Automated prediction of the Thoracolumbar Injury Classification and Severity Score from CT using a novel deep learning algorithm

Sophia A. Doerr, MSE, 1 Carly Weber-Levine, MS, 1 Andrew M. Hersh, AB, 1 Tolulope Awosika, BS, 1 Brendan Judy, MD, 1 Yike Jin, MD, 1 Divyaansh Raj, BS, 1 Ann Liu, MD, 1 Daniel Lubelski, MD, 1 Craig K. Jones, PhD, 2 Haris I. Sair, MD, 3 and Nicholas Theodore, MD 1

1Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore; 2Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore; and 3Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, Maryland

OBJECTIVE Damage to the thoracolumbar spine can confer significant morbidity and mortality. The Thoracolumbar Injury Classification and Severity Score (TLICS) is used to categorize injuries and determine patients at risk of spinal instability for whom surgical intervention is warranted. However, calculating this score can constitute a bottleneck in triaging and treating patients, as it relies on multiple imaging studies and a neurological examination. Therefore, the authors sought to develop and validate a deep learning model that can automatically categorize vertebral morphology and determine posterior ligamentous complex (PLC) integrity, two critical features of TLICS, using only CT scans.

METHODS All patients who underwent neurosurgical consultation for traumatic spine injury or degenerative pathology resulting in spine injury at a single tertiary center from January 2018 to December 2019 were retrospectively evaluated for inclusion. The morphology of injury and integrity of the PLC were categorized on CT scans. A state-of-the-art object detection region-based convolutional neural network (R-CNN), Faster R-CNN, was leveraged to predict both vertebral locations and the corresponding TLICS. The network was trained with patient CT scans, manually labeled vertebral bounding boxes, TLICS morphology, and PLC annotations, thus allowing the model to output the location of vertebrae, categorize their morphology, and determine the status of PLC integrity.

RESULTS A total of 111 patients were included (mean ± SD age 62 ± 20 years) with a total of 129 separate injury classifications. Vertebral localization and PLC integrity classification achieved Dice scores of 0.92 and 0.88, respectively. Binary classification between noninjured and injured morphological scores demonstrated 95.1% accuracy. TLICS morphology accuracy, the true positive rate, and positive injury mismatch classification rate were 86.3%, 76.2%, and 22.7%, respectively. Classification accuracy between no injury and suspected PLC injury was 86.8%, while true positive, false negative, and false positive rates were 90.0%, 10.0%, and 21.8%, respectively.

CONCLUSIONS In this study, the authors demonstrate a novel deep learning method to automatically predict injury morphology and PLC disruption with high accuracy. This model may streamline and improve diagnostic decision support for patients with thoracolumbar spinal trauma.

https://thejns.org/doi/abs/10.3171/2022.1.FOCUS21745

KEYWORDS machine learning; spine; trauma; injury; motor vehicle crash; TLICS; Thoracolumbar Injury Classification and Severity Score

THE most frequent form of traumatic spine injury involves damage to the thoracolumbar spine, due to the unique biomechanics of this region as a highly flexible and mobile transition zone between the kyphotic thoracic and lordotic lumbar spine. 1 Injury to the thoracolumbar spine most commonly results from motor vehicle accidents and high-energy falls and can confer significant morbidity, including permanent neurological deficits and mortality. 2 Thoracolumbar injuries comprise a heterogeneous group, with management ranging from conservative
treatment for small fractures to surgical intervention for cases of spinal instability.3

There are several thoracolumbar injury classification methods used for clinical decision support, including the Magerl, Denis, and McAfee classification systems.4 Nonetheless, the decision to operate can be controversial given difficulties in predicting the stability of spine injuries and a lack of clear treatment guidelines. The Thoracolumbar Injury Classification and Severity Score (TLICS) was proposed in 2005 to provide a classification system that accounted for neurological deficits and spinal stability and is intended to provide guidelines for operative and nonoperative decision-making. TLICS consists of 3 parameters: 1) injury morphology, defined as compression fracture, burst fracture, translational/rotational injury, or distraction injury; 2) integrity of the posterior ligamentous complex (PLC), defined as intact, suspected/indeterminate disruption, or definite disruption; and 3) neurological status, defined as intact neurological examination, nerve root injury, complete cord injury, incomplete cord injury, or cauda equina injury. A score ≥ 5 is an indication for operative management.5 The system has demonstrated relatively high levels of reliability and validity;6 however, cases that use MRI to evaluate PLC integrity generally yield higher interrater variability than those that do not.7

In clinical practice, TLICS requires a CT scan, neurological examination, and potentially MRI when PLC integrity is not clear from CT. In addition, a trained expert, either a radiologist or neurosurgeon, must interpret and classify the imaging, producing potential bottlenecks in patient care and increasing costs to patients. In recent years, deep learning, a subset of artificial intelligence, has demonstrated high performance in replicating medical imaging tasks, such as segmentation, classification, and detection. Convolutional neural networks, in particular, have shown success in imaging tasks, demonstrating the ability to retain salient contextual information surrounding a particular object in an image. Recently, machine learning has shown success in interpreting and quantifying imaging studies for neurosurgeons; however, few studies have applied state-of-the-art deep learning methods to automate neurosurgical decision support tools.8 Deep learning for computer-aided diagnosis can potentially streamline diagnostic and treatment planning by improving diagnostic confidence, the objectivity of metrics, and the teaching of diagnostic techniques.9

The aim of this work was to develop and validate a novel deep learning method to automatically classify the vertebral morphology and PLC integrity components of TLICS. The approach leverages the objectivity of a model-based method to supplement and streamline radiological workflows with consistent, data-driven methods in preoperative neurosurgical care.

Methods

Patient Population

The authors retrospectively reviewed all patients who underwent neurosurgical consultations at a single tertiary center from January 2018 to December 2019. Patients were included if they presented with traumatic injury to the spine or degenerative pathologies resulting in spine injury (e.g., degenerative spondylolisthesis) and obtained a preoperative CT scan of the spine injury site. The morphology of the injury and integrity of the PLC were classified according to TLICS categories. A minimum of 10 patients were included for each permutation of injury morphology and PLC integrity to address class imbalances associated with training a machine learning classification algorithm. Compression fractures were the most common patient injury presentation, while rotations and translations were the least common (Table 1).

Image Annotation and Preparation

For each vertebral level present in a training data set CT image, the morphology was classified as no injury, compression fracture, burst fracture, translational or rotational injury, or distraction injury. Using both CT and MRI, the PLC integrity was categorized as intact, suspected or indeterminate injury, or injured. Annotation was performed by neurological residents and a medical imaging specialist under the oversight of those residents. Although TLICS is typically computed as a single score per patient case, we annotated TLICS per vertebral level for model training. After training on these ground truth inputs, the resulting model can predict TLICS per vertebral level given unannotated sagittal CT image slices of novel images.

As an additional input for model training, we manually annotated the vertebral centroids in each CT image using 3D Slicer (https://www.slicer.org).10 By leveraging vertebral localization and TLICS as inputs to the network, the network training can maximize the contextual information between location and class to improve performance.

Network Model and Training

We selected a state-of-the-art object detection region-based convolutional neural network (R-CNN), Faster R-CNN (Torchvision), for the deep learning network architecture.11 This architecture performs with high accuracy in detecting and classifying objects in 2D images.12 Our network outputs the location of vertebrae in 2D slices of CT images and categorizes their morphology and PLC integrity based on the TLICS system. A multitask detection and classification network was chosen because studies have suggested that multitask networks, which enforce the learning of object location and classification, perform superior to single-task classification networks.13 Additionally, associating TLICS with a specific vertebra allows for improved model interpretability and diagnostic utility.

<table>
<thead>
<tr>
<th>Morphology</th>
<th>PLC, n (%)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intact</td>
<td>Suspected or Indeterminate Injury</td>
<td>Injured</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Compression fracture</td>
<td>24 (41)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Burst fracture</td>
<td>19 (32)</td>
<td>11 (39)</td>
<td>10 (24)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Translation or rotation</td>
<td>16 (27)</td>
<td>12 (43)</td>
<td>16 (38)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Distraction</td>
<td>NA</td>
<td>5 (18)</td>
<td>16 (38)</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

NA = not applicable.
Two separate networks were trained. The first localizes and classifies vertebrae into TLICS morphology classes (0–4), and the second localizes and classifies vertebrae into binary PLC integrity scores (0 and 1 for intact and suspected PLC injury, respectively). Vertebral localization is represented as four coordinates that define a bounding box, which surrounds a given vertebra. We computed bounding box labels by extrapolating a constant width and height box centered around the manually annotated vertebral centroids. Both networks were trained using 2D sagittal slices of patient CT scans, the corresponding manually labeled vertebral bounding boxes, their TLICS morphology, and the PLC annotations as ground truth inputs. For each patient case, we extracted multiple regions of approximately 3 to 7 vertebral bodies comprising sagittal slices around the mean coronal vertebral coordinate. This preparation created 50 images per patient CT scan. We applied standard data augmentation methods to further expand the training and testing data sets and ensure robustness against natural variations in clinical imaging. Data augmentation included randomized translations, rotations, and noise simulation applied at each iteration of training. The resulting 90%/10% train-test split corresponds to more than 5000 training and 500 testing images before augmentation.

Figure 1 depicts the network diagram flow for the architecture. Training was performed and optimized using a parameter sweep over the learning rate. The morphology network was trained first, and using transfer learning, the optimized parameters for the feature extractor layers were frozen for training of the PLC network. Training and testing were performed with a 90%/10% train-test split, ensuring that no patient case appeared in both training and testing data. Evaluation was performed on testing data via localization and classification metrics.

**Results**

**Patient Characteristics**

A total of 111 patients assessed from January 2018 to December 2019 were included, consisting of 56 male and 55 female patients with a mean ± SD age of 62 ± 20 years. A total of 129 separate injury classifications were recorded since the injuries were at multiple vertebral levels in patients (Table 2). A known traumatic cause was identified in 94.6% of injuries, including 29 patients who were involved in motor vehicle accidents, 67 patients who experienced a traumatic fall, 1 patient who experienced a pathologic injury from cancer, 1 patient who was assaulted, and 7 patients with spondylolisthesis. Fourteen patients with nontraumatic causes were included in the study to alleviate class imbalance. Twenty patients (18%) had cervical injuries, 35 patients (32%) had thoracic injuries, 49 patients (44%) had lumbar injuries, and 7 patients (6%) had injuries across multiple spine regions. Fifty-seven TLICSs (51%) were < 4, 11 (10%) were equal to 4, and 43 (39%) were > 4, indicating nonoperative, borderline, or operative treatment, respectively. Following injury, 43 patients (39%) received support braces, 13 patients (12%) underwent fusion, 2 patients (2%) underwent decompression, 31 patients (28%) underwent both decompression and fusion, and 22 patients (20%) received no further treatment. The mean time to surgery for patients undergoing surgical intervention was 3 days from presentation.

Network performance was evaluated using measures of vertebral localization and injury classification on 11 test cases. According to TLICS, 5 of these cases were stable and 6 were unstable, reflecting a representative sample of the training data set. Vertebral localization evaluation included the Dice score and Intersection over Union (IoU)
for predicting ground truth bounding box labels. Injury classification evaluation included accuracy, true positive rate, false positive rate, true negative rate, false negative (FN) rate, and a positive injury mismatch classification rate.

The morphology network demonstrated a Dice score of 0.92. Binary classification accuracy (between noninjured morphology and injured morphological scores) demonstrated 95.1% accuracy, with a TLICS morphology accuracy, true positive rate, and positive injury mismatch classification rate of 86.3%, 76.2%, and 22.7%, respectively, for CT sagittal slice prediction. High accuracy in binary classification compared with morphology-specific classification indicates a possible need for more samples of each morphology class to accurately model variation. Table 3 depicts classification performance results for morphology. The PLC network demonstrated a Dice score and IoU of 0.88 and 0.85, respectively. Classification accuracy between no injury and suspected PLC injury was 86.8%, and the true positive, FN, and false positive rates were 90.0%, 10.0%, and 21.8%, respectively. Output image predictions for morphology and PLC are illustrated in Fig. 2.

Discussion

Injuries to the thoracolumbar spine comprise a heterogeneous group so that management is controversial, with the decision for operative management contingent on the degree of spinal instability. TLICS provides a framework to assess the need for operative management and relies on a CT scan to determine vertebral stability, MRI for examining the integrity of the PLC, and a physical examination for neurological deficits. Herein, we have illustrated that a deep learning neural network can analyze CT spine images and predict injury morphology and PLC disruption with reasonable accuracy. Further validation and development of the model can have important implications for spine trauma, allowing physicians to decide on treatment more rapidly for patients presenting in an acute setting and reducing costs associated with imaging tests and the length of hospitalization. Furthermore, the algorithm is useful in low-resource settings that lack access to MRI machines or physicians skilled in TLICS use.

Other classification systems exist for thoracolumbar injury, including the Denis 3-column model, in which disruption of 2 spinal columns leads to instability, and the AO Spine Trauma Classification system, which incorporates patient-specific modifiers that vary depending on the location of injury. The proliferation of classification systems reflects long-standing difficulty and disagreement in defining and measuring spinal instability, which can result in inappropriate treatment of patients. The TLICS system focuses on categories of instability rather than the degree of instability, including a measure of immediate mechanical stability (injury morphology), long-term stability (PLC integrity), and neurological stability (presence of neurological deficits). Compared with other scoring systems, TLICS has shown high validity and intra- and interrater reliability. For example, Park et al. retrospectively reviewed 328 patients with thoracolumbar injuries and found that TLICS matched the decision for conservative treatment in 95% of cases and for operative treatment in 84% of cases. Interestingly, while Dawkins et al. found high interrater reliability using TLICS in a study of 81 pediatric patients, the reliability decreased when MRI was used by surgeons for injury assessment. Our study demonstrates that a data-driven algorithm can reliably detect a suspected disruption of the PLC, in most cases, without the need for MRI. Although the most severe injury is used to determine the course of treatment, all vertebral injuries are considered pertinent for surgical approach or course of treatment decision-making. Therefore, our model, which automatically predicts TLICS by vertebral level, has utility in both diagnostic and decision support domains.

Machine Learning in Spine Surgery

The use of deep learning in medical imaging has be-
come increasingly popular in recent years, enabled by advances in graphics processing units, data augmentation, transfer learning, large-scale open-source image data sets, and improved convolutional network architectures.18 Deep learning has demonstrated success in replicating complex imaging tasks in many domains of medicine but has seldom been applied to neurosurgery. Instead, conventional machine learning models have been applied to spine surgery, including the fields of spine oncology, degenerative spine surgery, and spine trauma. A systematic review by Lubelski et al. found that nearly 42% of prediction models for outcomes following elective degenerative spinal surgery were generated using machine learning algorithms—the same percentage as those using conventional logistic regression analysis.19 For example, Karhade et al. compared four machine learning algorithms to predict nonroutine discharge in 26,364 patients identified in the American College of Surgeons National Surgical Quality Improvement Program database, finding a C-statistic of 0.82 for their neural network model and deploying the calculator online.20 Additionally, several recently published deep learning methods have demonstrated success using radiographic spine surgery images, including the automatic detection of spine instrumentation and cervical spinal cord compression.21,22

These models have also found increasing utility in the field of neurotrauma. Huie et al. propounded that the unique complexity and heterogeneity of traumatic spinal injuries in the acute and chronic settings rendered algorithmic decision-making particularly important.23 Khan et al. reviewed 9 articles describing machine learning algorithms, mostly using supervised learning models, for spinal cord injury. Identified algorithms ranged from those predicting functional outcomes after injury to those predicting prolonged opioid use after surgery.24–26 Based on their identified trends, the authors proposed that machine learning will occupy an increasingly prominent role in understanding and managing spine trauma.26 However, Khan et al. identified relatively few studies focused on neuroimaging in spine surgery. One study by Tay et al. used machine learning to diagnose spinal cord injury in the C4–6 region using diffusion tensor imaging,27 while McCoy et al. deployed a convolutional neural network to study T2-weighted MR images of patients with spinal cord injury to provide metrics of motor impairment.28 Although these are important applications, no studies have used machine learning of radiographic imaging to decide on the question of conservative or surgical management. By outputting a TLICS, our algorithm can serve a crucial function in the trauma setting and aid clinical decision-making concerning optimal management.

A few studies in the spine neuroimaging literature have used machine learning to diagnose traumatic pathology of the vertebral spine. Frighetto-Pereira et al. used the k-nearest-neighbors neural network and naive Bayes classifier to diagnose compression fractures based on MRI, obtaining a model with a high area under the curve of 0.97. The model could also differentiate between benign and malignant fractures with high accuracy.29 Amitai et al. extended this work to CT scans using a convolu-
dutional neural network, obtaining nearly 90% accuracy,\textsuperscript{30} while Burns et al. used machine learning to measure bone density from CT imaging and to classify the severity of fractures using the Gennat classification system.\textsuperscript{31} Mehta and Sebro showed that a support vector machine could identify lumbar spine fractures from routine, dual-energy x-ray absorptiometry (DEXA) studies, without the need for additional imaging, and could even identify fractures missed by radiologists.\textsuperscript{32} Although diagnosis of pathology is vital, we have shown that deep learning can also be used to assist clinical decision-making. By developing an algorithm that can interpret sagittal CT spine images and predict TLICS morphology and PLC components, we demonstrate the ability to provide surgeons fast and simple access to information on the severity of injury and need for surgical intervention.

Future work includes expanding the cohort of data used for training—particularly in the rare translational and rotational spine trauma category—to improve algorithmic performance. Analysis of the mismatch injury rates in morphology and false positive rates in PLC suggests the possibility of overfitting when comparing training and testing metrics. For full clinical adoption, machine learning techniques are generally expected to maintain error rates ranging from 3% to 5% across diverse populations. Further inclusion of a larger cohort and regularization in the network may help to alleviate this problem; however, tools with higher error rates may still prove useful for clinical decision support if not complete automation. Additionally, modifications will be tested in the network to determine if including multiple 2D slices improves the prediction of traumatic injury, as a single 2D slice often does not capture the presence of all injuries for a typical human reader. Future studies will also be aimed at external validation of the algorithm with an interrater control study and improving its accuracy in larger cohorts.

Ultimately, the algorithm is intended for use in the clinical setting, where it can reduce the need for expensive MRI and provide physicians a risk assessment level regardless of their expertise in calculating TLICS. Integration of the algorithm with electronic medical records would allow for simple and widespread use. Nonetheless, the algorithm should not replace expert clinical judgment. A thorough history and physical examination remain valuable tools for the physician; indeed, the integration of novel machine learning algorithms with traditional patient-centered examinations is becoming an increasingly important and useful method.

Limitations

Limitations of this study include the small sample size for traumatic injuries resulting from translational or rotational injury, which have a low prevalence. To account for this limitation, we included several patients with degenerative spondylolisthesis and cervical injury. Although inclusion of these patients provided the algorithm with a more complete depiction of translational pathologies, these cases may have provided additional variation that skewed predictions. Additionally, the algorithm would benefit from external validation by peer institutions to determine its generalizability.

Conclusions

In this study, we demonstrated the success of leveraging deep learning for prediction of PLC and morphology components of TLICS. The morphology network can distinguish between injured and noninjured morphology and between subclasses of morphology with 95.1% and 86.3% accuracy, respectively. Meanwhile, the PLC network can predict suspected PLC disruption with 86.8% accuracy and a 90.0% true positive rate. Automation of this decision support tool is valuable for improving reader confidence, streamlining the diagnostic and decision-making process, and retrospectively reviewing TLICS validity.

References


**Disclosures**

Nicholas Theodore: royalties from Globus Medical and DePuy Synthes; stock ownership in Globus Medical; consultant for Globus Medical and Augmedics; and scientific advisory board/other office for Globus Medical.

**Author Contributions**

Conception and design: Theodore, Doerr, Liu, Jones. Acquisition of data: Doerr, Weber-Levine, Hersh, Awosika, Judy, Jin, Raj, Liu, Lubelski, Sair. Analysis and interpretation of data: Doerr, Awosika, Judy, Lubelski, Jones. Drafting the article: Doerr, Weber-Levine, Hersh, Awosika. Critically revising the article: Theodore, Doerr, Weber-Levine, Hersh, Judy, Jin, Raj, Liu, Lubelski, Jones, Sair. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Theodore. Statistical analysis: Doerr. Administrative/technical/material support: Theodore, Jones, Sair. Study supervision: Judy, Jin, Liu, Lubelski, Jones, Sair.

**Correspondence**

Nicholas Theodore: Johns Hopkins University, Baltimore, MD. theodore@jhmi.edu.