Scoring systems to suggest surgical intervention in degenerative spondylolisthesis - a retrospective study

Rajib Ahmed, Mr Francis Brooks

Objective

Find out if scoring systems can be utilized to suggest surgical interventions on degenerative spondylolisthesis in lumbar spine (DSLS) by peforming a retropspective study.

Introduction

Degenerative spondylolisthesis is a common spinal pathology where one vertebral body slips forward over the vertebral body below due to aging processes that weaken the spine[1]. Slippage can induce radiculopathy which leads to symptoms such as back pain, sciatica or even loss of bladder function.[3 6]

The primary surgical intervention available to DSLS patients is a decompression surgery, where segments of vertebra are removed to relieve pressure on nerves. Further interventions such as posterior fusions (PF) where posterior segments of the spine are fused together to prevent further slippage. The creation of minimally invasive lumbar interbody fusions (LIF) such as TLIF's and PLIF's are a part of a spinal surgeons arsenal to treat DSLS.[1 5 2]

Despite these advancements, there is no agreed system or guidelines to help surgeons in select which interventions are best to treat DSLS cases. The aim of this study was to find which scoring systems are available to suggest surgical interventions for DSLS cases and assess their potential for clinical use.

Method

Two potential scoring systems were identified through an initial review of the literature; the degenerative spondylolisthesis instability classficiation[7] (DSIC) and the University of California San Francisco Degenerative Spondylolisthesis[4] (UCSF DS) classification system. Both scoring systems use a combination of clinical and radiological information to categorize DSLS cases - both are able to suggest interventions ranging from decompression alone to decompression with PF and LIF

DSIC SCORING SYSTEM[7]

Parameter	Type 1 - Stable Decompression Alone	Type 2 - Potentially Unstable Decompression + PF	Type 3 - Unstable Decompression +PF+LIF
Low Back Pain	None or Very Mild	Primary/Secondary Complaint	Primary/Secondary Complaint
Disc Angle	Lordotic Disc Angle or <3mm of translation	Neutral Disc Angle or 3-5 mm of translation	Kyphotic Disc Angle or >5mm of translation
Joint Effusion	No Facet Joint Effusion	Facet Joint Effusion without joint distraction	Large facet joint effusion
Restabilisation	Grossly narrowed disc height	Reduced Disc Height	Normal to slightly reduced disc height

UCSF DS SCORING SYSTEM[4]

Parameter	Decompression Alone	Decompression + PF	Decompression +PF+-LIF
Primary Clinical Presentation	Primarily Leg Pain	Both Leg and Back Pain	Primarily Back Pain
Disc Angle	Lordotic Disc Angle or <3mm of translation	Neutral Disc Angle or 3-5 mm of translation	Kyphotic Disc Angle or >5mm of translation
Joint Effusion	No Facet Joint Effusion	Facet Joint Effusion without joint distraction	Large facet joint effusion
Restabilisation	Grossly narrowed disc height	Reduced Disc Height	Normal to slightly reduced disc height

89 DSLS cases affecting the L4/L5 vertebra were collated from the Cardiff Spinal Surgery Database. Both scoring systems were retrospectively applied to these cases and suggested interventions were compared with the actual intervention used. Statistical analysis was performed to produce results.

Results

F1 scores (harmonic mean of precision and accuracy) were calculated for each system. UCSF DS pulled ahead of DSIC with an F1 score of **0.438** against **0.371**. However a score of 0.438 is relatively low so further investigation was done by comparing suggested and actual surgeries.

We calculated the following for each scoring system in each case:

- Underestimations - Where suggested surgeries were less invasive than the actual surgery
- Overestimations - Where suggested surgeries were more invasive than the actual surgery
- Exact Matches - Where suggested surgeries were the same as the actual surgery

Scoring System	Underestimations	Overestimations	Exact Match
DSIC	23.6%	39.3%	37.1%
UCSF DS	41.6%	14.6%	43.8%

UCSF DS displays a high percentage of underestimations and exact matches. In the dataset we assume all surgeries had completely successful outcomes. This would indicate that the scoring system is suggesting less invasive surgeries that could be just as effective as the one performed.

Conclusions

Of the two scoring systems UCSF DS displayed the most potential as a scoring system to suggest surgical interventions in DSLS. It displayed the highest F1 score of the two tested scoring systems.

Furthermore, it tended to suggest less invasive surgical interventions in the dataset, which, given assumptions in the dataset, would mean UCSF DS could improve patient outcomes with less invasive interventions and save the NHS money on operations.

Further study is indicated to prove the efficacy of these scoring systems, this study shows there is significant potential for these scoring systems to be used in practise to improve patient outcomes and reduce NHS costs.

Improvements and Limitations

- As this was a retrospective study, we cannot look at the direct efficacy of these scoring system. Further study with prospective surgeries would be required
- The lack of QoL scoring pre and post operatively limited our ability to assess if a surgery in the dataset could be improved therefore it was assumed all surgeries had 100% success and improvement of patient QoL
- Potential improvements to the scoring systems were identified during the study, such as adding further clinical factors (i.e

References

1. Bydon M, Alvi MA, Goyal A. Degenerative Lumbar Spondylolisthesis: Definition, Natural History, Conservative Management, and Surgical Treatment. Neurosurg Clin N Am. 2019 Jul;30(3):299–304. doi: 10.1016/j.nec.2019.02.003
2. Gulgul P, Ferrero E. Surgical treatment of degenerative spondylolisthesis. Orthopaedics & Traumatology: Surgery & Research. 2017 Feb 1;103(1, Supplement):S11–20. doi: 10.1016/j.otsr.2016.06.022
3. Herkowitz HN. Spine update. Degenerative lumbar spondylolisthesis. Spine (Phila Pa 1976). 1995 May 1;20(9):1084–90. doi: 10.1097/00007632-199505000-00018
4. Rangwalla K, Filley A, El Naga A, Gendelberg D, Baldwin A, Mazlad A, et al. Degenerative lumbar spondylolisthesis: review of current classifications and proposal of a novel classification system. Eur Spine J. 2023 Aug 6; [accessed 16 Mar 2024] Available from: https://doi.org/10.1007/s00586-023-07818-x
5. Samuel AM, Moore HG, Cunningham ME. Treatment for Degenerative Lumbar Spondylolisthesis: Current Concepts and New Evidence. Curr Rev Musculoskelet Med. 2017 Oct 9;10(4):521–9. doi: 10.1007/s12178-017-9442-3
6. Sengupta DK, Herkowitz HN. Degenerative Spondylolisthesis: Review of Current Trends and Controversies. Spine. 2005 Mar 15;30(6S):S71. doi: 10.1097/01.brs.0000155579.88537.8e
7. Simmonds AM, Rampersaud YR, Dvorak MF, Dea N, Melnyk AD, Fisher CG. Defining the inherent stability of degenerative spondylolisthesis: a systematic review. Journal of Neurosurgery: Spine. 2015 Aug 1;23(2):178–89. doi: 10.3171/2014.11.SPINE1426

Acknowledgements - Many thanks to Mr Brooks for their guidance and insight throughout the project

Abstract

Author's – Rajib Ahmed, Mr Francis Brooks

Background

Degenerative spondylolisthesis (DS) is a common spinal pathology where one vertebral body slips forwards over the inferior vertebral body. It occurs because of degenerative change in the vertebra, commonly in areas of high load such as the L4/L5 vertebral level. Slippage usually induces a lumbar canal stenosis which leads to neurological symptoms such as back pain, sciatica, or bladder/bowel incontinence. Surgical intervention is a common, especially when conservative management fails, there are a variety of approaches that can be used to treat DS.

The common interventions are decompressions and fusions. There are different types of fusions available to spinal surgeons, such as posterior fusions or lumbar interbody fusions. However, there is no agreed upon algorithm for choosing which combination of interventions to use, despite DS being variable in severity.

There has been some effort to create scoring systems to categorize DS cases, and then using these to suggest surgical interventions. Using Cardiff's spinal surgery database, this project aims to perform a retrospective study to assess whether these scoring systems have a place in practice.

Methods

An initial literature review was conducted to identify two potential novel scoring systems that could be used to categorize DS in lumbar spine – DSIC and UCSF DS. 89 cases were pulled from the spinal surgery database, and clinical and surgical history was collected as well as extensive radiological information. All cases contained disease at the L4/L5 level with some spreading to L3 and S1. Scoring was retrospectively applied to all cases and then compared with the actual intervention used. F1-Scoring and estimation percentages were calculated for both scoring systems to assess accuracy, precision, and recall.

Result

UCSF DS had the higher F1 score of the two systems at 0.438 compared to DSIC at 0.371. Whilst both scores are less than 0.5, suggesting poor precision and recall, given the assumption that all cases had positive outcomes, surgical under and over estimation was calculated. It was found that with USCF DS, 41.6% of cases had a suggested surgical intervention that was less invasive than the actual intervention – meaning these could potentially have had positive outcomes with a less complex surgery. 43.8% of cases had matching suggested and actual surgeries, and only 14.6% of cases had a more complicated surgery suggested for its actual surgery – which given the assumption that all cases had positive outcomes – is a negative suggestion as it puts the patient under unnecessary surgery.

Conclusion

Given that over 80% of cases were either matched in suggestion or could possibly have been improved with a less invasive surgery, there is certainly a case to be made for the application of these scoring systems in DS treatment. However, further study will need to be completed to assess clinical viability. There were limitations in this study which reduces the ability to advocate for its clinical use; lack of QoL scoring pre and post operatively limits the ability to assess which surgeries led to better outcomes; and scoring was applied retrospectively so direct outcomes of the scoring was not measurable. Furthermore, through assessment of cases there could be suggested improvements to the scoring system to help guide treatment such as addition of further clinical features such as loss of bladder function/saddle anaesthesia; and further guidance on choice of surgery such as which type of LIF to use. The creation and use of a reliable scoring system in DS would be a valuable clinical tool in guiding surgical treatment, there is scope for extensive further study in this area.

Scoring systems to predict surgical treatment in Degenerative Spondylolisthesis – a retrospective review.

Abstract

Purpose – To retrospectively assess whether existing classification systems for degenerative spondylolisthesis of the lumbar spine (DSLS) can provide treatment options for patients and determine their use in clinical settings.

Methods – data on 89 cases of DSLS where surgical intervention was given was collected retrospectively. Two classification systems were applied, and the outcome was compared with the actual intervention used. Prediction scores were then calculated for each system.

Results – UCSF DS scoring system performed better than DSIC with an F1 score of 0.438 compared to 0.371. UCSF DS also predicted surgical intervention for 43.8% of cases whilst DSIC did 37.1% of cases. UCSF DS also tended to suggest less invasive procedure (41.6% of cases) whilst DSIC tended to suggest more invasive procedure (39.3% of cases).

Conclusion – UCSF DS displays potential as a possible scoring system to suggest surgical intervention for degenerative spondylolisthesis in lumbar spine cases. Further study must be performed to assess its clinical effectiveness.

Introduction

Degenerative spondylolisthesis is a common spinal pathology where one vertebral body slips forward over the vertebral body below it due to aging processes that weaken the spine. There is controversy concerning treatment of the condition, as there is no strong consensus or agreed pathway in determining treatment.

Slippage usually induces a lumbar canal stenosis, which leads to radiculopathy resulting in neurological symptoms such as back pain, sciatica or even bladder/bowel symptoms. Surgical intervention is often indicated to improve these symptoms if conservative management fails to relieve symptoms (Herkowitz 1995; Sengupta and Herkowitz 2005). Slippage tends to occur in the L3/L4, L4/L5 and L5/S1, although most commonly in the L4/L5 region (Guigui and Ferrero 2017).

The primary surgical intervention available to DSLS patients is a decompression surgery, whereby segments of the vertebra are removed to relieve pressure on the nerves, for example a laminectomy. Decompressions can be performed alone or with fusions. This is where 2 or more vertebrae are fused together to prevent further slippage and maintain spinal integrity. Posterior lumbar fusions have been a popular choice to do alongside decompressions however newer techniques known as lumbar interbody fusions (LIF) are also used.

The question for clinicians is which surgical approaches to use; this question is generally left up to an individual surgeon and is there is no agreed method of selecting which surgical approaches to use. Whilst there are currently simple classification systems such as the Meyerding classification or CARDS (Rock and Gaillard 2008; Kepler et al. 2015), these do not provide any useful information regarding which interventions to use and are simply used clinically to help categorise cases.

There have been attempts to create such a classification system that can help predict which surgical technique to use, two promising systems include the degenerative spondylolisthesis instability classification (DSIC) proposed by Simmonds et al (2017), and the University of California San Francisco Degenerative Spondylolisthesis (UCSF DS) classification system proposed by Rangwalla et al (2023), which use a combination of clinical and radiographic data to help suggest surgical interventions.

The primary aim of this paper is, using our extensive Cardiff spinal surgery database spanning 5000+ patients over the last 10 years, to retrospectively apply these scoring systems to successful DSLS surgeries to observe if these scoring systems can accurately suggest the correct surgery.

Methods

Classification Systems

Through an initial literature review of novel DSLS scoring systems to provide treatment suggestions, we have identified two promising candidates; DSIC which was proposed by Simmonds et al (2017) and the UCSF DS system proposed by Rangwalla et al (2023). Both systems aimed to improve upon existing scoring systems by integrating clinical and detailed radiological categories in order to predict outcomes.

Table 1 demonstrates the DSIC scoring system, it classifies DSLS into 3 types; stable, potentially unstable and unstable. The system utilises low-back pain as its only clinical category, and uses restabilisation/disc height, disc angle and joint effusion as its radiological parameters. It also suggests a surgical approach for each type, decompression alone, decompression with a posterior fusion and decompression with both a posterior and interbody fusion, respectively.

Table 2 demonstrates the UCSF DS scoring system has 4 categories like DSIC; segmental dynamic instability, location of spinal stenosis, sagittal alignment, and clinical presentation. In the paper where UCSF DS was first discussed, they discussed what potential surgical predictions the scoring system could be used for, however did not provide any strict criterion for this. Thus, we have taken it upon ourselves to test this scoring system using our own applied predictions – we have used the discussion and taken inspiration from DSIC, splitting the scoring into 3 categories where the same interventions are used: decompression alone, decompression with a posterior fusion and decompression with a posterior and interbody fusion (Table 3)

Table 1 - DSIC Scoring System (Simmonds et al, 2017)

Parameter	Type 1 – Stable	Type 2 – Potentially Unstable	Type 3 - Unstable
Low- Back Pain	None or very mild	Primary or Secondary Complaint	Primary or Secondary Complaint
Restabilisation	Restabilisation signs - Grossly narrowed disc height	Some restabilisation signs – reduced disc height	No restabilisation signs, normal to sightly reduced disc height
Disc Angle	Lordotic disc angle on flexion radiograph or <3mm of translation on dynamic films	Neutral disc angle on flexion radiographs or 3-5 mm of translation on dynamic films	Kyphotic disc angle on flexion radiographs or >5mm of translation on dynamic films
Joint Effusion	No facet joint effusions on MRI	Facet joint effusion on MRI without joint distraction	Large facet joint effusion on MRI
Surgical Approach	Decompression Alone	Decompression and Posterior fusion	Decompression and posterior fusion with interbody fusion

Table 2 - UCSF DS Scoring System

Segmental dynamic instability	Location of spinal stenosis	Sagittal alignment	Primary clinical presentation
<3mm translation	Central or lateral recess stenosis alone	Maintained segmental lordosis	Primary leg pain
3-5 mm translation	Foraminal stenosis without up/down stenosis	Segmental neutral or kyphotic alignment	Both leg and back pain
>5mm translation	Foraminal stenosis with up/down stenosis	Global malalignment (SVA >5cm or PT > 30°)	Primarily Back pain

Table 3 - UCSF DS with added surgical suggestion

Segmental dynamic instability	Location of spinal stenosis	Sagittal alignment	Primary clinical presentation	Surgical Intervention
<3mm translation	Central or lateral recess stenosis alone	Maintained segmental lordosis	Primary leg pain	Decompression alone
3-5 mm translation	Foraminal stenosis without up/down stenosis	Segmental neutral or kyphotic alignment	Both leg and back pain	Decompression with posterior fusion
>5mm translation	Foraminal stenosis with up/down stenosis	Global malalignment (SVA >5cm or PT > 30°)	Primarily Back pain	Decompression with posterior fusion and interbody fusion

Neither scoring system provided a guide on how to use their respective system. We undertook two approaches when calculating these scores called highest value (HV) or highest non-zero (HNZ). In HV we chose the scoring type based on which scored the highest, i.e. for DSIC if type 2 had ¾ filled categories then this was the chosen one. If two types had the same value, i.e. type 2 and type 3 both scoring 2, then the highest severity type was chosen, in this case type 3 could be chosen. For HNZ the highest severity type with a non-zero score would be chosen, i.e. if type 1 scored ¾ and type 3 scored ¼ then type 3 would be chosen. Using both methods allowed us to check how the efficacy of each scoring system changed with how they were used.

Data Collection

Data was collected from a database of spinal surgeries performed by Cardiff spinal surgeons over the last 10 years. Over 7000 cases were filtered through to identify cases of DSLS. Cases had to have DSLS in L4/L5 region, we chose to include cases that spanned more regions than this such as L3/L4/L5 or L4/L5/S1. Clinical, surgical, and radiological information was gathered from each of their respective portals, any data entries with incomplete information was then removed. Figure 1 demonstrates the filter process for data collection – after filtering we are left with 94 cases.

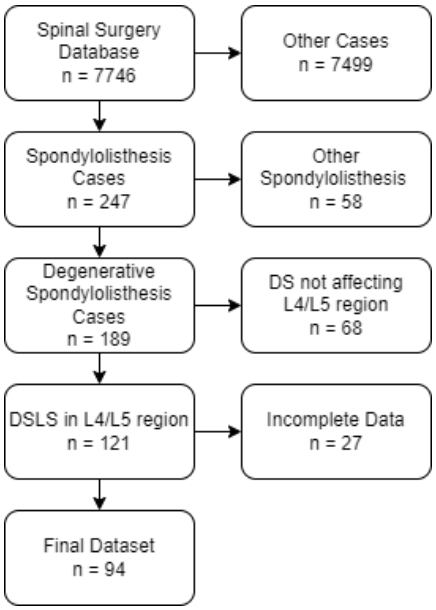

Figure 1 - Data collection process.

Clinical data was collected by reading referral information and clinical histories, data was classified by their presentation, primarily leg pain, primarily back pain or both back and leg pain. The severity of back pain was also recorded to be used with DSIC.

Radiological information collected included disc height in pathological regions, joint effusion, alignment in pathological regions, location of stenoses if present, aswell as the translation of vertebra.

Surgical information collected was divided into the interventions used, decompressions, posterior fusions, and interbody fusions. The type of lumbar interbody fusion (LIF) was also recorded.

Dataset

In the 94 cases, we first looked at the surgical interventions used. We found that 5 of our cases did not fit into the predictive categories set by the scoring systems. Table 4 demonstrates the different surgical interventions used, 5 used decompressions with LIFs only, which is not a suggested intervention. We have chosen to remove these from the dataset as to not confuse any confidence statistics.

Table 4 - Surgical Interventions across dataset

Decompression Alone	Decompression with Posterior Fusion	Decompression with LIF (Data removed from set)	Decompression with Posterior Fusion and LIF
11	56	5 (4 TLIF, 1 ALIF)	22 (16 TLIF, 5 PLIF, 1 ALIF)

With this data removed the dataset is now made up of 89 cases.

Figure 2 demonstrates the spread of ages in the dataset, with the minimum being 37 and the maximum being 83. The average age across all cases was 65.

Of our cases, 56 presented with DSLS in L4/L5 alone, whilst 22 presented with L3/L4 and L4/L5 disease, and 11 presented with L4/L5 and L5/S1 disease (Table 4).

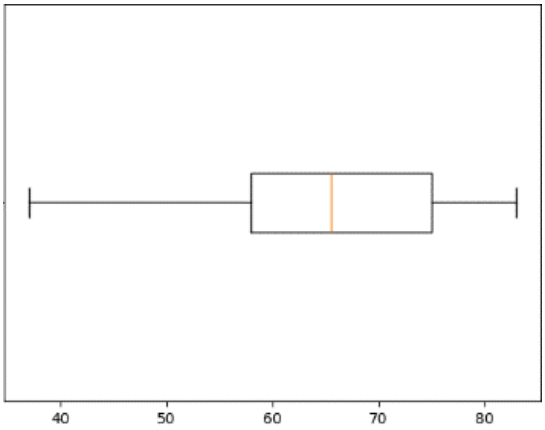

Figure 2 - Boxplot of ages in dataset.

L3/L3/L5	L4/L5	L4/L5/S1
22	56	11

Table 5 - Vertebral levels of disease across dataset

Table 6 shows the surgical interventions split by affected vertebral levels.

	Decompression Only	Decompression with Posterior Fusion	Decompression with Posterior Fusion and LIF
L3/L4/L5	4	16	2 (1 TLIF, 1 PLIF)
L4/L5	5	33	18 (14 TLIF, 4 PLIF)
L4/L5/S1	2	7	2 (1 TLIF, 1 ALIF)

Table 6 - Surgical Interventions for each Vertebral Category of Disease

When performing analysis, all types of LIF's were calculated as one, since scoring systems make no distinction or suggestion for which type of LIF should be used. All cases in the dataset were primary surgeries.

Statistical Analysis

Upon completion of data collection, cases were sub-divided into surgical intervention categories, decompressions alone, decompressions with posterior fusions and decompressions with poster and interbody fusions. There were some cases where decompressions with interbody fusions were performed, these were discarded as they did not fit into the prediction algorithm.

Using the collected information, each case was scored with both HV and HNZ methods as discussed above. Using Python, confusion matrices were created for each scoring system and method, from this F1-Scores were then calculated, these provide the harmonic mean between precision and recall for classification models.

Results

Confusion matrices were created for each scoring system using the 2 methods of scoring (HV and HNZ) as shown in Figures 3,4,5 and 6. In each matrix, the actual and predicted value are shown for each category (D – Decompression, D + PF – Decompression and Posterior Fusion, D + PF + IF – Decompression and Posterior Fusion and LIF). Precision, Recall and F1-Scores were then calculated for each which can be found in the Appendix (Table 8)

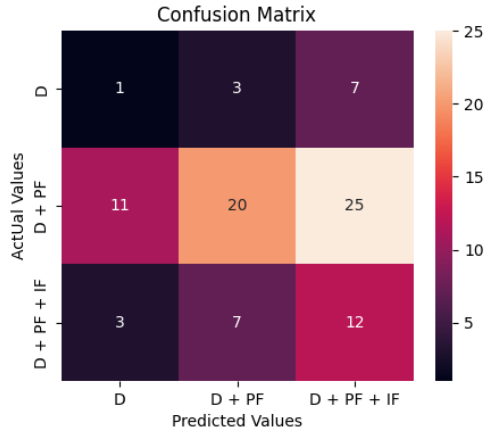

Figure 3 - DSIC using Highest Value Category Scoring

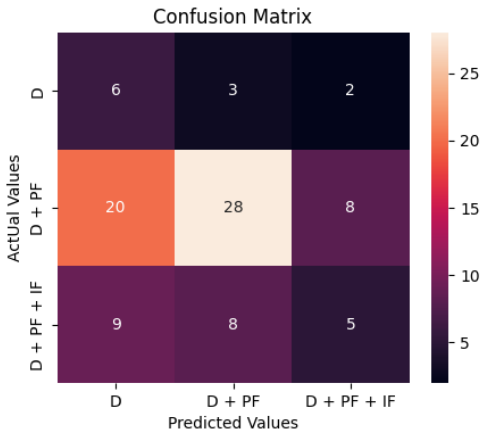

Figure 4 - UCSF DS using Highest Value Category Scoring

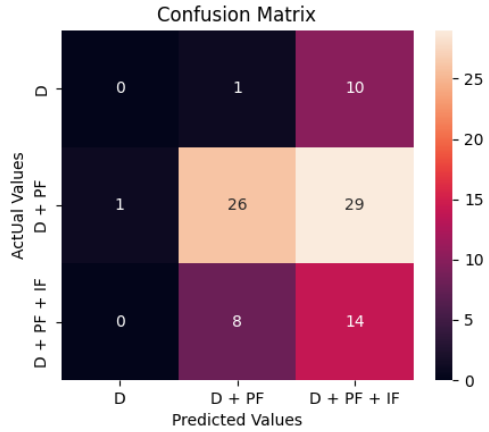

Figure 5 - UCSF DS using Highest Non-Zero Category Scoring

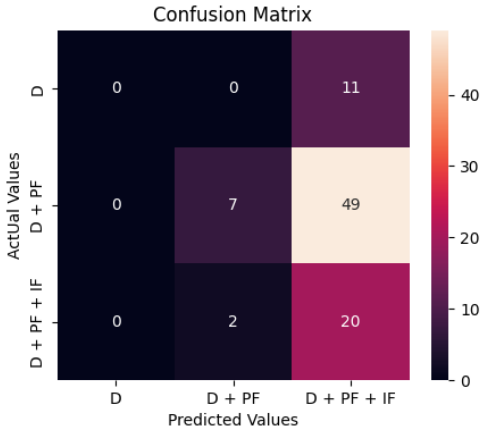

Figure 6 - DSIC using Highest Non-Zero Category Scoring

	DSIC	UCSF DS
HV	0.371	0.438
HNZ	0.303	0.449

Table 7 - Overall F1-Score for each scoring system and method

Following on from this we also calculated estimation percentages, we created three categories, underestimation – where the suggested surgical intervention was less invasive than the actual, matched estimation – where the suggested intervention matched the actual, then overestimation where the suggested intervention was more invasive than the actual. Since we have assumed all cases have a positive outcome on the patient, underestimation may suggest that the suggested surgery may have had the same effect. Overestimation would give a patient a more invasive surgery when a less invasive one would have performed just as well. Table 8 demonstrates this.

Table 8 - Estimation of each scoring system relative to actual surgical intervention

	Underestimations	Matches	Overestimations
UCSF DS HV	0.416	0.438	0.146
UCSF DS HNZ	0.101	0.449	0.449
DSIC HV	0.236	0.371	0.393
DSIC HNZ	0.022	0.303	0.674

Discussion

Scoring Method

Of the two scoring methods there was a clear selection, highest value category was best across both. In DSIC, HV scoring had a higher F1 score of 0.371 compared to 0.303. In UCSF DS, HNZ had a barely higher F1 score of 0.449 compared to 0.438, however when looking at the confusion matrix and the overestimation value of 0.449 for this scoring method it is clear to see that HNZ only managed this F1 score by placing nearly all the data in either the D+PF or D+PF+IF categories which shifted the values – this was also done in DSIC but with less success as most of the data was placed in one category rather than across two. Thus, we can say that using highest scoring category is the method of choice when using these scoring systems – HV scoring will be referenced only moving forward.

Scoring System

Of which scoring systems performed better, UCSF DS comes out ahead of DSIC. F1 scoring for UCSF DS was 0.438 whilst for DSIC was 0.371, indicating a better precision and recall – however given that these scores are still quite low it was important to investigate where each of the scoring systems went wrong. To do this, we calculated the underestimations (where suggested surgical interventions were less invasive than the actual surgery), matches, and overestimations (where suggested surgical interventions were more invasive that the actual surgery – the reason we do this is because one assumption we have made with the dataset is that all surgeries performed were completely successful. We do this because there are no QoL data pre and post operatively, so we chose surgeries that did not have any further direct revisions. Given this assumption, we prefer underestimations, as it means there could have been a less invasive surgery with similar outcomes, overestimations would mean that a more invasive surgery would have been performed when a less invasive surgery would've had the same outcome.

DSIC matched 37.1% of its suggestions to the actual surgeries, overestimated 39.3% of cases and underestimated 23.6% of cases. UCSF DS matched 43.8% of cases, overestimated only 14.6% of cases and underestimated 41.6% of cases. UCSF DS was able to match more cases and also underestimated far more cases than DSIC which tended to overestimation. This means UCSF DS has the potential to suggest less invasive surgeries which could lead to better patient outcomes as well as less expensive healthcare.

Improvements and Further Study

As mentioned, our dataset did not include any QoL or surgical success scoring, therefore we had to assume that all surgeries had a perfect outcome – further study will be needed to correlate surgical success with these scoring systems.

The fact this was a retrospective study also has some limitations as we cannot directly assess the scoring systems; further study will be required to assess the efficacy of these scoring systems in suggesting surgical intervention in prospective surgeries.

Further improvements to the scoring system themselves could be added aswell. During data collection some factors appeared routinely in some cases which may be helpful indicators. This included loss of bladder/bowel function in some patients and the number of spinal levels effected. Another further improvement that could be made is the suggestion of specific LIFs such as TLIF, ALIF, PLIF etc.

Conclusions

The UCSF DS scoring system showed the most promise as a viable system to suggest surgical interventions for DSLS patients; it has the potential to improve patient outcomes with less invasive and thus more cost-effective surgeries. Further study needs to be completed to assess its clinical effectiveness and how it can improve quality of life in patients. There are also some potential improvements that could be made to UCSF DS to help increase clinical effectiveness.

Bibliography

Guigui, P. and Ferrero, E. 2017. Surgical treatment of degenerative spondylolisthesis. *Orthopaedics & Traumatology: Surgery & Research* 103(1, Supplement), pp. S11–S20. doi: 10.1016/j.otsr.2016.06.022.

Herkowitz, H.N. 1995. Spine update. Degenerative lumbar spondylolisthesis. *Spine* 20(9), pp. 1084–1090. doi: 10.1097/00007632-199505000-00018.

Kepler, C.K. et al. 2015. Clinical and radiographic degenerative spondylolisthesis (CARDS) classification. *The Spine Journal* 15(8), pp. 1804–1811. doi: 10.1016/j.spinee.2014.03.045.

Rangwalla, K. et al. 2023. Degenerative lumbar spondylolisthesis: review of current classifications and proposal of a novel classification system. *European Spine Journal*. Available at: https://doi.org/10.1007/s00586-023-07818-x [Accessed: 16 March 2024].

Rock, P. and Gaillard, F. 2008. Spondylolisthesis grading system. In: *Radiopaedia.org*. Radiopaedia.org. Available at: http://radiopaedia.org/articles/2076 [Accessed: 16 March 2024].

Sengupta, D.K. and Herkowitz, H.N. 2005. Degenerative Spondylolisthesis: Review of Current Trends and Controversies. *Spine* 30(6S), p. S71. doi: 10.1097/01.brs.0000155579.88537.8e.

Simmonds, A.M., Rampersaud, Y.R., Dvorak, M.F., Dea, N., Melnyk, A.D. and Fisher, C.G. 2015. Defining the inherent stability of degenerative spondylolisthesis: a systematic review. *Journal of Neurosurgery: Spine* 23(2), pp. 178–189. doi: 10.3171/2014.11.SPINE1426.

Appendix

Table 1 - DSIC Scoring System (Simmonds et al, 2017)

Parameter	Type 1 – Stable	Type 2 – Potentially Unstable	Type 3 - Unstable
Low- Back Pain	None or very mild	Primary or Secondary Complaint	Primary or Secondary Complaint
Restabilisation	Restabilisation signs - Grossly narrowed disc height	Some restabilisation signs – reduced disc height	No restabilisation signs, normal to sightly reduced disc height
Disc Angle	Lordotic disc angle on flexion radiograph or <3mm of translation on dynamic films	Neutral disc angle on flexion radiographs or 3-5 mm of translation on dynamic films	Kyphotic disc angle on flexion radiographs or >5mm of translation on dynamic films

Joint Effusion	No facet joint effusions on MRI	Facet joint effusion on MRI without joint distraction	Large facet joint effusion on MRI
Surgical Approach	Decompression Alone	Decompression and Posterior fusion	Decompression and posterior fusion with interbody fusion

Table 2 - UCSF DS Scoring System

Segmental dynamic instability	Location of spinal stenosis	Sagittal alignment	Primary clinical presentation
<3mm translation	Central or lateral recess stenosis alone	Maintained segmental lordosis	Primary leg pain
3-5 mm translation	Foraminal stenosis without up/down stenosis	Segmental neutral or kyphotic alignment	Both leg and back pain
>5mm translation	Foraminal stenosis with up/down stenosis	Global malalignment (SVA >5cm or PT > 30°)	Primarily Back pain

Table 3 - UCSF DS with added surgical suggestion

Segmental dynamic instability	Location of spinal stenosis	Sagittal alignment	Primary clinical presentation	Surgical Intervention
<3mm translation	Central or lateral recess stenosis alone	Maintained segmental lordosis	Primary leg pain	Decompression alone
3-5 mm translation	Foraminal stenosis without up/down stenosis	Segmental neutral or kyphotic alignment	Both leg and back pain	Decompression with posterior fusion
>5mm translation	Foraminal stenosis with up/down stenosis	Global malalignment (SVA >5cm or PT > 30°)	Primarily Back pain	Decompression with posterior fusion and interbody fusion

Table 4 - Surgical Interventions across dataset

Decompression Alone	Decompression with Posterior Fusion	Decompression with LIF (Data removed from set)	Decompression with Posterior Fusion and LIF
11	56	5 (4 TLIF, 1 ALIF)	22 (16 TLIF, 5 PLIF, 1 ALIF)

Table 5 - Vertebral levels of disease across dataset

L3/L3/L5	L4/L5	L4/L5/S1
22	56	11

Table 6 - Surgical Interventions for each Vertebral Category of Disease

	Decompression Only	Decompression with Posterior Fusion	Decompression with Posterior Fusion and LIF
L3/L4/L5	4	16	2 (1 TLIF, 1 PLIF)
L4/L5	5	33	18 (14 TLIF, 4 PLIF)
L4/L5/S1	2	7	2 (1 TLIF, 1 ALIF)

Table 7 - Overall F1-Score for each scoring system and method

	DSIC	UCSF DS
HV	0.371	0.438
HNZ	0.303	0.449

Table 8 - Estimation of each scoring system relative to actual surgical intervention

	Underestimations	Matches	Overestimations
UCSF DS HV	0.416	0.438	0.146
UCSF DS HNZ	0.101	0.449	0.449
DSIC HV	0.236	0.371	0.393
DSIC HNZ	0.022	0.303	0.674

Table 9 – Precision, Recall and Fβ in each category of intervention for each scoring system. (HV Scoring Only)

	Decompression	Decompression + Posterior Fusion	Decompression + Posterior Fusion + Interbody Fusion
Precision	DSIC – 0.667 UCSF DS – 0.171	DSIC – 0.667 UCSF DS – 0.718	DSIC – 0.272 UCSF DS – 0.333
Recall	DSIC – 0.090 UCSF DS – 0.545	DSIC – 0.357 UCSF DS – 0.500	DSIC – 0.545 UCSF DS – 0.227
Fβ	DSIC – 0.078 UCSF DS – 0.261	DSIC – 0.465 UCSF DS – 0.589	DSIC – 0.364 UCSF DS – 0.270

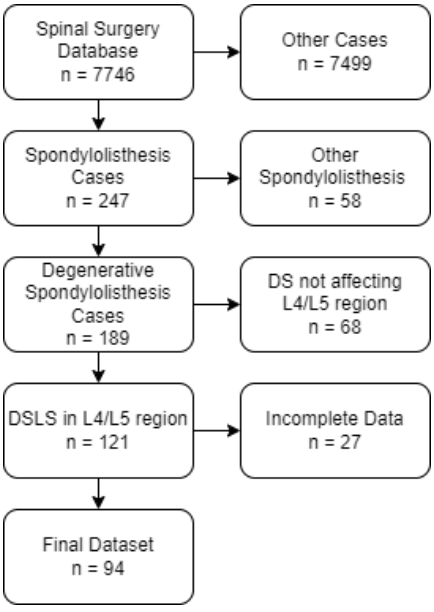

Figure 1 - Data collection process.

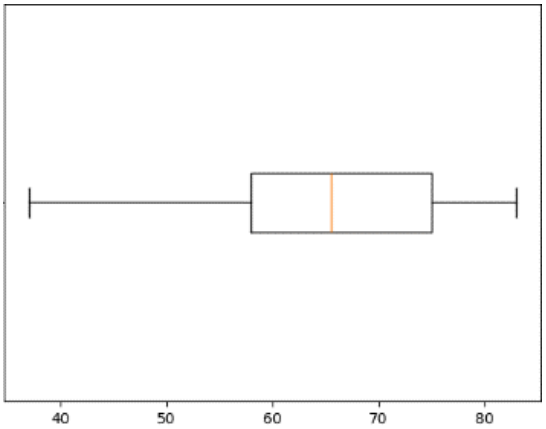

Figure 2 - Boxplot of ages in dataset.

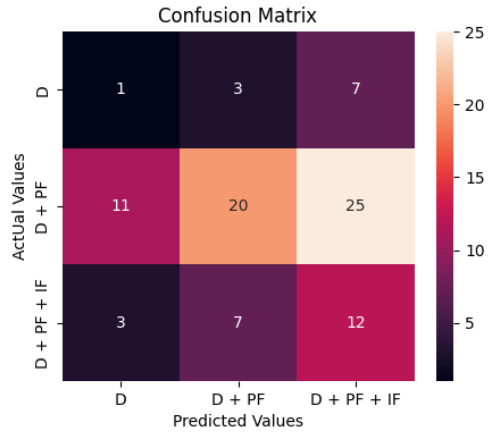

Figure 3 - DSIC using Highest Value Category Scoring

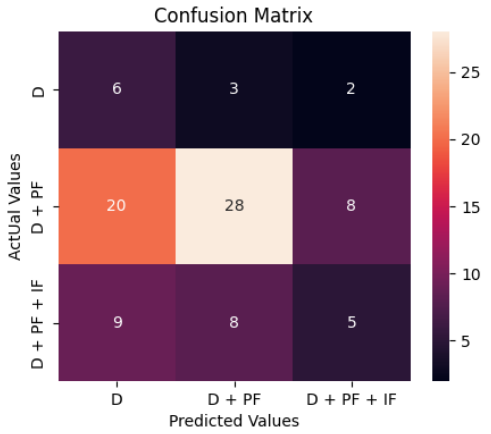

Figure 4- UCSF DS using Highest Value Category Scoring

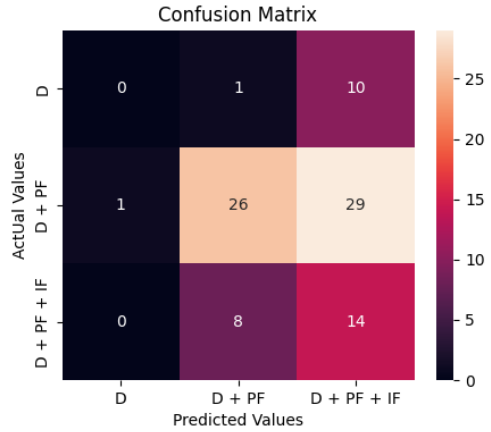

Figure 5 - UCSF DS using Highest Non-Zero Category Scoring

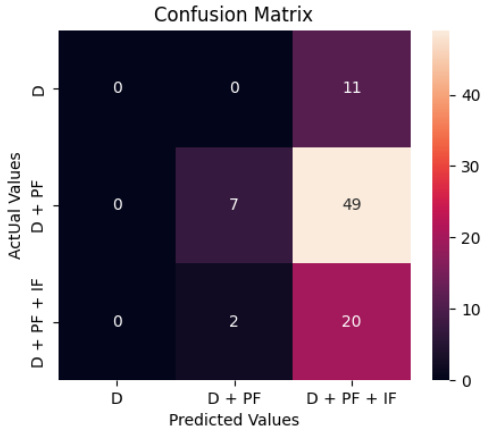

Figure 6 - DSIC using Highest Non-Zero Category Scoring

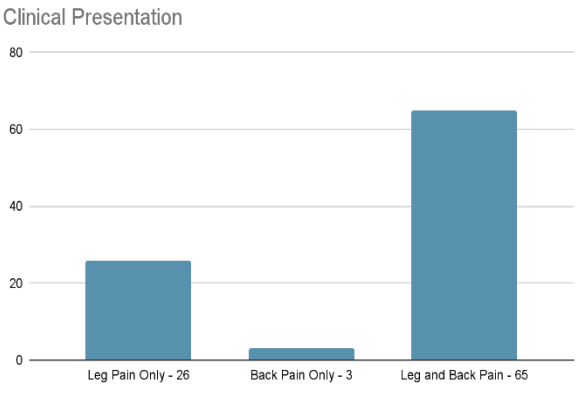

Figure 7 – Clinical Presentations in dataset

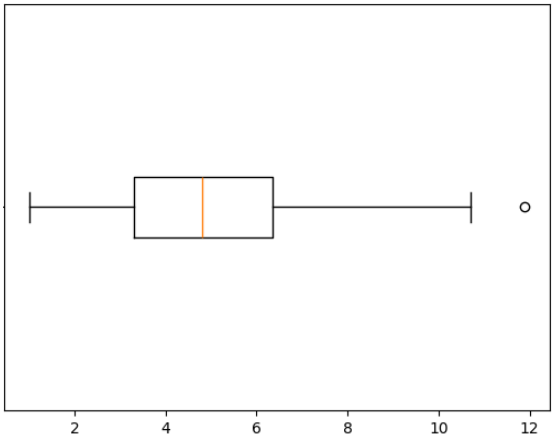

Figure 8 – Boxplot of segmental translation in L4/5 segment across dataset in mm