ICSD Diagnostic Criteria for Narcolepsy: Interobserver Reliability

Luca Vignatelli MD,1 Giuseppe Piazza MD,1 Leona Bassein CSTAT,2 Alfredo Barbato MD, PhD,3 Armando De Vincentiis MD,3 Elio Lugaresi MD,1 Roberto D’Alessandro MD1 on behalf of the “GIN-SEN” (see appendix)

1Institute of Clinical Neurology, University of Bologna, Italy; 2S.Orsola-Malpighi Hospital, Bologna, Italy; and 3Dompé Biotec spa, Italy

Study Objectives: To estimate the reliability of the diagnosis of narcolepsy after clinical interview and polysomnographic evaluation among sleep medicine doctors, before and after training in application of the International Classification of Sleep Disorders (ICSD).

Setting: Videotaped semi-structured interviews of 10 patients complaining of daytime sleepiness of different etiologies. Questions referred to ICSD criteria for narcolepsy. A further series of 10 cases of narcolepsy without cataplexy were simulated, with at least a random one to three of the ICSD polysomnographic criteria at pathological levels.

Participants and Design: Seventeen doctors were required to classify each videotaped case as “ascertained,” “possible,” or “excluded” narcolepsy, in two sessions: one before and one after discussion of ICSD criteria. The observers were invited to confirm or exclude the diagnosis of narcolepsy in the 10 simulated cases, according to the given polysomnographic findings, before and after an agreed proposal of the interpretation of ICSD polysomnographic criteria. Interobserver reliability was calculated using Kappa statistics.

Measurements and Results: Interobserver reliability of clinical judgement improved from “substantial” at baseline (Kappa 0.61) to “almost perfect” after training (Kappa 0.95). Interobserver reliability of polysomnographic findings was “fair” at baseline (Kappa 0.24), unanimous after the proposed interpretation of ICSD polysomnographic criteria.

Conclusions: Baseline reliability of diagnostic judgement in suspected narcolepsy was found satisfactory among Italian sleep medicine doctors. Educational training, based on discussion of ICSD criteria, further improved agreement. Diagnosis based on polysomnographic findings, not reliable at baseline, needed a strict interpretation of ICSD criteria to attain standardization.

Key words: Narcolepsy; videotape recording; diagnosis; reliability of results; Kappa statistics; polysomnography.

INTRODUCTION

INTEROBSERVER RELIABILITY (IR), THAT IS THE DEGREE OF AGREEMENT AMONG DIFFERENT OBSERVERS,1 is necessary to standardize diagnostic process when designing multicenter epidemiological studies and clinical trials.2 The American Academy of Sleep Medicine Task Force recently recommended applying diagnostic methods of proven reliability also in the sleep medicine field.3 Basic research on diagnostic methodology, given the scarcity of studies on this topic, was highly solicited.

To date International Classification of Sleep Disorders (ICSD) Revised4 represents the gold standard for the diagnosis of narcolepsy. The classical association of recurrent daytime naps and cataplexy is sufficient for a definite diagnosis, the latter symptom being specific. On the other hand, in the wide spectrum of cases without cataplexy, where daytime sleepiness and other associated features are evocative but not specific, polysomnographic studies are mandatory. When narcolepsy is suspected in a specialist setting, the IR of diagnostic judgement after clinical interview and after polysomnographic evaluation is not known.

Before embarking on a national multicentric epidemiological survey on narcolepsy (Gruppo Italiano Narcolessia - Studio Epidemiologico Nazionale “GIN-SEN”), this study aimed to estimate the IR of diagnosis of suspected narcolepsy among doctors of Italian sleep centers before and after training in the application of ICSD criteria, and to investigate the value of an interactive method of education, based on videotaped interviews.

METHODS

Subjects for the Videotaped Interview

Ten subjects (all men, mean age 44 yrs, range 23—67 yrs), consecutively referred to the Sleep Disorders Center of the Institute of Clinical Neurology of the University of Bologna for suspected primary hypersomniac disorder (seven subjects) or sleep disordered breathing with subjective sleepiness (three subjects), were included. Final diagnosis included narcolepsy (three subjects, one with cataplexy and two without cataplexy), narcolepsy with cataplexy and obstructive sleep apnea syndrome (OSAS) (one subject), idiopathic hypersomnia (two subjects), OSAS (two subjects), OSAS and REM behavior disorder (RBD) (one subject), nocturnal groaning (one subject).5

After signed informed consent, each patient underwent a videotaped semi-structured interview by one examiner (LV). Each question corresponds to a simple standard paraphrase of each ICSD criterion for the diagnosis of narcolepsy.4 Other simple questions on sleep habits, snoring, and interrupted breathing during snoring were formulated to investigate causes of daytime sleepiness other than narcolepsy. After cutting, a video totalling about 70 minutes duration was obtained.
Simulated Polysomnographic Cases

Ten series of four simulated polysomnographic results (Nocturnal sleep latency [NSL] in minutes, Nocturnal REM sleep latency [NRL] in minutes, Mean Sleep Latency [MSL] at MSLT in minutes, number of sleep onset REM periods [SOREMPs] at MSLT) for 10 simulated cases of narcolepsy without cataplexy (i.e., needing laboratory confirmation of the diagnosis) were devised. Each series randomly presented one, two, or three of the four polysomnographic characteristics required by ICSD criteria: E: 1) NSL less than 10 minutes; 2) NRL less than 20 minutes; 3) MSL less than 5 minutes; 4) Two or more SOREMPs; (see original text).

Raters and Reliability Procedures

Seventeen doctors (seven resident neurologists or psychiatrists and 10 trainees neurologists or neurophysiopathologists) each from the sleep centers of the Italian Association of Sleep Medicine acted as raters. They all attended a video session with the 10 interviews and were asked to classify independently (on a form) each case as "ascertained," "possible," or "excluded" narcolepsy. Each rater was unaware of the others' diagnostic classification. Then, as a training procedure, one of us (GP) formally presented the ICSD Revised clinical criteria for the diagnosis of narcolepsy, followed by a free discussion. On the basis of the ICSD clinical criteria, we proposed to adopt the following diagnostic classification: "ascertained narcolepsy" in the presence of criteria A ("a complaint of excessive sleepiness or sudden muscle weakness") and at least one of the point of criterion B (sleep paralysis or hypnagogic hallucinations or automatic behaviors or disrupted major sleep episode), and "excluded narcolepsy" when the previous conditions were not satisfied. After a repeat video session with the same interviews, the raters were invited to classify each patient according to the above diagnostic categories, specifying also the presence or the absence of each of the ICSD clinical criteria. All symptoms but sleepiness were dichotomized as "absent" or "present"; sleepiness was trichotomized as "absent," presence of "complaint of excessive sleepiness," and presence of "recurrent daytime naps or lapses into sleep that occur almost daily for at least three months." The 10 simulated polysomnographic cases of "possible" narcolepsy without cataplexy were classified as "ascertained" or "excluded" narcolepsy by the raters. Finally, after the proposal to consider the diagnosis of narcolepsy when two or more SOREMPs at MSLT together with at least one of the other polysomnographic findings present, the raters were invited to reconsider their diagnoses.

Statistical Analysis

The overall proportion of agreement and IR of the diagnosis in the two case series were evaluated in the whole group and in residents and trainees subgroups, before and after the training procedures. IR was calculated by the means of Kappa statistics, which is the ratio of the observed agreement beyond chance to the potential agreement beyond chance, according to the formula by generalization of Kappa for dichotomous or polychotomous data and more than two raters, proposed by Fleiss. Kappa values were interpreted according to conventional groups (0.0-0.20=slight agreement; 0.21-0.40=fair; 0.41-0.60=moderate; 0.61-0.80=substantial; 0.81-1.00=almost perfect).

To test for a possible difference in agreement before and after training procedures, "overall consensus for a patient" was defined as the mode of pooled judgments (before and after) separately for that particular patient, and a contingency table was set up, with "before training" vs. "after training" and the modalities "consensus judgements" and "non-consensus judgments." The exact McNemar's test was then performed.

RESULTS

The diagnostic classification of the 10 videotaped cases by the 17 raters, before and after training, is presented in Table 1A. Before training, the overall proportion of agreement on final diagnosis was 77% (73% in residents and 81% in trainees). The IR was "substantial" among all raters (Kappa 0.61, Standard Error - SE - 0.02), "moderate" among residents (Kappa 0.54, SE 0.05) and "substantial" among trainees (Kappa 0.67, SE 0.04). After training, the overall proportion of agreement was 97% (97% in residents and 96% in trainees). The IR was "almost perfect" among all raters (Kappa 0.95, SE 0.02), without differences between residents and trainees (Kappa, respectively, 0.95, SE 0.05, and 0.94, SE 0.03).

Overall consensus for each patient coincided with the prevailing judgment both before and after training. Before training, 143 out of 170 judgments coincided with the consensus modality, whereas after the second session, consensus rose to 167 out of 170; only two judgments with previous consensus and one previously discordant judgment were discordant in the final session. McNemar's test indicated that the increase in Kappa from 0.61 before training to 0.95 after training was significant (p<0.0001).

Considering data from the second video session, reliability of diagnostic judgement on cataplexy and sleep paralysis was unanimous, "substantial" for automatic behaviors and sleepiness (respectively, Kappa 0.80, SE 0.05, and Kappa 0.64, SE 0.02), "moderate" for disrupted major sleep episode (Kappa 0.58, SE 0.05), "fair" for hypnagogic hallucinations (Kappa 0.35, SE 0.05).

In the series of polysomnographic simulated cases (table 1B), after the first evaluation the overall proportion of agreement was 72% (72% in residents and 70% in trainees); the reliability was "fair" among all raters (Kappa 0.24, SE 0.03), "fair" among residents (Kappa 0.30, SE 0.07) and "slight" among trainees (Kappa 0.15, SE 0.05). After the proposed interpretation of ICSD polysomnographic criteria no disagreement was found.

DISCUSSION

Our study disclosed that, when hypersomniac patients are evaluated on the basis of a standardized clinical interview, diagnostic judgment of narcolepsy has a satisfactory baseline reliability, among doctors practicing in the sleep medicine field. In our case series, disagreement was probably generated by confounding levels of sleepiness in patients with narcolepsy without cataplexy, idiopathic hypersomnia, OSAS. In the presence of cat-
Cataplexy, disagreement affected the suspicion or certainty of narcolepsy, probably because of indecision on the diagnostic strength to attribute to cataplexy. After specific training for the application of ICSD diagnostic criteria, with elimination of "criterion variance" (i.e., the variability in diagnostic criteria), 9 IR became "almost perfect." Residual disagreement must be ascribed to "interpretation variance" (i.e., the variability in the interpretation of patient information and in the application of diagnostic rules.

Our study presents some limitations. First, because of the difficulty in grouping many videotaped interviews in a brief lapse of time, the number of cases proposed to observers is low. So, given the limited clinical variability of cases, reliability could be overestimated. Second, even if videotaped semi-structured interviews guarantee stable diagnostic information without influencing examined subjects with repeated identical questions, results may be limited by the artfulness of the setting and the impossibility to consider "information variance" (i.e., the variability in the information sought and in patient responses instability of clinical features). Third, the design did not include any control group of observers, so the improvement of reliability after training could be biased by a "practice effect."

As opposed to IR of clinical interview, IR on evaluating polysomnographic results was not satisfactory at baseline. Before training, unanimity on confirming the diagnosis was obtained only when two or more sleep onset REM periods at MSLT together with other two positive findings of the ICSD criterion E were present (cases 7 and 8, Table 1B). All other combinations of positive findings led to discordant conclusions. In this case, unreliability might have originated from "interpretation variance" 9, probably due to the wording of the text of the E criterion (see original text). After training, unanimous agreement could be obtained only proposing a strict interpretation of the criterion.

In conclusion, when diagnosis of narcolepsy relies on polysomnographic result, strict interpretation of criteria is necessary to attain standardization in multicentric studies.

ACKNOWLEDGMENTS

Carolina Cerrè and Eleonora Romagnoli of Bioikos srl for technical support and precious advices; Federica Pompei of Dompé Biotec spa; Anne Collins for revising the English; Carlo Grassi for video-editing.

REFERENCES

American Academy of Sleep Medicine, 2001.

ABBREVIATIONS

ICSD—International Classification of Sleep Disorders; IR—Interobserver reliability; OSAS—obstructive sleep apnea syndrome; RBD—REM sleep behaviour disorder; SE—standard error; MSLT—multiple sleep latency test; MSL—mean sleep latency; NRL—nocturnal REM sleep latency; NSL—nocturnal sleep latency; SOREMPs—sleep onset REM periods; GIN-SEN—Gruppo Italiano Narcolessia - Studio Epidemiologico Nazionale (Italian Narcolepsy Group - National Epidemiological Study)

APPENDIX

GIN-SEN Coordinating Members: R D’Alessandro, G Plazzi, L Vignatelli (Institute of Clinical Neurology, Bologna); A Barbato, A De Vincentiis (Dompé Biotec spa, Italy).
Participants from the Italian Association of Sleep Medicine Centres: S Misceo, F Puca (Bari); C Servalli, E Ubiali (Bergamo); R Vetrugno, G Plazzi (Bologna, Institute of Clinical Neurology); G Buzzi, F Cirignotta (Bologna, S.Orsola-Malpighi University Hospital); V Fassari, A Scrofani (Catania); M Beelke, F Ferrillo (Genova); C Costa, R Di Perri (Messina); C Landi, L Ferini-Strambi (Milano); M Rossi, MG Terzano (Parma); I Sartori, R Manni (Pavia); E Bonanni, L Murri (Pisa, Neuroscience Department); L Palagini, M Guazzelli (Pisa, Psychiatry Department); S Miano, O Bruni, G Antonini (Roma, La Sapienza University); B Farina, GF Mennuni (Roma, Institute of Neurology UCSC); F Cosentino, R Ferri (Troina); R Marinig, P Bergonzi (Udine, University Institute of Neurology); PL Dolso, GL Gigli (Udine, Neuroscience Department).