2017

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

Michael A. Mooney
Douglas A. Hardesty
John P. Sheehy
C. Roger Bird
Kristina Chapple

See next page for additional authors

Follow this and additional works at: https://scholar.barrowneuro.org/neurosurgery

Recommended Citation

This Article is brought to you for free and open access by the Neurosurgery at Barrow - St. Joseph's Scholarly Commons. It has been accepted for inclusion in Neurosurgery Articles by an authorized administrator of Barrow - St. Joseph's Scholarly Commons. For more information, please contact molly.harrington@dignityhealth.org, andrew.wachtel@dignityhealth.org.
Authors
Michael A. Mooney, Douglas A. Hardesty, John P. Sheehy, C. Roger Bird, Kristina Chapple, William L. White, and Andrew S. Little
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).


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 classification system has served as a descriptive tool for pituitary adenomas and is often utilized in research studies.4–13

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).
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. 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.

Resident raters [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.

<table>
<thead>
<tr>
<th>Grade/Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sellar invasion</td>
<td></td>
</tr>
<tr>
<td>Grade 0</td>
<td>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.</td>
</tr>
<tr>
<td>Grade I</td>
<td>The sella turcica is within normal limits in size (less than 16 × 13 mm; 208 mm²) but shows a lowering of the floor on one side or a bulging of the cortex.</td>
</tr>
<tr>
<td>Grade II</td>
<td>The sella turcica is enlarged to various degrees but the floor remains intact.</td>
</tr>
<tr>
<td>Grade III</td>
<td>The sella is more or less enlarged but there is a local erosion or destruction of the floor.</td>
</tr>
<tr>
<td>Grade IV</td>
<td>The entire floor of the sella is diffusely eroded or destroyed, giving a characteristic “phantom sella” with all the boundaries barely visible.</td>
</tr>
<tr>
<td>Suprasellar extension</td>
<td></td>
</tr>
<tr>
<td>Type 0</td>
<td>The tumor is entirely confined within the sella turcica.</td>
</tr>
<tr>
<td>Type A</td>
<td>The suprasellar expansion bulges into the chiasmatic cistern but does not reach the floor of the anterior third ventricle.</td>
</tr>
<tr>
<td>Type B</td>
<td>The tumor reaches the floor of the third ventricle, giving the image of an inverse cupula of the anterior recesses of the third ventricle.</td>
</tr>
<tr>
<td>Type C</td>
<td>A voluminous suprasellar expansion bulges largely into the third ventricle up to the foramen of Monro.</td>
</tr>
<tr>
<td>Type D</td>
<td>Rare aberrant expansions occur in temporal or frontal fossa.</td>
</tr>
</tbody>
</table>

Dichotomized Hardy classification

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sellar invasion</td>
<td></td>
</tr>
<tr>
<td>Grade 0–III</td>
<td>The sellar floor remains entirely or partially intact.</td>
</tr>
<tr>
<td>Grade IV</td>
<td>The sellar floor is completely eroded, with diffuse adenoma invasion into the sphenoid sinus.</td>
</tr>
<tr>
<td>Suprasellar extension</td>
<td></td>
</tr>
<tr>
<td>Type 0–C</td>
<td>The adenoma is confined to the sella or extends superiorly to the level of the foramen of Monro.</td>
</tr>
<tr>
<td>Type D</td>
<td>The adenoma has aberrant expansion into the temporal or frontal fossa.</td>
</tr>
</tbody>
</table>

*Added for the purpose of this study.
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

<table>
<thead>
<tr>
<th>Scale</th>
<th>Sellar invasion(^a)</th>
<th>Suprasellar extension(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reliability (95% CI)</td>
<td>Percent agreement</td>
</tr>
<tr>
<td>Full scale</td>
<td>0.69 (0.51–0.81)</td>
<td>8/50 (16%)</td>
</tr>
<tr>
<td>Intermediate scores</td>
<td>0.15 (– 0.28 to 0.53)</td>
<td>1/23 (4%)</td>
</tr>
<tr>
<td>Scale ends</td>
<td>0.82 (0.63–0.91)</td>
<td>7/27 (26%)</td>
</tr>
<tr>
<td>Dichotomous scale</td>
<td>0.62 (0.41–0.76)</td>
<td>32/50 (64%)</td>
</tr>
</tbody>
</table>

Abbreviation: CI, confidence interval.

\(^a\)Full scale: Grades 0–IV. Dichotomous scale: Grades 0–III versus Grade IV.

\(^b\)Full 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. transphenoidal 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 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 Intrarater reliability by training level

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

Abbreviation: CI, confidence interval.

\(^a\)Full scale: Grades 0–IV. Dichotomous scale: Grades 0–III versus Grade IV.

\(^b\)Full 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. 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. 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 (cranietomy 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. Another limitation of our study is the small number of Type D scans, as detailed earlier. We do not believe that this limitation affects our conclusions.

In conclusion, the Hardy classification system continues to be a mainstay of the preoperative assessment of pituitary adenomas, and our findings support the dichotomization of this scale to improve its reliability in the MRI era. Further studies are needed to extend these findings to reflect the current imaging environment and to determine the impact of dichotomization on surgical decision making and research.

References
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