Brain and Spine 5 (2025) 104325

Contents lists available at ScienceDirect

Brain and Spine

journal homepage: www.journals.elsevier.com/brain-and-spine

Diagnostic reliability of the FinSpine nationwide spine surgery registry: Comparison between clinical registry diagnoses and blinded imaging assessments

Jussi P. Repo [a,*], Katri Pernaa [b], Liisa Pekkanen [c], Juho Hatakka [b], Heikki Mäntymäki [d], Antti Malmivaara [e,f], Henri Salo [g], Eetu Suominen [h], Jyrki Kankare [i], Jukka Huttunen [j]

[a] Unit of Musculoskeletal Diseases, Department of Orthopedics and Traumatology, Tampere University Hospital and Tampere University, Tampere, Finland
[b] Department of Orthopaedics and Traumatology, Turku University Hospital, and University of Turku, Turku, Finland
[c] Department of Orthopaedics and Traumatology, Hospital Nova of Central Finland, Finland
[d] Department of Neurosurgery, Tampere University Hospital and Tampere University, Tampere, Finland
[e] Department of Data and Analytics, Finnish Institute for Health and Welfare, Helsinki, Finland
[f] Orton Orthopedic Hospital, Helsinki, Finland
[g] Datasets and Dataproducts, Finnish Institute for Health and Welfare, Helsinki, Finland
[h] The Primary Health Care Services, Wellbeing Services County of North Savo, Kuopio, Finland
[i] Department of Orthopedics and Traumatology, HUS Helsinki University Hospital, Finland
[j] Department of Neurosurgery, Kuopio University Hospital and Institute of Clinical Medicine, University of Eastern Finland, Finland

ARTICLE INFO

Handling Editor: Prof F Kandziora

Keywords:
FinSpine
Spine
Registry
Diagnosis
Diagnostic
Reliability

ABSTRACT

Introduction: The Finnish national spine surgery registry (FinSpine) collects data on patients, operative procedures, and outcomes.
Research question: Is the reliability of diagnosis collection for the FinSpine registry across various institutions sufficient?
Methods and methods: A random sample of 110 spine surgeries performed in Finland since 2017 was selected from the FinSpine registry. Details of the operative procedures, including characteristics of the hospitals, primary and secondary diagnoses, the anatomical level of the spine surgery, and the date of the surgery, were recorded. Three experienced spine surgeons assessed radiographic imaging independently on two occasions with a one-month interval. Cohen and Fleiss kappa values were calculated for inter- and intra-rater reliability with registry diagnoses as the standard.
Results: Altogether, 85 patients were included in the final analysis. The intra-rater reliability for primary diagnosis was moderate to almost perfect among all three evaluators, with Cohen's kappa values ranging from 0.765 to 0.847. In the inter-rater reliability analysis, there was moderate agreement between evaluators 1 and 2 (Cohen's $\kappa = 0.592$, SD 0.061) and evaluators 1 and 3 ($\kappa = 0.595$, SD 0.061). The agreement was substantial between evaluators 2 and 3 ($\kappa = 0.676$, SD 0.057). The Fleiss' kappa value was 0.622 (95 % CI from 0.570 to 0.673). The inter-rater reliability ranged from 0.596 to 0.652 when the evaluators' diagnoses were compared with those in the registry.
Discussion and conclusion: The primary diagnoses recorded in the FinSpine registry are reproducible and reliable, making the registry data valuable for further research and clinical decision-making.

1. Introduction

Besides evidence from randomized controlled trials, also register-based data is needed for interpretations of effectiveness of spinal surgery in real-life conditions. There is evidence that clinical register data from e.g., Norway by the NORspine (Austevoll et al., 2017) and Sweden by the SweSpine (Strömqvist et al., 2009) have provided results influencing clinical practice (Försth et al., 2013). By collecting

* Corresponding author. Unit of Musculoskeletal Diseases, Department of Orthopedics and Traumatology, Tampere University Hospital and Tampere University, Elämänaukio 2, 33520, Tampere. Finland.
E-mail address: mrjussirepo@gmail.com (J.P. Repo).

https://doi.org/10.1016/j.bas.2025.104325
Received 3 May 2025; Received in revised form 1 July 2025; Accepted 9 July 2025
Available online 9 July 2025

standardized data on procedures and outcomes, registries enable clinicians to benchmark performance, identify best practices, and drive quality improvement across institutions (van et al., 2015a). National spine surgery benchmarking registries are vital for improving patient outcomes, monitoring quality, and ensuring the safety of surgical devices (van et al., 2015a). With transparent register-based benchmarking, it could be possible to create more uniform standards among the surgeons and to improve the outcomes of the patients. They are essential for post-market surveillance of spinal implants, helping detect safety issues and track device performance. Incorporating patient-reported outcome measures (PROMs) provides valuable insights into recovery and patient satisfaction, supporting value-based care and personalized treatment (Marjamaa et al., 2023). Registries also facilitate large-scale research and evidence-based practice by allowing comparisons of surgical techniques and devices (Röder et al., 2006). Standardization and international collaboration are needed to maximize their value (Pascucci et al., 2023).

FinSpine is a national spine surgery registry in Finland that aims to systematically collect and analyze data on spine surgeries and their outcomes (Marjamaa et al., 2023). The registry is designed to improve the quality of spine surgery care, enable benchmarking between hospitals, and support scientific research (Marjamaa et al., 2023). Validation of the FinSpine registry is essential to ensure the diagnostic validity, reliability, and utility of the data in improving clinical outcomes and informing healthcare policies. Accurate and reliable data strengthen stakeholder confidence, fostering trust among surgeons, healthcare providers, policymakers, and patients (Marjamaa et al., 2023; Röder et al., 2006; van der et al., 2019a). A validated FinSpine registry enhances research by ensuring that studies based on registry data are robust, representative, and generalizable.

In FinSpine, diagnoses that indicate surgery are represented as specific details of the underlying pathology, with an emphasis on clinical relevance to mitigate classification bias. The aim of the present study is to confirm the reliability of data collection processes across institutions, ensuring that the registry data reflects true clinical practices.

2. Methods

The FinSpine research steering group developed the study methodology, following the Consensus-based standards for the selection of health measurement instruments Risk of Bias tool (Mokkink et al., 2020) guidelines. The study protocol was approved by the Finnish Institute for Health and Welfare (THL). The study was conducted from January 1, 2023, to December 31, 2024. All patients with available data on the anatomical level of spine surgery and radiographs, such as computed tomography, EOS® imaging (low dose weight-bearing radiographs technology), magnetic resonance imaging (MRI), or plain radiographs (X-ray) were eligible for the study. Patients who did not have surgery-related data (e.g., anatomical level of spine) in the FinSpine registry or who did not receive any radiological imaging from the index hospital were excluded from the analysis. The sociodemographic data and coverage of the FinSpine registry are presented elsewhere in more detail (Klimko et al., 2025; Hatakka et al., 2025).

The registry entails 22 different specific diagnoses (Table 1) and an additional section 'Other' for cases where the diagnosis is none of the predefined specific diagnoses. For the data analysis, we used the Researcher's Secure Workspace (T3) provided by Kuopio University Hospital (KUH). The Researcher's Secure Workspace (T3) is a secondary use environment compliant with the Finnish Secondary Use Act and fulfilling data security and data protection requirements. The workspace (T3) is used for the secure processing of a limited subset of research data from the Finnish Spine Surgery Registry, for which the Finnish Institute for Health and Welfare (THL) is the responsible data controller. The research data transferred to the T3 workspace includes all imaging-related information concerning patients undergoing spinal surgery in Finland, such as referrals, images, and reports, as well as pseudonymized personal identification numbers. A Weasis DICOM Viewer enabling the viewing of DICOM images was installed in the workspace.

Table 1
Specified diagnoses of the FinSpine registry.

Disc herniation
Central stenosis (displacement max 3 mm)
Central stenosis (displacement greater than 3 mm)
Recess stenosis (displacement max 3 mm)
Recess stenosis (displacement of more than 3 mm)
Nerve root canal stenosis
Spondylolysis
Disc degeneration
Postoperative instability
Pseudoarthrosis
Degenerative scoliosis
Non-degenerative scoliosis
Kyphosis
Myelopathy
Metastasis
Primary neoplasia
Postoperative hematoma
Postoperative infection
Postop cerebrospinal fluid leak
Primary infection
Fracture
Dislocation
Other

A random sample of 110 spine surgery patients was selected from the FinSpine registry, including comprehensive data on spine surgeries performed in Finland since 2016, by a THL statistician. Information about the details of the surgeries, such as the hospital where the surgery was performed, the primary diagnosis, the secondary diagnosis, the anatomical level of the spine surgery, and the date of the surgery, was obtained from the FinSpine Spine Register for the randomly selected group of 110 patients. Data on EOS®, computed tomography, MRI, and plain X-ray images of the patients who underwent spine surgery were collected into the T3 secure research environment at the request of the Finnish Institute for Health and Welfare (THL) from the hospitals where the spine surgeries were conducted.

Three spine surgeons independently evaluated the pseudonymized dataset of the 110 randomly selected patients and provided their expert opinions on the primary and secondary diagnoses. One of the raters is a specialist in neurosurgery, while the other two are orthopedic spine surgeons. The evaluators received information only on the anatomical level of the spine surgery and the date of spine surgery. The surgeons were naive to the cases and did not receive any additional information for identifying the patients. The raters assessed the imaging studies twice independently, with a one-month interval between the evaluations, and provided their diagnosis assessments on both occasions. The raters were asked to report any discrepancies, problems, or missing data that would deteriorate the quality of the assessment. All patients included in the registry were considered eligible for random selection to achieve a representative study population. Exclusion criteria from analysis were missing patient details or necessary data (e.g., level of surgery), duplicates produced by the patient social security number when obtaining the data, and missing imaging.

The assessments were made on 1) the degree of consistency and agreement made by the same rater evaluating the data on two different occasions (intra-rater reliability); 2) how well the diagnoses entered into the FinSpine registry by surgeons performing spine surgeries in Finland corresponded to the diagnoses assigned by the validators for a randomly

selected group of patients (validity), and 3) the degree of agreement and consistency of the primary and secondary diagnoses between three raters (inter-rater reliability).

2.1. Statistics

Calculating the sample size for Kappa with minimum acceptable kappa (κ0) of 0.6, expected kappa (κ1) value of 0.85, proportion of outcome at 0.5, significance level (α) at 0.05, power (1 - β) of 80 % and an expected dropout rate of 15 % gives us an sample size of 81 and (with 15 % dropout) of an estimated sample needed of 96 cases.

The intra-rater and inter-rater reliability was evaluated with an attribute gauge chart and defined using Cohen's kappa values derived from the analysis (Agresti, 2012). Agreement (κ-value) was classified as 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. The Fleiss' kappa value was used for assessing the reliability of agreement between all three raters.

3. Results

The randomly selected cases came from eight different centers. Altogether, 85 patients were included in the final analysis. The flowchart in Fig. 1 shows the patient selection in detail. Most patients had undergone surgery of the lumbar spine. The most frequent indication for surgery was disc herniation, involving 23 patients (27 %), followed by spinal stenosis (<3 mm) in 13 patients (15 %) and recess stenosis (<3 mm) in 10 patients (12 %) (Table 2).

According to the intra-rater reliability analysis, the reliability was substantial or almost perfect in all three evaluators, with Cohen's kappa values ranging from 0.765 to 0.847 (Table 3).

The inter-rater reliability analysis showed moderate agreement between evaluators 1 and 2 (Cohen's κ = 0.592, SD 0.061) and evaluators 1 and 3 (κ = 0.595, SD 0.061). The agreement was substantial between evaluators 2 and 3 (κ = 0.676, SD 0.057).

The Fleiss' kappa value was 0.622 (95 % CI from 0.570 to 0.673). The inter-rater reliability ranged from 0.596 to 0.652 when the evaluators' diagnoses were compared with those in the registry (Table 4).

Table 2
Diagnoses and level of spine surgery in the blinded registry data.

	N
Level of spine	
Cervical	27
Thoracal	3
Thoracolumbar	3
Lumbar	52
Register diagnosis	
Discus herniation	23
Spinal stenosis (<3 mm[a])	13
Recess stenosis (<3 mm[a])	10
Spinal stenosis (>3 mm[a])	7
Fracture	6
Spondylolysis	4
Disc degeneration	2
Postop instability	2
Myelopathy	2
Recess stenosis (>3 mm[a])	1
Metastasis	1
Primary neoplasm	1
Primary infection	1
Degenerative scoliosis	1
Other	1

[a] Extent of slippage; N, number of.

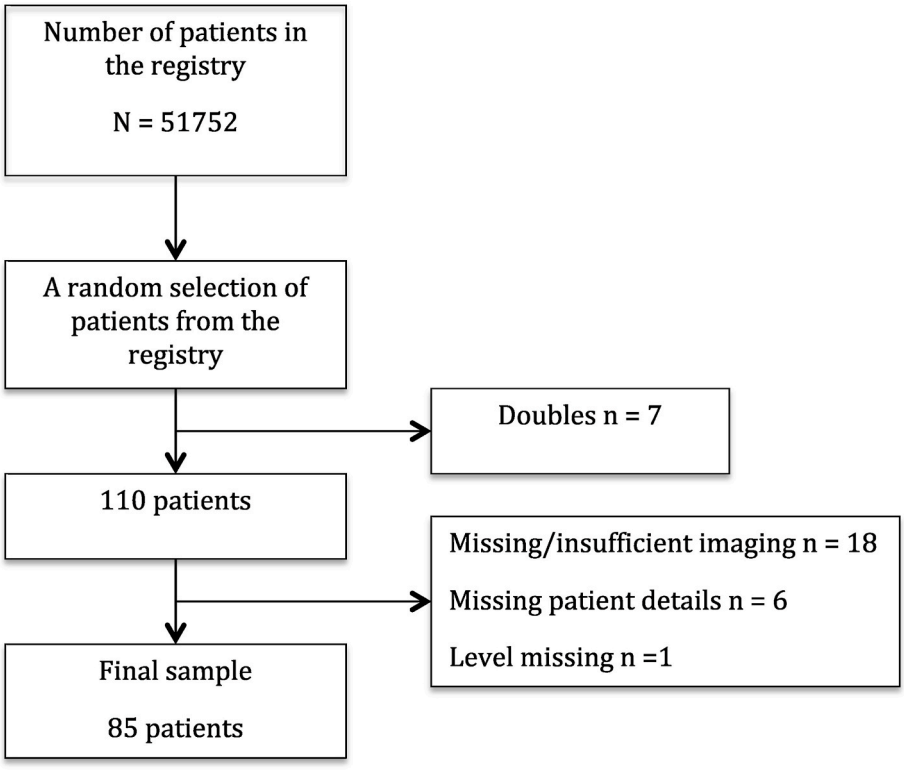

Fig. 1. Flow chart for patient selection.

Table 3
Primary diagnosis assessing the intra-rater reliability.

	Cohen's kappa (SD)
Rater 1	0.765 (0.050)
Rater 2	0.847 (0.043)
Rater 3	0.807 (0.047)

Table 4
Correlation of the primary diagnosis set by the evaluators to the registry data.

	Cohen's kappa (SD)
Rater 1	0.596 (0.061)
Rater 2	0.652 (0.057)
Rater 3	0.615 (0.057)

4. Discussion

The analysis indicates that the overall diagnostic reliability of the FinSpine registry, which covers nearly all spine surgeries in Finland, is high. The intra-rater reliability shows high reproducibility of the FinSpine registry diagnoses.

Inter-rater reliability in the present study is in line with those reported from the NORSpine registry (van der et al., 2019a), in which Cohen's kappa values ranged from 0.21 to 0.51, with a larger emphasis on higher values. Alhaug et al. showed that disc herniation and central spinal stenosis had higher intra-rater reliability compared to, for example, foraminal stenosis. In their study, the inter-rater analysis varied extensively depending on the underlying diagnosis (Alhaug et al., 2025).

The level of primary diagnoses reliability is good for identifying trends and overall patterns. The results of the present study also reflect the complexity of diagnosing spine conditions, where subjectivity or variability in imaging interpretation and clinical judgment can influence outcomes. There were quite different individual diagnoses for the assessment. For example, distinguishing primary neoplasia and metastasis from spine images without clinical data can potentially be quite difficult, unless there are several metastases in the images or the primary lesion is particularly atypical (e.g., osteoid osteoma). Similarly, postoperative instability requires a series of images from the past. The diagnosis class 'Other' also has challenges. The class 'Other' makes up around 6 % of all diagnoses, which could explain part of the inter- and intra-rater variation.

Previous studies have focused e.g. on data accuracy, completeness, or validation of content or assessment tools in the registries (Hoeijmakers et al., 2018; van der et al., 2019b; Holzer et al., 2021; Pop et al., 2019). To the best of our knowledge, this is the first study of a national spine registry for which reliability of its diagnoses is reported for a random sample of all hospitals collecting spine registry data in Finland. The random selection of 110 spine surgeries from the registry and the analysis of 85 after exclusions due to technical deficiencies provide a representative sample of the surgeries recorded in the registry. Larger research datasets in the future may improve the result of reproducibility by reducing the influence of random variability. In addition, a larger data sample could provide sufficient power to estimate the reliability of a single specific class diagnosis in the FinSpine registry, such as those of Alhaug et al. (2025) The small sample size could potentially enhance the possibility for misinterpretation.

In this study, the raters assessed diagnoses based only on the anatomical level of the spine surgery and the imaging studies. However, degenerative imaging findings in the spine are common also in asymptomatic populations, and the decision on surgery is always based on symptoms and clinical findings as well (Brinjikji et al., 2015). Considering the lack of clinical data, the repeatability of assessing the diagnosis is accurate. Additionally, there is a lack of consensus on pure radiological diagnostic criteria for most common degenerative spinal conditions like lumbar spinal stenosis and disc herniation (Steurer et al., 2021; Pojskic et al., 2024).

Researchers and policymakers can trust the registry data because it is derived from consistent evaluations, not affected by significant rater-to-rater variability. Inter-rater reliability confirms that evaluators provide consistent ratings, reducing bias and variability in assessments, ensuring the FinSpine measures what it intends to measure, enhancing the registry's utility. While there is room to refine processes and further improve reliability, this level of agreement already ensures that the registry can serve as a trustworthy source for benchmarking, research, and policymaking.

The present study diminishes concerns about FinSpine data reliability considered by several authors in other spine registries (van et al., 2015; Meyer et al., 2020). To improve reliability, it is important to ensure that all surgeons and healthcare providers follow standardized protocols for diagnosis, treatment, and data entry. Definitions of diagnoses and spinal surgical procedures, as determined by an expert panel, have been integrated into the data collection platform used by spine surgeons in FinSpine. Definitions can be utilized by surgeons when entering information into FinSpine while completing data entries. The purpose, in order to improve the reliability of the registry, is to provide feedback to hospitals or surgeons on data quality and reliability, fostering continuous improvement, and to offer periodic training to all contributors on how to correctly document and report cases in the registry.

5. Conclusions

The primary diagnoses recorded in the FinSpine registry are reproducible and show strong reliability. The FinSpine registry data is reliable and useful for further research and clinical decision-making.

Data availability statement

Finnish Law prohibits public sharing of datasets with personal data. The data that support the findings of this study are available from FinData with permission.

Authorship

All authors have made substantial contributions to all of the following: (1) the conception and design of the study, or acquisition of data, or analysis and interpretation of data, (2) drafting the article or revising it critically for important intellectual content, (3) final approval of the version to be submitted.

Funding

This study received funding from State funding for university-level health research, Orton Orthopaedic Hospital, Helsinki, Finland.

Competing interests

The authors declare no conflicts of interest.

Acknowledgements

The authors would like to thank the following members of the Finnish Society of Spine Surgery: Johan Marjamaa, Pietari Kinnunen and Raul Laasik, who offered insight for conducting this study.

References

Agresti, A., 2012. Categorical Data Analysis, 792. John Wiley & Sons, pp. 434–435.

Alhaug, O.K., Furunes, H., Kaur, S., et al., 2025. Reliability of surgeon-reported MRI findings to a national spine register. Acta Neurochir. 167, 105.
Austevoll, I.M., Gjestad, R., Brox, J.I., Solberg, T.K., Storheim, K., Rekeland, F., Hermansen, E., Indrekvam, K., Hellum, C., 2017. The effectiveness of decompression alone compared with additional fusion for lumbar spinal stenosis with degenerative spondylolisthesis: a pragmatic comparative non-inferiority observational study from the Norwegian Registry for Spine Surgery. Eur. Spine J. 26 (2), 404–413.
Brinjikji, W., Luetmer, P.H., et al., 2015. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations". AJNR Am. J. Neuroradiol. 36 (4), 811–816.
Försth, P., Michaelsson, K., Sanden, B., 2013. Does fusion improve the outcome after decompressive surgery for lumbar spinal stenosis?: a two-year follow-up study involving 5390 patients. Bone Joint Lett. J 95 (B7), 960–965.
Hatakka, J., Laaksonen, I., Kostensalo, J., Mäkelä, K.T., Salo, H., Pernaa, K., 2025. 1-year results of lumbar spinal stenosis surgery in Finland: a national FinSpine register study. Acta Orthop. 96, 154–160.
Hoeijmakers, F., Beck, N., Wouters, M.W.J.M., Prins, H.A., Steup, W.H., 2018. National quality registries: how to improve the quality of data? J. Thorac. Dis. 10 (Suppl. 29), S3490–S3499.
Holzer, E.M., Aghayev, E., O'Riordan, D., Fekete, T.F., Jeszenszky, D.J., Haschtmann, D., Porchet, F., Kleinstueck, F.S., Pigott, T., Munting, E., Luca, A., Mannion, A.F., 2021. Validation of a surgical invasiveness index in patients with lumbar spinal disorders registered in the Spine Tango registry. Eur. Spine J. 30 (1), 1–12.
Klimko, N., Danner, N., Salo, H., Kotkansalo, A., Leinonen, V., Huttunen, J., 2025. Outcome after anterior cervical decompression and fusion - a nationwide FinSpine register study of independent predictors of outcome at 12 Months after surgery for degenerative cervical spine. Spine. https://doi.org/10.1097/BRS.0000000000005323. Epub ahead of print. PMID: 40047144.
Marjamaa, J., Huttunen, J., Kankare, J., Malmivaara, A., Pernaa, K., Salmenkivi, J., Pekkanen, L., 2023. The Finnish spine register (FinSpine): development, design, validation and utility. Eur. Spine J. 32 (11), 3731–3743.
Meyer, B., Shiban, E., Albers, L.E., et al., 2020. Completeness and accuracy of data in spine registries: an independent audit-based study. Completeness and accuracy of data in spine registries: an independent audit-based study Eur. Spine J. 29, 1453–1461.
Mokkink, L.B., Boers, M., van der Vleuten, C.P.M., et al., 2020. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med. Res. Methodol. 20, 293.
Pascucci, S., Langella, F., Franzò, M., Tesse, M.G., Ciminello, E., Biondi, A., Carrani, E., Sampaolo, L., Zanoli, G., Berjano, P., Torre, M., 2023. National spine surgery registries' characteristics and aims: globally accepted standards have yet to be met. Results of a scoping review and a complementary survey. J. Orthop. Traumatol. 24 (1), 49.
Pojskic, M., Bisson, E., Oertel, J., Takami, T., Zygourakis, C., Costa, F., 2024. Lumbar disc herniation: epidemiology, clinical and radiologic diagnosis WFNS spine committee recommendations. World Neurosurg. X 22, 100279. https://doi.org/10.1016/j.wnsx.2024.100279.
Pop, B., Fetica, B., Blaga, M.L., Trifa, A.P., Achimas-Cadariu, P., Vlad, C.I., Achimas-Cadariu, A., 2019. The role of medical registries, potential applications and limitations. Med. Pharm. Rep. 92 (1), 7–14.
Röder, C., Müller, U., Aebi, M., 2006. The rationale for a spine registry. Eur. Spine J. 15 (1), S52–S56. Suppl 1.
Steurer, J., Roner, S., Gnannt, R., et al., 2021. Quantitative radiologic criteria for the diagnosis of lumbar spinal stenosis: a systematic literature review. BMC Muscoskelet. Disord. 12, 175.
Strömqvist, B., Fritzell, P., Hägg, O., Jönsson, B., 2009. Swedish Society of Spinal Surgeons the Swedish spine register: development, design and utility. Eur. Spine J. 18 (Suppl. 3), 294–304.
van der Werf, L.R., Voeten, S.C., van Loe, C.M.M., Karthaus, E.G., Wouters, M.W.J.M., Prins, H.A., 2019a. Data verification of nationwide clinical quality registries. BJS Open 3 (6), 857–864.
van der Werf, L.R., Voeten, S.C., van Loe, C.M.M., Karthaus, E.G., Wouters, M.W.J.M., Prins, H.A., 2019b. Data verification of nationwide clinical quality registries. BJS Open 3 (6), 857–864.
van Hooff, M.L., Jacobs, W.C., Willems, P.C., Wouters, M.W., de Kleuver, M., Peul, W.C., Ostelo, R.W., Fritzell, P., 2015a. Evidence and practice in spine registries. Acta Orthop. 86 (5), 534–544.
van Hooff, M.L., Jacobs, W.C., Willems, P.C., Wouters, M.W., de Kleuver, M., Peul, W.C., Ostelo, R.W., Fritzell, P., 2015b. Evidence and practice in spine registries. Acta Orthop. 86 (5), 534–544.