ANALYSIS OF INTER-EXPERT AGREEMENT WHEN WORKING WITH THE AOSPINE CLASSIFICATION (TLCS, 2013): OUR EXPERIENCE, QUESTIONS AND CONTRADICTIONS

A.V. Dydykin, K.S. Yashin, A.E. Bokov, S.G. Mlyavykh
Privolzhsky Research Medical University, Nizhny Novgorod, Russia

Objective. To assess inter-expert agreement among spine surgeons having different levels of clinical experience when working with the AOSpine classification (TLCS, 2013).

Materials and Methods. The study involved nine surgeons divided into three equal groups depending on work experience. All respondents were asked to classify the MSCT data of 50 patients with acute injuries to the thoracic and lumbar spine pursuant to TLCS (2013) classification. To evaluate inter-expert agreement, a Kappa coefficient interpreted according to Landis – Koch criteria was used.

Results. The overall coefficient of inter-expert agreement for all observations among all groups of respondents was 0.43, which reflects a moderate level of agreement. Moderate inter-expert agreement was revealed for injury types A (0.45) and C (0.56), and satisfactory – for type B (0.34). The highest levels of agreement were obtained for subtypes A1 (0.67) and A4 (0.80) in the group of advanced specialists and for type C (0.70) in the group of specialists with a basic level of experience.

Conclusion. The study demonstrated predominantly moderate level of inter-expert agreement when working with the AOSpine classification (TLCS, 2013). The accuracy of its use increases with a gain in practical experience of a surgeon.

Key Words: AO Spine TLCS 2013 classification, injuries to the thoracic and lumbar spine, inter-expert agreement.

Please cite this paper as: *Dydykin AV, Yashin KS, Bokov AE, Mlyavykh SG. Analysis of inter-expert agreement when working with the AOSpine classification (TLCS, 2013): our experience, questions and contradictions. Hir. Pozvonoc. 2019;16(4):13–20. In Russian.*

DOI: http://dx.doi.org/10.14531/ss2019.4.13-20.

Spinal injuries account for more than 20 % of total injuries [1]; of these, thoracic and lumbar spine injuries amount to 50–90 % [2, 3]. Over the past decade, the percentage of spinal cord injuries has increased tenfold, with severe injuries prevailing over minor ones, which is naturally explained by growth of industrial-transport parameters [4, 5].

In this regard, there is an obvious need for a spinal injury classification that optimizes the choice of a treatment approach and predicts treatment outcomes. Furthermore, the classification system should be useful for research purposes and be a tool for communication and training. To date, no uniform classification of spinal injuries has been finally developed [6, 7].

The first attempts to classify thoracic and lumbar spine injuries were made by Böhler in 1929. Since then, more than 10 different classifications of spinal injuries have been proposed, which are based on the anatomy (Böhler and Nicoll, 1949), morphology (Watson-Jones, 1938; AO/Magerl, 1994), injury mechanism (Böhler, Ferguson, Allen, 1984), two column stability concept (Holdsworth, 1963; Kell, Whitesides, 1968), three-column approach (Denis, 1983; Gertzbein, 1992; McAfee, 1993), and scoring assessment system (McCormack, Gaines, 1994; TLISS/TLICS, 2005) [7, 8].

Practice has demonstrated that each classification system has its own advantages and disadvantages. For example, the classifications proposed by Böhler, Watson-Jones, Nicoll, Holdsworth, Kelly, and Whitesides were, on the one hand, simple, easy to use, enabled to some extent choosing a therapeutic approach, but, on the other hand, were not validated and prognostic and were mostly descriptive. The advantage of the Denis classification is its simplicity, widespread use, and reflection of some relation between neurological disorders and stability principles. However, it does not enable clear distinguishing between stable and unstable injuries and predicting the outcome of treatment. Furthermore, the Denis classification is characterized by low inter-expert reliability. The Ferguson and Allen classification, being comprehensive, also does not predict the outcome of treatment and is not validated. The McAfee classification has not been widely used and is not validated. The recognized comprehensive AO/Magerl classification that enables determining the severity and stability of injury is, however, considered extremely complex and has moderate inter-expert reliability. The TLISS/TLICS classification can be used to assess stability of injury and predict the outcome of treatment. At the same time, indicators of inter-expert reliability upon application of TLISS/TLICS demonstrate moderate levels [8–10].

The AOSPine classification (TLCS, 2013) was developed by a large team of authors, with allowance for the advantages and disadvantages of previously used classification systems, primarily as a tool to standardize treatment approaches for spinal injuries. It was based on the two most popular recent classifications: AO/Magerl and AO TLISS/TLICS. The main features of the new AOSpine classification (TLCS) include a reduced number of spinal injury variants (9 patterns compared to 53 types proposed by Magerl) and the emphasis on integrity of the posterior ligamentous complex, rather than on the three column concept. The injury assessment algorithm is changed from more severe (type C) to simple (type A) damage, but not towards increasing complexity of structural damage (from A to C). The authors have also taken into account the fact that MRI, as a neuroimaging method, is not as available and widespread as CT. Therefore, assessment of injury morphology is based mainly on CT findings. Structurally, the TLCS classification includes assessment of injury morphology (compression — type A, distraction — type B, dislocation — type C), neurological status, and clinical modifiers [8, 11, 14].

The most controversial component of the AOSpine classification is assessment of injury morphology. In this case, assessment of the neurological status and use of clinical modifiers does not raise questions [12].

The results of several international studies do not provide a final conclusion about reliability of the AOSpine classification for systematization of spinal injuries [11–14]. In addition, Russian experts have not conducted a similar study yet.

This study objective was to assess practical application of the morphological component of the AOSpine classification (TLCS, 2013) for thoracic and lumbar spine injuries and analyze inter-expert agreement among spinal neurosurgeons with different levels of clinical experience.

Material and Methods

Study design. CT data (Dicom catalog) of patients with thoracic and/or lumbar spine injuries, who were treated in the Neurosurgical Department in 2014–2017, were retrospectively evaluated. At the first stage, respondent surgeons were asked to classify identified injuries according to the AOSpine classification (TLCS, 2013). At the second stage, a group analysis of the results and a re-analysis of clinical cases that caused disagreement in their classification were performed.

Respondents. The study involved 9 surgeons: 6 employees of the Neurosurgical Department and 3 first and second year resident surgeons. According to the level of training, the surgeons were divided into 3 equal groups: initial level – less than two years of clinical practice; basic level – 3–5 years of clinical practice; advanced level – more than seven years of clinical practice.

Clinical material. The study included MSCT data from 50 patients with acute high-impact injury to the thoracic and lumbar spine. All examinations were performed on a Philips Aquillion 32 multispiral computed tomograph (Japan). CT findings (Dicom data) were analyzed using freely distributed RadiAnt Dicom Viewer software enabling multiplanar (MPR) and 3D reconstruction.

Statistical analysis. The obtained data were analyzed using descriptive statistics methods. A Kappa coefficient (k) was used to assess inter-expert agreement. Kappa coefficient values were interpreted according to the Landis – Koch criteria: a value of less than 0.20 – poor agreement; 0.21 to 0.40 – satisfactory agreement; 0.41 to 0.60 – moderate agreement; 0.61 to 0.80 – good agreement; more than 0.81 – very good agreement [15].

Inter-expert agreement was analyzed in two stages: first, Kappa was determined for the type of fracture as a whole and separately within each group of respondents, and then, similarly, according to injury subtypes.

MS Office Excel 2007 and Statistica 10 software products were used.

Results

According to the MSCT data, 76 injuries were identified and classified in 50 patients included in the study: 51 injuries in the thoracic spine and 25 injuries in the lumbar spine. Each injury was further considered as a separate case.

There were no errors in determining the level of injury.

First of all, the percentage ratio of opinion agreements and disagreements within each group of respondents (Table 1) and the total percentage indicators (Table 2) of injury classification were calculated.

Preliminary data showed that the opinion on injury type classification within each respondent group coincided in more than 50 % of cases and reached 63.1 %. Complete divergence of opinions did not exceed 27.6%. However, unanimous identification of the injury type by all respondents amounted to 35.5 % only, and complete disagreement reached 47.3 %.

Next, we assessed inter-expert agreement using the Kappa coefficient.

The total Kappa agreement coefficient of inter-expert agreement for all cases among all groups of respondents was 0.43, which indicates moderate agreement. The Kappa coefficient was 0.45, 0.34, and 0.56 for type A, B, and C injuries, respectively. The obtained data demonstrate moderate inter-expert agreement for type A and C injuries and satisfactory agreement for type B injuries.

The total inter-expert agreement value increased depending on the level of specialist training. It amounted to 0.20 in the initial level group, which indicates poor agreement; it was 0.51 (moderate agreement) in the basic group and 0.63 (good inter-expert agreement) in the advanced group.

Upon assessing types of injury, agreement within respondent groups ranged from values indicating poor (0.03 – type B, initial group) agreement to good (0.75 – type A, advanced group) agreement.

Total Kappa values and injury type values are presented in Table 3.

Inter-expert agreement indicators for injury subtypes varied widely and depended more on the type of injury than on the respondent's qualification.

Poor agreement was observed upon classification of subtype A3 injuries in the initial and basic groups, subtype B1 injuries in the initial level group, and subtype B2 injuries in the initial and advanced groups. Satisfactory agreement was found for subtype A1 injuries in the initial group, subtype B1 injuries in the advanced user group, and subtype

Table 1

Agreement of opinions within each group of respondents, %

Classification of injury	Level of respondent's skills					
	initial		basic		advanced	
	agreement	disagreement	agreement	disagreement	agreement	disagreement
By type	57.9 (n = 44)	27.6 (n = 21)	56.5 (n = 43)	5.2 (n = 4)	63.1 (n = 48)	1.3 (n = 1)
By type and subtype	26.3 (n = 20)	35.5 (n = 27)	38.0 (n = 29)	18.4 (n = 14)	48.6 (n = 37)	6.5 (n = 5)

B2 injuries in the basic group. Moderate agreement was observed for type C and subtype A3 injuries in the advanced group, subtype A1 injuries in the initial and basic groups, subtype A4 injuries in the initial and basic groups, and subtype B1 injuries in the basic group. Good inter-expert agreement was revealed for subtype A1 and A4 injuries in the advanced group and type C injuries in the initial and basic groups of respondents.

Kappa values for all injury subtypes are presented in Table 4.

Thus, the highest inter-expert agreement indicators were associated with classification of A1, A4, and C injuries.

For example, upon classification of subtype A1 injuries, identification of typical injury to a single endplate without affecting the integrity and height of the posterior vertebral body did not cause difficulties and disagreements among respondents (Fig. 1).

Upon classification of subtype A4 injuries, identification of typical injury to both endplates with destruction of the posterior vertebral body was also not questioned by most respondents (Fig. 2).

Identification of displacement, dislocation, or translation in scans allowed all respondents to accurately classify type C injuries (Fig. 3).

Poor and moderate agreement was most often observed in cases of subtype A3, B1, and B2 injuries (inter-expert agreement coefficients were 0.24, 0.26, and 0.04, respectively). The inter-expert agreement indicator was found to be higher upon classification of polar injuries, which are less serious and most severe (Fig. 4).

The results of this study demonstrate that, in general, inter-expert agreement upon application of the AOSpine classification (TLCS, 2013) has an acceptable level that consistently elevates as the clinical experience of the surgeon increases (Fig. 5).

At the next stage of this study, MSCT data of the most controversial clinical cases were re-analyzed. In this case, all respondents jointly reviewed and analyzed Dicom data in strict accordance with the morphological classification algorithm proposed by AO experts (www.aospine.org/TLclassification). This revealed a number of errors made by respondents during classification of injuries. A vertebral body fracture, not injury as a whole, was most often evaluated; translation of the superjacent vertebra was considered as crushing of the subjacent vertebral body with loose fragments; the degree of spinous process separation and a reduction in the posterior endplate height were underestimated. Also, ignorance of the frontal projection in assessing the overall picture of injury was marked.

Discussion

The need to develop and implement a new classification of spinal injuries is dictated by the imperfection of previously used ones as evidenced by numerous publications. The most modern classification system is that of thoracolumbar injuries proposed by AOSpine in 2013. Despite the fact that the authors took into account the advantages and disadvantages of previous classification systems during development of the TLCS, there are still disputes in the medical community regarding its practical suitability.

We retrospectively examined 76 cases of thoracic and lumbar spine injuries and classified them according to the TLCS (2013). Practical applicability of the classification was based on assessment of inter-expert agreement (Kappa coefficient) in groups of specialists with different levels of clinical experience.

The total inter-expert agreement indicator was 0.45 for type A injuries, 0.34 for type B injuries, and 0.56 for type C injuries. These findings do not contradict previously published data.

The developers of TLCS (2013) demonstrated predominantly good and moderate inter-rater agreement indicators: 0.72 for type A, 0.58 for type B, and 0.70 for type C [11]. However, it should be borne in mind that the study involved medical experts in spinal surgery and the developers of this classification.

Later, primarily good inter-expert agreement was shown by Kepler et al. [14] in a large-scale multicenter study involving 100 leading spinal surgeons from around the world. The inter-expert agreement value was 0.80 for type A inju-

Table 2

Total indicators of respondent's opinion agreement, %

Classification of injury	Opinion agreement	Opinion disagreement
By type	35.5 (n = 27)	47.3 (n = 36)
By type and subtype	11.8 (n = 9)	47.3 (n = 36)

Table 3

Inter-expert agreement Kappa (k) in groups of respondents classifying the type of injury

Level of respondent's skills	Injury type			Total k for all injury types
	A	B	C	
Initial	0.19	0.03	0.61	0.20
Basic	0.54	0.40	0.70	0.51
Advanced	0.75	0.56	0.54	0.63
Total k for all groups of respondents	0.45	0.34	0.56	0.43

Table 4

Inter-expert agreement Kappa upon classification of injury subtypes in all groups of respondents

Injury type	Level of respondent's skills			Total indicator
	initial	basic	advanced	
A0	Not found			
A1	0.37	0.60	0.67	0.45
A2	Not found			
A3	0.20	0.12	0.50	0.24
A4	0.50	0.49	0.80	0.63
B1	0.14	0.53	0.22	0.26
B2	0.15	0.23	0.17	0.04
B3	Not found			
C	0.61	0.70	0.54	0.56

Fig. 1
CT scans of a 60-year-old male patient with a compression fracture of the L2 vertebral body (morphological subtype A1)

ries, 0.68 for type B injuries, and 0.72 for type C injuries.

On the other hand, in a study by Cheng et al. [13], which involved also young specialists, the total k values were generally lower: 0.38 for type A injuries, 0.29 for type B injuries, and 0.55 for type C injuries, which indicates satisfactory and moderate inter-expert agreement.

In a study by Kaul et al. [12], k values for injury types ranged from 0.43 to 0.59, which indicates moderate inter-expert agreement. The authors emphasize the need for further studies on the use of TLCS (2103) by doctors with different levels of training.

Therefore, the obtained data and the results of international studies demonstrate that inter-expert agreement increases depending on the clinical experience of the surgeon and is directly dependent on this fact. The percentage difference in opinions upon classification of injury types and subtypes is directly related to the level of respondents. The maximum indicators of opinion agreement were detected in the group of advanced specialists upon classification of injury types – 63.1 %, and upon classification of types and subtypes – 48.6 %. However, these indicators demonstrate that, even in this group, the greatest disagreements arose upon assessing injury subtypes, mainly type B, which does not contradict the Kappa coefficient values for type B (the coefficient was 0.03 in the initial group, 0.40 in the basic group, and 0.56 in the advanced group).

A joint group analysis of the entire sample of clinical cases revealed a number of typical errors made upon classification of injuries:

1) assessment of the fracture, but not injury as a whole: often, respondents considered segmental type B and C injuries as separate injuries of the vertebral bodies and, respectively, classified them as type A injuries;

2) initial translation of the superjacent vertebra was assessed as crushing of the subjacent vertebral body with loose fragments, resulting in underestimation of injury severity: type C injuries were classified as type A injuries;

3) underestimating the degree of spinous process separation, leading to classification of type B injuries as type A injuries;

4) underestimating the significance of posterior endplate height, which led to incorrect classification as subtype A injury;

5) ignoring the frontal projection, which led to incorrect classification of subtype B injuries.

An analysis of the typical errors showed that they were often caused by a premature conclusion based on analysis of standard scans, underestimation of the capabilities of multiplanar reconstruction, and a departure from the AOSpine-recommended algorithm for classification of the morphological type of injury.

In our opinion, the optimal assessment of CT data is a sequential analysis of the sagittal, and then axial and coronary planes.

Fig. 2
CT scans of a 24-year-old male patient with a L1 vertebral fracture (morphological subtype A4)

Fig. 3
CT scans of a 49-year-old female patient with injury at the T3–T4 level (morphological type C)

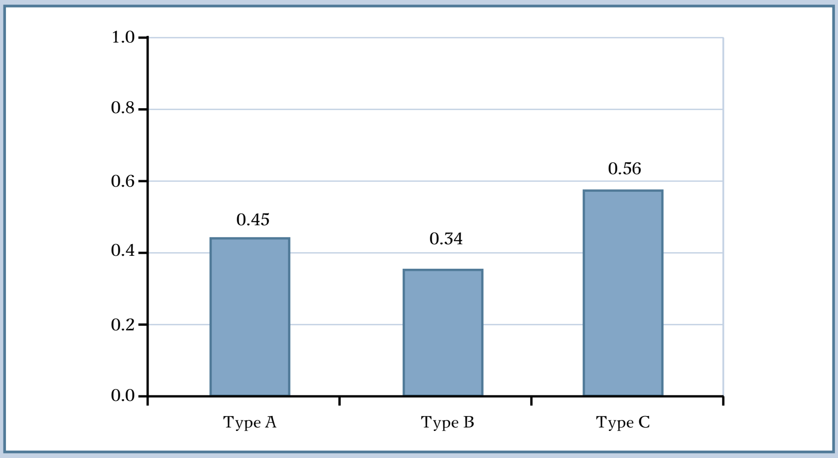

Fig. 4
Total value of inter-expert agreement (Kappa) in various injury types

Conclusion

The search for a spinal injury classification that is able to optimize the choice of a treatment approach, predict the outcome of treatment, and be useful for research purposes is the subject of discussion among spinal surgeons. Our study evaluated inter-expert agreement among spinal surgeons with different levels of practical experience upon application of the AOSpine classification (TLCS, 2013). Inter-expert agreement was moderate upon classification of type A and C injuries and satisfactory upon classification of type B injuries. The accuracy of classification implementation is significantly affected by practical experience of the specialist. Our findings suggest that the TLCS classification (2013) can be used in clinical practice, but only partially applied for research purposes.

The study was conducted without financial support.
The authors declare no conflict of interest.

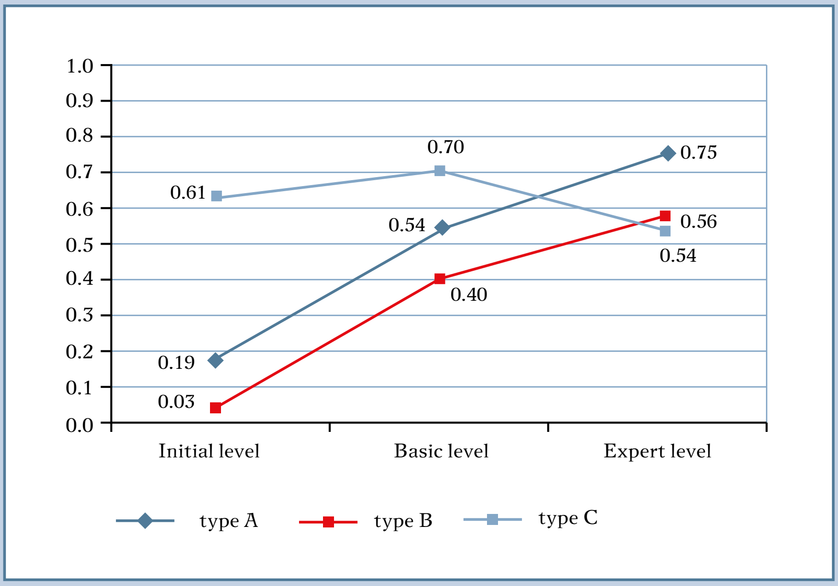

Fig. 5
Changes in inter-expert agreement (Kappa) indicators for different injury types depending on the level of respondent's skills

References

1. **Morozov IN, Mlyavykh SG.** The epidemiology of vertebral-cerebrospinal trauma: review. Meditsinskyi Almanakh. 2011;(4):157–159. In Russian.
2. **Alpantaki K, Bano A, Pasku D, Mavrogenis AF, Papagelopoulos PJ, Sapkas GS, Korres DS, Katonis P.** Thoracolumbar burst fractures: a systematic review of management. Orthopedics. 2010;33:422–429. DOI: 10.3928/01477447-20100429-24.
3. **DeWald RL.** Burst fractures of the thoracic and lumbar spine. Clin Orthop Relat Res. 1984;(189):150–161.
4. **Furlan, JC, Sakakibara BM, Miller WC, Krassioukov AV.** Global incidence and prevalence of traumatic spinal cord injury. Can J Neurol Sci. 2013;40:456–464. DOI: 10.1017/s0317167100014530.
5. **Singh A, Tetreault L, Kalsi-Ryan S, Nouri A, Fehlings MG.** Global prevalence and incidence of traumatic spinal cord injury. Clin Epidemiol. 2014;6:309–331. DOI: 10.2147/CLEP.S68889.
6. **Kolesov SV, Ptashnikov DA, Shvets VV.** Injuries to the Spinal Cord and Spine, ed. by S.P. Mironov. Moscow, 2018. In Russian.
7. **Cassar-Pullicino VN, Imhof H.** Spinal Trauma – An Imaging Approach. Transl. under the general edition of Prof. Sh. Sh. Shotemor. Moscow, 2009. In Russian.
8. **Schroeder GD, Harrop JS, Vaccaro AR.** Thoracolumbar trauma classification. Neurosurg Clin N Am. 2017;28:23–29. DOI: 10.1016/j.nec.2016.07.007.
9. **Patel AA, Vaccaro AR.** Thoracolumbar spine trauma classification. J Am Acad Orthop Surg. 2010;18:63–71.
10. **Sethi MK, Schoenfeld AJ, Bono CM, Harris MB.** The evolution of thoracolumbar injury classification systems. Spine J. 2009;9:780–788. DOI: 10.1016/j.spinee.2009.04.003.
11. **Vaccaro AR, Oner C, Kepler CK, Dvorak M, Schnake K, Bellabarba C, Reinhold M, Aarabi B, Kandziora F, Chapman J, Shanmuganathan R, Fehlings M, Vialle L.** AOSpine thoracolumbar spine injury classification system: fracture description, neurological status and key modifiers. Spine. 2013;38:2028–2037. DOI: 10.1097/BRS.0b013e3182a8a381.
12. **Kaul R, Chhabra HS, Vaccaro AR, Abel R, Tuli S, Shetty AP, Das KD, Mohapatra B, Nanda A, Sangondimath GM, Bansal ML, Patel N.** Reliability assessment of AOSpine thoracolumbar spine injury classification system and Thoracolumbar Injury Classification and Severity Score (TLICS) for thoracolumbar spine injuries: results of a multicentre study. Eur Spine J. 2017;26:1470–1476. DOI: 10.1007/s00586-016-4663-5.
13. **Cheng J, Liu P, Sun D, Qin T, Ma Z, Liu J.** Reliability and reproducibility analysis of the AOSpine thoracolumbar spine injury classification system by Chinese spinal surgeons. Eur Spine J. 2017;26:1477–1482. DOI: 10.1007/s00586-016-4842-4.
14. **Kepler CK, Vaccaro AR, Koerner JD, Dvorak MF, Kandziora F, Rajasekaran S, Aarabi B, Vialle LR, Fehlings MG, Schroeder GD, Reinhold M, Schnake KJ, Bellabarba C, Cumhur On**er F. Reliability analysis of the AOSpine thoracolumbar spine injury classification system by a worldwide group of naive spinal surgeons. Eur Spine J. 2016;25:1082–1086. DOI: 10.1007/s00586-015-3765-9.
15. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174. DOI: 10.2307/2529310.

Address correspondence to:
Dydykin Andrey Vladimirovich
Privolzhsky Research Medical University,
10/1, Minina I Pozharskogo sq., Nizhny Novgorod, 603950, Russia,
adydykin@yandex.ru

Received 01.04.2019
Review completed 26.06.2019
Passed for printing 01.07.2019

Andrey Vladimirovich Dydykin, MD, PhD, neurosurgeon, junior researcher in the Group of Spine and Spinal Cord Pathology, Privolzhsky Research Medical University, 10/1, Minina I Pozharskogo sq., Nizhny Novgorod, 603950, Russia, ORCID: 0000-0003-4824-9481, adydykin@yandex.ru;
Konstantin Sergeyevich Yashin, MD, PhD, neurosurgeon, Department of Oncology and Neurosurgery (Cerebral Unit), Privolzhsky Research Medical University, 10/1, Minina I Pozharskogo sq., Nizhny Novgorod, 603950, Russia, ORCID: 0000-0002-5723-7389, jashinkostya@gmail.com;
Andrey Evgenievich Bokov, MD, PhD, neurosurgeon, Head of Neurosurgery Department, Trauma and Orthopedic Institute, University Clinic, Privolzhsky Research Medical University, 10/1, Minina I Pozharskogo sq., Nizhny Novgorod, 603950, Russia, ORCID: 0000-0002-5203-0717, andrei_bokov@mail.ru;
Sergey Gennadievich Mlyavykh, MD, PhD, neurosurgeon, Head of the Trauma and Orthopedic Institute, University Clinic, Privolzhsky Research Medical University, 10/1, Minina I Pozharskogo sq., Nizhny Novgorod, 603950, Russia, ORCID: 0000-0002-6310-4961, spinedoc@bk.ru.