■ INSTRUCTIONAL REVIEW

The diagnostic and prognostic value of artificial intelligence and artificial neural networks in spinal surgery

A NARRATIVE REVIEW

J. M. McDonnell,
S. R. Evans,
L. McCarthy,
H. Temperley,
C. Waters,
D. Ahern,
G. Cunniffe,
S. Morris,
K. Synnott,
N. Birch,
J. S. Butler

From Mater Misericordiae University Hospital, Dublin, Ireland

In recent years, machine learning (ML) and artificial neural networks (ANNs), a particular subset of ML, have been adopted by various areas of healthcare. A number of diagnostic and prognostic algorithms have been designed and implemented across a range of orthopaedic sub-specialties to date, with many positive results. However, the methodology of many of these studies is flawed, and few compare the use of ML with the current approach in clinical practice. Spinal surgery has advanced rapidly over the past three decades, particularly in the areas of implant technology, advanced surgical techniques, biologics, and enhanced recovery protocols. It is therefore regarded an innovative field. Inevitably, spinal surgeons will wish to incorporate ML into their practice should models prove effective in diagnostic or prognostic terms. The purpose of this article is to review published studies that describe the application of neural networks to spinal surgery and which actively compare ANN models to contemporary clinical standards allowing evaluation of their efficacy, accuracy, and relatability. It also explores some of the limitations of the technology, which act to constrain the widespread adoption of neural networks for diagnostic and prognostic use in spinal care. Finally, it describes the necessary considerations should institutions wish to incorporate ANNs into their practices. In doing so, the aim of this review is to provide a practical approach for spinal surgeons to understand the relevant aspects of neural networks.

Cite this article: ***Bone Joint J*** **2021;103-B(9):1442–1448.**

Introduction

Artificial intelligence (AI) is a term originally coined by Dr John McCarthy.[1] It describes the inevitable progression of the functionality of computers that learn to perform tasks by pattern recognition, with minimal or no human input. Although similar, there are notable differences between AI and machine learning (ML). AI refers to a broad class of technological systems that are designed to simulate human behaviour.[2] The application of AI is evident across various industries, including image recognition (automated photo tagging on social media), speech-to-text (smartphone dictation), natural language processing (chatbots), recommendation systems (personalized advertisements), video classification (security cameras), and tabular systems (email spam filters).[2-4] The scope of AI is therefore vast. ML and deep learning (DL) are the two main subsets of AI (Figure 1). The premise of ML is that an external user or operator provides data to a machine or model, and allows that machine or model to learn and design algorithms for novel application.[2,3] ML has a narrower scope than AI, and can only perform tasks for which it is trained. This can be achieved through three different forms of training: supervised, unsupervised, and reinforcement.

ML models. The three most commonly employed ML models are logistic regression (LR), support vector machines (SVM), and artificial neural networks (ANNs), due to their ease of design and implementation, as well as their recognized predictive ability.[2,3,5] The most common method of distinguishing between these three models is their origin. LR was designed and developed by statisticians, while the latter two were developed by computer scientists concerned with novel ML algorithms and models. The premise of LR is the examination of the relationship between variables, referred to as inputs, and an outcome which can be continuous or categorical, referred to as an output.[6] LR can be designed in a backward (variables and outcome known) or forward (outcome unknown) selection manner.[7] Stepwise LR is a combination

Correspondence should be sent to J. M. McDonnell; email: jakemcdonnell@rcsi.ie

doi:10.1302/0301-620X.103B9.BJJ-2021-0192.R1 $2.00

Bone Joint J 2021;103-B(9):1442–1448.

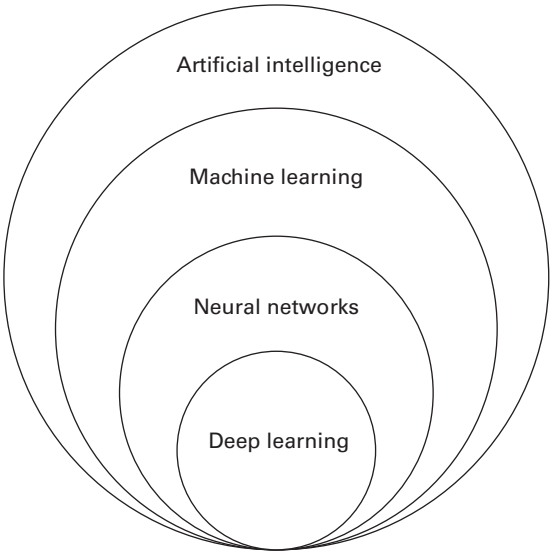

Fig. 1

Artificial intelligence and sub-categories.

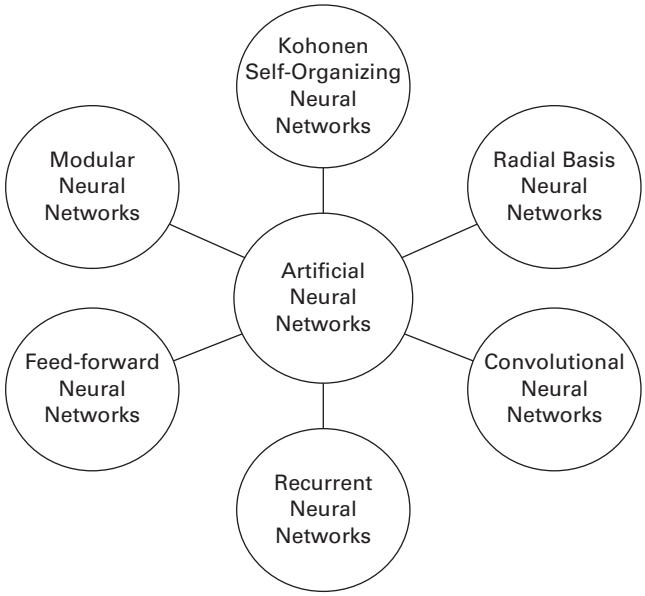

Fig. 2

Different types of artificial neural networks.

of backward and forward selection, in which variables are automatically selected to best fit the regression model.

SVMs are supervised linear models that function by finding an optimal separation line (referred to as a hyperplane) to differentiate between two classes, or two sets of data.[5] An optimal hyperplane is one defined as having maximum margin. Margin is calculated by taking the points closest to the hyperplane, known as support vectors, from each respective dataset, and calculating the distance from the hyperplane to support vectors. Important parameters in SVM models are "C" and "γ". C allows for control of error within the model. γ indicates the degree of spread influence of the support vectors.[5] Support vectors located close to the boundary are reported to have a high γ, while those far from the boundary have a low γ. High γ is typically represented by a line with a large degree of curvature and is often inferred as a model with a high degree of bias.

ANNs are computer models designed to mimic biological neural networks.[2-4] Their structure consists of multiple interconnected "nodes" arranged in three layers: input, hidden, and output.[2] Each node in the input layer represents an input variable. There is typically one node in the output layer, which represents the outcome of interest. Nodes in the hidden layer allow ANNs to model complex relationships between the input nodes and output node. Nodes in respective layers are connected by connection weights, referred to as arcs. These connection weights represent the relationship between variables, similar to coefficients in a LR model. ANNs function to learn these relationships and to develop prediction models by learning and training.[2] Although ANNs might appear to be merely automated LR models, this is not the case as the training algorithms are distinctly different. The most common method of ANN training is by backpropagation, which instructs the machine to adjust its internal parameters, allowing computation of the output of each layer of "neurones" comparing it to the previous one.[2,6,7] These intricate learned patterns, once set in motion, make ANN models largely autonomous. DL models are simply an extension of this concept, and serve to understand the complex relationship between larger datasets by transformation and extraction of a greater amount of data through additional nodes in the hidden layer.[3,4] The various types of ANNs are shown in Figure 2. The individual functionality and scope of each ANN is beyond the scope of this article.

Orthopaedic applications of AI. AI has shown great potential in the fields of image recognition, preoperative risk assessment, and clinical decision-making. A literature review by Cabitza et al[8] highlights the work that has been done to date. They analyzed 70 studies which had implemented ML in the field of orthopaedics over the previous 20 years, including the detection of spinal pathology, the identification of anterior and posterior cruciate ligaments, fractures, and the classification of osteoarthritis and cartilage imaging.

Although a number of studies have shown that ML and neural network applications are potentially effective in orthopaedics, including spinal surgery, few have actively compared those models with current diagnosticprognostic standard in clinical practice. To address this in terms of spinal care, the following sections review the literature on the application of ANN to spinal surgery, comparing the models described to the current diagnostic and prognostic standards to evaluate their efficacy, accuracy, and relatability. Additionally, this review aims to delineate certain limitations of ANN models and discuss how institutions can begin the process of introducing these models into clinical practice.

Diagnostic validity of ML in spinal surgery. ML shows promise in analyzing and identifying abnormalities in radiological images. A number of recently published studies describe the development of ML algorithms for analyzing spinal

images and comparing them with the analysis of expert radiologists and surgeons.[9-13]

Pan et al[9] assessed the ability of two Mask R-CNN (convolutional neural networks, a form of ANN) models to detect, segment, and measure the Cobb angles of 248 chest radiographs in patients with lung cancer and compared the results with those of two experienced radiologists. Throughout this study, the CNN models were referred to as the computer-aided method (CAM), and the radiologists were referred to as the manual method. For the radiologists, the intraclass correlation coefficients (ICCs) of intra- and interobserver reliability analysis were 0.941 and 0.887 with a mean absolute difference (MAD) between the two radiologists of < 3.5°. The MAD provides an indication of the mean variance or discrepancy between the values reported by each radiologist. The ICC between CAM and the manual method was 0.854 with a mean absolute difference of 3.32°. In a separate comparison between radiologist 1 and CAM, the ICC was 0.868 (95% confidence interval (CI) 0.819 to 0.902) with a MAD of 3.33°. The ICC for radiologist 2 and CAM was 0.812 (95% CI 0.723 to 0.868). The sensitivity, specificity, and accuracy for the CAM were 89.59%, 70.37%, and 87.50% respectively, showing that the CAM had the potential to diagnose scoliosis.

Zhang et al[10] trained a deep neural network (DNN, a DL model), to measure the Cobb angle in scoliosis patients from 275 PA radiographs of a spine model, with the aim of reducing variability of measurement. The results were compared to the manual measurements of 105 radiographs (40 model and 65 in vivo films) made by an experienced spinal deformity surgeon. The ICCs for intraobserver analysis of the model radiographs ranged from 0.937 to 0.986, with a MAD of < 3°. For the patient radiographs, the ICCs ranged from 0.901 to 0.953 with an MAD of 4.5°. Interobserver ICCs were 0.870 to 0.980 for the model films (MAD: 2.9°) and 0.862 to 0.889 for the in vivo films (MAD: 5.1°). Automatic measurements using the DNN were compared to manual measurements made by two examiners, one a spinal clinician with 21 years' experience in a scoliosis clinic and the other a software engineer with no previous radiological experience. The ICC for the model films was > 0.91 (95% CI 0.815 to 0.962) and for in vivo radiographs 0.771 to 0.835 (95% CI 0.602 to 0.914). The authors concluded that the automatic method of Cobb angle measurement showed good agreement with the manual measurement method and reduced variability of measurement, but future systems would need to include in vivo radiographs in the DNN training set to allow the system to measure Cobb angles accurately.

In 2017, Jamaludin et al[11] described a system using a convolutional neural network (CNN) to grade lumbar intervertebral discs and vertebral bodies for signs of degeneration using 12,018 MRI images from 2009 patients collected during the Genodisc Project. They assessed Pfirrmann grade[14] of each disc, disc narrowing, spondylolisthesis, central canal stenosis, and the presence of endplate changes. The performance of the model was compared with the intraobserver class mean accuracy of an expert spinal radiologist. The model produced an accuracy of 95.6% for labelling and disc detection, only failing on images of inadequate quality. The researchers found that the difference on average between the model and the intrarater (radiologist) agreement was around 0.4%. Such results indicate that the model is a close automated analogue of the radiologist in terms of the ability to analyze some features of MRI scans. Although the model could produce predictions of pathological gradings comparable to manual interpretation, it tended towards predicting more abnormal/pathological findings than the radiologist. Advantages of this "flaw" could be that a screening model is created providing a safety net for review by the radiologist. The authors concluded that automation of radiological grading was on par with human performance.

An automated CNN was developed by Weng et al[12] for measuring sagittal vertebral axis (SVA) on 990 standing whole spine lateral radiographs. The software, ResUNet (the name given to the CNN model), was developed for the detection of degenerative changes and deformities in the vertebral column. Following training of the software, the ICC of the inter-rater reliability of human experts and ResUNet was 0.946 to 0.993, indicating excellent consistency and reliability in detection and therefore its use in clinical settings.

Korez et al[13] designed and employed two DL models, RetineNet and U-Net, for the fully automated measurement of sagittal spinopelvic balance from radiographs of the spine, comparing the results with manual measurements. They assessed sacral slope, pelvic tilt, spinal tilt, pelvic incidence, and spinosacral angle. The MAD between the DL results and the manual measurements was 3.9° (1.2° to 5.5°) and the correlation coefficients ranged from 0.71 to 0.95. They concluded that apart from a few outlier images, the DL tool was equivalent to manual measurement. A summary of findings in all four studies are outlined in Table I.

Despite some limitations, many of the models developed show promise as diagnostic aids for surgeons and radiologists. These studies indicate that neural network models can provide a rapid and objective radiological analysis in the clinical setting, including the possible automation of diagnosis from plain radiographs that would reduce time to diagnosis. For example, the ResUNet algorithm demonstrated an inference time for one radiograph of 0.2 s, demonstrating the rapid screening capacity for large datasets in the clinical setting,[12] with potential beneficial repercussions in terms of optimizing future treatment strategies.

Prognostic utility of AI in spinal surgery. Predictive models, such as multivariate LR, can assist diagnostically in spinal conditions and can contribute to the optimization of treatment strategies for patients.[10] ML models, particularly ANNs, have the potential to add to, or even eventually replace, the prognostic use of LR for some aspects of spinal surgery including the prediction of postoperative complications, surgical satisfaction, rehabilitation needs, and the overall pathway of patient care.

For this review, seven studies were identified that investigated the predictive validity of ANN compared to LR across a range of spinal pathologies, including disc herniation and recurrence, lumbar spinal stenosis, spinal fusion, and the treatment of adult spinal deformity.[15–21] The prognostic focus included risk analysis for postoperative complications and patient satisfaction for both surgical outcomes and the process of patient care. All studies found ANN to be comparable to LR in predictive performance, with most results indicating

Table I. Summary of comparative diagnostic studies.

Author (year)	Measurement	Comparison	ICC* (95% CI)
Jamaludin et al (2017)[11]	Pfirrmann grade	Radiologist vs CNN	0.88 (N/R)
	Disc narrowing		0.89 (N/R)
Zhang et al (2017)[10]	Cobb angle	DNN vs radiologist	0.9 (0.811 to 0.991)
Pan et al (2019)[9]	Cobb angle	Radiologist 1 vs CAM	0.868 (0.819 to 0.902)
		Radiologist 2 vs CAM	0.812 (0.723 to 0.868)
Weng et al (2019)[12]	SVA	ResUNet vs rater	0.989 (0.984 to 0.993)
		ResUNet vs rater 2	0.946 (0.920 to 0.963)
		ResUNet vs rater 3	0.993 (0.989 to 0.995)
Korez et al (2020)[13]	SS	RetineNet and U-Net DL tools vs spine surgeon using SurgiMap Spine software	0.73 (N/R)
	PT		0.90 (N/R)
	ST		0.95 (N/R)
	PI		0.81 (N/R)
	SSA		0.71 (N/R)

*From publications comparing radiologist with machine learning algorithms.
CAM, computer-aided method; CI, confidence interval; CNN, convolutional neural network; DNN, deep neural network; ICC, intraclass correlation coefficient; N/R, not reported; PI, pelvic incidence; PT, pelvic tilt; SS, sacral slope; SSA, spinosacral angle; ST, spinal tilt; SVA, sagittal vertebral axis.

that ANN outperforms LR in at least one, if not all, performance measures.

Azimi et al[15] reported a risk analysis prediction model in 402 patients which aimed to determine whether ANN or LR was more accurate at predicting recurrence of a lumbar disc herniation. ANN outperformed LR with an AUC of 0.84 compared to 0.76 and had a greater accuracy than LR (94.1% vs 86.4%). Additionally, ANN outperformed LR in terms of specificity (46% vs 34%), positive predictive value (PPV) (69% vs 65%), and negative predictive value (NPV) (88% vs 82%). These results indicate that both LR and ANN can be used to predict recurrent lumbar disc herniation, and that ANN is potentially the more reliable model.

A group from Mount Sinai Hospital in New York, USA, have published three studies that compare ANN with LR using risk analysis based on data from the National Surgical Quality Improvement Program (NSQIP) database.[16–18] The first was a comparative risk analysis of postoperative complications after anterior cervical discectomy and fusion (ACDF).[16] The authors identified 20,879 patients who had an ACDF between 2010 and 2014. The ANN models were trained to predict the occurrence of venous thromboembolism (VTE), cardiac complications, wound complications, and mortality. ANN models were compared with American Society of Anesthesiologists[22] (ASA) physical status of the patients and their performance was represented as the area under the receiver operating characteristic (AUROC) curve, a popular measure of how well a model can predict the primary outcome. The ASA physical status classifiers were consistently outperformed by both LR and ANN. ANN performed better than LR when predicting complications with an AUC of 0.772 (LR 0.759) for cardiac complications, 0.656 (LR 0.639) for VTE, 0.518 (LR 0.501) for wound complications, and 0.979 (LR 0.974) for mortality. The findings of the study showed that both ANN and LR have the ability to accurately predict postoperative complications, with ANN proving to be more accurate than LR for postoperative VTE, wound complications, and mortality. Interestingly, Arvind et al[16] also reported findings for a SVM model. Similarly, ANN outperformed the SVM model in terms of predicting cardiac complications (AUC 0.772 vs 0.559), VTE (AUC 0.656 vs 0.430), wound complications (AUC 0.518 vs 0.422), and mortality (AUC 0.979 vs 0.214).

Their second study reported a risk analysis of postoperative complications in posterior lumbar fusion (PLF) using both ANN and LR.[17] A total of 22,629 patients were included in the dataset. Both ANN and LR outperformed ASA for all complications. LR outperformed ANN in predicting VTE (AUC 0.588 vs 0.567), wound complications (AUC 0.613 vs 0.606), and mortality (AUC 0.703 vs 0.680). ANN outperformed LR for predicting cardiac complications (AUC 0.710 vs 0.657). ANN was also found to be more sensitive than LR for detecting postoperative wound complications and mortality.

The third risk analysis study examined 4,073 patients undergoing correction of an adult spinal deformity (ASD).[18] LR and ANN outperformed ASA for all complications including VTE, cardiac complications, wound complications, and mortality. ANN outperformed LR in predicting cardiac (AUC 0.768), wound complications (AUC 0.606), and mortality (0.844). LR outperformed ANN in VTE predictions (AUC 0.547). ANN was once again more sensitive than LR for predicting wound complications and mortality.

These three linked studies highlight how ML models such as LR and ANN can be used to accurately predict postoperative outcomes in spinal surgery. ANN was consistently more sensitive than LR, an advantage in terms of clinical decision-making.

ML has not only been used to predict surgical outcome. LR and ANN have also been used to predict satisfaction with surgical outcomes and satisfaction with overall patient care. In 2014, Azimi et al[19] used an ANN to predict patient satisfaction after surgery for lumbar spinal stenosis in 168 patients. They compared the performance of LR to an ANN, showing that the ANN outperformed LR with an AUC of 0.80 compared with 0.76. Also, the ANN had a greater accuracy rate when compared with LR: 96.9% versus 88.4%. Similar to the previously mentioned study by Azimi et al,[19] ANN outperformed LR in terms of specificity (41% vs 34%), PPV (69% vs 63%), and NPV (89% vs 82%). These figures show that ANN has the ability to predict surgical satisfaction in patients postoperatively.

Staartjes et al[20] examined patient satisfaction in 422 patients after lumbar discectomy, using patient-reported outcome measures (PROMs). Improvement in the severity of leg pain was the main outcome measure. The minimal clinically important difference (MCID) was set at an improvement of more than 30% above baseline. ANN proved to be a better predictor of outcome than LR, with an AUC of 0.87 compared with 0.78. ANN was also more accurate, sensitive, and specific than LR (accuracy ANN 85%, LR 68%; sensitivity ANN 85%, LR 55%; specificity ANN 85%, LR 77%). It was noted that ANN also outperformed LR in terms of PPV (90% vs 60%) and NPV (79% vs 73%). Similar findings were reported for secondary outcomes, back pain, and functional disability.[20]

Matis et al[21] evaluated the predictive accuracy of ANN and LR for patient satisfaction in lumbar disc herniation. ANN marginally outperformed LR with an AUC of 0.985 compared to 0.97. ANN was also slightly more accurate and sensitive than LR with an accuracy rate of 96% and sensitivity of 98% compared with 94% and 96% for LR. ANN also marginally outperformed LR for specificity, (94% vs 92%), PPV (98% vs 97%), and NPV (94% vs 89%).

A complete summary of findings for all studies are reported in Supplementary Table i.

These studies show that there is great potential for ML models to help clinicians to optimize the management of surgical patients. Nevertheless, there is considerable scope for improvement.

Limitations of the clinical application of AI and ML. Regulation and sharing of stored patient information with AI providers remains a concern. No specific guidelines currently exist for the regulation of AI under the Health Insurance Portability and Accountability Act[23] (HIPAA). However, the European Commission published a white paper in 2020 which addressed this issue with the aim of achieving harmonization with the General Data Protection Regulation (GDPR) and Equality regulations in the coming years.[24]

For academic purposes, oversight is provided in a framework based on the Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines for diagnostic and prognostic studies involving AI and ML (TRIPOD-AI).[25] This framework goes a long way to providing assurance that such studies can be replicated, and their conclusions tested outside the originator institution. Such a framework negates the worries about potential publication bias in ML studies noted by Buchlak et al.[26]

Although such frameworks have led to significant improvements, existing studies cannot often be compared with the current gold standard in practice. The interpretation, reliability, and overall level of evidence of certain studies about ANN is questionable. Emphasis must be placed on improving the reporting of results. With regards to the performance of novel algorithms, AUROC is often favoured as a performance metric. However, its efficacy has been challenged for unbalanced datasets.[27] Furthermore, a single performance metric may not be sufficient to convey the attributes and ability of novel algorithms, and researchers should strive to report additional metrics, such as accuracy, sensitivity, specificity, and PPVs and NPVs.

Other potential inherent biases in ML have raised concerns that these technologies can make widespread systematic analytical mistakes.[20] For example, certain algorithms may not have the ability to discriminate confounders, particularly in a diagnostic sense. A study by Winkler et al[28] found that images with surgical skin markings had increased melanoma probability scores (sensitivity 100% vs 95.7%) and reduced specificity (45.8% vs 84.1%). Comparable AUROCs were also lower (0.922 vs 0.969) for images with skin markings. Thus, the presence of surgical skin markings was shown to falsely increase the melanoma probability score derived by the DL model. Other examples are training datasets with predominantly one specific demographic, whether it be age, sex, or ethnicity. Such biases divide opinion between those who favour the use of multicentre datasets that contain large numbers of patients from multiple hospitals and surgeons, and those arguing for centre-specific datasets. Staartjes et al[20] used a single-centre, single surgeon dataset of 422 patients to provide a coherent input to the model, which contrasts with the variability in patient demographics, surgical techniques, and selection criteria of a multicentre dataset. However, using multicentre data for training could expedite the development of an ANN model during a period of continuous advancements in spine surgery, allowing ML models to remain valid. Additionally, the speed of change in rapidly developing specialties could be an obstacle to the development of accurate ANN models. The need for large consistent datasets to train ML models may require retrospective collection of training data for an extensive period. During those years, advancements in surgical practice may occur, thereby rendering obsolete prediction models trained on historical data. Thus, a further important focus for future research will be on whether personalized prediction models can be created in a time-efficient manner, by either mechanism.

It remains to be understood which method would lead to improved generalizability, although one may assume the former is preferable. To date, generalization of novel algorithms has proven difficult, as indicated in a study by Hwang et al,[29] whose objective was to employ a DL model to detect abnormal chest radiographs. In this study of 54,221 radiographs, specificity was reported to vary considerably at a fixed operating point (0.566 to 1.000) across five independent datasets, highlighting the struggle to achieve reliability and reproducibility. Therefore, perhaps external validation on datasets from institutions other than those used for training is required to improve generalizability.

Implementation of neural networks into clinical practice. Internal discussion must determine whether an institution wishes to develop a novel algorithm and model for a particular purpose, or employ one already reported in the literature. This can be influenced by several factors relating to accessible resources. For example, institutions may not have sufficient personnel within their listed staff who are experienced in developing and maintaining ANN models. If this is not the case, institutions that wish to design and implement ANN models may need additional external expertise, often in the form of contracted experts or consultants. Through multidisciplinary discussion, researchers can develop a primary outcome (termed outcome X) and distinguish how many input variables they wish to incorporate (i.e. risk factors for developing X).

However, the feasibility of training an ANN model can be a barrier to successful design and implementation of the model. A major challenge to the adoption and translation of ML algorithms into clinical practice is accessibility of the data required for training. In traditional healthcare systems, data for a single patient may be stored in various locations, such as analogue medical records, digital imaging systems, and pathology archives. This can create significant difficulty and delay in collating the data needed to train a model. Electronic health records, often used by more modern healthcare systems, are not without certain difficulties and limitations, as discussed extensively by Hersh et al.[30] Therefore, institutions must establish whether they have the necessary labour resources to collect the data needed to train a model.

These difficulties can often be overcome by implementing a model already in existence which has proven efficacy. However, there may be an issue relating to generalizability. As a result, institutions are advised to validate adopted algorithms on their own patient populations and compare their results against the current standard in practice, preferably in a prospective manner. Additionally, institutions are encouraged to adhere with TRIPOD guidelines. These will provide guidance on model development and external validation.[31] By doing so, institutions should then be confident to introduce ANN into clinical practice.

The findings from the studies highlighted in this review show the potential applications of neural network models in spinal surgery. Diagnostically, they can be sensitive and accurate tools for measurement that could help to reduce workload and the time to diagnosis for radiological imaging. Prognostically, they have been shown to be capable of predicting surgical outcomes and patient satisfaction. However, more robust accurate models and methods of training are needed to provide a greater understanding of whether multicentre dataset prediction models or models individualized to specific demographics are more likely to provide accurate predictions. Such evolution will, in all likelihood, contribute significantly to the process of continuous improvement in spinal care that has gathered pace over past decades.

Take home message

- There is potential for the diagnostic and prognostic capacity of artificial neural networks in clinical practice; however, more robust models with scrupulous validation are needed.

Twitter

Follow J. M. McDonnell @JakeMc_D
Follow D. Ahern @dpahern
Follow G. Cunniffe @cunniffg
Follow National Spinal Injuries Unit Research Group @NSIURG

Supplementary material

Table summarizing comparative prognostic studies.

References

1. **McCarthy J**. What is artificial intelligence? 2007. http:// www-formal.stanford.edu/jmc/whatisai/ (date last accessed 28 July 2021).
2. **Hashimoto DA**, **Rosman G**, **Rus D**, **Meireles OR**. Artificial intelligence in surgery: Promises and perils. *Ann Surg*. 2018;268(1):70–76.
3. **Alzubi J**, **Nayyar A**, **Kumar A**. Machine learning from theory to algorithms: an overview. *J Phys Conf Ser*. 2018;1142:012012.
4. **KH Y**, **Beam AL**, **Kohane IS**. Artificial intelligence in healthcare. *Nature Biomedical Engineering*. 2018;2(10):719–731.
5. **Han T**, **Jiang D**, **Zhao Q**, **Wang L**, **Yin K**. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control*. 2017;40(8):2681–2693.
6. **Tu JV**. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Jour Clin Epid*. 1996;49(11):1225–1231.
7. **Bursac Z**, **Gauss CH**, **Williams DK**, **Hosmer DW**. Purposeful selection of variables in logistic regression. *Source Code Biol Med*. 2008;3:17.
8. **Cabitza F**, **Locoro A**, **Banfi G**. Machine learning in orthopedics: a literature review. *Front Bioeng Biotechnol*. 2018;6:75.
9. **Pan Y**, **Chen Q**, **Chen T**, **et al**. Evaluation of a computer-aided method for measuring the cobb angle on chest x-rays. *Eur Spine J*. 2019;28(12):3035–3043.
10. **Zhang J**, **Li H**, **Lv L**, **Zhang Y**. Computer-aided Cobb measurement based on automatic detection of vertebral slopes using deep neural network. *Int J Biomed Imaging*. 2017;2017:9083916.
11. **Jamaludin A**, **Lootus M**, **Kadir T**, **et al**. ISSLS prize in bioEngineering Science 2017: Automation of reading of radiological features from magnetic resonance images (MRIS) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur Spine J*. 2017;26(5):1374–1383.
12. **Weng C-H**, **Wang C-L**, **Huang Y-J**, **et al**. Artificial intelligence for automatic measurement of sagittal vertical axis using resunet framework. *J Clin Med*. 2019;8(11):1826.
13. **Korez R**, **Putzier M**, **Vrtovec T**. A deep learning tool for fully automated measurements of sagittal spinopelvic balance from X–ray images: performance evaluation. *Eur Spine J*. 2020;29(9):2295.
14. **Pfirrmann CWA**, **Metzdorf A**, **Zanetti M**, **Hodler J**, **Boos N**. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine*. 2001;26(17):1873–1878.
15. **Azimi P**, **Mohammadi HR**, **Benzel EC**, **Shahzadi S**, **Azhari S**. Use of artificial neural networks to predict recurrent lumbar disk herniation. *J Spinal Disord Tech*. 2015;28(3):E161-5.
16. **Arvind V**, **Kim JS**, **Oermann EK**, **Kaji D**, **Cho SK**. Predicting surgical complications in adult patients undergoing anterior cervical discectomy and fusion using machine learning. *Neurospine*. 2018;15(4):329–337.
17. **Kim JS**, **Merrill RK**, **Arvind V**, **et al**. Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. *Spine (Phila Pa 1976)*. 2018;43(12):853–860.
18. **Kim JS**, **Arvind V**, **Oermann EK**, **Kaji D**, **Ranson W**, **Ukogu C**, **et al**. Predicting surgical complications in patients undergoing elective adult spinal deformity procedures using machine learning. *Spine Deform*. 2018;6(6):762–770.
19. **Azimi P**, **Benzel EC**, **Shahzadi S**, **Azhari S**, **Mohammadi HR**. Use of artificial neural networks to predict surgical satisfaction in patients with lumbar spinal canal stenosis: Clinical article. *J Neurosurg Spine*. 2014;20(3):300–305.
20. **Staartjes VE**, **de Wispelaere MP**, **Vandertop WP**, **Schröder ML**. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling. *Spine J*. 2019;19(5):853–861.
21. **Matis GK**, **Chrysou OI**, **Silva D**, **et al**. Prediction of lumbar disc herniation patients' satisfaction with the aid of an artificial neural network. *Turk Neurosurg*. 2016;26(2):253–259.
22. **Saklad M**. Grading of patients for surgical procedures. *Anesthesiol*. 1941;2(5):281–284.
23. **Office for Civil Rights, HHS**. Standards for privacy of individually identifiable health information, 45 C.F.R. PTS. 160 and PT. 164. Subparts A and E. Final rule. *Fed Regist*. 2002;14:67(157):53181–53273.
24. **No authors listed**. On artificial intelligence – a European approach to excellence and trust. European Commission. 2020. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (date last accessed 15 July 2021).
25. **Collins GS**, **Moons KGM**. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):1577–1579.
26. **Buchlak QD**, **Esmaili N**, **Leveque J-C**, **et al**. Machine learning applications to clinical decision support in neurosurgery: An artificial intelligence augmented systematic review. *Neurosurg Rev*. 2020;43(5):1235–1253.

27. **Saito T**, **Rehmsmeier M**. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
28. **Winkler JK**, **Fink C**, **Toberer F**, **et al**. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol*. 2019;155(10):1135–1141.
29. **Hwang EJ**, **Park S**, **Jin K-N**, **et al**. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open*. 2019;2(3):e191095.
30. **Hersh WR**, **Weiner MG**, **Embi PJ**, **Logan JR**, **Payne PRO**, **Bernstam EV**, **et al**. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(8 Suppl 3):S30-7.
31. **Heus P**, **Damen JAAG**, **Pajouheshnia R**, **et al**. Uniformity in measuring adherence to reporting guidelines: The example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open*. 2019;9(4):e025611.

Author information:
J. M. McDonnell, BSc, MB BCh BAO, Academic Intern, School of Medicine, Royal College of Surgeons Ireland, Dublin, Ireland; National Spinal Injuries Unit, Mater Misericordiae University Hospital, Dublin, Ireland.

S. R. Evans, MSc, Medical Student
L. McCarthy, BSc, Medical Student
School of Medicine, University College, Dublin, Ireland.

H. Temperley, MB BCh BAO, Senior House Officer
C. Waters, MB BCh BAO, Senior House Officer
St. James' Hospital, Dublin, Ireland.

D. Ahern, MB BCh BAO, MRCS, Orthopaedic Registrar, National Spinal Injuries Unit, Mater Misericordiae University Hospital, Dublin, Ireland; Centre for Biomedical Engineering, Trinity College, Dublin, Ireland.

G. Cunniffe, PhD, Research Coordinator
S. Morris, MB BCh BAO, FRCSI, Consultant Orthopaedic Spine Surgeon
K. Synnott, MB BCh BAO, FRCSI, Consultant Orthopaedic Spine Surgeon
National Spinal Injuries Unit, Mater Misericordiae University Hospital, Dublin, Ireland.

N. Birch, MB BCh BAO, FRCS, Consultant Orthopaedic Spine Surgeon, Bragborough Hall Health and Wellness Centre, Daventry, UK.

J. S. Butler, PhD, FRCS, FACS, Consultant Orthopaedic Spine Surgeon, National Spinal Injuries Unit, Mater Misericordiae University Hospital, Dublin, Ireland; School of Medicine, University College, Dublin, Ireland.

Author contributions:
J. M. McDonnell: Conceptualization, Writing – original draft, Writing – review & editing.
S. R. Evans: Writing – original draft.
L. McCarthy: Writing – original draft.
H. Temperley: Writing – original draft.
C. Waters: Writing – original draft.
D. Ahern: Methodology.
G. Cunniffe: Methodology.
S. Morris: Writing – review & editing.
K. Synnott: Writing – review & editing.
N. Birch: Writing – review & editing.
J. S. Butler: Conceptualization, Writing – review & editing.

Funding statement:
No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

ICMJE COI statement:
N. Birch is an associate editor for *The Bone & Joint Journal,* and receives royalties from Medicrea International, unrelated to the study.

This article was primary edited by A. C. Ross