

Rater Reliability of the Hardy Classification for Pituitary Adenomas in the Magnetic Resonance Imaging Era

Michael A. Mooney¹ Douglas A. Hardesty¹ John P. Sheehy¹ C. Roger Bird² Kristina Chapple¹
William L. White¹ Andrew S. Little¹

¹ Department of Neurosurgery, Barrow Neurological Institute, St. Joseph's Hospital and Medical Center, Phoenix, Arizona, United States

² Departments of Neuroradiology, Barrow Neurological Institute, St. Joseph's Hospital and Medical Center, Phoenix, Arizona, United States

Address for correspondence Andrew S. Little, MD, Department of Neurosurgery, Barrow Neurological Institute, St. Joseph's Hospital and Medical Center, 350 W. Thomas Road, Phoenix, AZ 85013, United States (e-mail: Neuropub@dignityhealth.org).

J Neurol Surg B 2017;78:413–418.

Abstract

Objectives The Hardy classification is used to classify pituitary tumors for clinical and research purposes. The scale was developed using lateral skull radiographs and encephalograms, and its reliability has not been evaluated in the magnetic resonance imaging (MRI) era.

Design Fifty preoperative MRI scans of biopsy-proven pituitary adenomas using the sellar invasion and suprasellar extension components of the Hardy scale were reviewed.

Setting This study was a cohort study set at a single institution.

Participants There were six independent raters.

Main Outcome Measures The main outcome measures of this study were interrater reliability, intrarater reliability, and percent agreement.

Results Overall interrater reliability of both Hardy subscales on MRI was strong. However, reliability of the intermediate scores was weak, and percent agreement among raters was poor (12–16%) using the full scales. Dichotomizing the scale into clinically useful groups maintained strong interrater reliability for the sellar invasion scale and increased the percent agreement for both scales.

Conclusion This study raises important questions about the reliability of the original Hardy classification. Editing the measure to a clinically relevant dichotomous scale simplifies the rating process and may be useful for preoperative tumor characterization in the MRI era. Future research studies should use the dichotomized Hardy scale (sellar invasion Grades 0–III versus Grade IV, suprasellar extension Types 0–C versus Type D).

Keywords

- ▶ pituitary adenoma
- ▶ rater reliability
- ▶ transsphenoidal

Introduction

The operative approach to a pituitary adenoma is guided by the size and location of the tumor and its relation to surrounding anatomical structures. The large variation in sellar invasion and suprasellar extension of pituitary adenomas was recognized in the 1970s by Hardy and Vezina¹ and prompted the development of the Hardy classification criteria to better characterize these lesions (▶ **Table 1**).^{1–3} Since that time, the Hardy classification system has served as a descriptive tool for pituitary adenomas and is often utilized in research studies.^{4–13}

The Hardy classification comprises two subscales: one describes the integrity of the sellar floor and invasion into the sphenoid sinus (Grades 0–IV), whereas the other describes the degree of suprasellar extension of the tumor (Types A–D). Although these two subscales were described using lateral radiographs and encephalograms, respectively, the Hardy

received
February 14, 2017
accepted
April 30, 2017
published online
June 7, 2017

© 2017 Georg Thieme Verlag KG
Stuttgart · New York

DOI <https://doi.org/10.1055/s-0037-1603649>.
ISSN 2193-6331.

Table 1 Description of the Hardy and Vezina classification¹ and the proposed dichotomized scale

Grade/Type	Description
Hardy classification grade	
Sellar invasion	
Grade 0	The enclosed adenoma is described as a tumor that remains within the anatomical confines of the osteoaponeural sheath of the sella turcica. The floor of the sella is always intact.
Grade I	The sella turcica is within normal limits in size (less than 16 × 13 mm; 208 mm ²) but shows a lowering of the floor on one side or a bulging of the cortex.
Grade II	The sella turcica is enlarged to various degrees but the floor remains intact.
Grade III	The sella is more or less enlarged but there is a local erosion or destruction of the floor.
Grade IV	The entire floor of the sella is diffusely eroded or destroyed, giving a characteristic “phantom sella” with all the boundaries barely visible.
Suprasellar extension	
Type 0 ^a	The tumor is entirely confined within the sella turcica.
Type A	The suprasellar expansion bulges into the chiasmatic cistern but does not reach the floor of the anterior third ventricle.
Type B	The tumor reaches the floor of the third ventricle, giving the image of an inverse cupula of the anterior recesses of the third ventricle.
Type C	A voluminous suprasellar expansion bulges largely into the third ventricle up to the foramen of Monro.
Type D	Rare aberrant expansions occur in temporal or frontal fossa.
Dichotomized Hardy classification	
Sellar invasion	
Grade 0–III	The sellar floor remains entirely or partially intact.
Grade IV	The sellar floor is completely eroded, with diffuse adenoma invasion into the sphenoid sinus.
Suprasellar extension	
Type 0–C	The adenoma is confined to the sella or extends superiorly to the level of the foramen of Monro.
Type D	The adenoma has aberrant expansion into the temporal or frontal fossa.

^aAdded for the purpose of this study.

grading scale is still used to classify adenomas based on magnetic resonance imaging (MRI) scans. To determine the applicability of the Hardy scale in the era of using MRIs in pituitary practice, we evaluated the interrater and intrarater reliabilities of the scale. Furthermore, we evaluated dichotomized versions of the subscales that are clinically relevant and that may better guide future pituitary research studies.

Methods

Patients and Raters

The specific methodology used for this study type has been published previously.¹⁴ Briefly, 50 unique, preoperative, gadolinium-enhanced, dedicated pituitary MRI scans were selected from a prospectively maintained database of cases of biopsy-proven pituitary adenomas; cases represented a range of tumor sizes. Six independent raters participated in the study: three neurosurgery residents (M.A.M., D.A.H., and J.P.S.) and three faculty raters (two pituitary surgeons [W.L.W. and A.S.L.] and one neuroradiologist [C.R.B.]). Each rater was given a written and verbal description of the project and a copy of the original article describing the Hardy classification system for reference.¹ Three raters (one faculty [A.S.L.] and two

residents [D.A.H. and J.P.S.]) participated in the intrarater reliability portion of the study and were presented with 35 imaging studies a second time more than 4 weeks after their initial review. This study was approved by the institutional review board. Due to the retrospective nature of the report, informed consent was not required.

Statistics

We determined the number of scans to be analyzed and the number of raters to be included on the basis of a prestudy power analysis. Numerical values were assigned to roman numerals for grades of sellar invasion and to letters for grades of suprasellar extension. Scans without suprasellar extension were listed as not applicable for this subscale and counted as Type 0 for the analysis (→Table 1). Pairwise Spearman's correlation coefficients and phi coefficients were calculated between raters and then averaged to describe the reliability of both subscales. A scan was considered intermediate if the mean score of all raters for the scan was between the 25th and 75th percentiles of the mean scale score distribution. This resulted in a defined mean intermediate group of 2.00 to 3.15 for the sellar invasion scale and of 1.15 to 2.30 for the suprasellar extension scale. Intrarater reliability was calculated using Spearman's correlation coefficients.

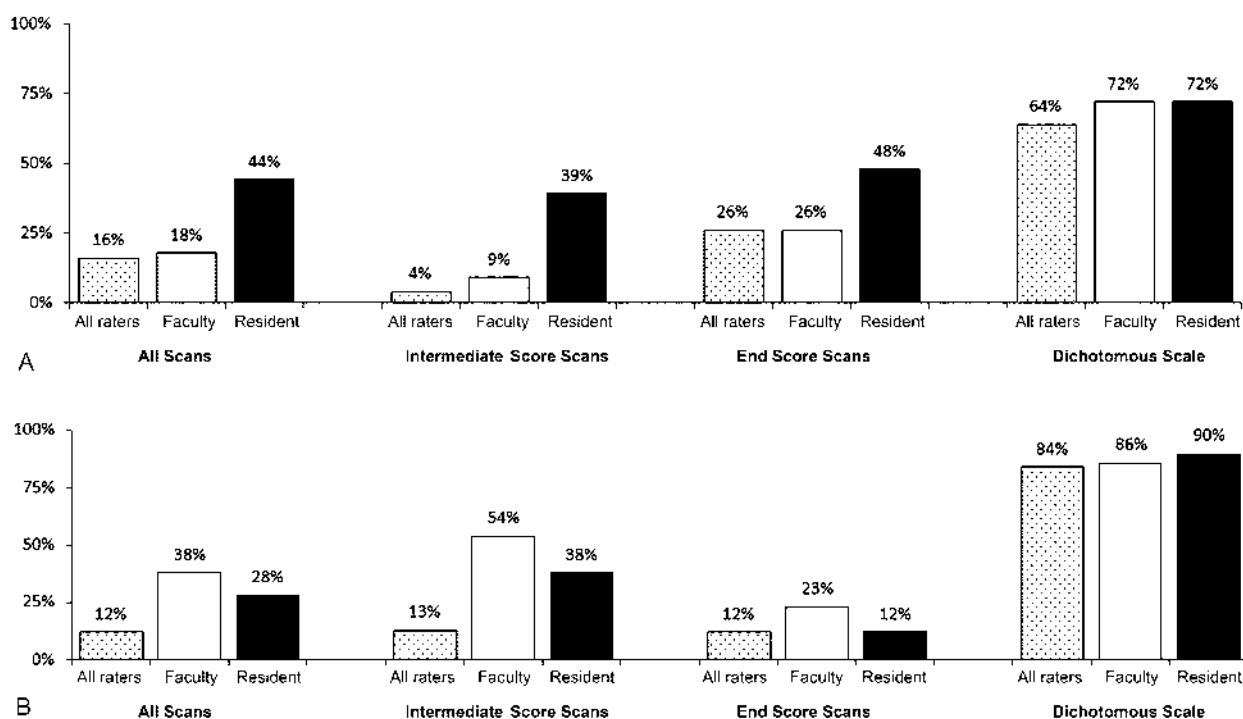


Fig. 1 Percent agreement for the (A) sellar and (B) suprasellar scales by training level. (Used with permission from Barrow Neurological Institute, Phoenix, Arizona, United States.)

We also tested dichotomized versions of both subscales that represent clinically relevant applications (sellar Grades 0–III vs. Grade IV; suprasellar Types 0–C vs. Type D). The mean of the pairwise phi coefficients between the six raters was used as an overall reliability for the dichotomous scale. Counts and percentages representing rater agreement are presented. Coefficients were interpreted as follows: 0.00 to 0.19 = “very weak”; 0.20 to 0.39 = “weak”; 0.40 to 0.59 = “moderate”; 0.60 to 0.79 = “strong”; and 0.80 to 1.00 = “very strong.” All calculations were performed by a dedicated biostatistician (K.C.) using SPSS Statistics for Windows, Version 22 (IBM Corp.).

Results

Interrater Reliability

Sellar Invasion

Overall interrater reliability for the full-scale sellar invasion rating was strong (0.69; 95% confidence interval [CI], 0.51–0.81)

(**Table 2**). When examined separately, reliability of the intermediate scores was very weak (0.15; 95% CI, – 0.28 to 0.53), whereas the reliability of the scale ends (i.e., mean Grades 0–IV) was very strong (0.82; 95% CI, 0.63–0.91). When the scale was dichotomized into clinically useful groups of the lesions likely to have completely eroded the sella (Grade IV) versus those with no erosion or lesser degrees of erosion (Grades 0–III), the reliability remained strong (0.62; 95% CI, 0.41–0.76) and the percent agreement among all raters improved from 16% (8/50 cases) for the full scale to 64% (32/50 cases) for the dichotomous scale (**Fig. 1**).

Suprasellar Extension

Overall interrater reliability for the full-scale suprasellar extension rating was strong (0.78; 95% CI, 0.65–0.87). When the intermediate scores were examined separately, the reliability was weak (0.35; 95% CI, – 0.06 to 0.66); however, the reliability of the scale ends was strong (0.86; 95% CI, 0.71–0.94). The scale was dichotomized into clinically useful groups of

Table 2 Interrater reliability: all raters

Scale	Sellar invasion ^a		Suprasellar extension ^b	
	Reliability (95% CI)	Percent agreement	Reliability (95% CI)	Percent agreement
Full scale	0.69 (0.51–0.81)	8/50 (16%)	0.78 (0.65–0.87)	6/50 (12%)
Intermediate scores	0.15 (– 0.28 to 0.53)	1/23 (4%)	0.35 (– 0.06 to 0.66)	3/24 (13%)
Scale ends	0.82 (0.63–0.91)	7/27 (26%)	0.86 (0.71–0.94)	3/26 (12%)
Dichotomous scale	0.62 (0.41–0.76)	32/50 (64%)	0.30 (0.03–0.54)	42/50 (84%)

Abbreviation: CI, confidence interval.

^aFull scale: Grades 0–IV. Dichotomous scale: Grades 0–III versus Grade IV.

^bFull scale: Types 0–D. Dichotomous scale: Types 0–C versus Type D.

lesions with extension into the frontal or temporal fossa (Type D) versus those with lesser degrees of extension (Types 0–C), which would influence approach selection (craniotomy vs. transsphenoidal approach). The percent agreement among all raters increased from 12% (6/50 cases) for the full scale to 84% (42/50) for the dichotomous scale. The reliability of the dichotomous scale was weak (0.30; 95% CI, 0.03–0.54); however, this was attributed to minimal variability in the sample (i.e., relatively few scans were rated D vs. A–C).

Training Level

Sellar Invasion

Overall, the differences by training level in interrater reliability for sellar invasion were not significant (►Table 3). Similar reliability emerged between faculty and resident raters for the full scale with all 50 MRI scans included (0.67 vs. 0.68, respectively; $p = 0.93$), for the intermediate scores only (0.14 vs. 0.13; $p = 0.97$), and for the end scores only (0.73 vs. 0.86; $p = 0.22$). For the dichotomous scale, the difference between faculty (0.58) and resident (0.62) reliability was also not statistically significant ($p = 0.76$).

Suprasellar Extension

Faculty versus resident reliability on suprasellar extension grading was not significantly different for the full scale (0.80 vs. 0.78, respectively; $p = 0.79$), for the intermediate scores only (0.27 vs. 0.49; $p = 0.41$), or for the scale-end scores only (0.85 vs. 0.86; $p = 0.90$). For the dichotomous scale, reliability for faculty raters (0.51) was significantly higher than resident raters (0.15) ($p = 0.046$).

Intrarater Reliability

Intrarater reliability was measured by having three raters—one faculty physician and two residents—provide sellar and

Table 4 Intrarater reliability of Hardy scales

Scale and raters	Reliability (95% CI)	Percent agreement
Sellar ^a full scale		
Faculty rater	0.78 (0.60–0.88)	25/35 (71%)
Resident rater 1	0.79 (0.61–0.89)	22/35 (63%)
Resident rater 2	0.68 (0.45–0.83)	24/35 (69%)
Sellar ^a dichotomous scale		
Faculty rater	0.88 (0.77–0.94)	34/35 (97%)
Resident rater 1	0.49 (0.19–0.71)	26/35 (74%)
Resident rater 2	0.65 (0.40–0.81)	30/35 (86%)
Suprasellar ^b full scale		
Faculty rater	0.86 (0.74–0.93)	23/35 (66%)
Resident rater 1	0.90 (0.80–0.95)	25/35 (71%)
Resident rater 2	0.85 (0.72–0.92)	29/35 (83%)
Suprasellar ^b dichotomous scale		
Faculty rater	0.80 (0.64–0.90)	34/35 (97%)
Resident rater 1	1 (1–1)	35/35 (100%)
Resident rater 2	0.36 (0.04–0.62)	32/35 (91%)

Abbreviation: CI, confidence interval.

^aFull scale: Grades 0–IV. Dichotomous scale: Grades 0–III versus Grade IV.

^bFull scale: Types 0–D. Dichotomous scale: Types 0–C versus Type D.

suprasellar ratings for 35 scans at two separate time points. Intrarater reliability was in the strong range for the full sellar invasion and suprasellar extension scales for all raters (►Table 4). When the scale was collapsed to a dichotomous measure, intrarater reliability remained in the strong range for the faculty physician. Intrarater reliability was

Table 3 Interrater reliability by training level

Scale and raters	Sellar invasion ^a		Suprasellar extension ^b	
	Reliability (95% CI)	Percent agreement	Reliability (95% CI)	Percent agreement
Full scale				
Faculty raters	0.67 (0.48–0.80)	9/50 (18%)	0.80 (0.68–0.88)	19/50 (38%)
Resident raters	0.68 (0.49–0.80)	22/50 (44%)	0.78 (0.64–0.87)	14/50 (28%)
Intermediate scores				
Faculty raters	0.14 (– 0.29 to 0.52)	2/23 (9%)	0.27 (– 0.15 to 0.61)	13/24 (54%)
Resident raters	0.13 (– 0.30 to 0.52)	9/23 (39%)	0.49 (0.11–0.75)	9/24 (38%)
Scale ends				
Faculty raters	0.73 (0.49–0.87)	7/27 (26%)	0.85 (0.70–0.93)	6/26 (23%)
Resident raters	0.86 (0.71–0.95)	13/27 (48%)	0.86 (0.70–0.95)	3/26 (12%)
Dichotomous scale				
Faculty raters	0.58 (0.36–0.74)	36/50 (72%)	0.51 (0.27–0.69)	43/50 (86%)
Resident raters	0.62 (0.41–0.77)	36/50 (72%)	0.15 (–0.22–0.48)	45/50 (90%)

Abbreviation: CI, confidence interval.

^aFull scale: Grades 0–IV. Dichotomous scale: Grades 0–III versus Grade IV.

^bFull scale: Types 0–D. Dichotomous scale: Types 0–C versus Type D.

stable to decreased for resident raters with dichotomization; however, percent agreement was improved.

Discussion

The preoperative characterization of pituitary adenomas helps guide surgical planning, approach selection, preoperative patient counseling, and intraoperative decision making. The Hardy classification of sellar invasion and suprasellar extension of pituitary adenomas, which is based on radiographs and encephalograms, is routinely cited in the literature and is used to characterize lesions preoperatively and for pituitary research studies.^{4–13} Despite tremendous advances in pituitary imaging, the Hardy system continues to be used in the MRI era.

The most important measures of any grading scale are its accuracy and its reproducibility among raters. An emphasis on these characteristics has emerged in the neurosurgical literature, and numerous scales have been evaluated over the past decade in an attempt to validate their use for clinical decision making and research.^{14–24} However, the reliability of the Hardy scale has never been evaluated. In this report, we provide the first analysis of both the interrater and intrarater reliabilities of the Hardy scale, and we provide an additional analysis of the effect of training level on reliability and percent agreement on grading cases.

Our results reveal three main findings regarding preoperative classification using the Hardy system. First, the reliabilities of the intermediate scores in the sellar invasion and suprasellar extension subscales were weak to very weak. Even in the MRI era, rater agreement regarding these characteristics has poor reliability. For the sellar invasion scale, raters demonstrated poor agreement on whether an adenoma simply expanded the sella (Grade II) or demonstrated local erosion of the sellar floor (Grade III). For the suprasellar extension scale, raters demonstrated poor agreement on whether an adenoma simply bulged into the chiasmatic cistern (Type A) or reached the floor of the third ventricle (Type B). As a result, pituitary research studies that use these grades for categorizing patients preoperatively must be interpreted with caution.

The second important finding of this study is that, in both subscales, the reliabilities of the scale ends were strong to very strong. This finding indicates that raters could reliably agree on characteristics of a nonerosive, well-contained adenoma (Grades 0–I, Types 0–A), as well as on a diffusely erosive adenoma with aberrant expansion into the frontal or temporal fossa (Grade IV, Type D). This finding is important because these characteristics might influence approach selection (craniotomy vs. transsphenoidal), likelihood of achieving biochemical remission or complete resection, and likelihood of recurrence; these issues may be important for tracking outcomes in the pituitary literature. Furthermore, these characterizing features determine the likelihood of gross total resection and may guide the preoperative strategy for approaching these lesions.

Because lesions that fall at the scale ends require unique consideration of approach and clinical outcomes, we proposed a dichotomized scale that may provide a better classification for pituitary adenomas in research studies moving forward.

Our results indicate that the percent agreement between raters significantly improves when the Hardy scale is dichotomized into clinically useful grades for the sellar invasion and suprasellar extension scales (►Fig. 1). Thus, raters are more likely to agree on the categorization of these lesions when it comes to preoperative grading for pituitary research studies. We propose the terminology “dichotomized Hardy scale” (sellar invasion Grades 0–III vs. Grade IV and suprasellar extension Types 0–C vs. Type D) to be used for classification of pituitary adenomas in future research studies.

Notably, the interrater reliability scores of the dichotomized scales were generally lower than the reliability scores of the full scale, which is a product of using the phi coefficient to measure the association between two dichotomous variables versus the Spearman’s coefficient to determine the association among three or more ordinal levels. The interrater reliability of the suprasellar extension dichotomous scale (0.30) was lower than the reliability of the full scale (0.78); however, this finding is attributable to the very low number of cases rated as Type D in this study and the low variability in our sample. Given the relatively rare occurrence of Type D lesions in practice, we believe that our sample remains representative and that the improved percent agreement among raters indicates that the dichotomized scale performs better for this classification.

The third important finding of this study is the preserved interrater reliability, as well as the intrarater reliability, of the Hardy classification across training levels. Interrater reliabilities for the full scale, for intermediate scores only, and for end scores only were not significantly different between resident raters and faculty raters (►Table 3). Although the interrater reliability of the dichotomized suprasellar extension scale was significantly improved for the faculty versus the resident raters, this analysis is limited by the same low number of Type D cases, as detailed earlier. We do not believe that this finding indicates a deficiency for the resident raters, and overall these findings can be used to argue that resident rating of preoperative MRI scans can be reliable for use in pituitary research studies moving forward.

This study provides the first assessment of the reliability of the Hardy classification in the MRI era; however, it is not without limitations. First and foremost, all studies examining interrater and intrarater reliabilities are limited by their selection of raters. We attempted to minimize this limitation by selecting both resident and faculty raters, and we compared the results of the two groups. We believe that including resident raters in studies of reliability is essential, since many studies may require resident evaluation of preoperative imaging. It should be noted that this study was performed at a high-volume pituitary center and included senior neurosurgery residents (postgraduate years 3–5). A neuroradiologist and two pituitary surgeons were included to represent the highest standard for preoperative imaging assessment.

The second major limitation of this study is the small number of suprasellar extension Type D scans. As detailed earlier, this small number limited the statistical analysis of the dichotomous suprasellar extension scale, and it resulted in artificially low reliability scores as measured by the phi

coefficient. Nonetheless, we believe that the results with the dichotomous scale still represent an improvement over the full scale, as evidenced by the substantial improvement in percent agreement. A larger cohort of Type D scans should be included in any future studies of this type.

Conclusion

Although the overall interrater and intrarater reliabilities of the Hardy classification are acceptable, this study raises concern about the poor reliability of the intermediates scores in this system. Dichotomizing the sellar and suprasellar subscales separates less invasive tumors (Grades 0–III, Types 0–C) from the most invasive tumors (Grade IV, Type D) and significantly improves the percent agreement across these scales in a way that is clinically meaningful. We believe a modification of the Hardy scale is still relevant in current pituitary practice and research, and authors of future pituitary studies are encouraged to use the dichotomized Hardy scale (sellar invasion Grades 0–III vs. Grade IV and suprasellar extension Types 0–C vs. Type D) to improve agreement among raters across different studies.

Disclosure

None.

Financial Support

None.

Acknowledgments

We thank the Neuroscience Publications staff of Barrow Neurological Institute for assistance in the preparation of this article.

References

- Hardy J, Vezina JL. Transsphenoidal neurosurgery of intracranial neoplasm. *Adv Neurol* 1976;15:261–273
- Hardy J. *Atlas of Transsphenoidal Microsurgery in Pituitary Tumors*. New York, NY: Igaku-Shoin Medical Publishers; 1991
- Hardy J. *Transsphenoidal Microsurgical Treatment of Pituitary Tumors*. New York: Raven Press; 1979
- Onoz M, Basaran R, Gucluer B, et al. Correlation between SPARC (osteonectin) expression with immunophenotypical and invasion characteristics of pituitary adenomas. *APMIS* 2015;123(03):199–204
- Hong JW, Ku CR, Kim SH, Lee EJ. Characteristics of acromegaly in Korea with a literature review. *Endocrinol Metab (Seoul)* 2013;28(03):164–168
- Ku CR, Kim EH, Oh MC, Lee EJ, Kim SH. Surgical and endocrinological outcomes in the treatment of growth hormone-secreting pituitary adenomas according to the shift of surgical paradigm. *Neurosurgery* 2012;71(2, suppl operative):ons192–ons203, discussion ons203
- D'Haens J, Van Rompaey K, Stadnik T, Haentjens P, Poppe K, Velkeniers B. Fully endoscopic transsphenoidal surgery for functioning pituitary adenomas: a retrospective comparison with traditional transsphenoidal microsurgery in the same institution. *Surg Neurol* 2009;72(04):336–340
- Gondim JA, Tella OI Jr, Schops M. Intracellular pressure and tumor volume in pituitary tumor: relation study. *Arq Neuropsiquiatr* 2006;64(04):971–975
- Alleyne CH Jr, Barrow DL, Oyesiku NM. Combined transsphenoidal and pterional craniotomy approach to giant pituitary tumors. *Surg Neurol* 2002;57(06):380–390, discussion 390
- Nam DH, Song SY, Park K, et al. Clinical significance of molecular genetic changes in sporadic invasive pituitary adenomas. *Exp Mol Med* 2001;33(03):111–116
- Kim K, Arai K, Sanno N, Osamura RY, Teramoto A, Shibasaki T. Ghrelin and growth hormone (GH) secretagogue receptor (GHSR) mRNA expression in human pituitary adenomas. *Clin Endocrinol (Oxf)* 2001;54(06):759–768
- Gökalp HZ, Deda H, Attar A, Uğur HC, Arasil E, Egemen N. The neurosurgical management of prolactinomas. *J Neurosurg Sci* 2000;44(03):128–132
- Bates AS, Farrell WE, Bicknell EJ, et al. Allelic deletion in pituitary adenomas reflects aggressive biological activity and has potential value as a prognostic marker. *J Clin Endocrinol Metab* 1997;82(03):818–824
- Mooney MA, Hardesty DA, Sheehy JP, et al. Interrater and intrarater reliability of the Knosp scale for pituitary adenoma grading. *J Neurosurg* 2017;126(05):1714–1719
- Ames CP, Smith JS, Eastlack R, et al; International Spine Study Group. Reliability assessment of a novel cervical spine deformity classification system. *J Neurosurg Spine* 2015;23(06):673–683
- Vachhrajani S, Sen AN, Satyan K, Kulkarni AV, Birchansky SB, Jea A. Estimation of normal computed tomography measurements for the upper cervical spine in the pediatric age group. *J Neurosurg Pediatr* 2014;14(04):425–433
- Cordova JS, Schreiber E, Hadjipanayis CG, et al. Quantitative tumor segmentation for evaluation of extent of glioblastoma resection to facilitate multisite clinical trials. *Transl Oncol* 2014;7(01):40–47
- Griessenauer CJ, Miller JH, Agee BS, et al. Observer reliability of arteriovenous malformations grading scales using current imaging modalities. *J Neurosurg* 2014;120(05):1179–1187
- Frisoli FA, Lang SS, Vossough A, et al. Intrarater and interrater reliability of the pediatric arteriovenous malformation compactness score in children. *J Neurosurg Pediatr* 2013;11(05):547–551
- Jiménez-Roldán L, Alén JF, Gómez PA, et al. Volumetric analysis of subarachnoid hemorrhage: assessment of the reliability of two computerized methods and their comparison with other radiographic scales. *J Neurosurg* 2013;118(01):84–93
- Gordon AS, Westrick AC, Falola MI, Shannon CN, Walters BC, Fisher WS. Reliability of postoperative photographs in assessment of facial nerve function after vestibular schwannoma resection. *J Neurosurg* 2012;117(05):860–863
- Thaler M, Lechner R, Gstöttner M, et al. Interrater and intrarater reliability of the Kuntz et al new deformity classification system. *Neurosurgery* 2012;71(01):47–57
- Harrop JS, Vaccaro AR, Hurlbert RJ, et al; Spine Trauma Study Group. Intrarater and interrater reliability and validity in the assessment of the mechanism of injury and integrity of the posterior ligamentous complex: a novel injury severity scoring system for thoracolumbar injuries. Invited submission from the Joint Section Meeting on Disorders of the Spine and Peripheral Nerves, March 2005. *J Neurosurg Spine* 2006;4(02):118–122
- Kulkarni AV, Riva-Cambrin J, Browd SR. Use of the ETV success score to explain the variation in reported endoscopic third ventriculostomy success rates among published case series of childhood hydrocephalus. *J Neurosurg Pediatr* 2011;7(02):143–146