner S, Liu KF, Hu XJ. Neurological deficit and extent of neuronal necrosis attributable to middle cerebral artery occlusion in rats. Statistical validation. *Stroke* 1995; 26: 627-635

2) Tomasello F, Mariani F, Fieschi C, Argentino C, Bono G, De Zanche L, Inzitari D, Martini A, Perrone P, Sangiovanni G. Assessment of inter-observer differences in the Italian multicenter study on reversible cerebral ischemia. *Stroke* 1982; 13: 32-35

3) Pantoni L, Bartolini L, Pracucci G, Inzitari D. Interrater agreement on a simple neurological score in rats. *Stroke* 1998; 29: 871-872

4) Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-174

5) Longa EZ, Weinstein PR, Carlson S, Cummins R. Reversible middle cerebral artery occlusion without craniectomy in rats. *Stroke* 1989; 20: 84-91

6) Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213-220

7) Fleiss JL, Levin BA, Paik MC. Statistical methods for rates and proportions. J. Wiley, Hoboken, N.J., 2003; p xxvii, 760

8) Schallert T, Kozlowski DA, Humm JL, Cocke RR. Use-dependent structural events in recovery of function. *Adv Neurol* 1997; 73: 229-238

9) Schallert T, Fleming SM, Leasuren JL, Tillerson JL, Bland ST. CNS plasticity and assessment of forelimb sensorimotor outcome in unilateral rat models of stroke, cortical ablation, parkinsonism and spinal cord injury. *Neuropharmacology* 2000; 39: 777-787

10) Olesen J, Simonsen K, Nørgaard B, Grønbæk M, Johansen OS, Krogsgaard A, Andersen B. Reproducibility and Utility of a Simple Neurological Scoring System for Stroke Patients (Copenhagen Stroke Scale). *Neurorehabilitation and Neural Repair* 1988; 2: 59-63

11) Buchanan GN, Halligan S, Taylor S, Williams A, Cohen R, Bartram C. MRI of fistula in ano: inter- and intraobserver agreement and effects of directed education. *AJR Am J Roentgenol* 2004; 183: 135-140

12) Depienne C, Stevanin G, Brice A, Durr A. Hereditary spastic paraplegias: an update. *Curr Opin Neurol* 2007; 20: 674-680

13) Gagliano ME. A literature review on the efficacy of video in patient education. *J Med Educ* 1988; 63: 785-792

14) Lyden P, Brott T, Tilley B, Welch KM, Mascha EJ, Levine S, Haley EC, Grotta J, Marler J. Improved reliability of the NIH Stroke Scale using video training. NINDS TPA Stroke Study Group. *Stroke* 1994; 25: 2220-2226

15) Todd KH, Braslow A, Brennan RT, Lowery DW, Cox RJ, Lipscomb LE, Kellermann AL. Randomized, controlled trial of video self-instruction versus traditional CPR training. *Ann Emerg Med* 1998; 31: 364-369

16) Katz RJ, Roth KA, Carroll BJ. Acute and chronic stress effects on open field activity in the rat: implications for a model of depression. *Neurosci Biobehav Rev* 1981; 5: 247-251

17) Tanaka M, Nakamura F, Mizokawa S, Matsumura A, Nozaki S, Watanabe Y. Establishment and assessment of a rat model of fatigue. *Neurosci Lett* 2003; 352: 159-162