Interobserver variability of the House-Brackmann facial nerve grading system for the analysis of a randomized multi-center phase III trial

Christian Scheller[1] · Andreas Wienke[2] · Marcos Tatagiba[3] · Alireza Gharabaghi[3] ·
Kristofer F. Ramina[3] · Konstanze Scheller[4] · Julian Prell[1] · Johannes Zenk[5] ·
Oliver Ganslandt[6] · Barbara Bischoff[6] · Cordula Matthies[7] · Thomas Westermaier[7] ·
Gregor Antoniadis[8] · Maria Teresa Pedro[8] · Veit Rohde[9] · Kajetan von Eckardstein[9] ·
Thomas Kretschmer[10] · Malte Kornhuber[11] · Fred G. Barker II[12] · Christian Strauss[1]

Abstract

Background Evidence of a high interobserver variability of the subjective House-Brackmann facial nerve grading system (HBGS) would justify cost- and time-consuming technological enhancements of objective classifications for facial nerve paresis.
Method A total of 112 patients were recruited for a randomized multi-center trial to investigate the efficacy of prophylactic nimodipine treatment in vestibular schwannoma (VS) surgery. For the present investigation both treatment groups were pooled for the assessment of facial nerve function preoperatively, in the early postoperative course and 1 year after the surgery. Facial nerve function was documented photographically at rest and in motion and classified according to the HBGS by three independent observers (neurosurgeon, neurologist, ENT) and by the investigator of each center.
Results Interobserver variability was considerably different with respect to the three time points depending upon the sever-

ity of facial nerve paresis. Preoperative facial nerve function was normal or only mildly impaired (HB grade I or II) and was assessed consistently in 97%. Facial nerve function deteriorated during the early postoperative course and was subsequently documented without dissent in only 36%, with one grade difference in 45%, two grade difference in 17% and three grade difference in 2%. One year after surgery, facial nerve function predominantly improved resulting in a consistent assessment in 66%. Differing ratings were observed in 34% with one grade deviation in 88% and of two grades in 12%. Patients with differing ratings of two or more grades exhibited considerably worse facial nerve function ($p < 0.001$).
Conclusions The HBGS produced comparable results between different observers in patients with normal or only mildly impaired facial nerve function. Interobserver variability increased depending on the severity of facial nerve paresis. The results suggest that the HBGS does not promote uniformity of

Christian Scheller
christian.scheller@uk-halle.de

[1] Department of Neurosurgery, Martin Luther University of Halle-Wittenberg, Ernst-Grube-Str. 40, 06097 Halle (Saale), Germany

[2] Institute of Medical Epidemiology, Biostatistics, and Informatics, University of Halle-Wittenberg, Halle (Saale), Germany

[3] Department of Neurosurgery, University of Tübingen, Tübingen, Germany

[4] Department of Oral and Maxillofacial Surgery and Facial Plastic Surgery, University of Halle-Wittenberg, Halle (Saale), Germany

[5] Department of Otorhinolaryngology, Head and Neck Surgery, University of Erlangen-Nuremberg, Erlangen, Germany

[6] Department of Neurosurgery, University of Erlangen-Nuremberg, Erlangen, Germany

[7] Department of Neurosurgery, Würzburg University Hospital, Würzburg, Germany

[8] Department of Neurosurgery, Bezirkskrankenhaus Günzburg, University of Ulm, Günzburg, Germany

[9] Department of Neurosurgery, University of Göttingen, Göttingen, Germany

[10] Department of Neurosurgery, Evangelisches Krankenhaus, University of Oldenburg, Oldenburg, Germany

[11] Department of Neurology, University of Halle-Wittenberg, Halle (Saale), Germany

[12] Department of Neurosurgery, Massachusetts General Hospital, Boston, MA, USA

reporting and comparison of outcomes in patients with moderate or severe facial nerve paresis.

Keywords House-Brackmann · Facial nerve · Vestibular schwannoma · Interobserver variability

Introduction

Since the House-Brackmann facial nerve grading system (HBGS) was introduced in 1983 and endorsed by the Facial Nerve Disorders Committee of the American Academy of Otolaryngology-Head and Neck Surgery in 1984 as the standard for reporting facial nerve function, there have been ongoing controversies about its suitability [11, 12]. Samii [25] and Kanzaki [15] reported that the HBGS is widely validated and generally accepted, and grading of facial nerve function is possible with only small interobserver variability [25], whereas other studies suggested only fair to moderate levels of interobserver agreement for precise functional outcome assessment [1, 2, 7, 29].

Multiple facial nerve grading systems were developed to improve the reliability and the interobserver agreement, such as the HBGS and its modifications [8, 11, 15], the Nottingham [20], the Sydney [5], the Sunnybrook Facial Grading Systems [24] and the MoReSS (standing for movement, rest, secondary defects and subjective scoring) [7].

However, it has been difficult to reach consensus at the Acoustic Neuroma Conferences since 1991 [15].

Considering the fact that facial nerve function serves as a quality control for the treatment modalities of vestibular schwannomas and taking into account the controversies about the HBGS and the availability of modern techniques for video-analysis by the strongly expanded use of high-resolution cameras, it is questionable, whether subjective clinical assessments are still up to date.

Despite preliminary publications [5, 9, 13, 17], this is the first report investigating the interobserver variability of the HBGS at defined pre- and postoperative time points by four independent investigators in a dedicated group of patients of a previously published multi-center trial [26].

Methods

The primary aim of the underlying multi-center trial was to investigate the efficacy and safety of prophylactic parenteral nimodipine and HES treatment in VS surgery [26]. For the present study the treatment and control group were pooled. Facial nerve function was investigated at defined time points (preoperative, during the in-patient stay and 1 year after surgery) and documented photographically at rest and in motion as described by House and Brackmann [12]. Photographs were evaluated by the investigator of the multi-center trial and additionally by three blinded reviewers experienced in facial nerve disorders (neurologist, ENT, neurosurgeon) and classified using the HBGS. Interobserver variability was investigated if data from at least two reviewers were present.

All patients underwent resection via a retrosigmoid approach. In all cases intraoperative neurophysiological monitoring including brainstem auditory evoked potentials (BAEP), continuous facial nerve electromyography (EMG) and direct facial nerve stimulation were applied, and the diagnosis of a schwannoma was confirmed histopathologically.

Results

Participant flow

A total of 112 patients were enrolled in the multicenter phase III trial [26]. There were nine dropouts before preoperative data acquisition. In one patient data concerning the preoperative facial nerve function was missing. Therefore, 102 patients were suitable for further investigation (Fig. 1).

Fig. 1 Participant flow diagram and preoperative assessment of facial nerve function
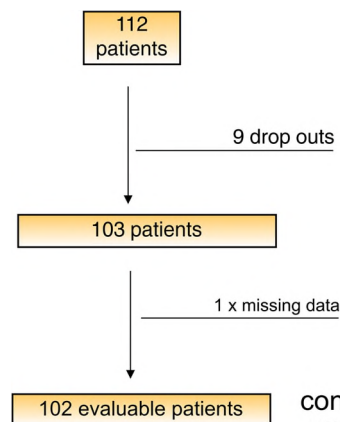
HB-grade	preoperative assessments			
	neuro-surgeon	neuro-logist	otolaryngo-logist	study investi-gator
I°	102	100	102	101
II°		2		1
III°				
IV°				
V°				
VI°				

consistent assessments: 99/102   97 %
differing assessments: 3/102   3 %

Preoperative function of the facial nerve was investigated by the four investigators in all 102 cases.

In the early postoperative course, 14 patients could not be analyzed because of missing photographs (n = 11) or assessment by only one investigator (n = 3) (Fig. 2). Therefore, 89 patients were suitable for further investigation in the early postoperative course. In 88 of these patients facial nerve function was rated by four and in one case by three investigators.

Facial nerve function 1 year after surgery was assessed in 103 patients. Photographs were classified by only one investigator in four cases and consequently were excluded (Fig. 3). The photographs of the resulting 99 patients were rated by 4 investigators in 92 patients, by 3 investigators in 4 patients and by 2 investigators in 3 patients.

Baseline data

The mean age of the patients was 49 years; 56% were female. Tumors predominantly were medium to large sized. Preoperative hearing was predominantly useful (Gardner-Robertson grade 1 or 2 in 63%). All patients revealed a normal (97%) or only a mildly impaired (3%) facial nerve function (Table 1).

Assessment of facial nerve function

Interobserver variability was considerably different with respect to the three time points depending upon the severity of facial nerve paresis. Preoperative facial nerve function that was normal or only mildly impaired (HB grade I or II) was equally assessed in 97% (Fig. 1). Facial nerve function that deteriorated in the early postoperative course was subsequently documented without dissent in only 36%, with one grade difference in 45%, two grade difference in 17% and three grade difference in 2% of the patients (Fig. 2). Within 1 year

after surgery in most cases facial nerve function had improved, resulting in a consistent assessment in 66%. Differing ratings were observed in 34% with deviation of one grade in 88% and of two grades in 12% (Fig. 3).

Significant differences concerning the severity of facial nerve paresis with a mean value of 2.4 [standard error of estimate (SE): 0.1, mean of all rated HB grades] in the early postoperative course compared to 1.5 (SE: 0.1) 1 year after surgery ($p < 0.001$, T-test for dependent samples) could be observed. Interobserver reliabilty was also significantly different between the early postoperative course and 1 year after surgery ($p < 0.001$, McNemar test).

Patients with two or more differing rating grades (n = 17) showed considerably worse facial nerve function ($p < 0.001$, Mann-Whitney U-test). In these patients the mean value of postoperative facial nerve paresis was 3.6 (SE: 0.2). Photographs of these 17 patients were assessed by 4 investigators. In contrast, the mean value of facial nerve paresis was 2.1 (SE: 0.2) in the 72 patients with no or only one grade of differently rated grades.

There was no tendency toward better ratings within the rater group of the operating neurosurgeons.

Discussion

This is the first study to show that there is a correlation between the severity of postoperative facial nerve paresis and the extent of interobserver variability using the HBGS. There were considerable limitations regarding interobserver agreements especially in patients with moderate and severe facial nerve paresis. Former studies reported only interrater agreements or interobserver variabilities of the HBGS without considering the degree of facial nerve paresis [5, 9, 13, 17]. The

Fig. 2 Participant flow diagram and assessment of facial nerve function in the early postoperative course
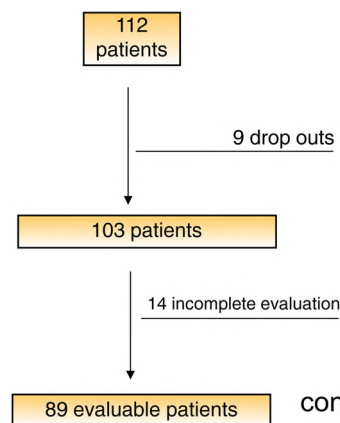
HB-grade	immediately postoperative assessments			
	neuro-surgeon	neuro-logist	otolaryngo-logist	study investi-gator
I°	31	23	41	33
II°	30	28	14	24
III°	8	10	14	8
IV°	9	11	7	15
V°	10	14	11	7
VI°		3	1	2

consistent assessments: 32/89   36 %
differing assessments: 57/89   64 %
                                    1 grade: 40/57   70 %
                                    2 grades: 15/57   26 %
                                    3 grades: 2/57   4 %

Fig. 3 Participant flow diagram and assessment of facial nerve function 1 year after surgery

	112 patients
	103 patients
	99 evaluable patients

9 drop outs

4 incomplete evaluation
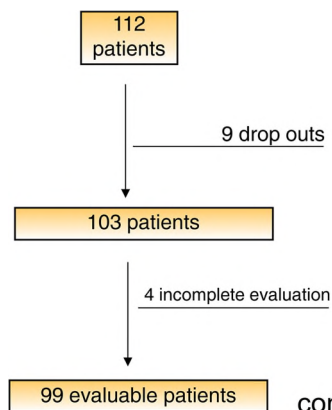
HB-grade	assessments 12 months after surgery			
	neuro-surgeon	neuro-logist	otolaryngo-logist	study investi-gator
I°	63	65	71	68
II°	18	14	5	15
III°	11	14	15	6
IV°	2	4	3	4
V°	1	2	1	1
VI°				1

consistent assessments: 65/99 66 %
differing assessments: 34/99 34 %
1 grade: 30/57 88 %
2 grades: 4/57 12 %

investigators of this study had acquired at least 15 years' experience in the assessment of facial nerve disorders. They are specialists for ENT, neurology and neurosurgery. Therefore, the study design (competence of the raters and the use of photographs) did not negatively influence the measure of agreement. In particular, there was no consistent tendency for a better rating of the operating center. This suggests an acceptable quality of monocentric studies reporting facial nerve functions following surgical and non-surgical procedures.

Video clips (instead of photographs) of patients at rest and with a series of facial expressions may improve the accordance of rating facial nerve function by different observers. However, Gordon et al. showed that photographs with standardized facial expressions are a reliable outcome measure for determining facial nerve function following vestibular schwannoma surgery [9].

General requirements of a facial nerve grading system are reporting the clinical assessment as objectively as possible, reflecting signs of recovery or changes in function following therapeutic intervention [8]. In principle, quantitative technology-based systems and subjective clinical assessments are described.

However, time restrictions in a purely clinical setting and the need for additional specialized equipment might be arguments in favor of the use of simple subjective standard rating scales. Even though it has been criticized as not being sufficiently sensitive to changes [20, 23] and regarding its interobserver variability [5, 6, 16, 20], the HBGS is the commonly used system and was adopted by the American Academy of Otolaryngology-Head and Neck Surgery as standard for grading facial nerve recovery in 1984 [12]. Measurement of facial nerve function could be performed quickly within a few minutes and does not require any equipment. However, there are problems with noninclusion of synkesis phenomena [10] and

the assessment of facial nerve function as HB grade III, because parasympathetic ("dry eye") and intermediate nerve functions are not addressed in the classification [15].

To improve the interobserver agreement of subjective clinical classification, further developments were proposed. The HBGS was modified at the Tokyo consensus meeting in 2001 [15] and the Facial Nerve Grading Scale 2.0 was developed [8]. In many studies [1, 7, 29], the HBGS showed only a fair to moderate level of agreement (according to the Landis and Koch guidelines: <0, no agreement; 0–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial;

Table 1 Baseline data

	Patients (n = 102)
Mean age in years (min–max)	49 (18–75)
Gender	
Female	57 (55.9%)
Male	45 (44.1%)
Koos grade (tumor size)	
I	2 (2%)
II	34 (33.3%)
III	42 (41.2%)
IV	24 (23.5%)
GR class (preoperative hearing)	
1	35 (34.3%)
2	29 (28.4%)
3	29 (28.4%)
4	5 (5%)
5	4 (4%)
HB grade (preoperative facial nerve function)	
1	99 (97.1%)
2	3 (2.9%)

0.81–1, almost perfect agreement) [17], which was in contrast to data of former validation studies (up to 0.77) [11, 23]. Burres and Fisch [3] suggested that assessment of facial nerve function should focus on motor function. The "Rough" facial nerve grading system is also based on facial motor function [2]. In 1994, the Nottingham System suggested the objective measurement of three facial expressions [20]. In 1995 and 1996, two similar objective scales were published: the Sydney System [4, 5] and the Sunnybrook Facial Grading System [24]. A comparative study between the Sunnybrook and HB facial grading scales investigating the repeatability and agreement demonstrated the superiority of the Sunnybrook classification [13].

However, the pitfall of all subjective classifications of facial nerve function is only a fair to moderate interobserver agreement, which lies in their very nature and cannot be improved by more sophisticated measurement methods. Quite in line with this, a comparison of the HBGS with other subjective grading systems showed no significant improvement in interobserver reliability [28].

Prospective computerized pixel change analysis [19, 21, 27] and facogram-based objective assessment of facial nerve function showed promising results and may eliminate clinicians' subjectivity to provide standardized measurement in the future [16, 18, 22].

Another method, the so-called "Moire topography," uses special cameras to measure facial contours. Investigations in 51 patients with facial palsy and 10 healthy volunteers showed high correlation between the results of the moire indexes and the HBGS [30]. Neely et al. reported an excellent test-retest reliability in 30 patients with a wide range of facial nerve paresis using image substraction techniques of digital video-recordings [21]. In the past, the lack of technical equipment was the greatest problem for computer analysis of facial nerve function [14]. In the future, easy access to high-resolution cameras and the development of evaluation software may provide objective assessment of facial nerve function in the clinical routine.

Conclusions

Further improvement of the assessment of facial nerve function will require an objective rating scale using modern techniques (video documentation and subsequent video analysis with motion analysis software), which nowadays may be easily applicable without involving extensive costs.

References

1. Ahrens A, Skarada D, Wallace M, Heung JY, Neely JG (1999) Rapid simultaneous comparison system for subjective grading scales for facial paralysis. Am J Otol 20:667–671

2. Alicandri-Ciufelli M, Piccinini A, Grammatica A et al (2013) A step backward: the "Rough" facial nerve grading system. J Craniomaxillofac Surg 41:175–179

3. Burres SA, Fisch U (1986) The comparison of facial nerve grading systems. Arch Otolaryngol Head Neck Surg 112:755–758

4. Coulson SE, Croxon GR (1995) Assessing physiotherapy rehabilitation outcomes following facial nerve paresis. Aust J Otolaryngol 2:20–24

5. Coulson SE, Croxon GR, Adams RD, O'Dwyer NJ (2005) Reliability of the "Sydney", "Sunnybrook", and "House Brackmann" facial grading systems to assess voluntary movement and synkenesis after facial nerve paralysis. Otolaryngol Head Neck Surg 132:543–549

6. Croxon G, May M, Mester SJ (1990) Grading facial nerve function: House-Brackmann versus Burres-Fisch methods. Am J Otol 11: 240–246

7. de Ru JA, Braunius WW, van Benthem PP, Busschers WB, Hordijk GJ (2006) Grading facial nerve function: why a new grading system, the MoReSS, should be proposed. Otol Neurotol 27:1030–1036

8. Fattah AY, Gurusinghe AD, Gavilan J et al (2015) Facial nerve grading instruments: systematic review of the literature and suggestion for uniformity. Plast Reconstr Surg 135:569–579

9. Gordon AS, Westrick AC, Falola MI et al (2012) Reliability of postoperative photographs in assessment of facial nerve function after vestibular schwannoma resection. J Neurosurg 117:860–863

10. Henstrom DK, Skilbeck CJ, Weinberg J, Knox C, Cheney ML, Hadlock TA (2011) Good correlation between original and modified House Brackmann facial grading systems. Laryngoscope 121: 47–50

11. House JW (1983) Facial nerve grading system. Laryngoscope 93: 1056–1069

12. House JW, Brackmann DE (1985) Facial nerve grading system. Otolarangol Head Neck Surg 93:146–147

13. Kanerva M, Poussa T, Pitkäranta A (2006) Sunnybrook and House-Brackmann facial grading systems: intrarater repeatability and interrater agreement. Otolaryngol Head Neck Surg 135:865–871

14. Kang TS, Vrabec FT, Giddings N, Terris DJ (2002) Facial nerve grading systems (1985–2002): beyond the House-Brackmann scale. Otol Neurotol 23:767–771

15. Kanzaki J, Tos M, Sanna M, Moffat DA (2003) New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. Otol Neurotol 24: 642–649

16. Kecskes G, Jori J, O'Reilly BF, Viharos L, Rovo L (2011) Clinical assessment of a new computerized objective method of measuring facial palsy. Clin Otolaryngol 36:313–319

17. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

18. Lee NL, Susarla SM, Hohman MH et al (2013) A comparison of facial nerve grading systems. Ann Plast Surg 70:313–316

19. Meier-Gallati V, Scriba H, Fisch U (1988) Objective evaluation scaling of facial nerve function based on area analysis (OSCAR). Otolaryngol Head Neck Surg 118:545–550

20. Murty GE, Diver JP, Kelly PJ, O'Donoghue GM, Bradley PJ (1994) The Nottingham system: objective assessment of facial nerve function in the clinic. Otolarangol Head Neck Surg 110:156–161

21. Neely JG, Wang KX, Shapland CA, Sehizadeh A, Wang A (2010) Computerized objective measurement of facial motion: normal variation and test-retest reliability. Otol Neurotol 31:1488–1492

22. O'Reilly BF, Soraghan JJ, McGrenary S, He S (2010) Objective method of assessing and presenting the House-Brackmann and regional grades of facial palsy by production of a facogram. Otol Neurotol 31:486–491

23. Rickenmann J, Jaquenod C, Cerenko D, Fisch U (1997) Comparative value of facial nerve grading systems. Otolaryngol Head Neck Surg 117:322–325

24. Ross BG, Fradet G, Nedzelski JM (1996) Development of a sensitive clinical facial grading system. Otolaryngol Head Neck Surg 114:380–386

25. Samii M, Gerganov V (2013) Surgery of cerebellopontine lesions, 1st edn. Springer, Heidelberg, pp 147–314

26. Scheller C, Wienke A, Tatagiba M et al (2016) Prophylactic nimodipine treatment for cochlear and facial nerve preservation after vestibular schwannoma surgery: a randomized phase III trial. J Neurosurg 124:657–664

27. Scriba H, Stoeckli SJ, Veraguth D et al (1999) Objective evaluation of normal facial nerve function. Ann Otol Rhinol Laryngol 108: 641–644

28. Smith IM, Murray JA, Cull RE et al (1992) A comparison of facial grading systems. Clin Otolaryngol 17:303–307

29. Vrabec JT, Backous D, Dyalilian H et al (2009) Facial nerve grading system 2.0. Otolaryngol Head Neck Surg 140:445–450

30. Yuen K, Inokuchi I, Maeta M, Kawakami SI, Masuda Y (1997) Evaluation of facial palsy by moire topography index. Otolaryngol Head Neck Surg 117:567–572