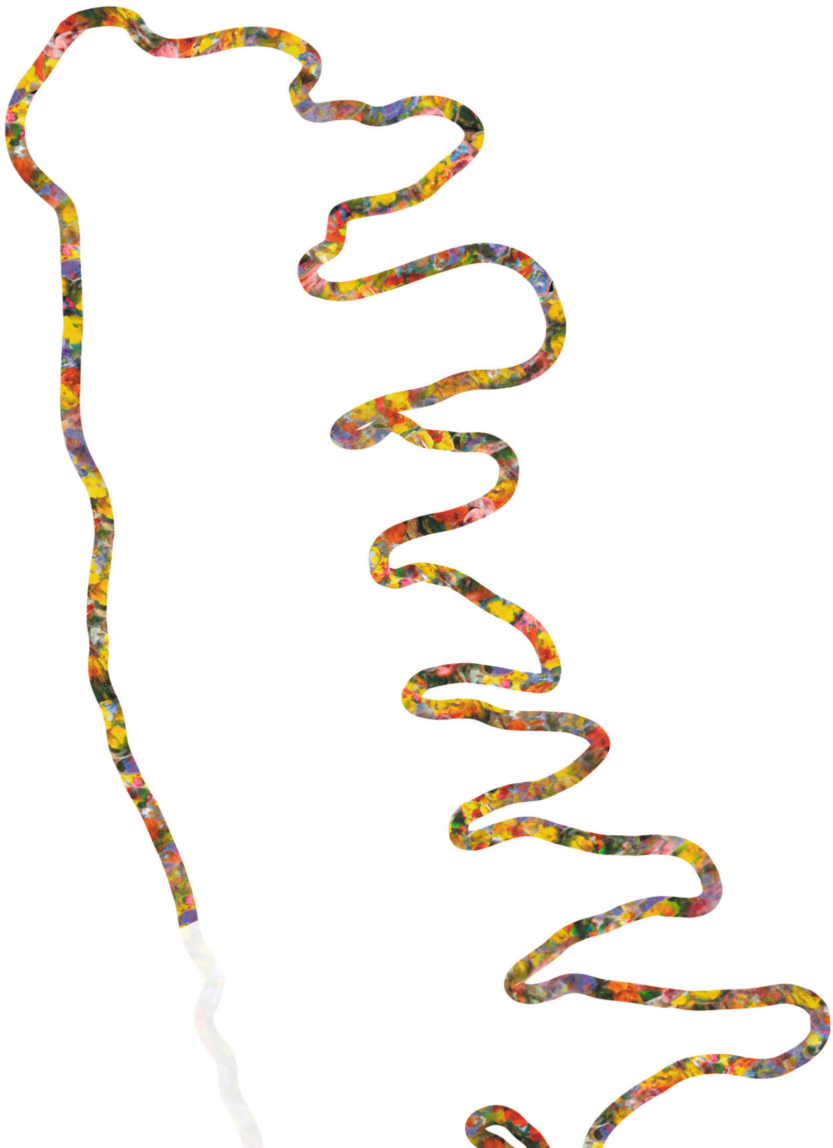
Outcome after anterior cervical discectomy: from inferential statistics to Machine Learning

Goedmakers, C.M.W.

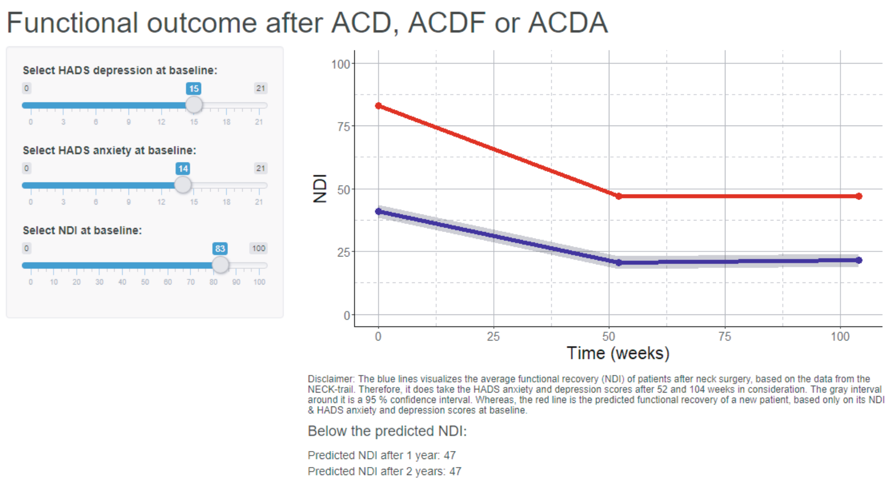


Chapter 11

Conclusion and discussion

Conclusion and discussion

In this thesis, classical statistical methods were applied to clinical data, and Machine Learning was used on medical imaging data, in an effort to determine the factors that contribute to the variation in, clinically important, functional outcomes following anterior cervical discectomy for radiculopathy and to explore how artificial intelligence can be utilized to improve the diagnostic and prognostic process.



Screenshots of the R shiny application illustrating how it functions. On the left the input can be given and on the right the predicted NDIs will be visualized in the graph (red line). On the left side the adjustable baseline measurements for NDI, HADS anxiety and HADS depression. On the right the graph visualizing (red line) the predicted NDI on the y-axis during the follow-up moment in weeks on the x-axis, and the marginal mean NDI (blue line) over time with a 95% confidence interval (in gray). Beneath the graph, the numerical values appear for the predicted NDI at week 52 and 104.

In a literature study, the alleged clinical superiority of the cervical disc prosthesis could not be confirmed in patients exclusively suffering from radiculopathy. Additionally, in mixed populations, consisting of both radiculopathy and myelopathy patients, the reported differences between prosthesis and fusion were considered to lack clinical relevance. Outcome data from a randomized controlled trial (RCT) conducted in the LUMC, demonstrated, both two and five years after surgery, that implanting a cervical disc prothesis after anterior cervical discectomy (ACDA) could not prevent ASD, and that clinical outcomes after ACDA were not superior to the current gold standard; anterior cervical discectomy and fusion (ACDF). Moreover, preliminary subgroup analysis on the two year outcome data could not indicate a certain type of patient that would benefit more from receiving the prosthesis. However, data from the same RCT showed that five years after single level ACD without implanting an intervertebral device, clinical outcome was worse compared to ACDF and ACDA, which was hypothesized to be caused by delayed fusion. In post-hoc analysis of the RCT data it was discovered that patients suffering from depression and anxiety before undergoing anterior cervical discectomy demonstrate significantly more neck disability after surgery and therefore do not benefit from surgery in the same way other patients do. Additionally, this study demonstrated that if, during follow-up, symptoms of depression and anxiety improved, patients' functional outcome improved as well. An R-shiny application was developed to facilitate an easier-to-interpret, visual communication to patients of the developed prediction models during a preoperative clinic visit. Using these type of applications can aid personalized treatment counselling and is a promising development for future shared decision-making healthcare.

In a time when our current healthcare systems are straining under burgeoning costs, a promising innovation is the use of artificial intelligence for the analysis of 'big data'. In order to benefit from the availability of big data, the reliance on computer algorithms has increased, while the direct human involvement and surveillance has declined. This shift raises ethical concerns in healthcare, as the machine-driven automation of processes, that were previously guided by human intelligence, creates new challenges. For the ethical application of Machine Learning in healthcare clinicians should balance the principles of beneficence, non-maleficence, justice and respect for autonomy to ensure patient safety, reliable diagnoses, and fair treatment.

Automating parts of the radiological image analysis process using Machine Learning could provide more accurate and consistent assessment, eliminate interobserver variability, and increase time efficiency. Utilizing Machine Learning in cervical spine analysis enables objectifying pattern recognition by statistical analysis of images, which surpasses the current reliance on human interpretation of patterns. Additionally, it provides valuable insights for future research, given its capacity to uncover the particular aspects of the image on which the prediction is based. Upon validation on larger imaging datasets, these models can ultimately support clinicians in decision-making and personalized treatment planning, offering promising prospects for the future of spine care.

11.1 Clinical outcome

Neck Disability Index (NDI)

The primary outcome measure used in the studies included in this thesis is the NDI. This decision was based on the use of NDI as the primary outcome metric in previous research on the cervical disc prosthesis (Chapter 2). Even though the NDI is widely used as the primary outcome measure in radiculopathy research, the questionnaire is focusing on neck pain and disability, instead of on cervical radiculopathy's primary symptom; arm pain. This gives rise to the question what studies that utilize NDI are truly evaluating - the efficacy of a surgery for radiculopathy or the effectiveness of addressing the total set of symptoms related to cervical degeneration, such as neck pain. In either case, considering the downstream accelerating effect of the surgical intervention on degeneration of other spinal anatomical features, like the facet joints, the NDI might be an excellent tool to quantify the neck disability originating from surgery induced cervical degeneration. Continuing that line of reasoning, the 25% of radiculopathy patients receiving ACDF that are considered to not be clinically successful, might just be patients suffering from cervical degeneration, which manifests as both neck and arm pain. And regardless of the surgery alleviating their arm pain, their continuously higher disability scores will really reflect their underlying cervical degenerative changes. In other words: the evaluation of arm pain reflects the surgeon's perception of success, while the NDI reflects the patient's perspective on success.

As research presented in this thesis has shown, NDI demonstrated highest discriminative ability for distinguishing between success and nonsuccess following neck surgery for radiculopathy, three months and one year postoperatively. This suggests that not the decrease in arm pain, but also the improvement in neck disability is what drives patient satisfaction after surgery. The fact that lower accuracy was found for the numeric rating scale for neck pain strengthens the conclusion that NDI as a PROM entails more than just neck pain. In Chapter 10, the NDI was used to identify the group of patients that shouldn't be operated on primarily because of a preoperative degenerative feature pattern on their imaging that correlates with clinical disability after surgery. I conclude that the NDI not only reflects cervical radiculopathy outcome, but also the symptoms accompanying degeneration of other anatomical features of the spine, like the facet joints. Distinguishing between arm pain, neck pain and neck pain that radiates to the arm, can be difficult for patients, and to complicate matters further these symptoms can present as a combined clinical manifestation. Therefore designing tools to aid clinicians in identifying these different groups of patients, automatically, is important.

Additionally, when using NDI for discriminating between clinical success, it was demonstrated that percentage change scores have improved discriminative ability, as compared to change scores that don't take into account baseline scores. This finding is in line with previous studies conducted on surgery for lumbar disc herniation and lumbar spinal stenosis [1-3]. Reporting change scores as a percentage of the baseline scores provides insight into the magnitude of improvement. The cut-off values for categorizing percentage change scores into the "success" or "no success" group in this thesis were calculated using the Global Perceived Effect scale (GPE) as an external criterion, as described in Chapter 3. Since the GPE is a self-reported scale, using it as an anchor could induce bias, as the answers will inherently be biased by the current health status of the patient. There are more objective anchors described in literature, however no golden standard currently exists [4, 5]. The possibility should be considered that the intrinsic strength of measures like 'quality of life' or 'satisfaction' is their subjectivity. In contrast, more objective measures, though maybe relevant to clinicians, may be of lesser importance to patients. Thus, the GPE has been recommended for research on this topic despite its limitations, mainly because of its good psychometric properties [6-11].

Minimal Clinically Important Difference (MCID)

In interpreting outcomes from patient reported outcome measures (PROMs), a similar, though different concept than the differentiation between success and nonsuccess, is that of the Minimal Clinically Important Difference (MCID). The prevailing trend in the current literature is to base conclusions on statistical significance of results, which, depending on the outcome measure, does not necessarily correspond to perceptible changes for patients. The MCID aims to describe the minimal change on a PROM questionnaire that is actually clinically important. The incorporation of the MCID in studies that report on PROMs is essential, because a statistically significant difference in PROMs, though mathematically correct, does not consistently correlate to a noticeable difference for patients nor clinicians [12-16]. In the studies that were included in this thesis, and that involved the analyses of NDI as the main outcome measure, results were consistently compared to the MCID, in order to analyze outcomes in such a way that is meaningful to both clinicians and patients. Nevertheless, MCID is a topic of ongoing scientific debate. The variable ways of calculating MCID values contribute to the differences between values that are reported in literature. Two of the most widely used methods for calculating the MCID are the distribution-based and the anchor-based method, the latter of which was used for the MCID calculated in Chapter 3 [17].

The distribution-based method typically involves calculating a standard deviation or effect size of the scores, and deriving the MCID based on a certain number of these units, such as half a standard deviation or a Cohen's d effect size of 0.2 or 0.5. In contrast, the anchor-based method uses an external criterion or reference point, such as a patient's global rating of change or a clinical expert's opinion, to anchor the MCID. This method involves comparing the difference in scores between two time points or groups on the outcome measure to the change in the anchor. The derived MCID is based on the minimal change in the outcome measure that corresponds to a meaningful change in the anchor. In Chapter 3 we used the anchor-based method, using the Global Perceived Effect scale (GPE) as the external criterion to anchor the MCID. It is possible that using the distribution-based method, relying on the statistical properties of the outcome measure, would have yielded slightly different MCID values. Before calculating the MCID from data in the NORspine registry and publishing on the topic, we determined values for MCID based on different publications, each using slightly different methods to approximate MCID, resulting in slightly different values [18-23]. At present, neither the anchor-based, nor the distribution-based method is universally accepted for MCID calculation, and there are valid arguments both for and against each method. However, I am thoroughly convinced that the anchor-based method, which enables the direct incorporation of clinical judgment and patient perspective into the MCID calculation, stands as the superior approach. The method allows for a more comprehensive and patient-centered calculation of MCID based on the magnitude of change that corresponds to a noticeable and clinically relevant improvement or deterioration within a specific patient population. This ability to tailor the MCID to the outcome measure and population of interest, while accounting for unique clinical considerations, sets it apart from distribution-based methods.

Anxiety and Depression

In Chapter 5 we demonstrated that patients who were suffering from depression or anxiety before anterior cervical discectomy reported statistically significant and clinically relevant higher disability, one and two years after surgery. Additionally, it was demonstrated that if, during follow-up, symptoms of depression and anxiety improved, patients' functional outcome, measured by NDI, improved as well. These findings raise the question whether decreased mental health status is either a cause of more disability, or an inherent effect of experiencing pain and disability from cervical disc disease. Establishing cause and effect is not possible in the current study set-up, nor are there any studies available in literature that have found a clear causal relationship. Interestingly, there are observational studies that have assumed the first hypothesis to be true, and therefore studied the effect of pre-operative pharmaceutical treatment for anxiety and depression, on pain and disability outcomes postoperatively [24, 25]. These studies demonstrated improved disability and less pain in the group that received pharmaceutical treatment. In addition to pharmaceutical intervention, preoperative counseling and cognitive behavioral therapy could be other potentially interesting strategies to investigate. However, instead of symptom control, future research should focus on causality and investigate the direction of the effect between mental health and disability scores after cervical spine surgery, as it may provide additional insights on how to manage patients with mental illnesses before and after spine surgery.

In order to gain deeper understanding of the association between mental health and neck disability, it is helpful to consider the ten dimensions of the NDI questionnaire separately. The ten items encompassed by the NDI include Pain Intensity, Personal Care, Lifting, Reading, Headaches, Concentration, Work, Driving, Sleeping, and Recreation. Although lifting, reading, and driving are directly influenced by neck pain, the remaining NDI items are recognized as being heavily influenced by anxiety and depression. Consequently, it is conceivable that pharmacological interventions and/or cognitive behavioral therapy could potentially ameliorate the symptoms of depression and anxiety while concurrently improving NDI scores.

Therefore, the previously stated argument comes full circle, underscoring that the NDI serves as a patient-centered and comprehensive outcome measure, encompassing clinically significant assessments for cervical spine disease that is patient centered and encompasses the broader aspects of well-being, rather than focusing solely on the mitigation of arm pain. Therefore, the previously stated argument comes full circle, underscoring that the NDI serves as a patient-centered and comprehensive outcome measure, encompassing clinically significant, patient-centered assessments for cervical spine disease. It encompasses the broader aspects of well-being for these patients, rather than focusing solely on the mitigation of arm pain.

Another approach to improve outcomes in patients with mental illness is to preoperatively inform them on the influence of anxiety and depression on neck disability. The previously mentioned R-shiny application (Chapter 5), offers the opportunity to effectively counsel patients in this area during their preoperative clinic visit. The application visualizes the predicted disability trajectory by the model, using a graph that adjusts based on the baseline scores of the individual patient. The direct implementation of predictive modelling results in this manner, improves preoperative counselling and becomes increasingly important with the rise of shared decision making in the current medical world, for which effective communication, of research results among other, is paramount.

Nevertheless, 25% of variance in the model described in Chapter 5 remains unexplained. To improve the accuracy of predictions and reduce the percentage of unexplained variance, future research should combine different types of outcome parameters. Additionally, to increase validity, the sample size of analyzed patients should be increased. The incorporation of radiological imaging data and histopathological parameters for example, could provide a more complete picture of the patient and result in higher accuracy in predicting outcomes for individual patients. To include these additional parameters in the most efficient manner, a fully automated deep learning system that extracts and subsequently analyses all parameters should be embedded in the electronic medical record.

11.2 The cervical disc prosthesis

In Chapter 2 it was discovered that the majority of studies comparing ACDF and ACDA aim to present results as favorable as possible for the prosthesis, are sponsored by industry and are not blinded. These factors are highly likely to induce bias in favor of the prosthesis, as the potential associated with an expensive, novel, and technologically advanced device is expected to exert an influence on experienced outcome by patients. Nevertheless, in the studies included in this thesis that evaluated the cervical disc prosthesis, the clinical advantage of the device was consistently absent one, two and five years after surgery, when compared to the current gold standard; fusion, with a stand-alone cage, in patients suffering from radiculopathy. Evidently, these studies were conducted in a double-blind manner.

In research comparing the cage and the prosthesis the distinction between myelopathy and radiculopathy patients should be made. Myelopathy patients are prone to have more severely degenerated cervical spines and perform different on outcome scales. Based on a comprehensive review of relevant literature presented in Chapter 2, there is suggestive evidence to indicate that the mean NDI two years post-operatively is lower (better) in the ACDA group including both myelopathy and radiculopathy patients, than in the ACDA group including only radiculopathy patients. However, the groups including only patients with radiculopathy were smaller and it is possible that they were too small to detect a difference. An alternative hypothesis could be that patients with advanced cervical spine degeneration, particularly those diagnosed with myelopathy, may have developed a higher pain tolerance and are therefore more inclined to report superior disability scores when compared to patients presenting solely with radiculopathy. Nevertheless, when comparing the prosthesis to fusion in either patients exclusively suffering from radiculopathy, or mixed myelopathy and radiculopathy patients, a clinically relevant benefit, in neck disability or neck pain, for the prosthesis was absent in both groups.

A clinical benefit for the prosthesis was also not found in subgroups of patients based on sex, age, BMI, baseline disc height and smoking history, in preliminary subgroup analysis included in Chapter 4. Nevertheless, it should be noted that results from these type of analysis should always be interpreted with caution as the combined RCT population of the study wasn't powered for a subgroup analysis. The minimal differences in clinical outcome metrics found between the different surgical techniques for each subgroup, should not be confused with evidence for the absence of a difference.

11.3 Anterior cervical discectomy without instrumentation

In Chapter 4 and 6, the randomized controlled trials investigate a third treatment option without instrumentation after anterior cervical discectomy (ACD), next to ACDA and ACDF. The ambivalence surrounding ACD is mainly caused by its alleged negative effect on local kyphosis, adjacent segment disease, recurrent cervical radiculopathy and neck pain on long-term clinical outcome. The rationale for placing an intervertebral device is that the original height of the removed disc should be restored, in order to keep the neuroforamen at its original height, prevent recurrent compression of the nerve root, and avoid segmental kyphosis. Previous research on ACD was not able to find high quality evidence of higher disability in the ACD group, when compared to ACDF [26-30], even though increased dissatisfaction rates were found, after longer follow-up with ACD [31]

The influence of foraminal height decrease and inducing segmental kyphosis were further explored in our research group. With regards to the preservation of foraminal heights, a study from our research group on the same patient population, as described in Chapter 4 and 6, demonstrated that mean foraminal height was only slightly decreased in the ACD group, and foraminal height did not correlate to radicular symptoms nor to general clinical outcome, one year after surgery [32]. Additionally, ACD is presumed to induce segmental kyphosis of the cervical column, and as a consequence induce neck pain. When segmental angulation was investigated in the NECK trial, of the six patients with a kyphotic cervical curvature at baseline, only 1 patient remained kyphotic; the other 6 patients recovered to a straight or lordotic spine two years and consistently five years after surgery. One patient that had received ACD recovered to a lordotic spine five years after surgery. The number of patients with a kyphotic spine was deemed too small to make a meaningful correlation to clinical data.

Consistent with the absence of both, a meaningful decrease in foraminal height, as well as kyphosis induction, the one and two year trial results demonstrated that ACD provided similar clinical results, when compared to ACDA and ACDF. However, clinical outcome after five years was statistically significantly worse after ACD, compared to ACDA and ACDF. We hypothesize that the difference in long-term clinical outcome is related to delayed fusion in the ACD group, in agreement with observations summarized in a previous Cochrane review [26]. If an intervertebral device is placed, this leads to immobilization of the vertebra, which stops the induction of osteophyte growth. However, if the vertebrae are allowed non-physiological movement (as is the case in ACD), even though it is minimal, this may allow osteophytes to grow, resulting in recurrent irritation of the nerve root. This would lead to more reoperations at the index level, which is indeed the observation in the five years follow-up results of the NECK-trial, where six patients were reoperated at the index level, of which four from the ACD group. It has to be noted, however, that the surgical incentive to reoperate is high after ACD, since this technique deviates from the normal routine ACDF. Interestingly, the number of reoperations at index level was not higher in the ACDA group, though the index level remained mobile too in this group. Presumably, the mobility allowed by the prosthesis was either capable of mimicking physiological movement, or fusion occurred (heterotopic ossification) and the segment was immobilized anyway.

In the studies included in this thesis anterior cervical discectomy and fusion was performed with a stand-alone PEEK cage without anterior plating. The rationale behind adding a plate is to increase postoperative stability, to avoid cage subsidence, and to accelerate fusion. However, several studies have shown that neither fusion rate, nor instability is a significant problem after performing a one level ACDF without a plate [33-35]. Some studies even report a favorable effect, with lower risk of postoperative dysphagia and ASD without plating [36]. Moreover, a recent large retrospective cohort study confirmed that cage subsidence in stand-alone cages after ACDF is minimal [37]. Nevertheless, in the United States cervical plating is considered routine after anterior cervical discectomy and fusion. This is potentially caused by a combination of the more defensive surgical practice pattern, and the medical compensation culture that pays more for an ACDF with plates and screws. Nevertheless, surgical instrumentation after a one level anterior discectomy remains a topic of scientific debate.

The primary objection to the effectiveness of RCTs in the surgical specialties is that the study design would be inadequate for addressing most surgical scientific inquiries [38-40]. While it is acknowledged that surgical outcomes can vary based on techniques and surgical experience, it can generally be expected that randomization within an RCT would evenly distribute these influential factors among the different treatment groups. However, in studies with a smaller number of patients, such as the NECK trial, it can be argued that the benefits of randomization are less evident, compared to studies with a larger patient population. Another frequently mentioned factor impacting the generalizability of RCTs for ACDF is the brand and type of implant used. This is despite the lack of available literature demonstrating statistically significant differences in clinical outcomes between different instrumentation types and brands. In the NECK trial three different types of PEEK cages were compared to the Activ®C prosthesis. By combining the results from both the NECK and Procon trials in Chapter 4, the number of different implant brands and types further increased. Using different types and brands of implants makes results more generalizable to the general population and therefore increases the external validity of our results.

11.4 Cervical spine degeneration

Cervical spine degeneration refers to the natural wear and tear that occurs in the bones, discs, and joints of the neck over time. As we age, the discs in our spine lose water content and become less flexible, which increases the load on the uncovertebral joints and can lead to the development of bone spurs (osteophytes). Moreover, the annulus degenerates concomitantly and damage to this annulus, with or without damage of the endplate, can lead to herniation of the disc. Finally, development of arthritic disease in the synovial-lined, diarthrodial joints of the spine adds to degenerative changes in the cervical spine upon aging. These changes can lead to symptoms such as pain, stiffness, and decreased range of motion in the neck. In some cases, cervical degeneration can also lead to nerve compression, which can cause numbness, tingling, and weakness in the arms and hands. Cervical spine degeneration is a normal part of aging, but it can be accelerated by certain factors such as genetics, injury, and lifestyle factors such as poor posture, obesity and smoking [41-45].

Surgery is one way to create controlled injury to the spine and surrounding tissue, and though successful, can therefore be a cause of accelerated degeneration at the levels adjacent to the index surgery. Adjacent segment disease (ASD) is defined as the development of new-onset clinical symptoms that correspond to degenerative radiographic changes on the spinal segments (vertebrae, discs, and joints) that are located next to a segment that has been previously surgically treated or stabilized. This can occur after spinal fusion surgery, where the movement in the treated segment is restricted. Due to the fused segments, the surrounding segments are forced to compensate, leading to increased stress and wear and tear on those adjacent segments over time. This can result in the development of symptoms such as pain, stiffness, and limited range of motion in the neck [46, 47]. Pain can be caused by a variety of factors, including nerve compression, muscle strain, and inflammation.

The presence, however, of degenerative changes on radiological imaging, such as osteophytes (bone spurs) or disc herniations, does not always correlate with the presence or severity of symptoms [48, 49]. Many people have degenerative changes in their cervical spine that do not cause any symptoms. This is known as asymptomatic degeneration. Additionally, some individuals may have relatively mild degenerative changes on imaging but experience significant symptoms, while others may have severe degenerative changes on imaging but have minimal symptoms. Additionally, other factors such as the patient's overall health, age, and the presence of other comorbidities may also play a role in determining the level of symptoms a patient experiences. Therefore, it is important to consider more than just the radiological imaging when assessing a patient's symptoms and determining the appropriate course of treatment.

In research on the cervical spine, authors repeatedly fail to make the distinction between clinically relevant, symptomatic degeneration, and not clinically relevant, asymptomatic degeneration, which leads to invalid conclusions. One example is in research on surgical technique, in Chapter 2 it was demonstrated that studies claiming superior radiological results in patients that underwent ACDA, fail to prove there is a correlation between the observed decline in degree of radiological degeneration, and clinical outcome. Another example, from in the discussion of Chapter 9, is that higher percentages of secondary surgical interventions for ASD were found in studies that used new radiological degenerative symptoms at the adjacent level, without corresponding clinical symptoms, as an indication for surgery [50].

A frequently cited annual incidence of ASD after fusion is 2.9% [51]. This is based on radiological aspects of degeneration, not on clinical symptoms of ASD. In order to evaluate clinically relevant ASD the number of reoperations for adjacent level radicular complaints can be assessed [52]. The results from our own study (Chapter 6) suggest that implanting a prosthesis does not prevent adjacent level disease, as the number of reoperations was comparable in patients undergoing ACDF to patients undergoing ACDA [53]. This is assumed to be due to the finding that a prosthesis loses its full range of mobility in the first year after implanting due to heterotopic ossification (HO) [49], which is in agreement with observations from other studies [54, 55]. Furthermore, in addition to the comparable number of reoperations between treatment groups, the long-term follow-up of the NECK trial revealed a significantly lower total reoperation rate of only 1.2% for adjacent level disease. This is in line with other, more recent studies that also report lower reoperation rates for adjacent level disease than the 2.9% reported by Hilibrand. Moreover, it was found that the annual reoperation rate for ASD decreased for ACDA and ACDF, with the increase in length of follow-up. This may be due to the fact that patients included in RCTs have more frequent follow-up visits in the first years after the intervention. During these visits, patients are actively asked if persisting or recurrent complaints are present, which lowers the threshold for reoperation. A similar trend was observed in the 10-years data of a Swedish RCT [56]. Therefore, the percentages for reoperation may even be lower in a usual care setting.

Previous research has estimated that a significant percentage of patients, current consensus is around 15%, requires a secondary surgery for adjacent segment disease (ASD) after an initial single-level procedure [46, 47, 57]. This suggests that in the United States alone, approximately 20,000 patients may require a second surgical intervention for ASD annually. This has a significant financial impact with the approximate cost per procedure estimated at $15,000, and minimizing secondary surgery rates could therefore save hundreds of millions of dollars annually [58]. Nevertheless, studies on ASD after cervical discectomy have shown that the development of the disease is dependent on time. The approach used in Chapter 9, which defined the ground truth based on the last available radiological and clinical follow-up, may not be accurate as a longer follow-up may reveal ASD in patients that were classified as non-ASD in the study [50]. Revalidating models over time, as the one developed in Chapter 9, is therefore an essential component of the research in ASD field. Additionally, studying conservative approaches, as well as the post-operative images is necessary to assess the effect of surgery in developing ASD.

Chapter 9 has demonstrated how deep learning can be harnessed to study adjacent segment disease. The CNN facilitates statistical testing on the imaging data by translating pixels into numerical series, enabling an analysis similar to regression to be conducted on these numerical sequences. In Chapter 10 we further explored the use of a CNN to identify image features for clinical success, as a way of gaining insight into what radiological features are important for recovery after surgery. The Grad-CAM heatmaps showed a significant influence of the facet joints, the only synovial-lined, diarthrodial joints in the spine, on the classification process. An interesting finding that suggests there might be more information in the cervical lateral radiograph than the human eye of clinicians has been able to appreciate so far. Currently, in the radiological analysis of cervical spine disease there is a focus on the intervertebral disc, shape of the vertebra and the cervical alignment, mainly because clinicians are able to appreciate those features more easily. The focus on the facets is limited. Results from Chapter 10 suggest future research should investigate the degenerative process of the facet joints in the cervical vertebrae and its indications as a diagnostic feature. Hence, a more thorough examination of the facet joint is warranted.

Nonetheless, this finding is remarkable since facet joints are believed not to be appropriately visualized on lateral radiographs. While a lateral radiograph of the cervical spine can be used to diagnose degeneration, to fully assess the integrity of major soft tissue structures of the spine such as intervertebral discs, nerves, the spinal canal, vertebral end-plates and the facet joints, magnetic resonance imaging (MRI) is considered the gold standard [59]. Thus, it would be interesting to repeat this study using MRI data, as it provides a three-dimensional image of the facet. This may lead to new insights regarding the assessment of degeneration in the cervical spine, as well as alternative treatment strategies, and potentially improve predictions. Nevertheless, both radiograph and MRI data needs to be interpreted by a qualified radiologist. Additionally, obtaining cervical MRI images is not only more time-consuming but also about 10 times more expensive when compared to radiographic images. Due to these factors, the decision to focus on radiographic images in Chapter 10 was made. Radiographic images are more accessible in terms of cost and time to acquire, and analyzing 2D images requires less computational power than 3D images, making the developed prediction tool more accessible.

11.5 Machine Learning (ML)

Ethical usage

Healthcare has enormous potential to take advantage of big data and artificial intelligence (AI), the two biggest technological innovations of the 21st century. AI, powered by Machine Learning, can generate new insights from big data. Machine Learning, a key component of AI, uses big data to enhance its performance and identify patterns to meet specific goals. With an overwhelming amount of healthcare data, including basic science, clinical, numerical, language-based, imaging, administrative and economic data, AI-powered programs are well-equipped to handle, analyze, and make sense of this data to improve healthcare delivery. However, the rise of big data and related technologies (e.g. Machine Learning, AI, and robotics) has led to an increased reliance on computer algorithms and a decline in direct human involvement and oversight. This shift raises ethical concerns in healthcare, as the machine-driven automation of processes, that were previously guided by human intelligence, creates new challenges that are discussed in Chapter 7.

To ethically use AI in healthcare it must balance the principles of beneficence, non-maleficence, justice and respect for autonomy to ensure patient safety, reliable diagnoses, and fair treatment. To guarantee beneficence and non-maleficence, AI needs to be primarily accurate and reliable. The data used to train AI algorithms must be carefully curated to avoid biases that could result in harm or injustice. Additionally, the healthcare provider must maintain patient autonomy by providing necessary information, regarding the underlying technology, for informed decision making.

A pathway to ethical AI in healthcare should closely adhere to the principles of transparency, responsibility and reproducibility. Transparency focuses on explainability of the model's decision-making process, and is crucial for assessing the risks and benefits involved, as well as identifying areas of improvement. Responsibility refers to accountability of the model's decisions. Who should be held accountable for AI systems is crucial due to their potential for causing harm in the real world and is a matter of concern as current legal systems only hold humans accountable. This leads to the question of whether AI systems in healthcare will ever have true independence, or always be connected by accountability to the clinicians using them. Reproducibility refers to the ability to reproduce computational results and is critical for integrating AI systems into real-world applications. The field of AI is facing unique reproducibility challenges and initiatives like the FDA's guidelines for clinical decision support software aim to address them, presenting non-binding guidelines. Ultimately, this type of academic guidelines for publication can ensure reproducibility by imposing rules and requirements on academic publications. Ideally, studies include an elaborate methods section for transparency, share data and code for reproducibility and clearly state accountability in its user interface.

Image analysis

With the current pressure on healthcare, the ageing population and the relatively high prevalence of spinal degeneration, there is an increasing demand on image analysis [41]. However, radiological analysis is a time consuming task which has to be done carefully, as diagnostic and treatment decisions are based on the outcomes, in addition to being subject to bias by inter-observer variability. Automating aspects of the radiological analysis process of the degenerative spine using AI could therefore improve the consistency and efficiency of the process, as well as alleviate pressure on healthcare providers [60].

Chapter 7 highlights the importance of transparency and reproducibility in Machine Learning research, and this is reinforced by the findings of Chapter 8. Results in the systematic review were too heterogeneously reported and therefore pooling the results was not possible. Reporting outcomes clearly and homogenously is an important requirement to compare performance among publications. It is imperative that there is a uniform approach to reporting outcomes, with a consistent set of outcome variables utilized for all studies in the domains of segmentation, classification, or prediction. This will enhance the external validity and reproducibility of the research. Several guidelines that describe the appropriate reporting process for Machine Learning studies have been published, but are not all-encompassing [61, 62]. Another essential recommendation is for authors to share code. The majority of publications included in Chapter 8 failed to share code. Establishing an academic environment in which code sharing is promoted is essential to keep improving the work in the image analysis field, and promotes transparency and reproducibility in the process. The concept of 'Grand Challenges' presents a promising initiative to encourage transparency and reproducibility, while simultaneously eliminating a range of biases. The aim of these public challenges is to let participants apply their algorithms to the provided Grand Challenge task, using the public test set of images provided by the challenge organizers [63].

Generating heatmaps that illustrate the output of the model is important for model interpretation and thus increases transparency. Utilizing Machine Learning in the cervical spine analysis process enables objectifying pattern recognition by statistical analysis of images, which surpasses the current reliance on human interpretation of patterns. Additionally, it provides valuable insights for future research. As seen in Chapter 10, the repeated focus of the model on the facet joints suggests there may be more to uncover in the image than previously thought, and the techniques outlined in Chapter 8, 9 and 10 can therefore be used to identify clinically important radiological image features for future research. However, it is important to note that the validity and reliability of heatmaps is still a subject of scientific debate, and model interpretability should never solely rely on them, but be supported by a comprehensive method section and code sharing [64, 65].

When conducting research with automated image analysis, several considerations must be taken into account, one of which is image quality. The quality of the images is a crucial factor for the accuracy of the results, as illustrated by the improved results on rescaled and cropped images in Chapter 10. In this chapter images converted into the Joint Photographic Experts Group (JPEG) format were used, for computational reasons. However, it should be noted that the JPEG format does not preserve meta information, such as voxel spacing, and may also introduce compression that can affect image quality. Digital Imaging and Communications in Medicine (DICOM) formatted images, on the other hand, retain meta data and the original file size, allowing for more granular data analysis, although this may increase the required computational power to run the model. It is important to note that converting images to JPEG from DICOM format may reduce the number of pixels per image, but this does not necessarily result in a decrease in accuracy, Dice scores, or error scores [66].

In addition to the quality of the images, the number of input images is another important consideration. In Chapter 9, 344 MR images were used, whereas in Chapter 10, only 70 unique radiographs were included, resulting in a smaller test set and, consequently, wider confidence intervals. The Convolutional Neural Network (CNN) model only evaluates the pixel values in relation to the neighboring pixels, and thus, even a minor rotation can provide a new viewpoint for the computer, making it a useful method for data augmentation in studies with limited data. This technique was implemented in Chapter 10 to augment the training and validation set images. However, some changes may have been too subtle, resulting in a training dataset that was not substantial enough to differentiate various features. To address this, it is suggested to train the model developed in Chapter 10 on a larger dataset, as this would enable higher model complexity and enhance its predictive power and external validity.

However, as the number of data samples and the scope of data collection expand, the importance of data protection grows exponentially. Chapter 7 outlines how privacy concerns, in relation to medical research, focus on how data is stored, processed, and transmitted with the consent of the owner, while security aims to safeguard sensitive data from breaches, leaks, and other security threats. The increasing utilization of machine learning algorithms in various tasks has propelled research in privacy and security regulations. Real-world controversies, such as facial recognition for authoritarian control, AI models determining sexual orientations without consent, and discriminatory hiring algorithms, have highlighted the potential challenges associated with this technology. In healthcare, privacy and security of digitized patient data are vital for medical ethics, justice, and autonomy, ensuring protection against harm caused by sharing information with individuals, organizations, or technologies. However, ensuring privacy and security has become challenging as current standards may not account for all AI capabilities. Existing privacy laws, like the GDPR in Europe and the HIPAA Privacy Rule in the United States, offer certain protections, but are not all-encompassing. Privacy and security laws worldwide aim to protect data providers, but potential threats arising from sharing supposedly secure and privacy-protected databases with powerful AI tools are often overlooked. Unchecked data sharing can lead to algorithmic biases, discrimination, and social disparities. Robust privacy and security audit policies, along with retrospective analysis of algorithms, are needed to mitigate risks and ensure fairness. Additionally, exploring alternative approaches, such as using generative models like GANs for record simulation, could provide solutions that preserve privacy and offer patients the choice to opt out of data sharing. These advancements in privacy protection models driven by AI-driven generative models offer promising avenues to reevaluate and refine current legislation.

The challenges for future research in image analysis of the spine are not limited to privacy and image quality or quantity, but mainly involve the integration of various models into a single, fully automated pathway. The integration of different models with various input and output parameters offers a more comprehensive view of the patient and can be expected to increase the accuracy of predicting outcomes for individual patients. This could involve incorporating not only radiological and clinical data, but also histological parameters, as recent studies have shown their association with clinical outcomes in sciatica [67]. Ideally, this would be achieved through a fully automatic model, embedded in the electronic healthcare record (EHR), where input data is passively collected without the need for human intervention. The preservation of accessibility is contingent upon the passive collection of data. Chapters 9 and 10 present algorithms that do not negatively impact accessibility as they are based solely on radiographical data, making them non-invasive, cost-effective, and easily accessible, thereby eliminating the need for, currently still, time-consuming questionnaires and costly histological analysis.

The models presented in Chapters 9 and 10 are not intended for immediate clinical use, but serve as a demonstration of the feasibility of using Convolutional Neural Networks (CNNs) to identify preoperative radiological image features that are relevant to clinically meaningful outcomes. Upon validation on larger imaging datasets, these models can ultimately support clinicians in decision-making and personalized treatment planning, offering promising prospects for the future of spine care.

References

1. Werner DA, Grotle M, Gulati S, Austevoll IM, Lønne G, Nygaard ØP, et al. Criteria for failure and worsening after surgery for lumbar disc herniation: a multicenter observational study based on data from the Norwegian Registry for Spine Surgery. European Spine Journal. 2017;26(10):2650-9.
2. Austevoll IM, Gjestad R, Grotle M, Solberg T, Brox JI, Hermansen E, et al. Follow-up score, change score or percentage change score for determining clinical important outcome following surgery? An observational study from the Norwegian registry for Spine surgery evaluating patient reported outcome measures in lumbar spinal stenosis and lumbar degenerative spondylolisthesis. BMC musculoskeletal disorders. 2019;20(1):1-15.
3. Hermansen E, Myklebust TÅ, Austevoll IM, Rekeland F, Solberg T, Storheim K, et al. Clinical outcome after surgery for lumbar spinal stenosis in patients with insignificant lower extremity pain. A prospective cohort study from the Norwegian registry for spine surgery. BMC Musculoskeletal Disorders. 2019;20(1):1-8.
4. Gatchel RJ, Mayer TG. Evidence-informed management of chronic low back pain with functional restoration. The Spine Journal. 2008;8(1):65-9.
5. Mayer TG, Gatchel RJ, Mayer H, Kishino ND, Keeley J, Mooney V. A prospective two-year study of functional restoration in industrial low back injury: an objective assessment procedure. Jama. 1987;258(13):1763-7.
6. Kamper SJ, Ostelo RW, Knol DL, Maher CG, de Vet HC, Hancock MJ. Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. Journal of clinical epidemiology. 2010;63(7):760-6. e1.
7. Bobos P, MacDermid J, Nazari G, Furtado R. Psychometric properties of the global rating of change scales in patients with neck disorders: a systematic review with meta-analysis and meta-regression. BMJ open. 2019;9(11):e033909.
8. Bobos P, Ziebart C, Furtado R, Lu Z, MacDermid JC. Psychometric properties of the global rating of change scales in patients with low back pain, upper and lower extremity disorders. A systematic review with meta-analysis. Journal of Orthopaedics. 2020;21:40-8.
9. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. Journal of Manual & Manipulative Therapy. 2009;17(3):163-70.
10. McLeod LD, Coon CD, Martin SA, Fehnel SE, Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert review of pharmacoeconomics & outcomes research. 2011;11(2):163-9.
11. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. The journal of pain. 2008;9(2):105-21.
12. Steinhaus ME, Iyer S, Lovecchio F, Khechen B, Stein D, Ross T, et al. Minimal clinically important difference and substantial clinical benefit using PROMIS CAT in cervical spine surgery. Clinical Spine Surgery. 2019;32(9):392-7.
13. Sedaghat AR. Understanding the minimal clinically important difference (MCID) of patient-reported outcome measures. Otolaryngology–Head and Neck Surgery. 2019;161(4):551-60.
14. Hung M, Saltzman CL, Kendall R, Bounsanga J, Voss MW, Lawrence B, et al. What are the MCIDs for PROMIS, NDI, and ODI instruments among patients with spinal conditions? Clinical orthopaedics and related research. 2018;476(10):2027.

15. Parker SL, Godil SS, Shau DN, Mendenhall SK, McGirt MJ. Assessment of the minimum clinically important difference in pain, disability, and quality of life after anterior cervical discectomy and fusion: clinical article. J Neurosurg Spine. 2013;18(2):154-60.
16. Narain AS, Hijji FY, Khechen B, Haws BE, Bohl DD, Yom KH, et al. Risk factors associated with failure to reach minimal clinically important difference in patient-reported outcomes following anterior cervical discectomy and fusion. International Journal of Spine Surgery. 2019;13(3):262-9.
17. Tunis SR, Stryer DB, Clancy CM. Practical Clinical TrialsIncreasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. JAMA. 2003;290(12):1624-32.
18. Young IA, Cleland JA, Michener LA, Brown C. Reliability, Construct Validity, and Responsiveness of the Neck Disability Index, Patient-Specific Functional Scale, and Numeric Pain Rating Scale in Patients with Cervical Radiculopathy. American Journal of Physical Medicine & Rehabilitation. 2010;89(10).
19. Young BA, Walker MJ, Strunce JB, Boyles RE, Whitman JM, Childs JD. Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. The Spine Journal. 2009;9(10):802-8.
20. Auffinger BM, Lall RR, Dahdaleh NS, Wong AP, Lam SK, Koski T, et al. Measuring Surgical Outcomes in Cervical Spondylotic Myelopathy Patients Undergoing Anterior Cervical Discectomy and Fusion: Assessment of Minimum Clinically Important Difference. PLOS ONE. 2013;8(6):e67408.
21. Cleland JA, Childs JD, Whitman JM. Psychometric Properties of the Neck Disability Index and Numeric Pain Rating Scale in Patients With Mechanical Neck Pain. Archives of Physical Medicine and Rehabilitation. 2008;89(1):69-74.
22. Parker SL, Godil SS, Shau DN, Mendenhall SK, McGirt MJ. Assessment of the minimum clinically important difference in pain, disability, and quality of life after anterior cervical discectomy and fusion. Journal of Neurosurgery: Spine. 2013;18(2):154-60.
23. Donk RD, Verbeek ALM, Verhagen WIM, Groenewoud H, Hosman ATF, Bartels RHMA. Substantial clinical benefit for Neck Disability Index revisited: establishing the goal for treatment? European Spine Journal. 2018;27(6):1262-5.
24. Adogwa O, Elsamadicy AA, Cheng J, Bagley C. Pretreatment of Anxiety Before Cervical Spine Surgery Improves Clinical Outcomes: A Prospective, Single-Institution Experience. World Neurosurg. 2016;88:625-30.
25. Elsamadicy AA, Adogwa O, Cheng J, Bagley C. Pretreatment of Depression Before Cervical Spine Surgery Improves Patients' Perception of Postoperative Health Status: A Retrospective, Single Institutional Experience. World Neurosurgery. 2016;87:214-9.
26. Jacobs W, Willems PC, van Limbeek J, Bartels R, Pavlov P, Anderson PG, et al. Single or double-level anterior interbody fusion techniques for cervical degenerative disc disease. Cochrane Database of Systematic Reviews. 2011(1).
27. Jacobs W, Willems PC, Kruyt M, van Limbeek J, Anderson PG, Pavlov P, et al. Systematic Review of Anterior Interbody Fusion Techniques for Single- and Double-Level Cervical Degenerative Disc Disease. Spine. 2011;36(14).
28. Matz PG, Ryken TC, Groff MW, Vresilovic EJ, Anderson PA, Heary RF, et al. Techniques for anterior cervical decompression for radiculopathy. Journal of Neurosurgery: Spine SPI. 2009;11(2):183-97.
29. van Middelkoop M, Rubinstein SM, Ostelo R, van Tulder MW, Peul W, Koes BW, et al. No additional value of fusion techniques on anterior discectomy for neck pain: A systematic review. PAIN. 2012;153(11).
30. Donk RD, Verbeek ALM, Verhagen WIM, Groenewoud H, Hosman AJF, Bartels RHMA. What's the best surgical treatment for patients with cervical radiculopathy due to single-level degenerative disease? A randomized controlled trial. PLOS ONE. 2017;12(8):e0183603.

31. Nandoe Tewarie RDS, Bartels RHMA, Peul WC. Long-term outcome after anterior cervical discectomy without fusion. European Spine Journal. 2007;16(9):1411-6.

32. Lafeber G, Goedmakers C, Arts M, Vleggeert-Lankamp CLA. The effect of discectomy on foraminal height in cervical surgery. Brain and Spine. 2021;1:100207.

33. Nambiar M, Phan K, Cunningham JE, Yang Y, Turner PL, Mobbs R. Locking stand-alone cages versus anterior plate constructs in single-level fusion for degenerative cervical disease: a systematic review and meta-analysis. European Spine Journal. 2017;26(9):2258-66.

34. Lu VM, Mobbs RJ, Fang B, Phan K. Clinical outcomes of locking stand-alone cage versus anterior plate construct in two-level anterior cervical discectomy and fusion: a systematic review and meta-analysis. European Spine Journal. 2019;28(1):199-208.

35. Zhao Y, Yang S, Huo Y, Li Z, Yang D, Ding W. Locking stand-alone cage versus anterior plate construct in anterior cervical discectomy and fusion: a systematic review and meta-analysis based on randomized controlled trials. European Spine Journal. 2020;29(11):2734-44.

36. Zhou J, Li J, Lin H, Li X, Zhou X, Dong J. A comparison of a self-locking stand-alone cage and anterior cervical plate for ACDF: Minimum 3-year assessment of radiographic and clinical outcomes. Clinical Neurology and Neurosurgery. 2018;170:73-8.

37. Schuermans VNE, Smeets AYJM, Wijsen NPMH, Curfs I, Boselie TFM, van Santbrink H. Clinical adjacent segment pathology after anterior cervical discectomy, with and without fusion, for cervical degenerative disc disease: A single center retrospective cohort study with long-term follow-up. Brain and Spine. 2022;2:100869.

38. Solomon MJ, McLeod RS. Should we be performing more randomized controlled trials evaluating surgical operations? Surgery. 1995;118(3):459-67.

39. Fung EK, Loré JM. Randomized Controlled Trials for Evaluating Surgical Questions. Archives of Otolaryngology–Head & Neck Surgery. 2002;128(6):631-4.

40. Mansouri A, Cooper B, Shin SM, Kondziolka D. Randomized controlled trials and neurosurgery: the ideal fit or should alternative methodologies be considered? Journal of Neurosurgery JNS. 2016;124(2):558-68.

41. Cohen SP. Epidemiology, diagnosis, and treatment of neck pain. Mayo Clin Proc. 2015;90(2):284-99.

42. Peng B, DePalma MJ. Cervical disc degeneration and neck pain. J Pain Res. 2018;11:2853-7.

43. Xing R, Liu W, Li X, Jiang L, Yishakea M, Dong J. Characteristics of cervical sagittal parameters in healthy cervical spine adults and patients with cervical disc degeneration. BMC Musculoskelet Disord. 2018;19(1):37.

44. Iyer S, Kim HJ. Cervical radiculopathy. Curr Rev Musculoskelet Med. 2016;9(3):272-80.

45. Davies BM, Mowforth OD, Smith EK, Kotter MR. Degenerative cervical myelopathy. Bmj. 2018;360:k186.

46. Xia X-P, Chen H-L, Cheng H-B. Prevalence of Adjacent Segment Degeneration After Spine Surgery: A Systematic Review and Meta-analysis. Spine. 2013;38(7).

47. Wang F, Hou H-T, Wang P, Zhang J-T, Shen Y. Symptomatic adjacent segment disease after single-lever anterior cervical discectomy and fusion: Incidence and risk factors. Medicine. 2017;96(47).

48. Rudy IS, Poulos A, Owen L, Batters A, Kieliszek K, Willox J, et al. The correlation of radiographic findings and patient symptomatology in cervical degenerative joint disease: a cross-sectional study. Chiropractic & Manual Therapies. 2015;23(1):9.

49. Yang X, Bartels RHMA, Donk R, Depreitere B, Walraevens J, Zhai Z, et al. Does Heterotopic Ossification in Cervical Arthroplasty Affect Clinical Outcome? World Neurosurgery. 2019;131:e408-e14.

50. Goedmakers CMW, Lak AM, Duey AH, Senko AW, Arnaout O, Groff MW, et al. Deep Learning for Adjacent Segment Disease at Preoperative MRI for Cervical Radiculopathy. Radiology. 2021;301(3):664-71.
51. Hilibrand AS, Carlson GD, Palumbo MA, Jones PK, Bohlman HH. Radiculopathy and myelopathy at segments adjacent to the site of a previous anterior cervical arthrodesis. J Bone Joint Surg Am. 1999;81(4):519-28.
52. Donk RD, Verhagen WIM, Hosman AJF, Verbeek A, Bartels RHMA. Symptomatic Adjacent Segment Disease After Anterior Cervical Discectomy for Single-level Degenerative Disk Disease. Clinical Spine Surgery. 2018;31(1).
53. Goedmakers CMW, de Vries F, Bosscher L, Peul WC, Arts MP, Vleggeert-Lankamp CLA. Long-term results of the NECK trial: implanting a disc prosthesis after cervical anterior discectomy cannot prevent adjacent segment disease. 5-years clinical follow-up of a double-blinded randomised controlled trial. The Spine Journal. 2022.
54. Nunley PD, Jawahar A, Cavanaugh DA, Gordon CR, Kerr 3rd EJ, Utter PA. Symptomatic adjacent segment disease after cervical total disc replacement: re-examining the clinical and radiological evidence with established criteria. The Spine Journal. 2013;13(1):5-12.
55. Suchomel P, Jurák L, Beneš V, Brabec R, Bradáč O, Elgawhary S. Clinical results and development of heterotopic ossification in total cervical disc replacement during a 4-year follow-up. European Spine Journal. 2010;19(2):307-15.
56. Kontakis M, Marques C, Löfgren H, Mosavi F, Skeppholm M, Olerud C, et al. Artificial disc replacement and adjacent-segment pathology: 10-year outcomes of a randomized trial. Journal of Neurosurgery: Spine. 2022;36(6):945-53.
57. Kong L, Cao J, Wang L, Shen Y. Prevalence of adjacent segment disease following cervical spine surgery: A PRISMA-compliant systematic review and meta-analysis. Medicine. 2016;95(27).
58. Robertson JT, Papadopoulos SM, Traynelis VC. Assessment of adjacent-segment disease in patients treated with cervical fusion or arthroplasty: a prospective 2-year study. Journal of Neurosurgery: Spine. 2005;3(6):417-23.
59. Kuo DT, Tadi P. Cervical Spondylosis. BTI - StatPearls.
60. Urrutia J, Zamora T, Yurac R, Campos M, Palma J, Mobarec S, et al. An Independent Inter- and Intraobserver Agreement Evaluation of the AOSpine Subaxial Cervical Spine Injury Classification System. Spine (Phila Pa 1976). 2017;42(5):298-303.
61. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016;18(12):e323.
62. Heil BJ, Hoffman MM, Markowetz F, Lee S-I, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nature Methods. 2021;18(10):1132-5.
63. Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, et al. BACH: Grand challenge on breast cancer histology images. Med Image Anal. 2019;56:122-39.
64. Arun NA-O, Gaw NA-O, Singh PA-O, Chang KA-O, Aggarwal M, Chen B, et al. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. (2638-6100 (Electronic)).
65. Thumbavanam Arun N, Gaw N, Singh P, Chang K, Viktoria Hoebel K, Patel J, et al. Assessing the validity of saliency maps for abnormality localization in medical imaging. arXiv e-prints. 2020:arXiv:2006.00063.
66. Al Arif SMMR, Knapp K, Slabaugh G. Fully automatic cervical vertebrae segmentation framework for X-ray images. Computer Methods and Programs in Biomedicine. 2018;157:95-111.
67. Djuric N, Lafeber GCM, Vleggeert-Lankamp CLA. The contradictory effect of macrophage-related cytokine expression in lumbar disc herniations: a systematic review. European Spine Journal. 2020;29(7):1649-59.

11