Endocrine (2019) 63:171–176
https://doi.org/10.1007/s12020-018-1752-8

ENDOCRINE METHODS AND TECHNIQUES

# Are volume measurements of non-functioning pituitary adenomas reliable?

Kristin Astrid Berland Øystese[1,2,3] · Sheren Hisanawi[4] · Manuela Zucknick[5] · Jens Bollerslev[1,2] · Geir Ringstad[2,4]

Received: 31 January 2018 / Accepted: 6 September 2018 / Published online: 26 September 2018


## Abstract

**Purpose** Precise radiological assessment of tumour volume is important in the follow-up of non-functioning pituitary adenomas (NFPAs). We compared the reliability of two methods for tumour volume measurements in the pre- and post-operative setting.

**Methods** We assessed the volume of 22 NFPAs at magnetic resonance imaging (MRI) scans before surgery and the first and third postoperative MRI obtained after submission from hospital. Volumetric assessments were performed both by summation of slices (SOS) and by diameter measures. All volumes were calculated independently by two readers.

**Results** The preoperative intra- and inter-rater reliability was good for both the SOS and the diameter method (intraclass correlation coefficient (ICC) 0.996 and 0.990, and ICC: 0.982 and 0.967, respectively). The first postoperative investigation showed poorer intra- and inter-rater reliability for both methods (ICC: 0.872 and 0.791 and ICC: 0.792 and 0.810, respectively). The third postoperative MRI showed good intra-rater reliability (ICC: 0.961 and 0.962, respectively), but poorer inter-rater reliability for both methods (ICC: 0.759 and 0.703, respectively). Volume assessment by SOS presented overall slightly higher reliability than the diametric method. Overall, the reliability between the two methods was good when measured by the same reader (ICC: 0.988, 0.945 and 0.962, for the preoperative, first and third postoperative MRI, respectively).

**Conclusion** The preoperative intra- and inter-rater reliabilities were satisfactory for both the SOS and diametric method. Postoperative MRI scans showed poorer reliability, suggesting that measurements at these time points should be interpreted with care. For each MRI scan, reliability between methods was satisfactory when investigated by the same reader.

**Keywords** Non-functioning pituitary adenomas · Growth · Tumour volume · Reliability · Magnetic resonance imaging

## Introduction

Pituitary adenomas (PAs) are common intracranial neoplasms, arising from epithelial cells in the anterior pituitary [1]. Approximately half of the tumours are non-functioning pituitary adenomas (NFPAs), not presenting symptoms of hormone overproduction [2]. The main treatment for NFPAs is surgery, with decompression of the optic pathways as the primary indication [3].

A substantial portion of NFPAs regrow after surgery, in particular when residual tumour is present [4]. The tailoring of the postoperative follow-up is for most cases determined by signs of growth, or regrowth, in radiological imaging series [5].

Adenoma size may simply be characterised by its largest diameter in one or more imaging planes [6]. Moreover, tumour volume may be calculated based on diameter measurements or by stereological methods [7, 8]. Only a few

✉ Kristin Astrid Berland Øystese
k.a.b.oystese@medisin.uio.no

1 Section of Specialized Endocrinology, Department of Endocrinology, Oslo University Hospital Rikshospitalet, P.b. 4950 Nydalen, 0424 Oslo, Norway

2 Faculty of Medicine, University of Oslo, Oslo, Norway

3 Research Institute for Internal Medicine (IMF), OUS Rikshospitalet, Postboks 4950 Nydalen, 0424 Oslo, Norway

4 Department of Radiology, Oslo University Hospital Rikshospitalet, P.b. 4950 Nydalen, 0424 Oslo, Norway

5 Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway

series report volumetric measurements of NFPAs [4]. The manual summation of slices (SOS) method of measuring tumour volume, also known as Cavalieri's method, has been considered to be the gold standard of tumour volume measurement, but is more time consuming [9–11].

The issue of tumour growth is frequently encountered in everyday clinical practice. In this study, we aimed to determine the intra- and inter-rater reliability of volumetric tumour measurements based on a diametric and the SOS approach both before and after resection of NFPAs.

## Methods

Twenty-two patients surgically treated for non-functioning pituitary macroadenomas were retrospectively and randomly selected from a pool of patients investigated with serial postoperative tumour volume measurements [12]. Inherent with the retrospective study design, the MRI data were collected from imaging scans obtained both at a tertiary hospital and at other referring hospitals. Forty-three scans were performed with 1.5 Tesla, while 16 were performed with 1.0 T. One and two scans were performed with 0.5 and 3 T, respectively. Acquisition parametres for the T1-weighted scans were typical: repetition time (TR) = 512 ms, echo time (TE) = 12 ms, FoV 178 mm × 220 mm, pixel size = 0.47 and Flip angle 150°. The majority of scans containing residual tumour tissue ($N = 42$) had a slice thickness of 3.3 mm, where the distance between slices amounted 0.3 mm. For the latter scans ($N = 16$), the slice thickness varied between 2.0 and 3.85 mm and distance between slices ranged from 0.0 to 1.65 mm. One scan had a slice thickness of 5 mm, this was however not included in any of the reliability calculations.

All post-scan analyses were done directly in the radiological picture archiving and communication system (PACS) (IDS7, Sectra, Sweden). Volumes were primarily calculated by the SOS method. The tumours were defined and delineated manually by a region of interest (ROI) in all MRI slices where tumour tissue was visible. All ROI areas were summed up and multiplied by the distance between the slices. Diametric measurements were retrieved from the largest diameter in coronal plane, and the two largest perpendicular diameters (height and length) in the sagittal plane. Volume was calculated by the formula width × height × length × 0.5 [8]. Cystic and haemorrhagic tumour components were included in the volume. Tumour fragments were summed in cases where the residual tumour was discontinuous.

Tumour volumes retrieved from MRI at the preoperative (MRI0), the first postoperative after submission from hospital (MRI1) and the third postoperative exam (MRI3) were calculated for all patients. A total of 62 MRI scans were investigated, a total of 58 of these scans were evaluated to have tumour tissue available for volume assessment by one of the readers. Four patients lacked preoperative MRI scans. The median (range) time intervals between MRI and surgery (defined as time point zero) were −3.6 (−0.1 to −11.9), 3.1 (2.1–10.0) and 35.5 (15.3–70.0) months for the MRI0, MRI1 and MRI3, respectively. All tumour volumes were investigated twice by the same reader (KAØ), and once by the second reader (SH).

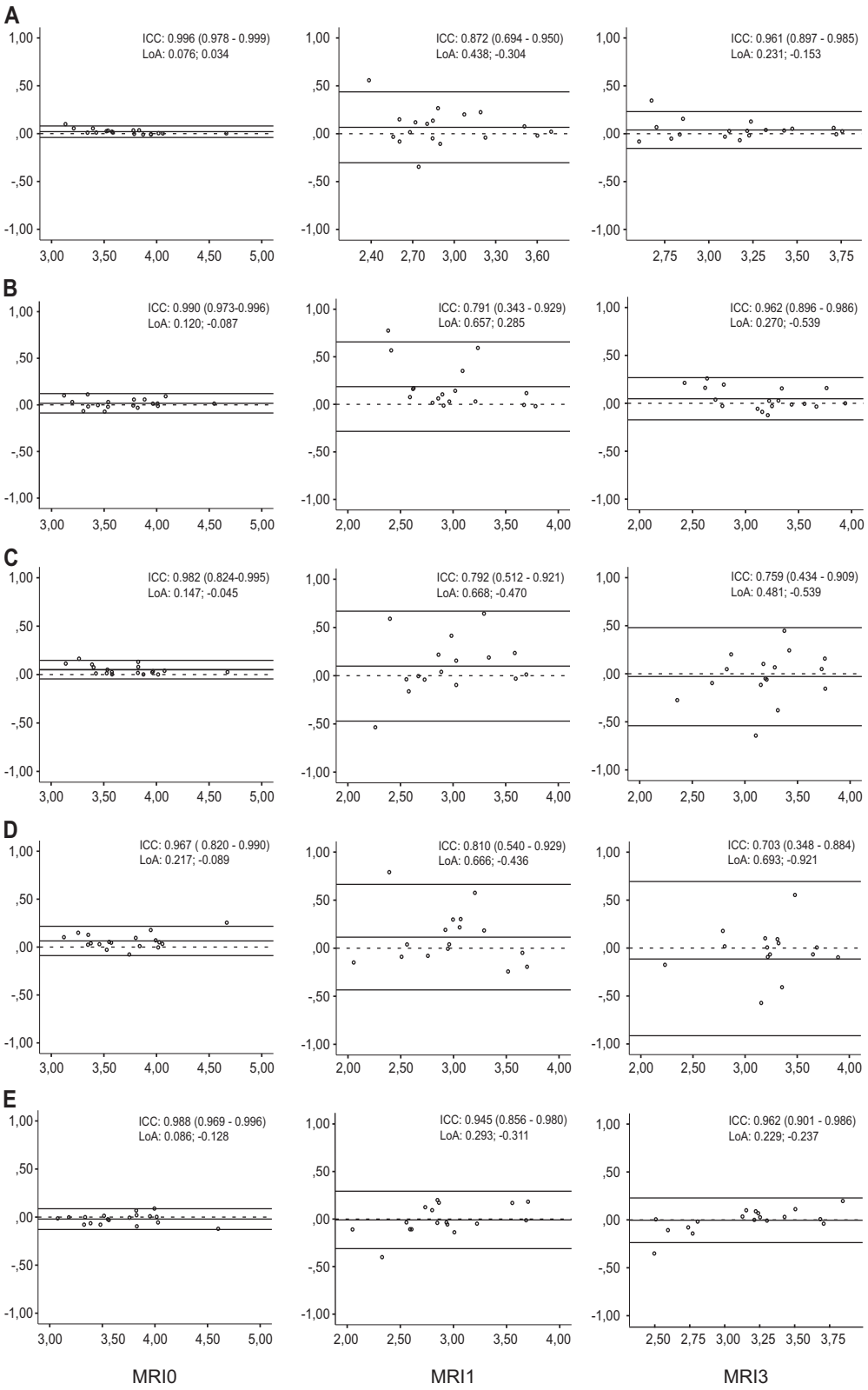
The intraclass correlation coefficient (ICC) is a measure of reliability, taking into account both the degree of correlation and agreement between measurements based on the research model used, the type of measurement protocol and the definition of the relationship (consistency or absolute agreement) [13, 14]. A single measurement mixed-effect model was used to calculate the ICC for the intra-rater comparison, and a single measurement random-effect model was used to calculate the ICC for the inter-rater and the comparison between methods. The ICC was given for the absolute agreement of the log-transformed volumes. Based on the lower bound of the 95% CI of the ICC values, values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9 and above 0.9 were defined as poor, moderate, good and excellent reliability, respectively [15]. The Bland–Altman plot illustrates the bias and the degree of agreement between the two measures compared, agreement intervals (limits of agreement), where 95% of the differences between the first and second measurement fall [16]. All analyses were performed in SPSS software version 24.

## Results

Of 22 patients, four and five patients were considered by reader 1 to not have residual tumour at MRI1 in the first and second measurement, respectively. Four patients were considered not to have residual tumour at MRI1 by reader 2. Readers 1 and 2 agreed on the complete absence of residual tumour in three of these patients. These three patients were not included in the reliability analyses.

### Preoperative MRI (MRI0)

The intra-rater reliability for both SOS and diametric method for the preoperative investigation was excellent (ICC: 0.996 (95% CI: 0.978–0.999), and ICC: 0.990 (95% CI: 0.973–0.996), respectively). The inter-rater reliability gave similar results for the two methods (0.982 (95% CI: 0.824–0.995) and ICC: 0.967 (95% CI: 0.820–0.990), respectively), though with slightly wider 95% limits of agreement (Fig. 1, Column 1).


A
ICC: 0.996 (0.978 - 0.999)
LoA: 0.076; 0.034
ICC: 0.872 (0.694 - 0.950)
LoA: 0.438; -0.304
ICC: 0.961 (0.897 - 0.985)
LoA: 0.231; -0.153
B
ICC: 0.990 (0.973-0.996)
LoA: 0.120; -0.087
ICC: 0.791 (0.343 - 0.929)
LoA: 0.657; 0.285
ICC: 0.962 (0.896 - 0.986)
LoA: 0.270; -0.539
C
ICC: 0.982 (0.824-0.995)
LoA: 0.147; -0.045
ICC: 0.792 (0.512 - 0.921)
LoA: 0.668; -0.470
ICC: 0.759 (0.434 - 0.909)
LoA: 0.481; -0.539
D
ICC: 0.967 ( 0.820 - 0.990)
LoA: 0.217; -0.089
ICC: 0.810 (0.540 - 0.929)
LoA: 0.666; -0.436
ICC: 0.703 (0.348 - 0.884)
LoA: 0.693; -0.921
E
ICC: 0.988 (0.969 - 0.996)
LoA: 0.086; -0.128
ICC: 0.945 (0.856 - 0.980)
LoA: 0.293; -0.311
ICC: 0.962 (0.901 - 0.986)
LoA: 0.229; -0.237
MRI0
MRI1
MRI3

◀ **Fig. 1** Bland–Altman plots showing intra-rater, inter-rater and intra-method variation. Column to the left shows the preoperative MRI scans (MRI0), the middle column shows the earliest MRI scan (MRI1) while the right column shows the third postoperative MRI scan (MRI3). • Row A: Intra-rater variability for summation of slices (SOS) volume measurements. • Row B: Intra-rater variability for diametric volume measurements. • Row C: Inter-rater variability for SOS volume measurements. • Row D: Inter-rater variability for diametric volume measurements. • Row E: Variability between SOS and diametric volume measurements performed by the same reader at the same time point. *X*-axis shows the mean of the two log-transformed volumes, while *Y*-axis shows the difference between the two measurements presented for all Bland–Altman plots. Upper and lower 95% limits of agreement (LoA) is given for all Bland–Altman plots. The stapled lines show a log-transformed volume difference of 0, while the solid lines show the mean difference of the measured volumes for each comparison. The intraclass correlation coefficient (ICC) with 95% confidence interval is given for all comparisons. A single measurement mixed-effect model was used to calculate the ICC for the intra-rater comparison, while a single measurement random-effect model was used to calculate the ICC for the inter-rater and inter-method comparisons [13, 14]

### First postoperative MRI (MRI1)

For the intra-rater reliability, MRI1 was the investigation with the least reliability for both the SOS and diametric method (ICC: 0.872 (95% CI: 0.694–0.950) and ICC: 0.791 (95% CI: 0.343–0.929). The inter-rater reliability was also lower for both methods at this time point (ICC: 0.792 (95% CI: 0.512–0.921) and ICC: 0.810 (95% CI: 0.540–0.929), respectively) than for the preoperative MRI, with wide 95% limits of agreement (Fig. 1, Column 2).

### Third postoperative MRI (MRI3)

The intra-rater reliability was good for the third postoperative MRI, for both measurement methods (ICC: 0.961 (95% CI: 0.897–0.985) and 0.962 (95% CI: 0.897–0.985), respectively), though with wider 95% limits of agreements than at MRI0 (Fig. 1, Column 3).

For the inter-rater comparison, this was the least reliable investigation for both methods (ICC: 0.759 (95% CI: 0.434–0.909) and ICC: 0.703 (95% CI: 0.348–0.884), respectively), with wide 95% limits of agreement (Fig. 1, Column 3).

### Reliability according to method

The SOS method showed approximately equal ICC for most comparisons, with slightly narrower 95% limits of agreement (Fig. 1, Row A vs B, and Row C vs D). An exception was the inter-rater comparison for MRI1, which showed a slightly lower ICC for the SOS compared to the diametric method (ICC: 0.792 (95% CI: 0.512–0.921) and 0.810 (95% CI: 0.540–0.929), respectively. The reliability between the methods when performed during the same investigation by the same reader was excellent, for all three time points (ICC: 0.988 (95% CI: 0.969–0.996), ICC: 0.945 (95% CI: 0.856–0.980) and ICC: 0.962 (95% CI: 0.901–0.986) for MRI0, MRI1 and MRI3, respectively) (Fig. 1, Row E). For the log-transformed volumes at MRI0, MRI1 and MRI3, the mean differences between the diametric and the SOS method was −0.021, −0.01 and −0.003, respectively. This suggests that the SOS method provides slightly larger volume estimates than the diametric method.

## Discussion

Preoperatively, the reliability of the volume assessments presented satisfactory intra- and inter-rater reliability for both SOS and diametric volume measurements. The early postoperative volume measurements (3 months) had the lowest reliability for the intra-rater comparisons, while the third postoperative volume measurements demonstrated the poorest reliability for the inter-rater comparisons. The SOS method provided approximately equal or slightly higher intra- and inter-rater reliability than the diameter-based method for most volume comparisons. There was a high reliability between the SOS and the diametric method.

Most studies reporting intra- or inter-rater comparisons of volume measurements in NFPAs have investigated preoperative MRI investigations, or do not differ between pre- or postoperative investigations [17–19]. Monsalves et al. reported the pre- and postoperative inter-observer consistency for an SOS-related method in pituitary adenomas [20]. They found the preoperative investigations to be more consistent than the postoperative, though both with high consistency [20]. However, the values of consistency were not directly comparable to values of absolute agreement used in the present ICC calculation, while they compared average scores in a group and not the scores of each subject [14]. Our results thus add to these previous reports showing that both the SOS and diametric methods are reliable for volumetric tumour analysis at preoperative MRI investigations.

The intra-rater reliability of volumetric measurements was poorer for both measurement methods at the first postoperative MRI compared to the later postoperative MRI in our study. The blood and secretions from surgery are resorbed during the first 2–3 months after surgery, hence the first postoperative MRI scan may typically be advised carried out after such early postoperative changes have subsided [21]. However, some investigators have found postoperative volume assessments in other intracranial tumours to have robust intra- and inter-rater reliability though with semi-automated methods [22]. The literature on

the precision of volume measurements for postoperative investigations in NFPAs is sparse. The fact that there was disagreement between the readers about the residual tumour presence in some cases, but also within the same investigator, indicates that interpretation of this early postoperative MRI scan is challenging. The ICC for the diametric method in MRI1 had a wider CI than the other intra-rater comparisons. This variation might have been caused by a limited number of tumours in the analysis (Fig. 1, Row B, Column 2).

A substantial portion of NFPAs regrow postoperatively [4, 12]. We therefore assumed that the tumours would be easier to delineate and the reliability better at MRI3 than at MRI1. This was the case for the intra-rater comparison for both the SOS and the diametric method. However, this was not the case for the inter-rater comparison (Fig. 1, Column 3). Monsalves et al. also found the same pattern, when reporting inter-rater reliability in PAs measured before and after surgery [20]. In our study, each reader investigated MRI scans from the three time points (MRI0, MRI1 and MRI3) serially, and therefore the tumour delineation at MRI1 has probably affected the delineation at MRI3.

The SOS method has been shown to have less retest error than other measures of size [10, 23], and in our study, the repeatability of this test (ICC) was slightly higher than for the diametric method. For the intra-rater comparison, the SOS method demonstrated narrower 95% limits of agreement compared to the diametric method for both methods. However, the ICC scores were quite similar for most analyses. The SOS method was superior, or approximately equal, to the diametric approach for the inter-rater comparisons, though the postoperative analyses appeared to be challenging. Our results demonstrated good agreement between the methods within the same reader during the same investigation session. The diametric method could therefore be used for serial investigations performed by the same reader when detection of tumour volume change is the main issue; however, the most optimal method in regard to reliability seems to be the SOS.

### Limitation

The study design was retrospective, and hence lacked a standardisation of the modalities and timing of the MRI scans, this might have introduced bias in the results. The lack of tumour tissue in four and five (in accordance to readers 1 and 2, respectively) of the postoperative MRI scans reduced the cohort size and thus accuracy of the reliability estimation. The variation of the measurements was greater on the postoperative MRI scans than on the preoperative MRI scans, and hence a larger number of investigation subjects would have improved the precision of the estimates. All measurements by reader 1 were done within a time span of 2 months. There was however a shorter period between the two measurements of MRI0, than the two measurements of MRI1 and MRI3, which possibly could affect the intra-rater comparisons. However, all image annotations from the first measurements were erased before the second measurements were started.

## Conclusion

Non-functioning pituitary adenoma volume measurements were highly reliable in the preoperative setting when assessed by both SOS and the diametric approach. The reliability of both methods was poorer for the postoperative measurements, particularly at the first postoperative scan, suggesting that these investigations should be interpreted with caution. The SOS method showed equal, or slightly better, intra- and inter-reliability than the diametric volume measurements for most comparisons. However, the reliability between the methods, when performed by the same reader, showed good reliability.

**Acknowledgements** The study was approved by the regional ethics committee and hospital authority. The work was funded by the South-Eastern Norway Regional Health Authority, Award number 2016026.

### Compliance with ethical standards

**Conflict of interest** J.B. is a member of the advisory board of Endocrine. The remaining authors declare that they have no conflict of interest.

**Informed consent** Informed consent was obtained from all living patients.

## References

1. S.L. Asa, Practical pituitary pathology: what does the pathologist need to know? Arch. Pathol. Lab. Med. **132**(8), 1231–1240 (2008). https://doi.org/10.1043/1543-2165(2008)132[1231:pppwdt]2.0.co;2
2. A. Tjornstrand, K. Gunnarsson, M. Evert, E. Holmberg, O. Ragnarsson, T. Rosen, H. Filipsson Nystrom, The incidence rate of pituitary adenomas in western Sweden for the period 2001-2011. Eur. J. Endocrinol. **171**(4), 519–526 (2014). https://doi.org/10.1530/eje-14-0144
3. P.U. Freda, A.M. Beckers, L. Katznelson, M.E. Molitch, V.M. Montori, K.D. Post, M.L. Vance, Pituitary incidentaloma: an endocrine society clinical practice guideline. J. Clin. Endocrinol. Metab. **96**(4), 894–904 (2011). https://doi.org/10.1210/jc.2010-1048
4. Y. Chen, C.D. Wang, Z.P. Su, Y.X. Chen, L. Cai, Q.C. Zhuge, Z. B. Wu, Natural history of postoperative nonfunctioning pituitary adenomas: a systematic review and meta-analysis. Neuroendocrinology **96**(4), 333–342 (2012). https://doi.org/10.1159/000339823

5. C. Cortet-Rudelli, J. F. Bonneville, F. Borson-Chazot, L. Clavier, B. Coche Dequeant, R. Desailloud, D. Maiter, V. Rohmer, J. L. Sadoul, E. Sonnet, P. Toussaint, P. Chanson, Post-surgical management of non-functioning pituitary adenoma. Ann. Endocrinol. (2015). https://doi.org/10.1016/j.ando.2015.04.003
6. E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, J. Verweij, New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur. J. Cancer **45**(2), 228–247 (2009). https://doi.org/10.1016/j.ejca.2008.10.026
7. H.J. Gundersen, T.F. Bendtsen, L. Korbo, N. Marcussen, A. Moller, K. Nielsen, J.R. Nyengaard, B. Pakkenberg, F.B. Sorensen, A. Vesterby et al. Some new, simple and efficient stereological methods and their use in pathological research and diagnosis. APMIS Scand. **96**(5), 379–394 (1988)
8. G. Di Chiro, K.B. Nelson, The volume of the sella turcica. Am. J. Roentgenol. Radium Ther. Nucl. Med. **87**, 989–1008 (1962)
9. G.A. Ringstad, K.E. Emblem, D. Holland, A.M. Dale, A. Bjornerud, J.K. Hald, Assessment of pituitary adenoma volumetric change using longitudinal MR image registration. Neuroradiology **54**(5), 435–443 (2012). https://doi.org/10.1007/s00234-011-0894-7
10. J.K. Varughese, T. Wentzel-Larsen, F. Vassbotn, G. Moen, M. Lund-Johansen, Analysis of vestibular schwannoma size in multiple dimensions: a comparative cohort study of different measurement techniques. Clinical otolaryngology: official journal of ENT-UK. Off. J. Neth. Soc. Oto-Rhino-Laryngol. Cervico-Facial Surg. **35**(2), 97–103 (2010). https://doi.org/10.1111/j.1749-4486.2010.02099.x
11. S. Bauer, R. Wiest, L.P. Nolte, M. Reyes, A survey of MRI-based medical image analysis for brain tumor studies. Phys. Med. Biol. **58**(13), R97–R129 (2013). https://doi.org/10.1088/0031-9155/58/13/r97
12. K.A. Oystese, M. Zucknick, O. Casar-Borota, G. Ringstad, J. Bollerslev, Early postoperative growth in non-functioning pituitary adenomas; a tool to tailor safe follow-up. Endocrine **57**(1), 35–45 (2017). https://doi.org/10.1007/s12020-017-1314-5
13. P.E. Shrout, J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. **86**(2), 420–428 (1979)
14. K.O. McGraw, S.P. Wong, "Forming inferences about some intraclass correlations coefficients": correction. Psychol. Methods **1**(4), 390–390 (1996). https://doi.org/10.1037/1082-989X.1.4.390
15. T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. **15**(2), 155–163 (2016). https://doi.org/10.1016/j.jcm.2016.02.012
16. J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement. Lancet **1**(8476), 307–310 (1986)
17. B.M. Davies, E. Carr, C. Soh, K.K. Gnanalingham, Assessing size of pituitary adenomas: a comparison of qualitative and quantitative methods on MR. Acta Neurochir. **158**(4), 677–683 (2016). https://doi.org/10.1007/s00701-015-2699-7
18. J. A. Balogun, E. Monsalves, K. Juraschka, K. Parvez, W. Kucharczyk, O. Mete, F. Gentili, G. Zadeh, Null cell adenomas of the pituitary gland: an institutional review of their clinical imaging and behavioral characteristics. Endocrine pathology (2014). https://doi.org/10.1007/s12022-014-9347-2
19. N. Lenders, S. Ikeuchi, A. W. Russell, K. K. Ho, J. B. Prins, W. J. Inder, Longitudinal evaluation of the natural history of conservatively managed non-functioning pituitary adenomas. Clin Endocrinol (2015). https://doi.org/10.1111/cen.12879
20. E. Monsalves, S. Larjani, B. Loyola Godoy, K. Juraschka, F. Carvalho, W. Kucharczyk, A. Kulkarni, O. Mete, F. Gentili, S. Ezzat, G. Zadeh, Growth patterns of pituitary adenomas and histopathological correlates. J. Clin. Endocrinol. Metab. **99**(4), 1330–1338 (2014). https://doi.org/10.1210/jc.2013-3054
21. J. F. Bonneville, F. Bonneville, E. Barrali, G. Jacquet, F. Cattin, *Functional and Morphological Imaging of the Endocrine System*. ed. by W. W. De Herder (Springer US, Boston, 2000) pp. 3–33
22. T. Sankar, N.Z. Moore, J. Johnson, L.S. Ashby, A.C. Scheck, W. R. Shapiro, K.A. Smith, R.F. Spetzler, M.C. Preul, Magnetic resonance imaging volumetric assessment of the extent of contrast enhancement and resection in oligodendroglial tumors. J. Neurosurg. **116**(6), 1172–1181 (2012). https://doi.org/10.3171/2012.2.jns102032
23. A.G. Sorensen, S. Patel, C. Harmath, S. Bridges, J. Synnott, A. Sievers, Y.H. Yoon, E.J. Lee, M.C. Yang, R.F. Lewis, G.J. Harris, M. Lev, P.W. Schaefer, B.R. Buchbinder, G. Barest, K. Yamada, J. Ponzo, H.Y. Kwon, J. Gemmete, J. Farkas, A.L. Tievsky, R.B. Ziegler, M.R. Salhus, R. Weisskoff, Comparison of diameter and perimeter methods for tumor volume calculation. J. Clin. Oncol. **19**(2), 551–557 (2001). https://doi.org/10.1200/jco.2001.19.2.551