01100001 - 01101110 -
"...bien qu'elles fissent plusieurs
choses aussi bien, ou peutestre mieux qu'aucun de nous, elles
manqueroient infalliblement en quelques autres, par lesquelles
on découvriroit qu'elles n'agiroient pas par connoissance,
mais seulement par la disposition du leurs organes."

01101001 - 01110100 - 01100001 - 00100000 - 01110111 - 01100001 -
01110011 - 00100000 - 01101000 - 01100101 - 01110010 - 01100101

Evolving Applications of Machine Intelligence in Neurosurgery

Victor E. Staartjes

EVOLVING APPLICATIONS OF MACHINE INTELLIGENCE IN NEUROSURGERY

Victor E. Staartjes

Department of Neurosurgery & Clinical Neuroscience Center
University Hospital Zurich, University of Zurich, Zurich, Switzerland

Department of Neurosurgery
Bergman Clinics, Naarden, The Netherlands

The studies described in this thesis were carried out at the Department of Neurosurgery & Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland, as well as the Department of Neurosurgery, Bergman Clinics, Naarden, The Netherlands.

Research described in this thesis was entirely supported by academic funding. No benefits in any form have been or will be received from any commercial parties related directly or indirectly to the subject or content of this manuscript.

Thank you to the following sponsors for supporting the printing of this thesis:

MRIGuidance B.V. ● mriguidance.com ● info@mriguidance.com

Qmediq ● qmediq.com ● md@qmediq.com

Newspine B.V. ● newspine.nl ● info@newspine.nl

ISBN:
Layout: Victor E. Staartjes
Cover Design and Illustrations: Anita M. Klukowska
Printing: Ridderprint, Alblasserdam, The Netherlands ● ridderprint.nl

VRIJE UNIVERSITEIT

EVOLVING APPLICATIONS OF MACHINE INTELLIGENCE IN NEUROSURGERY

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op woensdag 1 juni 2022 om 13.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Victor Egon Staartjes

geboren te Amsterdam

promotoren:	prof.dr. W.P. Vandertop prof.dr. L. Regli
copromotoren:	dr. M.L.J.F. Schröder dr. C. Serra
promotiecommissie:	prof.dr. P.C. de Witt Hamer prof.dr. W.C. Peul prof.dr. C.B.L.M. Majoie prof.dr. V. de Groot dr. H.A. Marquering drs. A. Stadhouder

"Knowledge is not a series of self-consistent theories that converges toward an ideal view; it is rather an ever increasing ocean of mutually incompatible (and perhaps even incom

mensurable) alternatives, each single theory, each fairy tale, each myth that is part of the collection forcing the others into greater articulation and all of them contributing, via this process of competition, to the development of our consciousness."

Paul Feyerabend

Against Method: Outline of an Anarchistic Theory of Knowledge (1975)

Table of Contents

[Chapter 1]

General Introduction and Outline

General Introduction

Brief History of Machine Intelligence in Clinical Neuroscience

The beginnings of a fascination with "artificial minds", automation, and the idea of inanimate "machines" tackling mundane problems have existed for millennia. "Robots" (originating from the Slavic word for "labour") as automated, task-performing machines have been imagined and automatons have been constructed in ancient Greece, China, and Egypt.[1] Mythological and alchemical ideas on making inanimate objects think were equally popular until the dawn of the 20th century but persist until today.[2]

Initially, the seeds that can be regarded as the first rudimentary learning techniques, came from the idea of mechanizing human thought by attempting to deduce formal rules of reasoning in many ancient cultures – resulting in logic, mathematical algorithms, and then foundational epistemological theories, most of which have persisted until today. One of the most well-known and simple of these formal rules of reasoning is the law of parsimony (or Occam's Razor), often cited as "*Entities should not be multiplied beyond necessity*"[3] and is still a frequently used heuristic to minimize unnecessary assumptions.

In the 17th century, Leibniz and Descartes further developed these seeds and had their attempts at creating a systematic theory of thought. Descartes specifically discusses thinking machines and their limitations in his *Discours* [4], first published in 1637, where he states that:

"*[…] even though some machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they are acting not from understanding, but only from the disposition of their organs. For whereas reason is a universal instrument, which can be used in all kinds of situations, these organs need some particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act."*

On a side note, these early thoughts on distinguishing artificial from human minds are not far removed from the later *Turing's Test.*[5]

In the 20th century, the same Alan Turing eventually demonstrated that at least some forms of rational thought – in the form of mathematical algorithms – can indeed be generated by simple, though abstract, machines.[6] These thoughts, coupled with the development of the modern computer after the second world war, led to the first approaches that are recognizable to us today as "machine intelligence": Pitts, McCulloch, Minsky, and Edmonds developed the first neural networks, and soon researchers were applying similar concepts to compete in boardgames, analyse text, and control robots.[7] Formally, the word "Artificial Intelligence" (AI) was first coined in 1956, when a group of mathematicians brainstormed on the topic of complex information processing at Dartmouth College in New Hampshire.[8]

After this initial period of excitement and optimism about AI in the 1950s and 1960s, government funding and interest in the field dropped off during a period in the late 1970s that has been coined the "AI Winter." Later, coinciding with the exponential development of computing power, the fascination with intelligent machines was re-kindled and many of the architectures that are used today were developed.

The field of data science has seen an immense rise in popularity in the past few years, due to advances in statistical modelling techniques, the wide accessibility of computing power, and the availability of online resources and software packages that enable powerful analyses, even with little domain knowledge.[9] This has been one of the main driving forces for the widespread emergence of machine intelligence – including AI and machine learning (ML) – and its increasing inclusion in modern medicine.

The prevalence of ML application in neurosurgery has been growing rapidly, which is demonstrated by the sharp increase in publications on machine learning in clinical neuroscience in the past years: The number of publications on ML in neurosurgery, neurology, neuroradiology, neurorehabilitation, and neurointensive care medicine have grown exponentially, partially attributable to the greater availability of "big data" – which means nothing more than the inevitable increase in the average sample size of medical research datasets that has followed introduction of electronic health records and sometimes even automated data collection.[10–13] Early instances of the development and application of ML in the clinical neurosciences can be traced back as early as the late 1980s[14] and has seen a steady incline in the following decades. Especially around the start of the 2010s, the field of neurosurgery saw a tremendous increase in interest surrounding ML and its uses in clinical practice and research, as popular learning libraries such as Tensorflow, Keras, and MXNet became freely available.[15–18] A recent global survey regarding the application of ML-based algorithms in neurosurgery showed that 28.5% of surveyed neurosurgical professionals use ML in clinical practice, and that 31.1% utilize it in clinical research, although this survey most definitely overestimated application of ML due to sampling bias.[19]

This "democratization" of ML also carries with it an increasing number of publications on the topic with poor methodology, which reviewers of expert medical journals cannot be realistically expected to pick up. Even if one can nowadays get started with e.g. training a clinical prediction model in an hour or two, a solid methodological basis is still required in order to avoid some very common pitfalls. In addition, in medical research specifically, the principles of epidemiology and biostatistics must still be considered. The subject of machine intelligence is vast, and there are many sub-fields and methodological intricacies that would deserve explanation. While this general introduction aims to convey the general intuition and the necessary fundamental knowledge surrounding AI and ML, it barely scratches the surface of this subject's extent. To go further into detail would go beyond the scope of this introductory chapter.

Definitions

With the growing interest and demand of machine intelligence in medicine, some confusion regarding definitions has been introduced – and there is a lack of clear definitions. "Machine Intelligence" can be used as an umbrella term for AI and ML. Furthermore, ML can be seen as a sub-domain of AI. In ML, a form of "narrow intelligence", an algorithm learns to tackle a specific task by looking at prior observations, without being specifically programmed.[20,21] "Learning techniques" simply indicate the ability of an algorithm to learn from data without specific instructions. In contrast, AI is philosophically much more extensive than ML. It can be defined as an aspiration to emulate human "wide" intelligence, and thus to solve multiple, more complex problems and to make sophisticated decisions.[10]

Many ML methods exist, which are typically divided into three categories: Supervised and unsupervised learning methods as well as reinforcement learning (which will not be discussed here). Supervised and unsupervised learning methods are nowadays most commonly applied. In supervised learning, the algorithm is presented with labelled data in form of a training set, comprised of input variables (e.g. age, gender, functional neurological status) and a known target variable, also known as "ground truth" or "label" (e.g. survival). The goal of the algorithm is to generate a predicted value based on the patterns that the algorithm recognizes through connecting the input variable and the known target variable. If the algorithm is then presented with a new set of data (test set), it should be able to predict the target variable through the generalizable rules it has learned. Since the target value is included in the training set, the data set is considered "labelled" and therefore the ML paradigm is supervised. In unsupervised learning, on the other hand, the ML algorithm is presented with unlabelled data: the target variable is not included in the data set or is wholly unknown. The algorithm then aims to find patterns or "clusters" within the given data. These clusters can then be interpreted post-hoc and may lead to the discovery of previously unknown associations in highly dimensional datasets.[10,22,23]

Specifically, "deep learning" refers to ML methods that are architecturally organized in multiple layers that are thought to represent multiple levels of abstraction – These usually very large models, such as deep neural networks, can e.g. interpret images more accurately by first recognizing edges and corners, then putting these together to simple structures, which are then further processed to recognize complex objects.[24]

The most salient other terms are parameters and hyperparameters: While hundreds of ML architectures exist, they all include parameters. The parameters are those operators that govern how the input data is processed to eventually arrive at an output. Thus, the parameters of a model are also those that need to be iteratively adjusted during model training to arrive at accurate predictions, which is explained in more detail below. Hyperparameters on the other hand are hierarchically one step up from parameters: The hyperparameters of a model govern how the parameters are learned and need to be set by the data scientist before training. For example, the rate at which a parameter is changed during every iteration is a hyperparameter. Even the method with which the data are pre-processed before being fed into the algorithm can be considered a hyperparameter.

Optimization

One of the best ways to understand ML initially and intuitively is to understand the concept of optimization – the central dogma of learning techniques. The concept of optimization lies at the core of the inner workings of ML and other statistical modelling techniques. Optimization can be defined as the iterative adjustment of parameters to improve some objective function (error function). This objective function can be minimized (in the case of e.g. an error rate), or maximized (e.g. accuracy or sensitivity). In optimization, usually, random values are initially assigned to all model parameters. Using this setting of parameters, the algorithm generates predictions based on the training set and calculates the error function. During the next iteration, the parameters are then (randomly) adjusted in a certain direction, and the error function is assessed again. If the error increases, the parameters are adjusted in the other direction, and the error function is evaluated again. If the error decreases, the algorithm knows that it is probably "on the right path", and the parameters are adjusted again in the same direction at each iteration until a global minimum of the error is reached.[10]

Overfitting

The goal in prediction modelling in ML is to produce accurate and generalizable predictions. Overfitting is one of the most common and prevalent problems of ML. This major pitfall happens when a model adjusts too closely to the training data and demonstrates poor performance when applied to the testing set.[10] Initially, the model performs admirably on the training set, however, it fails to make accurate predictions if applied to new data. This can happen when complex models with a large number of features are applied to small datasets, for example. Data leakage and the use of too complex models in relation to data complexity – apart from often producing "black box" models that are not interpretable – can also lead to overfitting.[10,25] While the model may minimize training error, it ultimately generalizes poorly.[26] One of the various reasons for the occurrence of overfitting is "memorization" of the training dataset.[27,28] Instead of learning generalizable interactions among variables, the algorithm simply remembers observations within the training set. By recalling the memorized observations, the model delivers minimal training error, which emulates the appearance of a good model fit. When introducing new data, these memorized patterns do not hold any value because they are too tightly fit to the training observations. Overfitting can be identified through a considerable difference in performance between training and testing. While adequate prediction models can often fare slightly worse in their testing performance compared to their training performance, an extensive difference may indicate relevant overfitting.[10] The gold standard in combatting overfitting is to use resampling methods during training.[29] While there are various ways resampling can be achieved during training, at its core lies the principle of splitting the training data. This allows fitting of multiple models on subsets of the training data, which are then already validated on the subsets of the training data that were not used by that specific model in training. Thus, "out-of-sample performance" on new data can already be estimated during the training process, which helps in selecting the optimal hyperparameters. After the optimal hyperparameters have been identified using a resampling method (such as cross validation or bootstrapping), a final model is usually trained on the entire dataset, or an ensemble (pooling of multiple model outputs) of the various trained models is utilized.[10]

Bias and the Importance of External Validation

External validation is an important step in the development and improvement of ML models. Bias can come in the form of centre bias, which includes variations in treatment protocols, surgical techniques or level of experience, a different magnetic resonance imaging (MRI) scanner, as well as sampling/selection bias, which refers to a systematically different data collection process of the patient cohort compared to the data that the model will ultimately be applied to. To address these issues empirically, external validation is necessary: Models are tested using an unrelated, external dataset. If the model performs adequately and performance is similar or slightly worse than during training, many forms of bias can be ruled out and generalizability of the model can be confirmed.[10]

Feature Selection

Especially in the era of "big data", selecting an optimal set of inputs for a certain model has become crucial. Often clinical researchers are faced with "wide" datasets with a large number of different features in relation to the number of observations. In the endeavour to continually improve performance, it has become quite commonplace to use a high number of features for the development of more complex

models. These models require a very large amount of training data and are at an increased risk of overfitting. In an effort to achieve easier application in clinical practice (because less variables need to be collected and entered to arrive at an output) and to prevent overfitting, less complex models with fewer features are often desired in real-world applications – Parsimonious models are needed.[30] Feature selection can play an important role in the development of these "simpler" algorithms.

Feature selection has been observed to improve model interpretation and yield shorter training times. Additionally, it can even have a positive impact on model performance. Because most algorithms estimate parameters for each term of the model, non-informative or redundant features can add uncertainty to the predictions and reduce overall performance. The main goal of feature selection is the removal of non-informative or redundant features.[31] This includes features that are known to correlate poorly with the endpoint or that are highly correlated among themselves, that are unreliable in their capturing, sparse features, and features such as patient ID and names.

Various methods of feature selection exist and can be split into two fundamental groups: supervised selection methods and unsupervised selection methods. In supervised methods, the outcome is considered, whereas in unsupervised selection methods the outcome is ignored. One standard method for selecting features in a supervised way is recursive feature elimination (RFE), in which models are built with a decreasing number of inputs, and variables with low importance measures are removed iteratively. Resampled performance is tracked, and the highest-performing combination of inputs is selected.[31,32]

Model Evaluation – Discrimination and Calibration

Model discrimination describes the accuracy with which the model predicts a binary outcome in a binary way.[10] In other words, it describes the ability of a model to accurately predict the occurrence of an outcome as a class – e.g. complication "yes" or "no" in binary classification problems. To evaluate the discriminative ability of a model, the predicted classes – which often need to be dichotomized from a predicted probability (ranging between 0% and 100%) as the standard output of most model architectures – are contrasted with the ground truth. Commonly, a confusion matrix is generated, too – a simple table with four fields, comparing predicted and observed classes. A list of common discrimination metrics can be found in Table 1.

Discrimination Metric	
Area under the curve (AUC)	
Accuracy	$Accuracy = \frac{TP + TN}{P + N}$
Sensitivity	$Sensitivity = \frac{TP}{P}$
Specificity	$Specificity = \frac{TN}{N}$
Positive Predictive Value	$PPV = \frac{TP}{TP + FP}$
Negative Predictive Value	$PPV = \frac{TP}{TP + FP}$
F1 Score	$F1 = 2 \times \frac{PPV \times Sensitivity}{PPV + Sensitivity}$

Table 1: List of common discrimination metrics

Model calibration, however, describes the degree to which a model's predicted probability correlates to the real-world incidence of the true outcome ("true posterior"). Model calibration is often omitted in many publications, even though it is generally more valuable in clinical practice compared to discrimination, as physicians and patients are interested in their risk instead of a binary prediction.[33] Calibration curves, slope, and intercept should therefore always be assessed in model evaluation and subsequently reported.

Apart from model discrimination and calibration, there are several other points that are important to note regarding the development of clinical prediction models. First, the sample size of the data set should be large enough to allow for adequate model training – although there is little consensus on how to calculate necessary sample sizes a priori.[10,34] Secondly, the occurrence of class imbalance must be recognized and adjusted for.[35,36] Missing data has to be reported and, if necessary, imputed.[37] Finally, in binary classification problems, the cut-off to transform the predicted probabilities into a dichotomous outcome should be reported.[10]

Current Applications of Machine Intelligence in Clinical Neuroscience

Clinical prediction models are by far the most common and most widely used ML based algorithms in clinical neuroscience.[19] Their objective and individualized predictions can in theory contribute to more accurate patient counselling, clinical decision-making, resource allocation, and more – Although there is little real-world evidence on their true impact.[38]

The field of neuroimaging has become increasingly popular in applications of ML and AI in recent years.[19] Neuroimaging has seen a tremendous influx of data due to the increasing incidence of imaging that is carried out.[39] Furthermore, data in neuroimaging are of complex nature and often have high resolutions, which fits perfectly within the realm of deep learning.[24] An important part in the application of ML in neuroimaging is the use of radiomics.[40] Radiomic analysis entails the extraction of a large number of features from medical images through the application of ML algorithms. These "radiomic features" may hold additional information, that can be utilized in image characterization or to perform classification.[39,41,42] Other applications of ML in neuroimaging include simple direct classification, image segmentation, image super-sampling (increasing resolution) and image conversion.[39]

Finally, ML has been successfully applied to electronic health records as so-called "natural language processing" (NLP) and to so-called "time series analysis", which indicated the use of temporally distributed data such as blood pressure, intracranial pressure, or electrocardiographic curves. In terms of NLP, ML can summarize or structure medical records and even help gather data for medical research in an automated fashion.[13] Especially in the neurointensive care unit, time series analysis has been applied successfully, for example to analyse intracranial pressure curves to forecast adverse events.[43]

Outline

Brief History of Machine Intelligence in Clinical Neuroscience

The overarching goal of this thesis is to delineate and potentially expand the limits of current applications of ML with a primary focus on spinal neurosurgery, although cranial neurosurgical applications are also discussed. This goal will be pursued by applying state-of-the-art ML methods to common clinical situations and existing, relatively standardized learning problems such as clinical prediction modelling, but also by

piloting novel approaches to ML in neurosurgery that may eventually lead to tangible benefits for patients, if further development. Special care will also be taken to demystify ML as a „magical“, revolutionary, or omnipotent method. Instead, the limitations of ML for specific clinical use-cases will be specifically discussed, and the applications of learning techniques in clinical medicine will be presented as what they truly are - an evolution from biostatistical and epidemiological principles, rather than a revolution. Although technicalities and jargon will be limited to a minimum, some technical terms will be necessary from time to time.

After the brief introduction to machine intelligence in clinical neuroscience presented in **chapter 1**, this thesis is structured in three separate parts that demonstrate different applications of ML in neurosurgery. The domains that will be covered are clinical patient assessment (**Part I**), operative imaging (**Part II**), and clinical prediction modelling (**Part III**).

Part I of this thesis revolves around clinical assessment of patients with degenerative disease of the lumbar spine. Objective functional tests have been introduced as an additional dimension of patients assessment in clinical practice and research, supplementing questionnaires on subjective functional impairment.[44] The five-repetition-sit-to-stand test (5R-STS) is an objective functional test that has been previously used in many other disciplines including chronic pulmonary or Parkinson's disease.[45,46] In **chapter 2**, we formally validate the 5R-STS for patients with degenerative disease of the lumbar spine, and calculate normative values as well as a baseline severity stratification to grade the extent of objective functional impairment (OFI).

Chapter 3 focuses on adding ML to objective functional testing to allow for truly personalized assessment of patients with lumbar degenerative disease. Normally, for objective functional tests or for any measurement (e.g. D-dimers or thyroid-stimulating hormone) for that matter, a single cut-off or a grading based on an entire normative population is used to distinguish "normal" from "abnormal". This approach does not consider the differences in test properties among different patients, or their expected normal values for their specific age, gender, body weight, and so forth. We set out to develop an unsupervised clustering model to grade the severity of OFI without relying on a normative population.

Finally, **chapter 4** links the findings from the previous two chapters. Here, we develop a regression model that is able to predict percentile-wise expected performance based on an individual patient's demographics. Together with the validation and the clustering model developed in the previous two chapters, this enabled the creation of a web-app that enables quantification of objective functional impairment in a way that is specific to each individual patient.

Part II focuses on novel approaches to improve operative imaging using advanced ML techniques. First, in **chapter 5**, we pilot a novel approach to surgical planning and intraoperative navigation in lumbar spine surgery by allowing fast generation of synthetic computed tomography (CT) images from magnetic resonance imaging (MRI) of the lumbar spine. This avoids the logistic, financial, and ionizing hazards of a separate CT scan.

Chapter 6 introduces the concept of real-time intraoperative anatomical navigation in cranial surgery. Current approaches to intraoperative navigation, such as frame-based or frameless neuronavigation, are based on preoperative imaging and are often unreliable after e.g. brain shift.[47] Other approaches such as

intraoperative MRI or ultrasound either are time-consuming and logistically heavy, or have limited sensitivity. In a proof-of-concept study, we evaluate the potential of navigating – like a master surgeon – purely based on endoscopic footage in endonasal surgery using machine vision.

Part III revisits clinical prediction modelling, which is still the most popular application of ML in clinical medicine. In two multicentre, multinational studies, we attempt to develop and thoroughly externally validate clinical prediction models for outcomes that are known as hard to predict. In **chapter 7**, we predict the risk of new functional impairment after brain tumour surgery, which is demonstrably hard to predict even for seasoned experts.[48] In **chapter 8**, an attempt was made to preoperatively determine which patients with degenerative disease of the lumbar spine are likely and which patients are unlikely to benefit from lumbar fusion surgery – a task that is known to be difficult.[49,50]

Chapter 9 discusses the main findings of this thesis and possible future directions of the field. Finally, **chapters 10 and 11** contain summaries in English and in Dutch, respectively.

References

1. Bassetti WHC. *Robotics and Automation Handbook*. Taylor & Francis; 2005.
2. Cave S, Dihal K. Hopes and fears for intelligent machines in fiction and reality. *Nat Mach Intell*. 2019;1(2):74-78. doi:10.1038/s42256-019-0020-9
3. Ariew R. Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony. Published online 1976.
4. Descartes R. *Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences*. Sutherland and Knox; 1850.
5. Turing AM. Computing Machinery and Intelligence. *Mind*. 1950;59(October):433-460. doi:10.1093/mind/LIX.236.433
6. Gonçalves B. "Can machines think?" The missing history of the Turing test. :24.
7. Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach*. 2nd ed. Prentice Hall/Pearson Education; 2003.
8. Solomonoff RJ. The time scale of artificial intelligence: Reflections on social effects. *Hum Syst Manag*. 1985;5(2):149-153. doi:10.3233/HSM-1985-5207
9. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med*. 2001;23(1):89-109. doi:10.1016/S0933-3657(01)00077-X
10. Kernbach JM, Staartjes VE. Machine learning-based clinical prediction modelling -- A practical guide for clinicians. *ArXiv200615069 Cs Stat*. Published online June 23, 2020. Accessed September 12, 2021. http://arxiv.org/abs/2006.15069
11. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181
12. Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health Care. *JAMA*. 2013;309(13):1351-1352. doi:10.1001/jama.2013.393
13. Staartjes VE, Stienen MN. Data Mining in Spine Surgery: Leveraging Electronic Health Records for Machine Learning and Clinical Research. *Neurospine*. 2019;16(4):654-656. doi:10.14245/ns.1938434.217
7. Mathew B, Norris D, Mackintosh I, Waddell G. Artificial Intelligence in the Prediction of Operative Findings in Low Back Surgery. *Br J Neurosurg*. 1989; Published online July 6, 2009. doi:10.3109/02688698909002791
15. Schilling AT, Pavan SP, Feghali J, Jimenez AE, Tej AD. A Brief History of Machine Learning in Neurosurgery. In: *Machine Learning in Clinical Neuroscience: Foundations and Clinical Applications*. Acta Neurochirurgica Supplement. Springer International Publishing [in press]; 2022. doi:10.1007/978-3-030-85292-4
16. Chollet F. Keras: Deep learning library for Theano and TensorFlow. *https://keras.io/*. 2015;7:8.
17. Chen T, Li M, Li Y, et al. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *ArXiv151201274 Cs*. Published online December 3, 2015. Accessed September 14, 2021. http://arxiv.org/abs/1512.01274
18. Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Published online 2015. http://tensorflow.org/
19. Staartjes VE, Stumpo V, Kernbach JM, et al. Machine learning in neurosurgery: a global survey. *Acta Neurochir (Wien)*. 2020;162(12):3081-3091. doi:10.1007/s00701-020-04532-1
20. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media; 2013.

21. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*. Published online 2015. doi:10.1126/science.aaa8415
22. Glaser JI, Benjamin AS, Farhoodi R, Kording KP. The roles of supervised machine learning in systems neuroscience. *Prog Neurobiol*. 2019;175:126-137. doi:10.1016/j.pneurobio.2019.01.008
23. Senders JT, Zaki MM, Karhade AV, et al. An introduction and overview of machine learning in neurosurgical care. *Acta Neurochir (Wien)*. 2018;160(1):29-38. doi:10.1007/s00701-017-3385-8
24. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539
25. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206. doi:10.1038/s42256-019-0048-x
26. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593
27. Zhang C, Vinyals O, Munos R, Bengio S. A Study on Overfitting in Deep Reinforcement Learning. *ArXiv180406893 Cs Stat*. Published online April 20, 2018. Accessed September 14, 2021. http://arxiv.org/abs/1804.06893
28. Arpit D, Jastrzębski S, Ballas N, et al. A Closer Look at Memorization in Deep Networks. In: *International Conference on Machine Learning*. PMLR; 2017:233-242. Accessed September 14, 2021. https://proceedings.mlr.press/v70/arpit17a.html
29. Staartjes VE, Kernbach JM. Letter to the Editor Regarding "Investigating Risk Factors and Predicting Complications in Deep Brain Stimulation Surgery with Machine Learning Algorithms." *World Neurosurg*. 2020;137:496. doi:10.1016/j.wneu.2020.01.189
30. Fu Y, Liu C, Li D, et al. Exploring Structural Sparsity of Deep Networks via Inverse Scale Spaces. *ArXiv190509449 Cs Math Stat*. Published online April 20, 2021. Accessed September 14, 2021. http://arxiv.org/abs/1905.09449
31. Staartjes V, Kernbach JM, Stumpo V, van Niftrik CHB, Serra C, Regli L. Foundations of feature selection in clinical prediction modelling. In: *Machine Learning in Clinical Neuroscience: Foundations and Clinical Applications*. Acta Neurochirurgica Supplement. Springer International Publishing [in press]; 2022. doi:10.1007/978-3-030-85292-4
32. Kuhn M, Johnson K. *Applied Predictive Modelling*. Springer New York; 2013. doi:10.1007/978-1-4614-6849-3
33. Staartjes VE, Kernbach JM. Letter to the Editor. Importance of calibration assessment in machine learning–based predictive analytics. *J Neurosurg Spine*. 2020;32(6):985-987. doi:10.3171/2019.12.SPINE191503
34. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Stat Med*. 2019;38(7):1262-1275. doi:10.1002/sim.7993
35. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw*. 2018;106:249-259. doi:10.1016/j.neunet.2018.07.011
36. Staartjes VE, Schröder ML. Letter to the Editor. Class imbalance in machine learning for neurosurgical outcome prediction: are our models valid? *J Neurosurg Spine*. 2018;29(5):611-612. doi:10.3171/2018.5.SPINE18543
37. Kowarik A, Templ M. Imputation with the R Package VIM. *J Stat Softw*. 2016;74(1):1-16. doi:10.18637/jss.v074.i07
38. Steinmetz MP, Mroz T. Value of Adding Predictive Clinical Decision Tools to Spine Surgery. *JAMA Surg*. Published online March 7, 2018. doi:10.1001/jamasurg.2018.0078
39. Stumpo V, Kernbach JM, van Niftrik CHB, et al. Machine Learning Algorithms in Neuroimaging: An Overview. In: *Machine Learning in Clinical Neuroscience: Foundations and Clinical Applications*. Acta Neurochirurgica Supplement. Springer International Publishing [in press]; 2022. doi:10.1007/978-3-030-85292-4

40. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights Imaging*. 2020;11(1):91. doi:10.1186/s13244-020-00887-2
41. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036
42. Avanzo M, Wei L, Stancanello J, et al. Machine and deep learning methods for radiomics. *Med Phys*. 2020;47(5):e185-e202. doi:10.1002/mp.13678
43. Shaw M, Hawthorne C, Moss L, et al. Time Series Analysis and Prediction of Intracranial Pressure Using Time-Varying Dynamic Linear Models. *Acta Neurochir Suppl*. 2021;131:225-229. doi:10.1007/978-3-030-59436-7_43
44. Stienen MN, Ho AL, Staartjes VE, et al. Objective measures of functional impairment for degenerative diseases of the lumbar spine: a systematic review of the literature. *Spine J Off J North Am Spine Soc*. 2019;19(7):1276-1293. doi:10.1016/j.spinee.2019.02.014
45. Jones SE, Kon SSC, Canavan JL, et al. The five-repetition sit-to-stand test as a functional outcome measure in COPD. *Thorax*. 2013;68(11):1015-1020. doi:10.1136/thoraxjnl-2013-203576
46. Duncan RP, Leddy AL, Earhart GM. Five Times Sit to Stand Test Performance in Parkinson Disease. *Arch Phys Med Rehabil*. 2011;92(9):1431-1436. doi:10.1016/j.apmr.2011.04.008
47. Iversen DH, Wein W, Lindseth F, Unsgård G, Reinertsen I. Automatic Intraoperative Correction of Brain Shift for Accurate Neuronavigation. *World Neurosurg*. 2018;120:e1071-e1078. doi:10.1016/j.wneu.2018.09.012
48. Sagberg LM, Drewes C, Jakola AS, Solheim O. Accuracy of operating neurosurgeons' prediction of functional levels after intracranial tumor surgery. *J Neurosurg*. 2017;126(4):1173-1180. doi:10.3171/2016.3.JNS152927
49. Staartjes VE, Vergroesen P-PA, Zeilstra DJ, Schröder ML. Identifying subsets of patients with single-level degenerative disc disease for lumbar fusion: the value of prognostic tests in surgical decision making. *Spine J*. 2018;18(4):558-566. doi:10.1016/j.spinee.2017.08.242
50. Willems P. Decision making in surgical treatment of chronic low back pain: the performance of prognostic tests to select patients for lumbar spinal fusion. *Acta Orthop Suppl*. 2013;84(349):1-35. doi:10.3109/17453674.2012.753565

[Part I]

Personalized Assessment of Lumbar Degenerative Disease

[Chapter 2]

The five-repetition sit-to-stand test: Evaluation of a simple and objective tool for the assessment of degenerative pathologies of the lumbar spine

Victor E. Staartjes
Marc L. Schröder

Published in: *J Neurosurg Spine. 2018 Oct;29(4):380-387.*

[Abstract]

Objective

Recently, objective functional tests have generated interest, since they can supplement an objective dimension to clinical assessment. The five-repetition sit-to-stand test (5R-STS) is a quick and objective tool that tests movements frequently used in everyday life. The aim of this prospective study was to evaluate the validity and reliability of the 5R-STS in degenerative pathologies of the lumbar spine.

Methods

Patients and healthy volunteers completed the standardized 5R-STS, Roland-Morris Disability Questionnaire (RMDQ), Oswestry Disability Index (ODI), Visual Analogue Scales (VAS) for back and leg pain, and EQ-5D for health-related quality of life (HRQOL). To assess convergent validity, the 5R-STS test times were correlated with these questionnaires.

Results

157 patients and 80 volunteers were enrolled. Direct correlation with RMDQ ($r = 0.49$), ODI (0.44), VAS for back pain (0.31), and indirect correlation with EQ-5D index (-0.41) was observed ($p < 0.001$). The 5R-STS showed no correlation with VAS for leg pain and EQ-5D VAS ($p > 0.05$). In 119 individuals, the 5R-STS demonstrated excellent test-retest reliability with an intraclass correlation coefficient of 0.98. The upper limit of normal, distinguishing patients with and without objective functional impairment, was identified as 10.35 seconds. A severity stratification classifies patients with test times of 10.5 – 15.2, 15.3 – 22.0, or greater than 22.0 seconds as having mild, moderate or severe functional impairment, respectively.

Conclusions

The 5R-STS test is a simple and effective tool to describe objective functional impairment. A patient able to perform the test in 10.4 seconds can be considered to have no relevant objective functional impairment. (ClinicalTrials.gov Identifier: NCT03303300)

Introduction

Validated patient-reported outcome measures (PROM) have been the gold standard for evaluating patients in spine surgery for decades.[1–4] Even so, objective tools have recently gained substantial importance.[5–7] Some objective functional tests have been introduced, and have even found their way into clinical practice, like the Six-Minute Walk test (6MWT), Timed Up and Go (TUG) test, and accelerometer-based tools.[6,8] Their proposed advantages include quick execution, high repeatability, straightforward interpretation of test results for patients, while supplementing an objective dimension to clinical assessment.[8] Moreover, it has been shown that patients frequently show preference of an objective functional test over questionnaires.[9] For these reasons, the development, standardization and validation of objective functional tests as adjuncts to conventional PROM is essential.

Surgical decision-making is based on clinical history, characterized by measures of pain along with impaired function and health-related quality of life (HRQOL), neurological, and radiological evidence of disease.[8] Commonly used PROM are the Visual Analogue Scale (VAS) for pain, Oswestry Disability Index (ODI) and Roland-Morris Disability Questionnaire (RMDQ) for functional impairment, and EQ-5D for HRQOL. Because accurate measurement of these parameters is critically important in spine surgery, and considering that the surgeon's assessment may considerably diverge from the patient's self-rating, adding an objective dimension to clinical decision-making may prove useful.[8,10] Furthermore, some patients may present with symptoms, or improvements after treatment, that simply cannot be captured by standardized questionnaires (e.g. painless symptoms like tingling, foot drop, limping).[8,11] Combined with traditional PROM, objective functional tests can extract novel and clinically helpful information.
Primary care, outpatient, and inpatient settings often do not allow for the use of objective functional tests like the 6MWT owing to restrictions in space, time and resources. Therefore, alternative options that are simpler to conduct have been sought. Sit-to-stand movements are commonly performed in everyday life, and are an indicator of physical activity, low back pain and muscle strength.[12–14]

The five-repetition sit-to-stand test (5R-STS) is a quick and convenient standardized test that is clinically useful and validated for various diseases including chronic obstructive pulmonary disease and Parkinson's disease (**Figure 1**).[13,15] However, the standardized 5R-STS has not been evaluated in patients with degenerative spinal pathologies. We aim to add a simple and objective tool to the spine surgeon's armamentarium by evaluating the 5R-STS for lumbar degenerative pathologies. In a prospective study, we assess its correlation with validated PROM, and propose an upper limit of normal (ULN) and a severity stratification.

Materials and Methods

Study Design and Oversight

Between October and December of 2017, patients were seen at a specialized outpatient spine surgery clinic. In addition, a representative population of healthy volunteers was enrolled as a control group. Convergent validity was assessed by correlating the 5R-STS results with validated PROM, namely VAS scores for back and leg pain, ODI, RMDB, and EQ-5D index and VAS.[2–4] In addition, we screened a range of demographic baseline variables, and propose reference values to facilitate interpretation of the 5R-STS. Participants filled in the questionnaires right after performing the test. This prospective trial (ClinicalTrials.gov Identifier: NCT03303300*)* was approved by the local institutional review board (Medical Research Ethics Committees United, Registration Number: W17.107), and was conducted according to the Declaration of Helsinki. Informed consent was obtained from all participants.

Study Population

All enrolled patients were candidates for surgery, and were assessed during outpatient consultations. Inclusion criteria were the presence of lumbar disc herniation (LDH), lumbar stenosis, lumbar spondylolisthesis, degenerative disc disease (DDD), or synovial facet cysts, requiring surgical treatment. Patients with hip or knee prosthetics, and those requiring walking aides were excluded to eliminate these confounders.

The control group comprised healthy individuals of all ages, and were either volunteers or employees of the department. Most volunteers were the patients' partners, and thus show comparable demographics. Volunteers disclosing spinal conditions, hip- or knee replacements, other lower extremity-related complaints, or that required walking aides were excluded.

Figure 1. Scheme of the five-repetition sit-to-stand (5R-STS) test.

The 5R-STS Test

The test was performed according to the protocol described by Jones et al.[13] The participants were asked to sit down on an armless chair of standard height (48 cm) and with a hard seat, firmly placed against a wall. The participants were instructed to fold their arms across their chest and to keep their feet flat on the ground. Participants were required to wear stable shoes for the test. To familiarize with the movement, the participants were asked to stand up fully and sit back down again once without using their upper limbs. If assistance was required, or if the maneuver could not be completed, the test was abandoned. Otherwise, the patients were asked to, starting on the command "go", stand up fully and sit down again, landing on the seat firmly, five times as fast as possible. Using a stopwatch, the five repetitions were timed from the initial command to the completed fifth stand. This time was recorded as

the participant's score. If the patient was unable to perform the test in 30 seconds, or not at all, this was noted down and the test score was recorded as 30 seconds.

Statistical Analysis

Data were reported as mean ± standard deviation for continuous and numbers (percentages) for categorical data. Analyses were carried out using R version 3.4.2 (The R Foundation for Statistical Computing, Vienna, Austria).[16] Groups were compared using Welch's two-sample *t* and χ^2 tests. 5R-STS test times were log_{10} transformed to achieve normal distribution. Pearson correlation was used to assess the correlation between log_{10} transformed 5R-STS test times and validated questionnaires. A Wilcoxon test was used to identify any learning effect between measurements. Intraclass correlation coefficients (ICC) for consistency and absolute agreement, along with their 95% confidence interval (CI) and standard error of the means (SEM), were used to examine test-retest reliability. The effect of age, gender, weight, height, BMI, and underlying pathology was assessed using linear regression. Standard and adjusted *z* scores, the ULN and the corresponding zone of indifference (ZOI) were calculated (Appendix 1).[17] A nonparametric severity stratification was created.[8,18] A two-sided $p \leq 0.05$ was considered significant.

Results

Cohort

One hundred fifty-seven patients and 80 healthy volunteers were enrolled (**Table 1**). Compared with the study group, participants in the control group were younger, showed a higher rate of smokers, and presented with higher body weight and BMI (all $p < 0.05$). The HRQOL data were comparable to the Dutch population.[19] Participants in the control group (6.44 ± 1.68 seconds) had significantly lower 5R-STS test times than those in the study group (13.32 ± 7.87, $p < 0.001$). Three patients (2%) were unable to perform the 5R-STS independently and abandoned the maneuver.

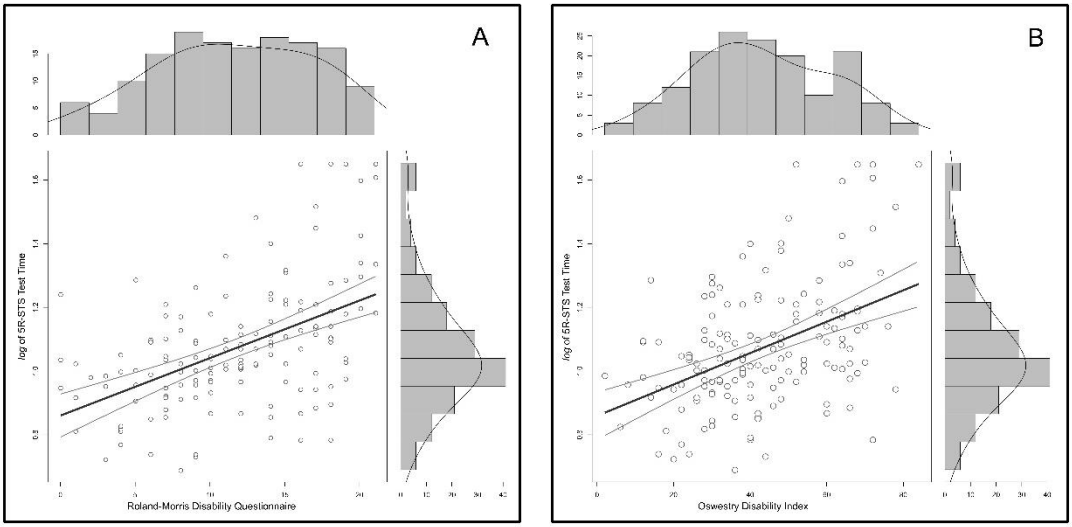


Figure 2. Scatterplots with marginal histograms showing logarithmically transformed five-repetition sit-to-stand (5R-STS) test times versus measures of functional impairment. There was a direct correlation with Roland-Morris Disability Questionnaire (A) (r = 0.49)

Convergent Validity

We observed a direct correlation of logarithmically transformed 5R-STS test times and functional impairment (**Figure 2**) as measured by RMDQ (r = 0.49, 95% CI: 0.36 – 0.60) and ODI (r = 0.44, 95% CI: 0.30

– 0.56), as well as with VAS back pain severity (r = 0.31, 95% CI: 0.16 – 0.45, all $p < 0.001$). There was no relevant correlation with VAS leg pain severity ($p = 0.207$, Figure 3). The 5R-STS also demonstrated indirect correlation with HRQOL (**Figure 4**) as measured by EQ-5D index (r = -0.41, 95% CI: -0.53 – 0.27, $p < 0.001$), but not with EQ-5D VAS ($p = 0.091$).

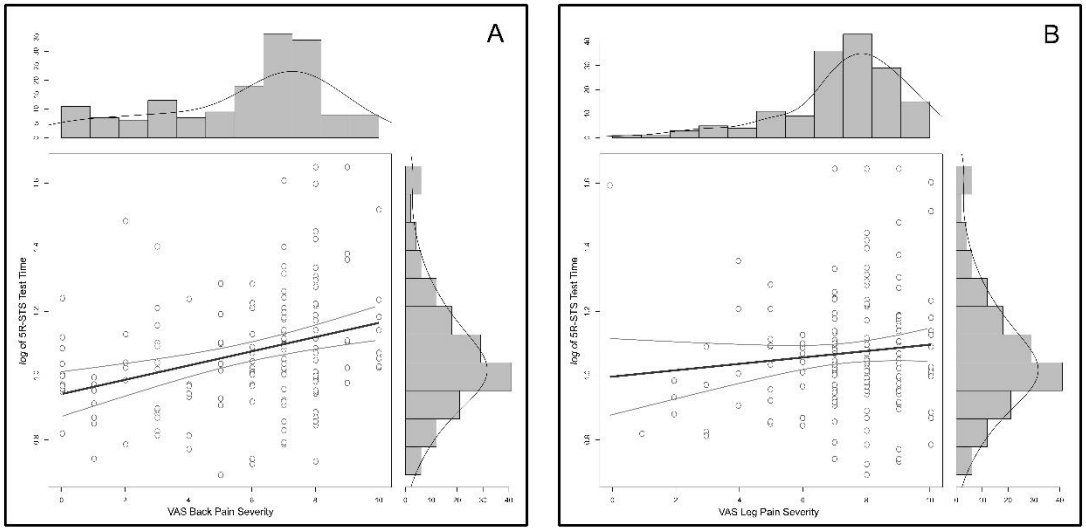


Figure 3. Scatterplots with marginal histograms showing logarithmically transformed five-repetition sit-to-stand (5R-STS) test times versus measures of pain severity. There was a direct correlation with the visual analogue scale (VAS) for back pain severity (A) (r = 0.31), but not with the VAS for leg pain severity (B) (r = 0.10).

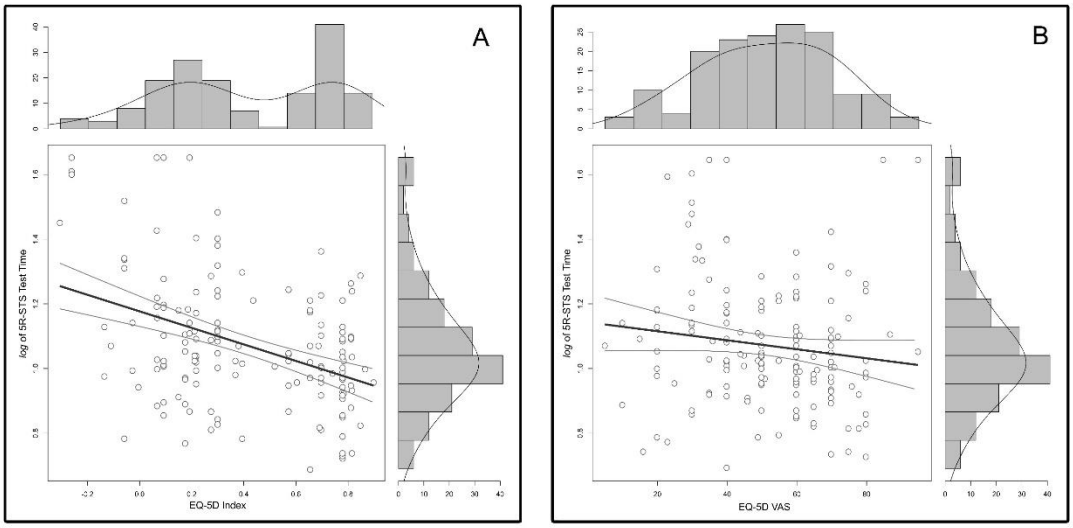


Figure 4. Scatterplots with marginal histograms showing logarithmically transformed five-repetition sit-to-stand (5R-STS) test times versus measures of health-related quality of life. There was a direct correlation with EQ-5D index (A) (r = -0.41), but not with EQ-5D visual analogue scale (VAS) scores (B) (r = -0.14).

Test-Retest Reliability

In 66 patients and 53 healthy volunteers, a second 5R-STS was performed after a 30-minute interval. No learning effect was detected between first (mean 9.35 seconds) and second (mean 9.25 seconds) measurements ($p = 0.204$). The ICC for consistency and for agreement was 0.98, indicating excellent test-retest reliability according to the Landis-Koch criteria.[20] Reliability was marginally better in the study group

(**Table 2**). A Bland-Altman plot (**Figure 5**) illustrates the test-retest bias of 0.1 seconds, with a 95% limit of agreement of -2.0 to 2.1 seconds.

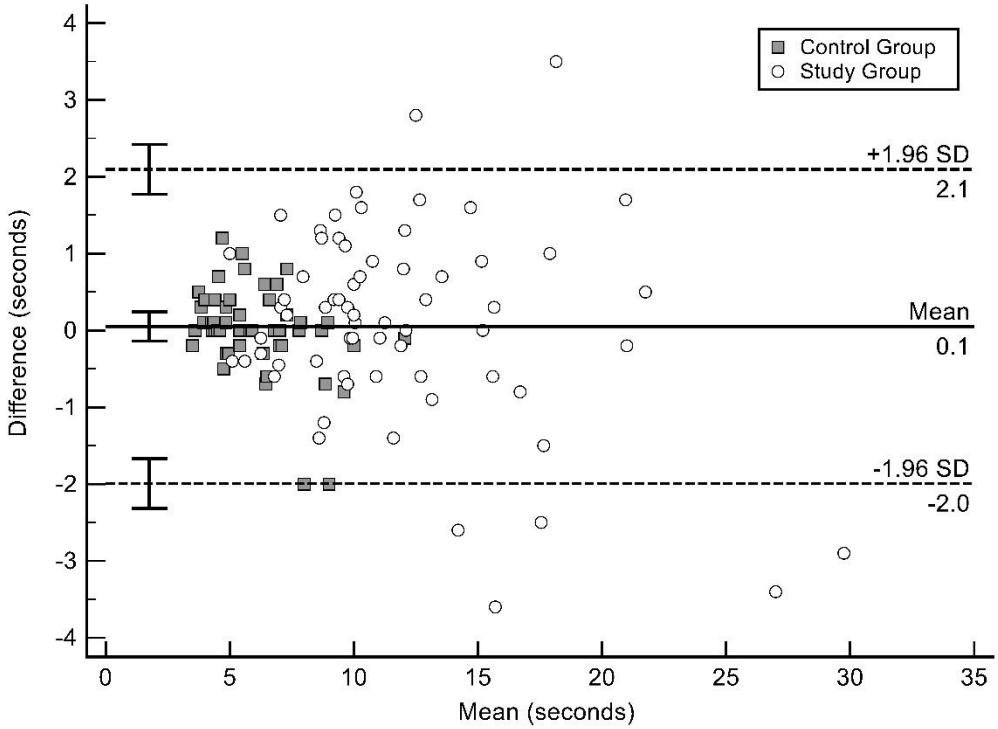


Figure 5. Bland-Altman plot for test-retest reliability. Mean differences between measurements are plotted against the mean measurements of 5R-STS test times. The interrupted lines represent the 95% limits of agreement, and the uninterrupted line demonstrates the mean retest bias of 0.1 seconds.

Upper limit of Normal

The ULN was identified as 10.35 seconds with a ZOI ranging between 9.32 and 11.38 seconds. Values in this "gray zone" cannot be unambiguously classified as either healthy or pathological owing to measurement error.

Severity Stratification

A severity stratification was developed by partitioning 5R-STS results of the study group into three grades of severity in a nonparametric fashion (**Table 3**). In this cohort, the ULN corresponded roughly to the 50th percentile. According to this severity stratification, patients with 5R-STS times lower or equal to the ULN can be considered without relevant functional impairment. Mild functional impairment was graded between the ULN (~50th percentile) and the 75th percentile, whereas moderate and severe functional impairment were between the 76th and 90th, and 91st and 100th percentile, respectively.

Patient-Specific Adjustment

Age (β = 0.037, p < 0.001), BMI (β = 0.841, p = 0.007), height (β = 0.249, p = 0.008), and weight (β = -0.248, p = 0.025) significantly influenced 5R-STS performance in healthy individuals (Constant = -40.9, r^2 = 0.376). Gender (p = 0.14) and smoking status (p = 0.37) had no effect. The following formula can be used to make a simple prediction of a patients expected normal test time (t_a):

$$t_a = 0.03\ \text{Age} + 0.15\ \text{BMI} + 1.7$$

This estimate could be used as targeted 5R-STS performance after successful treatment (r^2 = 0.292).

Table 1. Baseline characteristics of the study and control groups.

	Study Group N = 157	**Control Group N = 80**	***P Value***
Male gender	80 (51)	43 (54)	0.68
Age [yrs.]	49.90 ± 14.10	43.03 ± 18.68	0.004
Height [cm]	175.62 ± 10.44	173.40 ± 9.29	0.096
Weight [kg]	78.50 ± 13.51	72.40 ± 13.58	0.001
Body Mass Index [kg/m^2]	25.38 ± 3.28	24.05 ± 4.25	0.016
Smoking Status			< 0.001
Active smoker	44 (28)	14 (18)	
Ceased smoking	59 (38)	17 (21)	
Never smoked	54 (34)	49 (61)	
Ability to work			
Full	50 (32)	76 (95)	< 0.001
Limited	33 (21)	2 (3)	
Unable	74 (47)	2 (3)	
Prior spine surgery	25 (16)	6 (8)	0.07
History of pain			< 0.001
None – 6 weeks	8 (5)	73 (91)	
6 weeks – 6 months	27 (17)	1 (1)	
6 months – 1 year	42 (27)	0 (0)	
> 1 year	80 (51)	6 (8)	
Analgesic drug use			< 0.001
Daily	120 (76)	10 (13)	
Weekly	15 (10)	7 (9)	
Not regularly	22 (14)	63 (79)	
Indication			-
Disc herniation	109 (69)	-	
Stenosis	32 (20)	-	
Spondylolisthesis	9 (6)	-	
DDD	5 (3)	-	
Synovial facet cyst	2 (1)		
Index Level			-
L2 – L3	4 (3)	-	
L3 – L4	21 (13)	-	
L4 – L5	64 (41)	-	
L5 – S1	68 (43)	-	
PROM			
RMDQ	11.65 ± 5.34	0.78 ± 1.45	< 0.001
ODI	43.04 ± 17.60	2.55 ± 7.39	< 0.001
VAS back pain	5.83 ± 2.76	0.95 ± 1.73	< 0.001
VAS leg pain	7.35 ± 1.97	0.40 ± 1.10	< 0.001
EQ-5D index	0.41 ± 0.32	0.94 ± 0.15	< 0.001
EQ-5D VAS	51.44 ± 18.98	86.58 ± 11.62	< 0.001

DDD = degenerative disc disease, ODI = Oswestry disability index, PROM = patient-reported outcome measure, RMDQ = Roland-Morris disability questionnaire, VAS = visual analogue scale

Continuous variables are presented as mean ± standard deviation, and categorical variables as frequency (percentage).

Table 2. Measures of test-retest reliability and measurement error.

	Overall N = 119		Study Group N = 66		Control Group N = 53	
	ICC	**95% CI**	**ICC**	**95% CI**	**ICC**	**95% CI**
Consistency	0.98	0.97-0.98	0.97	0.94 – 0.98	0.96	0.92 – 0.97
Absolute Agreement	0.98	0.97-0.98	0.97	0.94 – 0.98	0.96	0.92 – 0.97
SEM	1.03		1.47		0.34	

ICC = intraclass correlation coefficient, SEM = standard error of the means, 95% CI = 95% confidence interval

Discussion

In this prospective study, we demonstrated that the 5R-STS can objectively assess functional impairment in degenerative pathologies of the lumbar spine, with moderate convergent validity and excellent test-retest reliability. Based on our data, we propose an upper limit of normal of 10.4 seconds. This threshold distinguishes between patients with and without relevant functional impairment. The zone of indifference around this threshold value of 10.4 seconds ranged from 9.3 to 11.4 seconds. In this zone of indifference, patients cannot unambiguously be determined to have functional impairment or not. Rather, functional impairment in these "gray zone" patients should be judged according to their clinical history, or re-evaluated using a different objective functional test, e.g. the TUG test.[8] Furthermore, a severity stratification was put in place to help grade test results. Patients can be ranked as having either mild, moderate or severe functional impairment if their 5R-STS test times equate to 10.5 – 15.2, 15.3 – 22.0, or greater than 22.0 seconds, respectively. A simple calculation using age and BMI effectively estimates a patient's target 5R-STS performance.

The TUG test has been thoroughly studied in the context of degenerative spinal pathologies. Pre-and postoperative correlation, the minimum clinically important difference, and patient preference have been assessed, and a baseline severity index has been validated.[8,18,21,22] In this light, the TUG test currently represents the most clinically applicable option for objective functional testing in lumbar spine surgery. Accordingly, the 5R-STS test does not intend to compete with the TUG test, but rather to expand the spine surgeon's arsenal for objective functional testing. In some cases, such as when a patient's test result falls within the "gray zone" of the TUG test, re-evaluating the patient with the 5R-STS, or vice-versa, may be appropriate. Surgeons and patients may also prefer one particular test over others due to various reasons, ranging from personal inclinations to plain restrictions in space.

The 5R-STS exhibited solid correlation with the RMDQ and ODI as measures of functional impairment, VAS for back pain severity, and the EQ-5D index for HRQOL. Interestingly, there was no consistent correlation with leg pain on the VAS, and EQ-5D VAS. This indicates that the 5R-STS is not simply an objectification of pain and pain-related symptoms, which undeniably affect HRQOL, but rather a novel dimension in the assessment of patients, coined "objective functional impairment" (OFI). OFI correlates with, but is not identical to functional impairment as measured subjectively using PROM like the RMDQ and ODI. The excellent correlation among RMDQ and ODI (r = 0.63), as opposed to the good correlation of 5R-STS performance with RMDQ (0.49) or ODI (0.44), exemplifies this fact.

A patient able to perform the 5R-STS in 10.4 seconds can be considered to have no relevant OFI. Using this threshold value, around half of the patients in our study group was without relevant OFI, as found with the TUG test.[8] This is of particular interest since all included patients were candidates for surgery with long-standing pain symptoms and failed conservative treatment, demonstrating that OFI is distinctly

different from subjective functional impairment and pain, which are frequent indications for surgery. Using a severity stratification (Table 3), the severity of OFI can be graded as mild, moderate or severe.
Performance on the 5R-STS is affected by demographic factors such as age, height, weight, and BMI, but not by the patient's gender. It has also been shown that seat height confounds 5R-STS performance, underlining the importance of a standardized seat height.[23] In combination with measurement error, especially those test results that range close to cut-off values such as the ULN, must be interpreted cautiously. The emergence of further clinical data in the future will permit the establishment of patient-adjusted cut-offs to enhance interpretation.

Table 3. Severity stratification for the 5R-STS test. The patient's objective functional impairment can be graded according to their 5R-STS test time in seconds. The approximate prevalence of each grade is given.

5R-STS Severity Stratification		
Objective Functional Impairment	**Test Time**	**Prevalence**
No significant	≤ 10.4	50%
Mild	10.5 – 15.2	25%
Moderate	15.3 – 22.0	15%
Severe	> 22.0	10%

5R-STS = five-repetition sit-to-stand test

In the clinical routine, the 5R-STS is useful in multiple ways. It allows for a simple, quick and objective estimate of the patient's basic functioning by testing a physical activity that is commonly performed in daily life. As such, it also summarizes the state of the patient's relevant neuromusculature related to the lumbar spinal pathology. These functions are not only crucial to objectively assess to what extent patients are impaired before, but also after surgery. They enable the patient to resume activities of daily living, and might be an indicator of when patients can be discharged home safely. It has been demonstrated that these objective functional tests are sensitive to change after surgical treatment, which may also be the case for the 5R-STS.[21,22] Using simple formulae, age- and BMI-adjusted expected test times can be calculated for any patient to illustrate the current degree of objective functional impairment, and to create an expectation of what a positive outcome after surgery may look like. Minor advantages include the fact that the 5R-STS can easily be administered by physiotherapists and other health care personnel, and that these tests are not dependent on language, making them suitable for illiterate patients. It has been shown that objective functional tests are more robust against the influence of mental health status than PROM.[24,25] Moreover, patients prefer performing an objective functional test over completing a battery of PROM, perhaps because they provide a direct and tangible feedback, and are less time-consuming.[9] Considering these factors, the 5R-STS may constitute an excellent follow-up tool that could even be performed by patients themselves at home, either unsupervised or with televised supervision.[26] It is important to stress that objective functional tests should not replace PROM, since they convey different types of information. Rather, both should be used complementarily in the clinical setting

Limitations

The implications of this study may be limited by sample size. While the present sample size appeared to be sufficient to reach a power of 0.97 according to a post-hoc analysis (Appendix 1), the estimation of correlations relies on large samples, and thus further and larger studies are indicated to give more precise estimates of 5R-STS validity. Since we included patients with various degenerative spinal pathologies to achieve a broad assessment of the 5R-STS, we cannot make any claims towards the validity of this test for each of the specific indications (LDH, stenosis, spondylolisthesis, DDD, synovial cyst)[27]. However,

particularly for LDH, which comprised the majority of patients, the results of this study indicate that there is a strong correlation with the relevant outcome measures that merits further research with larger sample sizes for each specific indication. Confounders which may not have been captured in this study may also potentially influence performance. For example, patient motivation may play a big role in 5R-STS performance. It is to be noted that the control group was younger, and included health professionals. As such, our control group may be healthier than the normal population, skewing the ULN. However, most participants in the control group were the patients' partners, and exhibited demographics similar to their diseased counterparts. Lastly, we have not assessed interrater agreement of the 5R-STS, which might importantly influence reliability. While there is a need for a specific analysis of interrater agreement in a spinal population, Jones et al. and Mong et al. have demonstrated excellent interrater agreement for the 5R-STS in different populations.[13,14] Future studies should focus on validating the 5R-STS as a follow-up tool, on improved interpretation through the development of highly accurate adjustments, validating the proposed ULN and severity stratification. Of particular clinical relevance is to assess the prognostic value of the 5R-STS overall and in specific degenerative spinal pathologies. While the data presented in this study are derived from a relatively small sample, we believe that at present they are accurate enough to fundamentally interpret baseline 5R-STS results and to estimate OFI in a clinical setting.

Conclusions

The 5R-STS appears to be a valid and reliable measure of objective functional impairment. The relevant values for the interpretation of 5R-STS results were determined. Based on our data, we propose an upper limit of normal of 10.4 seconds, meaning that patients with a 5R-STS time equal to or lower than this threshold can be considered without functional impairment. A severity stratification was also proposed, indicating that patients with test times of 10.5 to 15.2, 15.3 to 22.0, and greater than 22.0 seconds can be considered to have mild, moderate, and severe objective functional impairment, respectively.

Appendix 1

Logarithmic Transformation of Raw Test Times

Raw five-repetition sit-to-stand (5R-STS) test times were right-skewed, and were logarithmically transformed for the correlation analyses in order to reduce skew. The transformation was done as follows:

$$x_t = log_{10}\,(x)$$

x_t = transformed test time, x = raw test time

Z Scores

Standard *Z* scores were calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

z = standard *z* score, *x* = observed 5R-STS test time, μ = mean 5R-STS test time in the normal population, σ = standard deviation of 5R-STS test times in the normal population

Upper Limit of Normal

The upper limit of normal (ULN) was calculated as described before: The ULN is calculated by constructing a one-sided 99% confidence interval using the critical z-value 2.33 as follows:

$$ULN = \mu + (2.33 \times \sigma)$$

ULN = upper limit of normal, μ = mean 5R-STS test time in the normal population, σ = standard deviation of 5R-STS test times in the normal population

Zone of Indifference

Because every test shows retest variability, simple thresholds should not be used to classify between functionally normal and functionally disabled patients. Rather, it is possible to account for this uncertainty by creating a zone of indifference (ZOI) around the ULN, not unlike a confidence interval. The ZOI was calculated by use of the standard error of the means and intraclass correlation coefficient of absolute agreement, according to Stratford and Goldsmith:[2]

$$ZOI = ULN \pm \sqrt{\sigma^2 \times (1 - ICC)}$$

ZOI = zone of indifference, ULN = upper limit of normal, σ = standard deviation of 5R-STS test times in the normal population, ICC = intraclass correlation coefficient of absolute agreement

Severity Stratification

A nonparametric severity stratification was constructed. This stratification helps to interpret test results by grouping the possible range of test results (observed test times on the 5R-STS test) into three grades according to the severity of objective functional impairment (OFI). A nonparametric method is of particular importance in objective functional tests, as the raw test times usually show considerable skewing.

The severity stratification was constructed as follows: OFI was classified into three grades: mild, moderate and severe. Only data from the study (disease) group were included for construction of the stratification. Raw test times were divided into percentiles. Patients with test times under or equal to the ULN were considered to be without OFI. In all other cases, patients with 5R-STS percentiles ≤ 75 were considered to have mild, those with percentiles > 75 and ≤ 90 were considered to have moderate, and those with percentiles > 90 were considered to have severe OFI.

Patient-Specific Adjustment

A simple multivariate linear regression model that predicts 5R-STS test times using age and BMI as independent variables was constructed. Data were taken from a representative population of 80 healthy individuals. This estimate could be used as targeted 5R-STS performance after successful treatment (r2 = 0.292).

$$t_a = 0.03\ \text{Age} + 0.15\ \text{BMI} + 1.7$$

t_a = adjusted expected 5R-STS test time

Post-Hoc Power Analysis

A power analysis was performed using the "pwr" package in R version 3.4.2 (The R Foundation for Statistical Computing, Vienna, Austria). With an attained sample size of n = 157, α = 0.05, and threshold for detection of a pearson correlation coefficient of r = 0.300, the estimated power 1 - β was calculated to be 0.97 for the reported values.

Acknowledgements

The authors are grateful to all participating volunteers, and to Femke Beusekamp, BSc and Nathalie Schouman for study coordination and data collection. We also thank Marlies P. de Wispelaere, MSc for her efforts in clinical informatics, and Anna M. Nikitin for her artwork.

Disclosures

Conflict of Interest: The authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Grants and Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Deyo RAM, Battie M, Beurskens AJHM, et al. Outcome Measures for Low Back Pain Research: A Proposal for Standardized Use. [Miscellaneous Article]. *Spine*. 1998;23(18):2003-2013.
2. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy*. 1980;66(8):271-273.
3. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine*. 1983;8(2):141-144.
4. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*. 2001;33(5):337-343.
5. Gautschi OP, Corniola MV, Schaller K, Smoll NR, Stienen MN. The need for an objective outcome measurement in spine surgery—the timed-up-and-go test. *Spine J*. 2014;14(10):2521-2522. doi:10.1016/j.spinee.2014.05.004
6. Guyatt GH, Sullivan MJ, Thompson PJ, et al. The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. *Can Med Assoc J*. 1985;132(8):919-923.
7. Mobbs RJ, Phan K, Maharaj M, Rao PJ. Physical Activity Measured with Accelerometer and Self-Rated Disability in Lumbar Spine Surgery: A Prospective Study. *Glob Spine J*. 2016;6(5):459-464. doi:10.1055/s-0035-1565259
8. Gautschi OP, Smoll NR, Corniola MV, et al. Validity and Reliability of a Measurement of Objective Functional Impairment in Lumbar Degenerative Disc Disease: The Timed Up and Go (TUG) Test. *Neurosurgery*. 2016;79(2):270-278. doi:10.1227/NEU.0000000000001195
9. Joswig H, Stienen MN, Smoll NR, et al. Patients' Preference of the Timed Up and Go Test or Patient-Reported Outcome Measures Before and After Surgery for Lumbar Degenerative Disk Disease. *World Neurosurg*. 2017;99:26-30. doi:10.1016/j.wneu.2016.11.039
10. Porchet F, Lattig F, Grob D, et al. Comparison of patient and surgeon ratings of outcome 12 months after spine surgery: presented at the 2009 Joint Spine Section Meeting. *J Neurosurg Spine*. 2010;12(5):447-455. doi:10.3171/2009.11.SPINE09526
11. Gvozdyev BV, Carreon LY, Graves CM, et al. Patient-reported outcome scores underestimate the impact of major complications in patients undergoing spine surgery for degenerative conditions. *J Neurosurg Spine*. 2017;27(4):397-402. doi:10.3171/2017.3.SPINE161400
12. Sánchez-Zuriaga D, López-Pascual J, Garrido-Jaén D, de Moya MFP, Prat-Pastor J. Reliability and validity of a new objective tool for low back pain functional assessment. *Spine*. 2011;36(16):1279-1288. doi:10.1097/BRS.0b013e3181f471d8
13. Jones SE, Kon SSC, Canavan JL, et al. The five-repetition sit-to-stand test as a functional outcome measure in COPD. *Thorax*. 2013;68(11):1015-1020. doi:10.1136/thoraxjnl-2013-203576
14. Mong Y, Teo TW, Ng SS. 5-Repetition Sit-to-Stand Test in Subjects With Chronic Stroke: Reliability and Validity. *Arch Phys Med Rehabil*. 2010;91(3):407-413. doi:10.1016/j.apmr.2009.10.030
15. Duncan RP, Leddy AL, Earhart GM. Five Times Sit to Stand Test Performance in Parkinson Disease. *Arch Phys Med Rehabil*. 2011;92(9):1431-1436. doi:10.1016/j.apmr.2011.04.008
16. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. https://www.R-project.org/
17. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther*. 1997;77(7):745-750.
18. Stienen MN, Smoll NR, Joswig H, et al. Validation of the baseline severity stratification of objective functional impairment in lumbar degenerative disc disease. *J Neurosurg Spine*. Published online March 3, 2017:1-7. doi:10.3171/2016.11.SPINE16683
19. Hoeymans N, Lindert H van, Westert GP. The Health Status of the Dutch Population as Assessed by the EQ-6D. *Qual Life Res*. 2005;14(3):655-663.

20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
21. Gautschi OP, Joswig H, Corniola MV, et al. Pre- and postoperative correlation of patient-reported outcome measures with standardized Timed Up and Go (TUG) test results in lumbar degenerative disc disease. *Acta Neurochir (Wien)*. 2016;158(10):1875-1881. doi:10.1007/s00701-016-2899-9
22. Gautschi OP, Stienen MN, Corniola MV, et al. Assessment of the Minimum Clinically Important Difference in the Timed Up and Go Test After Surgery for Lumbar Degenerative Disc Disease. *Neurosurgery*. Published online June 24, 2016. doi:10.1227/NEU.0000000000001320
23. Ng SSM, Cheung SY, Lai LSW, Liu ASL, Ieong SHI, Fong SSM. Association of seat height and arm position on the five times sit-to-stand test times of stroke survivors. *BioMed Res Int*. 2013;2013:642362. doi:10.1155/2013/642362
24. Carreon LY, Djurasovic M, Dimar JR, et al. Can the anxiety domain of EQ-5D and mental health items from SF-36 help predict outcomes after surgery for lumbar degenerative disorders? *J Neurosurg Spine*. 2016;25(3):352-356. doi:10.3171/2016.2.SPINE151472
25. Stienen MN, Smoll NR, Joswig H, et al. Influence of the mental health status on a new measure of objective functional impairment in lumbar degenerative disc disease. *Spine J*. 2017;17(6):807-813. doi:10.1016/j.spinee.2016.12.004
26. Ejupi A, Brodie M, Gschwind YJ, Lord SR, Zagler WL, Delbaere K. Kinect-Based Five-Times-Sit-to-Stand Test for Clinical and In-Home Assessment of Fall Risk in Older People. *Gerontology*. 2015;62(1):118-124. doi:10.1159/000381804
27. Crawford CH, Carreon LY, Bydon M, Asher AL, Glassman SD. Impact of preoperative diagnosis on patient satisfaction following lumbar spine surgery. *J Neurosurg Spine*. 2017;26(6):709-715. doi:10.3171/2016.11.SPINE16848

[Chapter 3]

Identifying clusters of objective functional impairment in patients with degenerative lumbar spinal disease using unsupervised learning

Victor E. Staartjes
Anita M. Klukowska
Vittorio Stumpo
W. Peter Vandertop
Marc L. Schröder

[submitted]

[Abstract]

Background Context

The five-repetition sit-to-stand (5R-STS) test was designed to capture objective functional impairment, and thus provides an adjunctive dimension in patient assessment. It is conceivable that there are different subsets of patients with objective functional impairment (OFI) and degenerative lumbar disease.

Purpose

We aim to identify clusters of OFI in objectively functionally impaired individuals based on 5R-STS and unsupervised machine learning methods.

Study Design/Setting

Analysis of data from two prospective cohort studies on the 5R-STS in a Dutch spine center.

Patient Sample

We included all patients with disc herniation, spinal stenosis, spondylolisthesis, or discogenic chronic low back pain and a 5R-STS test time of 10.5 seconds or greater – indicating the presence of OFI.

Outcome Measures

The 5R-STS, along with questionnaires for quality of life, pain severity, and subjective functional impairment.

Methods

K-means clustering – an unsupervised machine learning algorithm – was applied to identify clusters of OFI. Hallmarks of these clusters were then identified using descriptive and inferential statistical analyses.

Results

We included 173 patients (mean age [standard deviation]: 46.7 [12.7] years, 45% male), and identified three types of OFI. OFI Types 1 (57 pts., 32.9%), Type 2 (81 pts., 46.8%), and Type 3 (35 pts., 20.2%) exhibited mean 5R-STS test times of 14.0 (3.2), 14.5 (3.3), and 27.1 (4.4) seconds, respectively. The grades of OFI according to the validated Baseline Severity Stratification of the 5R-STS increased significantly with each Type of OFI, as did extreme anxiety and depression symptoms, issues with mobility and daily activities. Types 1 and 2 are characterized by mild to moderate OFI – with female gender, lower body mass index, and less smokers as hallmarks of Type I.

Conclusions

Unsupervised learning techniques identified three distinct clusters of patients with OFI that represent a more holistic clinical classification of patients with OFI than test times alone.

Introduction

The clinical assessment of patients suffering from back and leg pain due to lumbar degenerative disease has recently been supplemented by tests for objective functional impairment (OFI)[1–6] Tests that have been well-validated include the timed-up-and-go, 6-minute-walk, and five-repetition sit-to-stand (5R-STS) tests.[1,7,8] In addition to clinical examination and questionnaire measures for pain and subjective functional impairment, these tests have been shown to be robust to mental status as a confounder and add the ability to capture deficits and complications, such as foot drop or limping.[2,9] Patients also prefer objective tests over a battery of questionnaires to assess functional impairment.[10,11] When applied together with questionnaires for pain severity, subjective functional impairment and health-related quality of life, these tests provide a holistic capture of a patient's health state for scientific and clinical purposes.[12–14]

A 5R-STS test time of 10.5 seconds or greater has been shown to correspond to a diagnosis of OFI based on normative data.[11,2,13–15] Baseline severity stratifications have also been constructed, specifying cut-offs for mild, moderate, and severe OFI.[15,16] However, these cut-offs assume a similar performance among normative populations across all sociodemographic groups. In reality, older patients, those with higher BMI, active smokers, taller patients, and many other groups do worse on the 5R-STS. Cut-offs should be calculated from normative data across all of these groups, but the cut-offs should be flexible and adjustable to an individual's characteristics.

Achieving such "personalized" cut-offs for OFI can be achieved by calculating cut-offs for specific populations, e.g. cut-offs for >/<65 years of age and for male and female individuals.[1] However, this would result in a great number of different cut-offs that would be hard to implement in clinical practice. In the era of "personalized/precision medicine", a more elegant option is to predict the expected upper limit of normal (ULN) for individual patients, based on their sociodemographic characteristics, in order to diagnose OFI.[17,18] This works well for single cut-offs, e.g. for the binary presence or absence of OFI, based on normative data, but is not suitable for identifying mild, moderate, and severe impairment. These subgroups should instead be defined according to real-world data of patients with established OFI, and should reflect specific hallmarks of these subgroups. Unsupervised machine learning techniques, such as clustering, are well-suited for identifying clusters of observations that exhibit high similarity, without providing labels (e.g. "mild", "moderate", "severe").[19–22] Clusters defined by a machine learning algorithm would not be based on disease-specific parameters, and could then be used to classify new patients into relevant subsets that may also exhibit differences in treatment response. We aimed to identify clusters of OFI in objectively functionally impaired individuals based on 5R-STS and unsupervised machine learning methods.

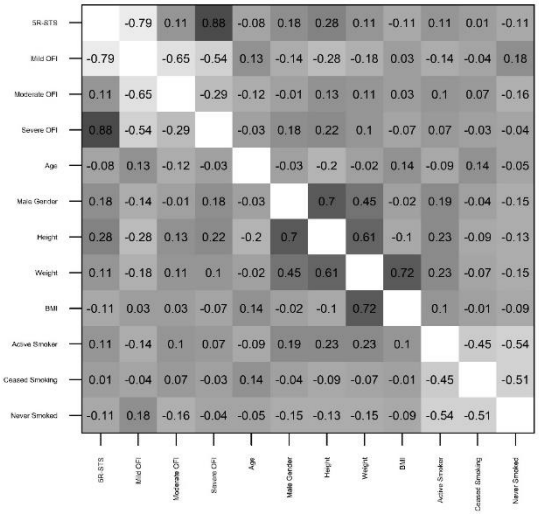


Figure 1. Correlation matrix for 5R-STS test time, baseline severity stratification (BSS), age, gender, height, weight, body mass index (BMI), and smoking status. Pearson's product-moment correlation is demonstrated.
5R-STS, five-repetition sit-to-stand test; OFI, objective functional impairment; BMI, body mass index;

Materials and Methods

Study Design

Pooled data from two prospective studies were used: ClinicalTrials.gov Identifiers: NCT03303300 and NCT03321357*)*.[1,23] Both studies were approved by the local institutional review board (Medical Research Ethics Committees United, Registration Numbers: W17.107 and W17.134), and were conducted according to the Declaration of Helsinki. Informed consent was obtained from all participants.

Patients scheduled for lumbar spine surgery for degenerative disease at a specialized short stay spine clinic were included between October 2017 and June 2018, and were assessed during outpatient consultations. Participating patients completed a variety of questionnaires, as well as the 5R-STS test. The pooled data from both studies was used to train an unsupervised machine learning model to automatically identify clusters of OFI. Subsequently, we compared the identified clusters to identify their hallmarks for further interpretation.

Inclusion and Exclusion Criteria

Inclusion criteria were the presence of lumbar disc herniation, lumbar spinal stenosis, spondylolisthesis, or discogenic chronic low back pain. Patients with synovial facet cysts causing radiculopathy, hip or knee prosthetics and those requiring walking aides, were excluded to eliminate these confounders. We also excluded all healthy volunteers, who were recruited in the control group. In addition, we excluded all patients without OFI (i.e. a 5R-STS test time of < 10.5 seconds, as defined by Staartjes et al.[1]) in order to cluster only those patients with established OFI.

Data Collection

The 5R-STS was performed according to the protocol described by Jones et al.[5] and Staartjes et al.[1] If the patient was unable to perform the test in 30 seconds, or not at all, this was noted and the test score was recorded as 30 seconds.[1] The baseline severity stratification for the 5R-STS, validated by Klukowska et al.[15], was used. Patients also filled in questionnaires containing baseline sociodemographic data including age, gender, smoking status, body mass index (BMI), prior spine surgery, indication and index level, history

of complaints, education, work type and ability, analgesia, symptom satisfaction, as well as numeric rating scales for back and leg pain severity, and validated Dutch versions of the Oswestry Disability Index (ODI), Roland-Morris Disability Questionnaire (RMDQ), and EuroQOL-5D-3L (EQ-5D) to capture subjective functional impairment, as well as HRQOL.[24–26] The EQ-5D included its single domains as well as the composite EQ-5D index and the EQ-5D thermometer on current subjective health status.[26] Participants filled out the questionnaires right after initially performing the test during outpatient consultation. For the EQ-5D, it has been established that the mood component of the EQ-5D correlates well with clinical depression.[27]

Analytical Methods

Data were reported as mean (standard deviation) for continuous and numbers (percentages) for categorical data. Variables with missingness over 25% were not included in the analysis. When data was assumed to be missing at random (MAR) or missing completely at random (MCAR), imputation was performed using a *k*-nearest neighbor (KNN) imputation with $k = 5$.[28] Pearson's product-moment correlation was applied to provide an overview of correlations within the dataset. One-way analysis of variance (ANOVA) or Pearson's Chi-Square tests were performed to test for differences among the identified clusters, for continuous and categorical variables, respectively. A $p \leq 0.05$ on two-sided tests was considered significant. Analyses were carried out using R version 4.0.3 (The R Foundation for Statistical Computing, Vienna, Austria).[29]

We chose *k*-means clustering to carry out unsupervised clustering of patients with OFI. The optimal number of clusters was chosen using the "elbow method" based on within-cluster sum of squares. Briefly, this method identifies the number of *k* clusters from which onwards the increase in similarity of observations within clusters becomes linear. The version of the *k*-means clustering algorithm described by Hartigan and Wong was used.[30] Pre-processing included centering and scaling (standardization), as well as one-hot encoding of categorical variables. The algorithm was run for a maximum number of iterations of 1000, with 100 initial configurations. Only the 5R-STS test time, 5R-STS baseline severity stratification, patient age, gender, height, weight, BMI, and smoking status were provided as inputs to the model, as sociodemographic variables unspecific to disease, as opposed to e.g. back pain severity or index level. A KNN algorithm with $k = 5$ was subsequently trained to classify new patients into the corresponding clusters.

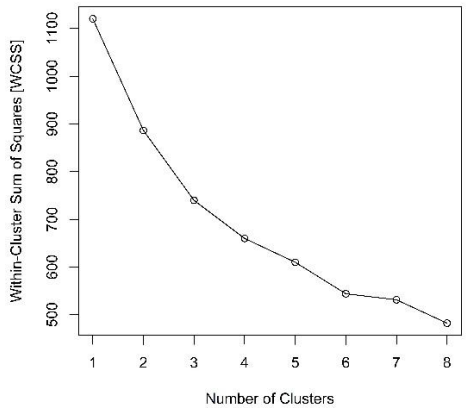


Figure 2. Plot of within-cluster sum of squares (WCSS) against number of clusters. The number of clusters at which the decrease in WCSS becomes linear ought to be chosen as the number of clusters for k-means clustering based on the "elbow" method.

Results

Patient Cohort

We included 173 patients with OFI fulfilling the inclusion criteria. Detailed characteristics are provided in **Table 1**. Data missingness was 3.5%. Mean age was 46.72 years (12.65), and 78 patients (45.1%) were male. According to the validated baseline severity stratification, 95 patients (54.9%) had mild, 45 (26.0) had moderate, and 33 (19.1%) had severe OFI. A correlation matrix of all variables included in the model is shown in **Figure 1**.

Clustering Analysis

A plot of Within-cluster sum of squares against the number of clusters **(Figure 2**) indicated that a number of clusters between 3 and 6 would constitute the optimal *k*, as this is the point from which onwards the similarity among observations within the clusters only increases marginally. For the analysis, *k* = 3 was chosen.

The three identified clusters (Types 1 to 3) contained 57 (32.9%), 81 (46.8%), and 35 (20.2%) patients, respectively. Within-cluster sum of squares values were 209, 363, and 167, respectively. The ratio of between-cluster sum of squares and total sum of squares was 34.1%.

Cluster Hallmarks

Clustered Variables

Table 2 provides an overview of the differences between the three clusters in terms of the variables that were included in the model. The clusters of impairment are illustrated in **Figure 3** for continuous variables and **Figure 4** for categorical variables. In terms of raw test times, Type 1 and Type 2 were comparable with mean test times between 14 and 15 seconds, while Type 3 demonstrated a mean test time of 27.1 (4.4) seconds. The distribution of mild, moderate, and severe OFI groups according to the validated 5R-STS baseline severity stratification increased steadily from Type 1 to Type 3.[15] Age was constant across all clusters. When comparing Type 1 and Type 2 OFI, the rate of smokers and males was significantly lower in Type 1, as were mean BMI and body height.

Unclustered Variables

To further characterize types of OFI, those variables not included in the clustering analysis ought to be analyzed (**Table 3**). There were marked differences in all EQ-5D domains, as well as the EQ-5D index and EQ-5D thermometer and the ODI and RMDQ. Specifically, the rate of patients with extreme anxiety and depression increased steadily from 3.5% in Type 1, 7.4% in Type 2, to 14.3% in Type 3, with statistical significance. In addition, mobility and ability to perform activities of daily life (ADL) was reduced in Type 3, with corresponding increases in subjective functional impairment scores (ODI, RMDQ).

The proportion of patients who had undergone prior spine surgery increased steadily from Type 1 with 14.0% to Type 3 with 28.6%, although this progression was not statistically significant. There were no differences in back or leg pain severity in the three identified clusters. Similarly, indications for surgery, history of complaints, index levels, education, work type and ability, analgesic medication use, and satisfaction also remained constant across all three clusters. A qualitative overview of the hallmarks of each type is provided in **Table 4**.

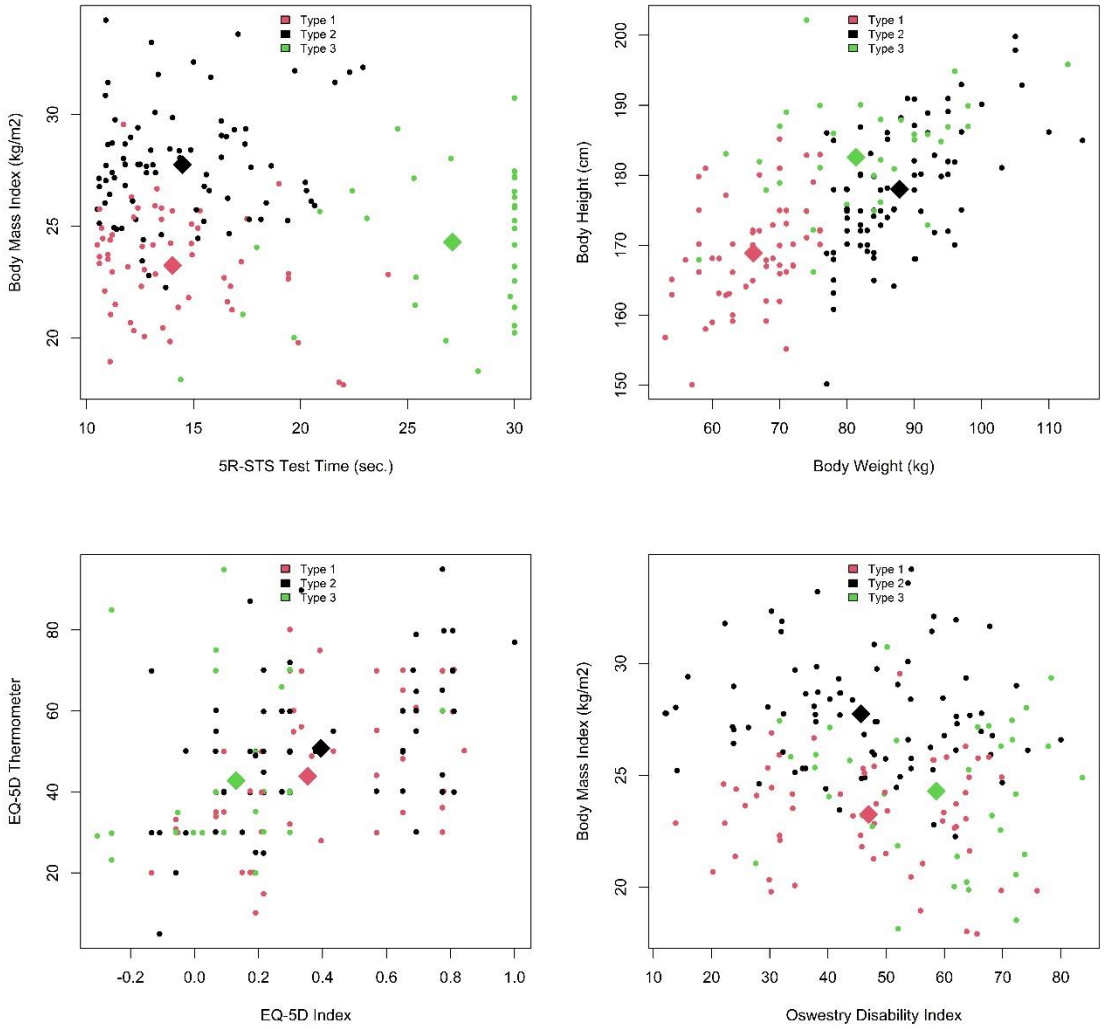


Figure 3. Scatterplots demonstrating the hallmarks of the three different clusters (Type 1 to 3) of objective functional impairment (OFI) in terms of continuous variables.
EQ-5D, EuroQOL five-dimensions questionnaire

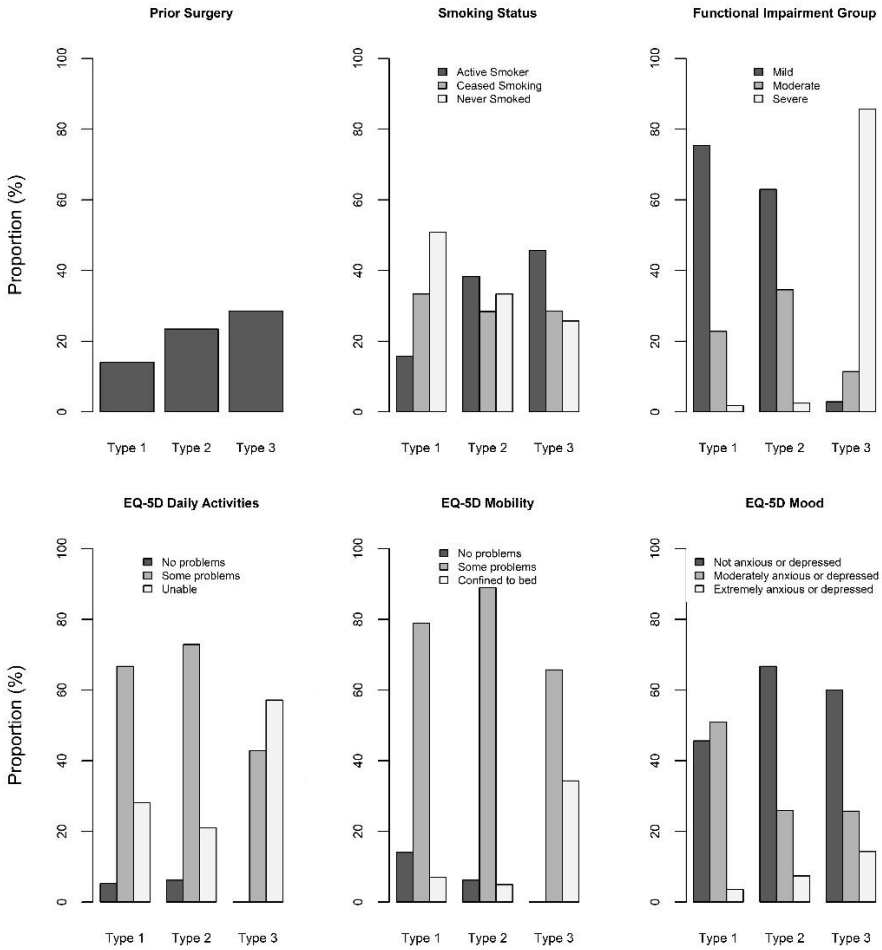


Figure 4. Boxplots demonstrating the hallmarks of the three different clusters (Type 1 to 3) of objective functional impairment (OFI) in terms of categorical variables.
EQ-5D, EuroQOL five-dimensions questionnaire

Discussion

Three characteristic clusters of patients with OFI were identified through unsupervised analysis. The clusters were termed Type 1, 2 and 3, and roughly correspond to mild, moderate, and severe impairment (**Table 4**).

Type 1 OFI was present in around a third of patients, and was characterized by a relatively rapid performance of the 5R-STS, and was only seldomly associated with problems in performing ADL, mobility, and clinical depression – indicating mild impairment. This is also supported by the low levels of subjective functional impairment found in these patients. As mentioned in the results section, concerning demographics, the vast majority of Type 1 patients were female nonsmokers, with a low BMI. The female gender also explains the lower average height in this group. It has been argued that female patients have a higher pain tolerance than male patients, and are likely to also present later for surgical treatment for degenerative spinal conditions.[31–34] This could partially explain that this largely female group experiences low subjective and OFI. The low incidence of active smoking demonstrably has no effect on 5R-STS performance [1] – and for that matter also not on other short-duration objective functional tests [35]. This likely indicates that smoking, while not a significant predictor of 5R-STS performance, was picked up by the clustering algorithm as a confounder associated with other, possibly psychosocial factors that in turn influence performance.

Type 2 OFI occurred in half of our cohort, and was linked with overweight in both genders, although test times were slightly elevated compared to Type 1. This indicates mild impairment, also corroborated by mild subjective functional impairment. As stated in the results, the incidence of extreme anxiety and depression symptoms was over twice as high as in Type 1, and statistically significantly so. The rate of smokers corresponded to that of our patient cohorts and indeed the Dutch population.[36] In addition, both genders were equally represented in this cluster. Type 2 likely indicates low levels of true functional impairment, but with a higher susceptibility for mood changes due to the mild or moderate impairment that is present.

In contrast to the mild/moderate levels of OFI observed in Types 1 and 2, Type 3 indicated extreme impairment with sequalae such as bedriddenness, high subjective functional impairment, mobility issues, and high rates of discomfort. Overall, patients with Type 3 impairment were of average BMI, mostly of male gender, and exhibited a significantly higher rate of active smoking. We also observed a doubling of the rate of extreme depression and anxiety compared to Type 2, and a quadrupling compared to Type 1.

Still, levels of pain were comparable among the three clusters, with the exception of the EQ-5D "pain and discomfort" domain, which also includes discomfort. Back and leg pain severity did not differ among the three clusters, demonstrating that the clusters represent true subgroups of impairment (including objective and subjective impairment), and are not influenced by pain severity as such. This is similar for sociodemographic factors, which could be assumed to influence the perception of impairment, such as level of education, work type, work ability, and age.

Up to now, grading of OFI was based on a fixed cut-off of 10.5 seconds on the 5R-STS test, though, realistically, obese and elderly, but otherwise healthy, individuals cannot be expected to perform equally well as younger individuals with a BMI in the normal range.[1] Ideally, an otherwise healthy, but obese, 75-year-old person and a 22-year-old athlete should not have their level of impairment rated by the same static cut-off. As one potential solution, Gautschi et al.[2] calculated a range of cut-offs for patients of certain gender and ages for the timed-up-and-go test, but clinical implementation of a larger amount of cut-offs that need to be remembered is cumbersome. Machine learning-based methods have the potential to suggest personalized "expected" cut-offs for each individual patient, based on socio-demographics, as has been alluded to in the initial validation of the 5R-STS in the spinal population.[1,37,38] Once a personalized cut-off has been established for the binary presence or absence of OFI, some form of impairment grading, that again takes into account socio-demographics, should be carried out, which the clustering algorithm developed in this study can do. Furthermore, machine learning methods in combination with motion tracking-based 5R-STS assessment[39] could lead to more intuitive and automated integration of objective functional testing in clinical practice. In the future, it may become possible to immediately calculate OFI, in contrast to other technological advances such as robotics, imaging, or neuronavigation, as algorithms can run server-side and even be applied on mobile devices, applications are far-reaching even in rural areas where patients cannot easily travel for in-person appointments.[40]

Limitations

Although we used prospectively collected data exclusively, this presents only single-center data. Therefore, generalizability of our findings and specifically of the identified clusters of OFI require further external validation before making the model available (e.g. in a web-app) and applying it in clinical practice elsewhere. However, the data that were used (preoperative socio-demographic parameters and 5R-STS testing) are not center-specific (such as e.g. surgical treatment or length of stay), and the 5R-STS has been established as having extremely high inter-rater reliability.[1,23] Inclusion of further parameters from patient history and clinical examination could possibly increase the distinctness of the clusters even further.

However, this would come at the cost of clinical usability and parsimony of the algorithm and derived classification of OFI. Currently, only variables that are easily and objectively assessable such as age, BMI, and gender are included in the model, which enables clinical application in under one minute. Although we included a comparatively large and homogenous cohort of patients with OFI, a larger number of patients would likely also lead to an increase in generalizability and distinctness of the clusters. Lastly, the model was not tested in separate diagnoses such as chronic low back pain and spondylolisthesis due to lacking statistical power for such subgroup analyses. However, the classification of OFI based on our model is independent of diagnosis (i.e. it is not a factor considered in this cluster analysis), and in addition, our analysis of unclustered parameters demonstrated that there is no interaction between diagnosis and cluster assignment, indicating robustness against different diagnostic categories.

Conclusions

In this study, we demonstrate that unsupervised machine learning techniques, in combination with the 5R-STS, identified three distinct clusters of patients with OFI that represent a more holistic and objective clinical classification of patients than test times and baseline severity stratifications alone. These findings may in the future be integrated with higher levels of automation into clinical practice, and may then also have diagnostic, prognostic, and predictive implications for surgical and nonsurgical treatment of degenerative spinal conditions.

Table 1. Baseline characteristics of the overall patient cohort.

Parameter	Value (N = 173)
5R-STS Test Time, mean (SD)	16.88 (6.24)
Functional Impairment Group, n (%)	
Mild (10.5 – 15.2 sec.)	95 (54.9)
Moderate (15.3 - 22.0 sec.)	45 (26.0)
Severe (> 22.0 sec.)	33 (19.1)
Age, mean (SD)	46.72 (12.65)
Male gender, n (%)	78 (45.1)
Height, mean (SD)	175.91 (9.81)
Weight, mean (SD)	79.36 (12.88)
Body Mass Index, mean (SD)	25.57 (3.32)
Smoking Status, n (%)	
Active Smoker	56 (32.4)
Ceased	52 (30.1)
Never Smoked	65 (37.6)
Prior Surgery, n (%)	37 (21.4)
Indication for Surgery, n (%)	
Lumbar Disc Herniation	127 (73.4)
Stenosis	29 (16.8)
Spondylolisthesis	7 (4.0)
Chronic Low Back Pain	10 (5.8)
History of Complaints, n (%)	
< 6 wks.	7 (4.0)
6 wks. – 3 months	29 (16.8)
6 months – 1 year	49 (28.3)
> 1 year	88 (50.9)
Index Level, n (%)	
L2-L3	6 (3.5)
L3-L4	10 (5.8)

L4-L5	75 (43.4)
L5-S1	82 (47.4)
Highest Level of Education, n (%)	
Elementary School	3 (1.7)
High School	80 (46.2)
Higher Education	84 (48.6)
(Post-)Doctoral	6 (3.5)
Type of Work, n (%)	
Student	1 (0.6)
Houseworker	7 (4.0)
Employed	106 (61.3)
Self-employed	28 (16.2)
On benefits	7 (4.0)
Retired	14 (8.1)
Jobless	10 (5.8)
Analgesic Medication, n (%)	
Not regularly	21 (12.1)
At least weekly	17 (9.8)
Daily	135 (78.0)
Satisfied with Current Symptoms, n (%)	
Yes	2 (1.2)
Neutral	3 (1.7)
No	168 (97.1)
Ability to Work, n (%)	
Fully able	42 (24.3)
Limited	28 (16.2)
Unable	103 (59.5)
EQ-5D Mobility, n (%)	
No problems	13 (7.5)
Some problems	140 (80.9)
Confined to bed	20 (11.6)
EQ-5D Selfcare, n (%)	
No problems	83 (48.0)
Some problems	88 (50.9)
Unable	2 (1.2)
EQ-5D Daily Activities, n (%)	
No problems	8 (4.6)
Some problems	112 (64.7)
Unable	53 (30.6)
EQ-5D Pain, n (%)	
No pain or discomfort	5 (2.9)
Moderate pain or discomfort	54 (31.2)
Extreme pain or discomfort	114 (65.9)
EQ-5D Mood, n (%)	
Not anxious or depressed	101 (58.4)
Moderately anxious or depressed	59 (34.1)
Extremely anxious or depressed	13 (7.5)
EQ-5D Index, mean (SD)	0.33 (0.30)
EQ-5D Thermometer, mean (SD)	46.89 (17.51)
NRS Back Pain, mean (SD)	6.57 (2.36)
NRS Leg Pain, mean (SD)	7.57 (1.82)
Oswestry Disability Index, mean (SD)	48.73 (16.42)
Roland-Morris Disability Questionnaire, mean (SD)	13.67 (5.05)

5R-STS, five repetition sit-to-stand test; SD, standard deviation; EQ-5D, EuroQOL five-dimensions questionnaire; NRS, Numeric Rating Scale;

Table 2. Comparative analysis of the three types of objective functional impairment identified in the clustering analysis by means of those variables included in the clustering analysis (Clustered Parameters).

Parameter	Type 1 Impairment	Type 2 Impairment	Type 3 Impairment	P
N, (%)	*57 (32.9)*	*81 (46.8)*	*35 (20.2)*	
Clustered Parameters				
5R-STS Test Time, mean (SD)	14.02 (3.23)	14.48 (3.29)	27.09 (4.42)	<0.001*
Age, mean (SD)	46.47 (12.63)	48.55 (12.98)	42.86 (11.27)	0.082
Male gender, n (%)	9 (15.8)	44 (54.3)	25 (71.4)	<0.001*
Height, mean (SD)	168.86 (7.48)	178.01 (8.89)	182.54 (8.35)	<0.001*
Weight, mean (SD)	66.11 (6.30)	87.82 (8.23)	81.34 (12.05)	<0.001*
Body Mass Index, mean (SD)	23.25 (2.36)	27.75 (2.50)	24.29 (3.14)	<0.001*
Smoking Status, n (%)				0.015*
Active Smoker	9 (15.8)	31 (38.3)	16 (45.7)	
Ceased	19 (33.3)	23 (28.4)	10 (28.6)	
Never smoked	29 (50.9)	27 (33.3)	9 (25.7)	
Functional Impairment Group, n (%)				<0.001*
Mild (10.5 – 15.2 sec.)	43 (75.4)	51 (63.0)	1 (2.9)	
Moderate (15.3 - 22.0 sec.)	13 (22.8)	28 (34.6)	4 (11.4)	
Severe (> 22.0 sec.)	1 (1.8)	2 (2.5)	30 (85.7)	

5R-STS, five repetition sit-to-stand test; SD, standard deviation;
** $p \leq 0.05$*

Table 3. Comparative analysis of the three types of objective functional impairment identified in the clustering analysis by means of the variables that were not considered within the clustering analysis (Unclustered Parameters).

Parameter	Type 1 Impairment	Type 2 Impairment	Type 3 Impairment	P
N, (%)	*57 (32.9)*	*81 (46.8)*	*35 (20.2)*	
Unclustered Parameters				
Prior Surgery, n (%)	8 (14.0)	19 (23.5)	10 (28.6)	0.211
Indication for Surgery, n (%)				0.153
Lumbar Disc Herniation	42 (73.7)	58 (71.6)	27 (77.1)	
Stenosis	10 (17.5)	17 (21.0)	2 (5.7)	
Spondylolisthesis	3 (5.3)	3 (3.7)	1 (2.9)	
Chronic Low Back Pain	2 (3.5)	3 (3.7)	5 (14.3)	
History of Complaints, n (%)				0.714
< 6 wks.	2 (3.5)	4 (4.9)	1 (2.9)	
6 wks. – 3 months	9 (15.8)	13 (16.0)	7 (20.0)	
6 months – 1 year	21 (36.8)	19 (23.5)	9 (25.7)	
> 1 year	25 (43.9)	45 (55.6)	18 (51.4)	
Index Level, n (%)				0.964
L2-L3	1 (1.8)	4 (4.9)	1 (2.9)	
L3-L4	3 (5.3)	5 (6.2)	2 (5.7)	
L4-L5	24 (42.1)	36 (44.4)	15 (42.9)	
L5-S1	29 (50.9)	36 (44.4)	17 (48.6)	
Highest Level of Education, n (%)				0.649
Elementary School	0 (0.0)	3 (3.7)	0 (0.0)	
High School	29 (50.9)	36 (44.4)	15 (42.9)	
Higher Education	26 (45.6)	39 (48.1)	19 (54.3)	
(Post-)Doctoral	2 (3.5)	3 (3.7)	1 (2.9)	
Type of Work, n (%)				0.179
Student	1 (1.8)	0 (0.0)	0 (0.0)	
Houseworker	5 (8.8)	2 (2.5)	0 (0.0)	
Employed	34 (59.6)	47 (58.0)	25 (71.4)	
Self-employed	7 (12.3)	16 (19.8)	5 (14.3)	
On benefits	1 (1.8)	3 (3.7)	3 (8.6)	

Retired	7 (12.3)	6 (7.4)	1 (2.9)	
Jobless	2 (3.5)	7 (8.6)	1 (2.9)	
Analgesic Medication, n (%)				0.452
Not regularly	9 (15.8)	9 (11.1)	3 (8.6)	
At least weekly	3 (5.3)	11 (13.6)	3 (8.6)	
Daily	45 (78.9)	61 (75.3)	29 (82.9)	
Satisfied with Current Symptoms, n (%)				0.248
Yes	1 (1.8)	0 (0.0)	1 (2.9)	
Neutral	0 (0.0)	3 (3.7)	0 (0.0)	
No	56 (98.2)	78 (96.3)	34 (97.1)	
Ability to Work, n (%)				0.467
Fully able	17 (29.8)	20 (24.7)	5 (14.3)	
Limited	8 (14.0)	12 (14.8)	8 (22.9)	
Unable	32 (56.1)	49 (60.5)	22 (62.9)	
EQ-5D Mobility, n (%)				<0.001*
No problems	8 (14.0)	5 (6.2)	0 (0.0)	
Some problems	45 (78.9)	72 (88.9)	23 (65.7)	
Confined to bed	4 (7.0)	4 (4.9)	12 (34.3)	
EQ-5D Selfcare, n (%)				0.010*
No problems	29 (50.9)	44 (54.3)	10 (28.6)	
Some problems	28 (49.1)	37 (45.7)	23 (65.7)	
Unable	0 (0.0)	0 (0.0)	2 (5.7)	
EQ-5D Daily Activities, n (%)				0.003*
No problems	3 (5.3)	5 (6.2)	0 (0.0)	
Some problems	38 (66.7)	59 (72.8)	15 (42.9)	
Unable	16 (28.1)	17 (21.0)	20 (57.1)	
EQ-5D Pain, n (%)				0.003*
No pain or discomfort	2 (3.5)	3 (3.7)	0 (0.0)	
Moderate pain or discomfort	22 (38.6)	30 (37.0)	2 (5.7)	
Extreme pain or discomfort	33 (57.9)	48 (59.3)	33 (94.3)	
EQ-5D Mood, n (%)				0.012*
Not anxious or depressed	26 (45.6)	54 (66.7)	21 (60.0)	
Moderately anxious or depressed	29 (50.9)	21 (25.9)	9 (25.7)	
Extremely anxious or depressed	2 (3.5)	6 (7.4)	5 (14.3)	
EQ-5D Index, mean (SD)	0.35 (0.27)	0.40 (0.30)	0.13 (0.23)	<0.001*
EQ-5D Thermometer, mean (SD)	43.89 (16.79)	50.77 (16.77)	42.80 (18.91)	0.022*
NRS Back Pain, mean (SD)	6.54 (2.67)	6.36 (2.16)	7.11 (2.21)	0.284
NRS Leg Pain, mean (SD)	7.86 (1.41)	7.35 (1.82)	7.63 (2.34)	0.261
Oswestry Disability Index, mean (SD)	46.98 (15.90)	45.70 (15.81)	58.57 (15.27)	<0.001*
Roland-Morris Disability Questionnaire, mean (SD)	13.07 (5.49)	12.89 (4.86)	16.46 (3.74)	0.001*

SD, standard deviation; EQ-5D, EuroQOL five-dimensions questionnaire; NRS, Numeric Rating Scale;
** $p \leq 0.05$*

Table 4. Qualitative overview of the hallmarks of the three types of impairment that were identified through unsupervised analysis.

Domain	Type 1 Impairment	Type 2 Impairment	Type 3 Impairment
5R-STS Test Time	↓	↓	↑
Body Mass Index	↓	↑	↔
Gender	♀	♀	♂
Smoking	↓	↔	↑
Subjective Functional Impairment	↔	↔	↑
Depression & Anxiety	↔	↑	↑↑
Mobility, ADL	↔	↔	↓↓
Age	↔	↔	↔
Pain	↔	↔	↔
History of Complaints	↔	↔	↔
Work Status & Type	↔	↔	↔

5R-STS, five-repetition sit-to-stand test; ADL, activities of daily life

Acknowledgements

The authors are grateful to all participating patients and to Femke Beusekamp, BSc and Nathalie Schouman for study coordination and data collection. We also thank Marlies P. de Wispelaere, MSc for her efforts in clinical informatics.

Disclosures

Conflict of Interest: The authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Grants and Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Staartjes VE, Schröder ML. The five-repetition sit-to-stand test: evaluation of a simple and objective tool for the assessment of degenerative pathologies of the lumbar spine. *J Neurosurg Spine*. 2018;29(4):380-387. doi:10.3171/2018.2.SPINE171416
2. Gautschi OP, Smoll NR, Corniola MV, et al. Validity and Reliability of a Measurement of Objective Functional Impairment in Lumbar Degenerative Disc Disease: The Timed Up and Go (TUG) Test. *Neurosurgery*. 2016;79(2):270-278. doi:10.1227/NEU.0000000000001195
3. Gautschi OP, Corniola MV, Schaller K, Smoll NR, Stienen MN. The need for an objective outcome measurement in spine surgery—the timed-up-and-go test. *Spine J*. 2014;14(10):2521-2522. doi:10.1016/j.spinee.2014.05.004
4. Stienen MN, Ho AL, Staartjes VE, et al. Objective measures of functional impairment for degenerative diseases of the lumbar spine: a systematic review of the literature. *Spine J*. 2019;19(7):1276-1293. doi:10.1016/j.spinee.2019.02.014
5. Jones SE, Kon SSC, Canavan JL, et al. The five-repetition sit-to-stand test as a functional outcome measure in COPD. *Thorax*. 2013;68(11):1015-1020. doi:10.1136/thoraxjnl-2013-203576
6. Crook S, Büsching G, Schultz K, et al. A multicentre validation of the 1-min sit-to-stand test in patients with COPD. *Eur Respir J*. 2017;49(3):1601871. doi:10.1183/13993003.01871-2016
7. Gautschi OP, Joswig H, Corniola MV, et al. Pre- and postoperative correlation of patient-reported outcome measures with standardized Timed Up and Go (TUG) test results in lumbar degenerative disc disease. *Acta Neurochir (Wien)*. 2016;158(10):1875-1881. doi:10.1007/s00701-016-2899-9
8. Guyatt GH, Sullivan MJ, Thompson PJ, et al. The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. *Can Med Assoc J*. 1985;132(8):919-923.
9. Stienen MN, Smoll NR, Joswig H, et al. Influence of the mental health status on a new measure of objective functional impairment in lumbar degenerative disc disease. *Spine J*. 2017;17(6):807-813. doi:10.1016/j.spinee.2016.12.004
10. Joswig H, Stienen MN, Smoll NR, et al. Patients' Preference of the Timed Up and Go Test or Patient-Reported Outcome Measures Before and After Surgery for Lumbar Degenerative Disk Disease. *World Neurosurg*. 2017;99:26-30. doi:10.1016/j.wneu.2016.11.039
11. Sosnova M, Zeitlberger AM, Ziga M, et al. Patients undergoing surgery for lumbar degenerative spinal disorders favor smartphone-based objective self-assessment over paper-based patient-reported outcome measures. *Spine J Off J North Am Spine Soc*. Published online December 17, 2020. doi:10.1016/j.spinee.2020.11.013
12. Jakobsson M, Gutke A, Mokkink LB, Smeets R, Lundberg M. Level of Evidence for Reliability, Validity, and Responsiveness of Physical Capacity Tasks Designed to Assess Functioning in Patients With Low Back Pain: A Systematic Review Using the COSMIN Standards. *Phys Ther*. 2019;99(4):457-477. doi:10.1093/ptj/pzy159
13. Stienen MN, Maldaner N, Joswig H, et al. Objective functional assessment using the "Timed Up and Go" test in patients with lumbar spinal stenosis. *Neurosurg Focus*. 2019;46(5):E4. doi:10.3171/2019.2.FOCUS18618
14. Tosic L, Goldberger E, Maldaner N, et al. Normative data of a smartphone app-based 6-minute walking test, test-retest reliability, and content validity with patient-reported outcome measures. *J Neurosurg Spine*. Published online May 29, 2020:1-10. doi:10.3171/2020.3.SPINE2084
15. Klukowska AM, Schröder ML, Stienen MN, Staartjes VE. Objective functional impairment in lumbar degenerative disease: concurrent validity of the baseline severity stratification for the five-repetition sit-to-stand test. *J Neurosurg Spine*. 2020;33(1):4-11. doi:10.3171/2019.12.SPINE191124

16. Stienen MN, Smoll NR, Joswig H, et al. Validation of the baseline severity stratification of objective functional impairment in lumbar degenerative disc disease. *J Neurosurg Spine*. 2017;26(5):598-604. doi:10.3171/2016.11.SPINE16683
17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
18. Snyderman R. Personalized health care: from theory to practice. *Biotechnol J*. 2012;7(8):973-979. doi:10.1002/biot.201100297
19. Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Front Comput Neurosci*. 2019;13. doi:10.3389/fncom.2019.00031
20. Miller DJ, Wang Y, Kesidis G. Emergent unsupervised clustering paradigms with potential application to bioinformatics. *Front Biosci J Virtual Libr*. 2008;13:677-690. doi:10.2741/2711
21. Ficiarà E, Boschi S, Ansari S, et al. Machine Learning Profiling of Alzheimer's Disease Patients Based on Current Cerebrospinal Fluid Markers and Iron Content in Biofluids. *Front Aging Neurosci*. 2021;13. doi:10.3389/fnagi.2021.607858
22. Ames CP, Smith JS, Pellisé F, et al. Artificial Intelligence Based Hierarchical Clustering of Patient Types and Intervention Categories in Adult Spinal Deformity Surgery: Towards a New Classification Scheme that Predicts Quality and Value. *Spine*. 2019;44(13):915-926. doi:10.1097/BRS.0000000000002974
23. Staartjes VE, Beusekamp F, Schröder ML. Can objective functional impairment in lumbar degenerative disease be reliably assessed at home using the five-repetition sit-to-stand test? A prospective study. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc*. Published online January 24, 2019. doi:10.1007/s00586-019-05897-3
24. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy*. 1980;66(8):271-273.
25. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine*. 1983;8(2):141-144.
26. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*. 2001;33(5):337-343.
27. Peasgood T, Brazier J, Papaioannou D. A systematic review of the validity and responsiveness of EQ-5D and SF-6D for depression and anxiety. *HEDS Discuss Pap 1215*. Published online 2012. Accessed November 15, 2020. http://eprints.whiterose.ac.uk/74659/
28. Templ M, Kowarik A, Alfons A, Prantner B. *VIM: Visualization and Imputation of Missing Values*.; 2019. Accessed January 5, 2020. https://CRAN.R-project.org/package=VIM
29. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. https://www.R-project.org/
30. Hartigan JA, Wong MA. A k-means clustering algorithm. *JSTOR Appl Stat*. 1979;28(1):100-108.
31. Siccoli A, Staartjes VE, de Wispelaere MP, Schröder ML. Gender differences in degenerative spine surgery: Do female patients really fare worse? *Eur Spine J*. Published online August 21, 2018. doi:10.1007/s00586-018-5737-3
32. Kim H-J, Suh B-G, Lee D-B, et al. Gender difference of symptom severity in lumbar spinal stenosis: role of pain sensitivity. *Pain Physician*. 2013;16(6):E715-723.
33. Racine M, Tousignant-laflamme Y, Kloda LA, Dion D, Dupuis G, Choinière M. A systematic literature review of 10 years of research on sex/gender and pain perception – Part 2: Do biopsychosocial factors alter pain sensitivity differently in women and men? *Pain*. 2012;153(3):619-635. doi:10.1016/j.pain.2011.11.026
34. Pochon L, Kleinstück FS, Porchet F, Mannion AF. Influence of gender on patient-oriented outcomes in spine surgery. *Eur Spine J*. 2016;25(1):235-246. doi:10.1007/s00586-015-4062-3

35. Joswig H, Stienen MN, Smoll NR, et al. Effects of Smoking on Subjective and Objective Measures of Pain Intensity, Functional Impairment, and Health-Related Quality of Life in Lumbar Degenerative Disk Disease. *World Neurosurg*. 2017;99:6-13. doi:10.1016/j.wneu.2016.11.060
36. Brathwaite R, Smeeth L, Addo J, et al. Ethnic differences in current smoking and former smoking in the Netherlands and the contribution of socioeconomic factors: a cross-sectional analysis of the HELIUS study. *BMJ Open*. 2017;7(7):e016041. doi:10.1136/bmjopen-2017-016041
37. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181
38. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med*. 2019;380(14):1347-1358. doi:10.1056/NEJMra1814259
39. Ejupi A, Brodie M, Gschwind YJ, Lord SR, Zagler WL, Delbaere K. Kinect-Based Five-Times-Sit-to-Stand Test for Clinical and In-Home Assessment of Fall Risk in Older People. *Gerontology*. 2015;62(1):118-124. doi:10.1159/000381804
40. Munakomi S. Letter to the Editor. Reappraising role of clinical evaluations in degenerative lumbar spine pathologies. *J Neurosurg Spine*. 2019;30(6):860-861. doi:10.3171/2018.10.SPINE181282

[Chapter 4]

Machine learning-augmented objective functional testing in the degenerative spine: Quantifying impairment using patient-specific five-repetition sit-to-stand assessment

Victor E. Staartjes
Anita M. Klukowska
Moira Vieli
Christiaan H. B. van Niftrik
Martin N. Stienen
Carlo Serra
Luca Regli
W. Peter Vandertop
Marc L. Schröder

Published in: *Neurosurg Focus. 2021 Nov;51(5):E8. [in press]*

[Abstract]

Objective

What is considered "abnormal" in clinical testing is normally defined by simple thresholds derived from normative data. For instance, when testing using the five-repetition sit-to-stand (5R-STS) test, the upper limit of normal (ULN) from a population of healthy volunteers (10.5 seconds) is used to identify objective functional impairment (OFI) – This fails to consider different properties of e.g. taller and shorter or older and younger individuals. We developed a personalized testing strategy to quantify patient-specific OFI using machine learning.

Methods

We included patients with disc herniation, spinal stenosis, spondylolisthesis, or discogenic chronic low back pain, and a population of healthy volunteers from two prospective studies. A machine learning model was trained on normative data to predict personalized "expected" test times and their confidence intervals (CIs) and ULNs (99th percentiles) based on simple demographics. OFI was defined as a test time greater than the personalized ULN. OFI was categorized into Types 1 to 3 based on a clustering algorithm. A web-app (https://neurosurgery.shinyapps.io/5RSTS/) was developed to deploy the model clinically.

Results

We included 288 patients and 129 healthy individuals. Our model predicted "expected" test times with a mean absolute error of 1.18 (95% CI: 1.13-1.21) seconds and R2 of 0.37 (95% CI: 0.34-0.41). Based on our personalized testing strategy, 191 patients (66.3%) exhibited OFI. Of these, 64 (33.5%), 91 (47.6%), and 36 (18.8%) were recognized as Type 1, 2 and 3, respectively. Increasing detected levels of OFI were associated with statistically significant increases in subjective functional impairment, extreme anxiety & depression symptoms, bedriddenness, extreme pain or discomfort, inability to carry out activities of daily life, and limited ability to work.

Conclusion

In the era of "precision medicine", simple population-based thresholds may eventually not be adequate anymore to monitor quality and safety in neurosurgery. Individualized assessment integrating machine learning techniques provides more detailed and objective clinical assessment. The personalized testing strategy demonstrated concurrent validity with measures of quality of life, and the freely accessible web-app enables clinical application.

Introduction

Standardized outcome assessment has evolved from radiological and physician-rated outcomes towards patient-reported outcome measures – not only in clinical practice, but importantly also in quality and safety improvement programs and in scientific research.[1–5] Accurate capture of clinical outcomes is a necessary step toward monitoring trends in neurosurgical quality and safety improvement programs, e.g. to detect trends or spikes in poor outcomes and infection or complication rates. In addition, standardized outcome measurement enables setting benchmarks for surgical quality among individual centers and surgeons, assessment of the efficacy of new interventions, checklists, and protocols, and identification of systematic human errors.[3,6]

Up to now, both patient-reported and objective outcome measures have relied on single, fixed thresholds derived from normal populations to distinguish between healthy and unhealthy individuals, or between a good and a bad outcome. For example, in degenerative lumbar spine disease, the presence of objective functional impairment (OFI) is normally determined by comparing the 5R-STS test time of a particular patient with the upper limit of normal (ULN) of test times in a spine-healthy population (10.5 seconds).[7–12] If the patient takes longer than these 10.5 seconds to complete the 5R-STS, OFI can be diagnosed and further classified based on fixed thresholds.[8,12] Advantages of such thresholds are their simplicity, generalizability, ease of derivation and validation, and simple anchoring to a representative normal population. There are however inherent disadvantages: Differences in test properties among individuals become obvious when considering the example of body height – One of the most powerful determinants of 5R-STS performance, as tall patients need to cover a longer distance standing up and sitting down from a chair with standardized height.[8,13,14]

Instead of fixed thresholds, dynamic thresholds that respect a patients' demographics could allow for a more accurate grading of OFI. Some developments in this direction have been made, such as the introduction of tables reporting fixed grading thresholds for e.g. male and female or younger and older than 65 years.[15,16] However, memorizing a range of fixed thresholds makes clinical application cumbersome. A still more detailed and more personalized testing strategy could improve upon fixed thresholds by enabling grading of disease tailored to a particular patient, instead of groups or subgroups. The future of medicine is moving towards ever more personalized healthcare analytics in the era of "personalized" or "precision medicine".[17] We aim to implement this rationale by developing a machine learning-based personalized testing strategy to quantify impairment using patient-specific five-repetition sit-to-stand assessment.

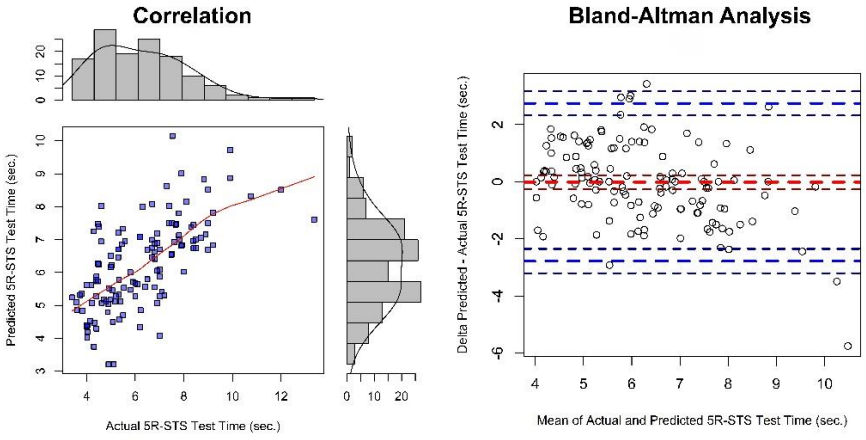


Figure 1. Performance of the quantile regression model. The actual 5R-STS performance of the healthy population (N = 129) is compared to the corresponding predictions (tau = 0.50, 50th percentile). Correlation was 0.61 (95% CI: 0.58 to 0.64). Bland-Altman analysis revealed a mean bias of −0.02 sec., with a 95% limit of agreement of −2.77 to 2.74 sec.

Materials and Methods

Study Design

To train and validate the patient-specific objective functional testing model, data from two prospective studies including both patients and healthy individuals were pooled.[8,9] Between October 2017 and June 2018, patients were seen at a specialized outpatient spine surgery clinic.

We trained a machine learning model to predict a personalized "expected" or "normal" test time from basic demographic data including age, height, weight, body mass index (BMI), gender, and smoking status. This individually predicted 5R-STS test time can be used as a "benchmark" of the performance that a patient would be expected to achieve without disease, or in case of full recovery after e.g. surgery for lumbar disc herniation.[8]

Subsequently, individualized thresholds such as the personalized ULN can be calculated, representing the 99th percentile of the 5R-STS test time that would be expected among individuals with the same demographics in the normative population. If patients can perform the 5R-STS within their personalized ULN, presence of OFI can be ruled out. Instead, if patients perform more slowly than this personalized ULN, the presence of OFI can be diagnosed, and the type of OFI can then be assessed using a clustering method *[V.E. Staartjes et al., unpublished data]*: This method applies unsupervised clustering using a *k*-means matching algorithm and classifies patients with OFI into three clinically distinct "OFI Types". Type I and II represent relatively mild to moderate impairment, with Type II additionally representing a higher likelihood of extreme anxiety and depression symptoms, bedriddenness and inability to work. Type III OFI corresponds to severe impairment that is associated with an even higher magnitude of the aforementioned accompanying symptoms.

This report was compiled according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.[18]

Ethical Approval

The two prospective studies (ClinicalTrials.gov Identifiers: NCT03303300 and NCT03321357*)* were approved by the local institutional review board (Medical Research Ethics Committees United, Registration Numbers: W17.107 and W17.134). Informed consent was obtained from all participants.

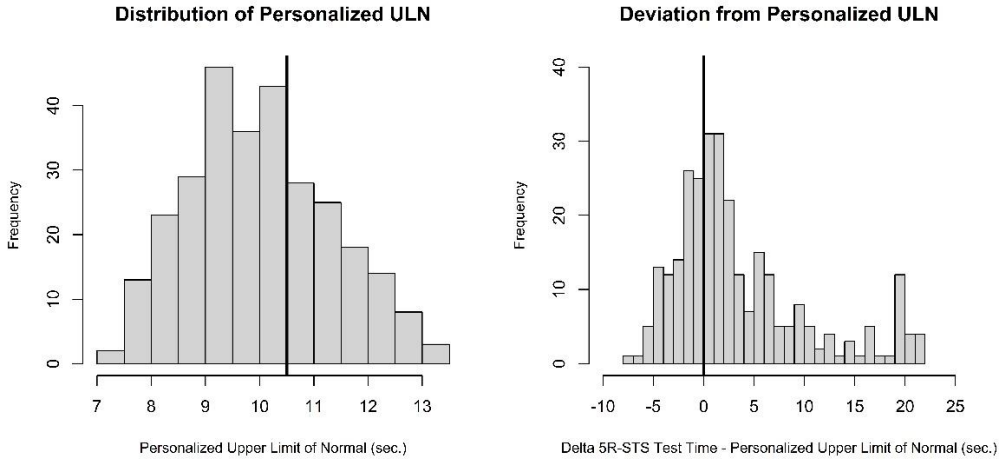


Figure 2. Histograms of the personalized ULNs generated for the entire patient cohort (n = 288) as well as the personalized performance of the patient cohort, expressed as the deviation of the actual test time from each patient's personalized ULN.

Study Population

All enrolled patients were scheduled for surgery and were assessed during outpatient consultations. Inclusion criteria were the presence of lumbar disc herniation, lumbar spinal stenosis, spondylolisthesis, or discogenic chronic low back pain. Patients with synovial facet cysts causing radiculopathy were not included. Patients with hip or knee prosthetics, and those requiring walking aides were excluded to eliminate these confounders. Individuals with missing 5R-STS data were excluded. We also included spine-healthy individuals as a normative reference population, most of whom were partners of the patients with similar demographics, employees of the department, or other volunteers.

Measurements and Data Collection

The 5R-STS was performed according to a previously published testing protocol.[8,9,19] Most importantly, an armless, hard-seated chair of standard height (48 cm) was firmly placed against a wall, stable shoes were worn, and patients were instructed and motivated to perform the test "as fast as possible". The 5 repetitions were timed from the "go" command to the completed fifth stand (5R-STS test time). If the patient was unable to perform the test in 30 seconds, or not at all, this was noted and the test score was recorded as 30 seconds.[8] Some patients and healthy individuals performed the test twice, in which case the mean test time was used.

A range of questionnaires were additionally used. Patients provided information on baseline sociodemographic data, as well as numeric rating scales for back and leg pain severity, and validated Dutch versions of the Oswestry Disability Index (ODI), Roland-Morris Disability Questionnaire (RMDQ), and EuroQOL-5D-3L (EQ-5D) to capture subjective functional impairment as well as health-related quality of life.

Statistical Analysis

Analyses were carried out using R version 4.0.5 (The R Foundation for Statistical Computing, Vienna, Austria).[20] A $p \leq 0.05$ on two-tailed tests was considered statistically significant. Data were reported as mean ± standard deviation for continuous and numbers (percentages) for categorical data. Variables or patients with missingness over 25% were excluded from the analysis. Missing data that were assumed to be missing (completely) at random were imputed using *k*-nearest neighbor (KNN) imputation, with $k = 5$.[21] Baseline characteristics of study and control group individuals were compared using Pearson's χ^2 tests or Welch's two-sample *t*-test. Patients without OFI and those with the three types of OFI were compared using Pearson's χ^2 tests or one-way analysis of variance (ANOVA).

Model Development

To predict personalized "expected" 5R-STS test times along with their 95% confidence intervals and ULN (99th percentile), a quantile regression model with a least absolute shrinkage and selection operator (Lasso) penalty was trained for the 2.5th, 50th, 97.5th, and 99th quantiles (tau).[22,23] This machine learning algorithm was trained on data from a representative cohort of healthy individuals of all ages. The model was internally validated using repeated 5-fold cross-validation with 10 repeats to assess out-of-sample performance. Resampled root mean square error (RMSE), mean absolute error (MAE), and R^2, along with their 95% confidence intervals (CI) were obtained using 1000 repetitions of a bootstrap with replacement. Agreement of predicted and actual test times in the normative population was further evaluated using Bland-Altman analysis.[24]

If patients were able to perform the 5R-STS within their personalized ULN (actual 5R-STS test time ≤ personalized ULN), OFI was ruled out. Whenever OFI was diagnosed (actual 5R-STS test time > personalized ULN), we applied a clustering algorithm *[V.E. Staartjes et al., unpublished data]* to identify OFI Types 1 to 3.

A web app allowing for measurement of the 5R-STS and automatizing the prediction and clustering process was constructed. The calculations run server-side, and the web app can thus easily be applied on mobile devices, too.

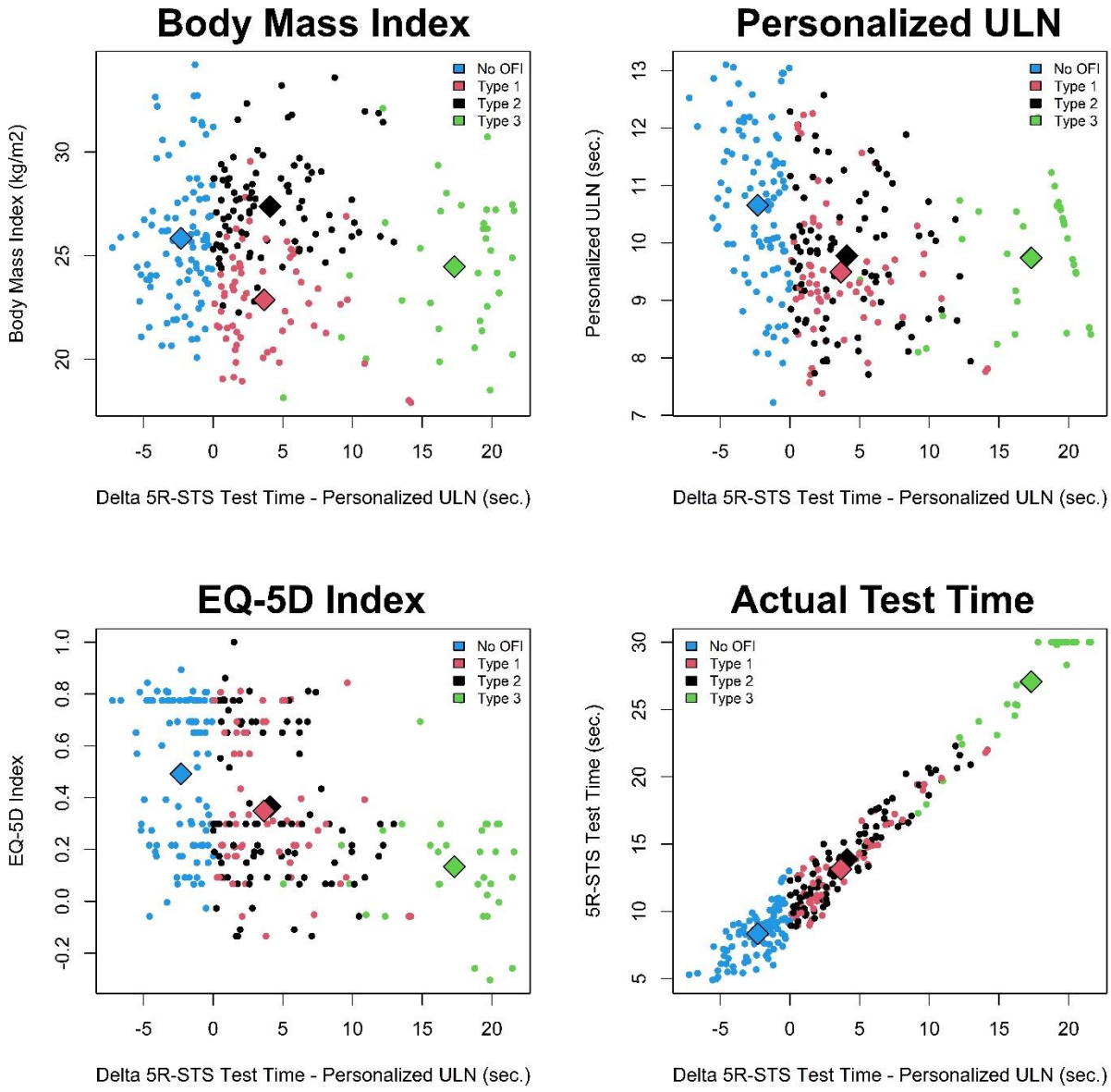


Figure 3. Scatterplots demonstrating clusters of functional impairment among the patient cohort (n = 288) in terms of selected continuous variables.

Results

Cohorts

Detailed characteristics of the normative and patient cohorts are provided in **Table 1**. Among the healthy population of 129 volunteers, 167 of 3354 (5.0%) data fields were missing. Similarly, among the 288 patients included, 215 of 7488 (2.9%) data fields were missing. On average, healthy individuals were aged 40 ± 19 years, and patients were aged 47 ± 13 years old ($p < 0.001$). Sixty healthy individuals (47%) and

141 patients (49%) were male (p = 0.722). Mean 5R-STS test time recorded in the normative population was 6.3 ± 1.8 seconds, while mean test time was 13.5 ± 6.4 seconds among patients (p < 0.001).

Personalized Test Time Quantiles

Expected Test Times

To assess model fit at internal validation, we compared actual test times and predicted (tau = 0.50, 50th percentile) test times during cross-validation (**Table 2**). In terms of classical performance measures, RMSE was 1.48 (95% CI: 1.43 to 1.53), MAE was 1.18 (95% CI: 1.13 to 1.21), and R^2 was 0.37 (95% CI: 0.34 to 0.41). Correspondingly, the correlation R of actual and predicted test times was 0.61 (95% CI: 0.58 to 0.64). Bland-Altman analysis (**Figure 1**) revealed a mean bias of −0.02 sec., with a 95% limit of agreement of −2.77 to 2.74 sec.

Personalized Upper Limit of Normal (ULN)

The average personalized ULN – derived through prediction of the 99th percentile of the expected test time – for the entire patient population was 10.0 ± 1.3 sec. and ranged from 7.2 to 13.1 sec (**Figure 2**).

In-Silico Application of Personalized Testing Strategy

Test Performance

All 288 patients were run through the web app to evaluate the results of the personalized testing strategy. Average 5R-STS test time was 13.5 ± 6.4, ranging from 4.9 to 30.0 seconds. On average, the deviation of actual test time from a particular patient's personalized ULN (**Figure 2**) was 3.5 ± 6.7 (range: -7.2 to 21.6 sec.), leading to a diagnosis of OFI in 191 patients (66.3%).

Cluster Assignment

Among the 191 patients with OFI, 64 (34%) were recognized as Type 1 OFI, while 91 (48%) and 36 (19%) patients were recognized as Type 2 and Type 3 OFI, respectively.

Test Interpretation

Table 3 demonstrates the final classification of all 288 patients using the machine learning-augmented testing strategy. Subjective functional impairment (ODI, RMDQ) increased with severity of OFI, as did rates of extreme anxiety & depression symptoms, bedriddenness, extreme pain or discomfort, and inability to carry out activities of daily life (all p ≤ 0.003). Limited ability or inability to work also increased steadily with OFI severity (p = 0.012). Analgesic drug use was similar among all classifications (p = 0.499).

Back pain severity correlated with severity of OFI (p < 0.001) while leg pain did not (p = 0.173). Chronic low back pain was by far the most common among patients with Type 3 OFI, while patients without OFI had a significantly higher rate of lumbar spinal stenosis (p < 0.001).

Average age was significantly higher among patients without OFI, while there were no significant differences in age among the three types of OFI (p < 0.001). Average BMIs (around 25 kg/m^2) were observed in patients without OFI and those with Type 3 OFI, while Type 1 and Type 2 OFI are clearly demographically distinguished by normal-weight and overweight patients, respectively (**Figure 3**). The rate of active smokers increased steadily with severity of OFI (p = 0.028).

Deployment

A web app containing detailed testing instructions and providing capabilities for testing (either measuring the 5R-STS test time using an integrated stopwatch or entering a previously measured test time), automated generation of personalized "expected" test time as well as personalized ULN, and automated

interpretation (Presence and Type of OFI) was constructed. Results of five exemplary patients from our cohort are presented in **Table 4**. The web app is freely available at https://neurosurgery.shinyapps.io/5RSTS/.

Discussion

Using data from two prospective cohort studies, we have developed and internally validated a personalized testing strategy based on machine learning. Based on age, gender, height, weight, BMI, and smoking status, precise predictions of personalized "expected" test times and their ULNs can be generated for each patient. Patients requiring longer to complete the 5R-STS than their personalized ULN are deemed to be objectively functionally impaired. The extent of OFI can then be further classified using a clustering process. All steps of the testing process have been implemented in a freely accessible web app.

What is considered "abnormal" in clinical testing is usually defined by simple thresholds derived from normative data.[25] For instance, when testing using the five-repetition sit-to-stand (5R-STS) test, the ULN from a population of healthy volunteers (10.5 seconds) is used to identify OFI.[8] This approach is simple and effective, yet it fails to consider the radically different 5R-STS testing properties of different individuals. For instance, body height is known to influence 5R-STS performance significantly.[8,13,14] Since chairs of standardized height are used, the distance that needs to be covered with each sit-to-stand action is proportional to body height. Thus, a tall individual with the same health status as a comparable shorter individual will usually still require significantly longer to complete the 5R-STS. Apart from such obvious differences in testing properties, what is considered "normal" should optimally be based on a normative population that is as similar to the test subject as possible. One would expect a completely healthy 21-year-old rugby player to perform the 5R-STS more quickly than an otherwise healthy 78-year-old obese retiree, although both performances could be seen as "normal" for their specific situations. For this reason, "upper limits of normal" should be derived from individuals without functional impairment of different age ranges, nutritional status, et cetera. Of course, one could simply calculate multiple ULNs for younger and older, normal-weight and obese, male and female, or tall and short individuals. This would require generating an exponentially growing number of different thresholds for each subset, eventually also running into sample size limitations. Memorization and clinical application would also be increasingly cumbersome. A more elegant and detailed way of arriving at a personalized threshold for each patient is to model the effects of the most important demographic parameters for different quantiles of the normative population. Some machine learning methods such as quantile regression enable this approach, and are able to generate precise ULNs for each individual.[22,23]

Our model demonstrated its capacity to predict personalized "expected" test times (50th percentile) with an accuracy of within 1.2 seconds of the actual test time, as well as predicting individualized ULNs (99th percentile).[8] When defining the presence of OFI as an actual 5R-STS performance that is slower than the personalized ULN, we observed that a slightly higher percentage of around two thirds of the spinal patient population are deemed to be impaired. This compares to approximately 50% to 60% of the spinal patient population that are deemed to be objectively functionally impaired using the standard ULN of 10.5 seconds.[8,12] Those patients that were additionally classified as having OFI by our personalized testing strategy – and not by the usual fixed 10.5 second cut-off – were mostly younger and shorter patients who would indeed realistically be expected to complete the test in 7 or 8 seconds. Conversely, some very tall patients that normally would initially have been classified as impaired were now deemed not to have OFI, because a test time of e.g. 13 seconds is still considered "normal", given their height. Hence, one can argue that using personalized cut-offs for objective tests of function seems to increase the diagnostic yield of these tests, which is of obvious value for both clinical care and research.

Whenever OFI was diagnosed, it was classified as OFI Types 1, 2, or 3 using a clustering algorithm. As discussed previously, the three groups roughly correspond to different levels of impairment, with Type 3 indicating severe OFI. Types 1 and 2 often show similar levels of impairment – especially when considering 5R-STS test time only – but Type 2 carries slightly higher likelihood of extreme anxiety and depression symptoms, bedriddenness and inability to work. In addition, Type 2 patients virtually all are overweight individuals (BMI ≥ 25 kg/m^2) and on average taller and more likely to be male and actively smoking than Type 1. These differences may underline the practical applicability of this grading versus just looking at the 5R-STS test time alone: Type 1 and 2 patients exhibit the same test times and report virtually the same level of symptoms, yet Type 1 patients appear to be slightly less troubled by their symptoms than patients with Type 2 OFI.

Concurrent validity of an outcome measurement or classification is assessed by comparing a certain measurement of interest to other relevant parameters that one would expect to differ between the levels of that measurement.[26] Our personalized testing strategy demonstrated that multiple relevant anchors of health-related quality of life change steadily from "no OFI" to "OFI Type 3", indicating concurrent validity. For instance, increasing levels of OFI were associated with increases in subjective functional impairment, extreme anxiety & depression symptoms, bedriddenness, extreme pain or discomfort, inability to carry out activities of daily life, and limited ability to work. Differences were particularly pronounced between patients classified as being without impairment vs. with Type 1/2 OFI, and between patients with Type 1/2 vs. Type 3 OFI. It is also known that low back pain can lead to relatively more impairment in activities of daily life than radiculopathy, particularly when performing the 5R-STS.[27–30] Correspondingly, back pain severity increased with each level of OFI, while leg pain severity was not affected.[12]

As machine learning methods become more broadly adopted in many fields of medicine[17,31–33], it is feasible that clinical and scientific patient assessment – including laboratory studies, radiological studies, and physical examination – will move from simple fixed thresholds (e.g. a ULN for D-dimer of < 250 ng/mL[25]) to personalized cut-offs based on comparable individuals from a normative population – such as e.g. with age-adjusted D-dimers.[34] We also expect that integration of other machine learning techniques will enable even more automated testing: The 5R-STS could be automatically rated using machine vision or accelerometers for motion tracking[35], and demographic data about a particular patient could be pulled from electronic health records.[36] At an even higher level of abstraction, OFI could potentially be graded based on how patients walk into the examination room and sit down or get up from their chair. Nonetheless, the applications of personalized cut-offs and other extremely personalized measures in actual clinical practice and in quality and safety improvement – apart from their applications in research – are currently few and far between, and there is not yet enough evidence to support their adoption as standard of care. Even if clear prognostic subgroups can be defined and outcome measurements become more granular and specific, it does not necessarily follow that this would lead to any real-world benefits to patients.

Limitations

Data originated from two prospective studies, but were collected at a single Dutch center. Although we collected data from a normative population of all age ranges, the models developed on Dutch individuals may not necessarily generalize to other populations. However, the data that were used (demographics such as age, gender, and BMI as well as 5R-STS testing) are not center-specific. Furthermore, the 5R-STS has demonstrably high inter-rater reliability.[9,10] An external validation study would enable a definite statement on generalizability. Similarly, although out-of-sample error was properly assessed using cross validation in this study, a prospective validation study would provide further evidence on the out-of-

sample performance (overfitting) of the quantile regression model. Patients with hip or knee prosthetics and those requiring walking aides were not enrolled, and other comorbidities such hip or knee arthritis and non-spinal neuropathies (e.g. diabetic polyneuropathy) were not systematically assessed. It is plausible that such comorbidities may skew 5R-STS performance towards higher test times. We could have included further input parameters into the quantile regression model to make its predictions even more accurate, but this would have come at the cost of ease-of-use. Perhaps even more importantly, the predictive value of OFI and its classification on outcomes after surgery must also be assessed. Lastly, we have not validated the personalized testing strategy in specific subgroups such as lumbar disc herniation or lumbar spinal stenosis, but it hence serves as a general model for frequent degenerative lumbar spine conditions. Both the prediction of personalized ULN and the clustering algorithm are independent of diagnosis or other clinical characteristics.

Conclusions

In the era of "precision medicine", simple thresholds or even multiple thresholds for certain demographic subgroups – which may be hard to implement clinically – may eventually not be adequate anymore to monitor quality and safety in neurosurgery. Individualized assessment integrating machine learning techniques provides more detailed and objective clinical assessment. We have developed and internally validated a method for generation of personalized reference ranges for the 5R-STS that allows for patient-specific quantification of impairment. If impairment is present, it can be further classified using a clustering algorithm. The personalized testing strategy demonstrated concurrent validity with measures of quality of life. A freely accessible web-app (https://neurosurgery.shinyapps.io/5RSTS/) enables clinical application of this personalized testing strategy.

Table 1. Baseline characteristics of the healthy individuals and patients with lumbar degenerative disease, pooled from two prospective studies.

Parameter	Healthy Individuals (N = 129)	Patients (N = 288)	P Value
5R-STS test time [sec], mean ± SD	6.27 (1.84)	13.50 (6.44)	<0.001*
Age [yrs.], mean ± SD	40.48 (18.80)	47.12 (13.38)	<0.001*
Male gender, n (%)	60 (46.5)	141 (49.0)	0.722
Height [cm], mean ± SD	171.90 (9.99)	175.83 (10.14)	<0.001*
Weight [kg], mean ± SD	71.06 (13.93)	78.40 (13.67)	<0.001*
Body Mass Index [kg/m2], mean ± SD	24.01 (4.04)	25.27 (3.34)	0.001*
Smoking status, n (%)			<0.001*
Active smoker	19 (14.7)	81 (28.1)	
Ceased smoking	27 (20.9)	88 (30.6)	
Never smoked	83 (64.3)	119 (41.3)	
Prior spine surgery, n (%)	7 (5.4)	55 (19.1)	0.001*
Indication for surgery, n (%)			0.001*
Disc herniation	-	201 (69.8)	
Spinal stenosis	-	57 (19.8)	
Spondylolisthesis	-	15 (5.2)	
Discogenic chronic low back pain	-	15 (5.2)	
Highest level of education, n (%)			0.225
Elementary school	4 (3.1)	4 (1.4)	
High school	44 (34.1)	122 (42.4)	
Higher education	77 (59.7)	149 (51.7)	
(Post-)doctoral	4 (3.1)	13 (4.5)	
Analgesic drug use, n (%)			<0.001*
Not regularly	108 (83.7)	50 (17.4)	
Weekly	9 (7.0)	26 (9.0)	
Daily	12 (9.3)	212 (73.6)	
Ability to work, n (%)			<0.001*
Full	122 (94.6)	76 (26.4)	
Limited	5 (3.9)	64 (22.2)	
Unable	2 (1.6)	148 (51.4)	
EQ-5D index, mean ± SD	0.95 (0.14)	0.38 (0.30)	<0.001*
EQ-5D thermometer, mean ± SD	84.78 (12.37)	49.46 (17.81)	<0.001*
NRS back pain severity, mean ± SD	0.96 (1.82)	5.95 (2.64)	<0.001*
NRS leg pain severity, mean ± SD	0.52 (1.36)	7.47 (1.88)	<0.001*
Oswestry Disability Index, mean ± SD	2.53 (6.72)	45.12 (17.02)	<0.001*
Roland-Morris Disability Questionnaire, mean ± SD	0.64 (1.86)	12.06 (5.35)	<0.001*

Data are presented after imputation for missing data. Continuous variables are presented as mean (standard deviation), and categorical variables as frequency (percentage).

SD, standard deviation; DDD, degenerative disc disease; ODI, Oswestry disability index; RMDQ, Roland-Morris disability questionnaire; VAS, visual analogue scale;

*p ≤ 0.05

Table 2. Performance measures of the quantile regression model during repeated five-fold cross validation. The actual 5R-STS performance of the healthy population (N = 129) is compared to the corresponding predictions of the expected median test time (tau = 0.50, 50th percentile).

Performance Measure	Cross-Validation Performance (95% CI)
Root Mean Square Error (RMSE)	1.48 (1.43 to 1.53)
Mean Absolute Error (MAE)	1.18 (1.13 to 1.21)
R^2	0.37 (0.34 to 0.41)
Bland-Altman Analysis	**Result [sec.]**
Mean bias [sec.]	-0.02
95% limit of agreement	−2.77 to 2.74

Table 3. Personalized classification of all 288 patients according to personalized upper limits of normal and cluster assignment.

Parameter	No OFI (N = 97)	Type 1 OFI (N = 64)	Type 2 OFI (N = 91)	Type 3 OFI (N = 36)	P Value
Upper Limit of Normal (ULN), mean ± SD	10.66 (1.37)	9.50 (1.15)	9.78 (1.19)	9.74 (1.03)	<0.001*
5R-STS test time [sec], mean ± SD	8.35 (1.78)	13.15 (3.19)	13.86 (3.48)	27.07 (4.32)	<0.001*
Age [yrs.], mean ± SD	53.58 (14.14)	42.85 (12.36)	44.51 (11.62)	43.88 (10.95)	<0.001*
Male gender, n (%)	56 (57.7)	13 (20.3)	47 (51.6)	25 (69.4)	<0.001*
Height [cm], mean ± SD	175.71 (10.74)	169.86 (7.83)	178.92 (8.14)	181.92 (8.66)	<0.001*
Weight [kg], mean ± SD	79.81 (12.90)	65.74 (6.08)	87.48 (8.74)	80.97 (12.38)	<0.001*
Body Mass Index [kg/m2], mean ± SD	25.83 (3.16)	22.86 (2.41)	27.36 (2.35)	24.47 (3.38)	<0.001*
Smoking status, n (%)					0.028*
Active smoker	16 (16.5)	16 (25.0)	34 (37.4)	15 (41.7)	
Ceased smoking	36 (37.1)	20 (31.2)	23 (25.3)	9 (25.0)	
Never smoked	45 (46.4)	28 (43.8)	34 (37.4)	12 (33.3)	
Prior spine surgery, n (%)	12 (12.4)	11 (17.2)	22 (24.2)	10 (27.8)	0.099
Indication for surgery, n (%)					<0.001*
Disc herniation	52 (53.6)	49 (76.6)	72 (79.1)	28 (77.8)	
Spinal stenosis	33 (34.0)	11 (17.2)	11 (12.1)	2 (5.6)	
Spondylolisthesis	7 (7.2)	2 (3.1)	5 (5.5)	1 (2.8)	
Discogenic chronic low back pain	5 (5.2)	2 (3.1)	3 (3.3)	5 (13.9)	
History of symptoms, n (%)					0.749
6 weeks or less	2 (2.1)	2 (3.1)	5 (5.5)	1 (2.8)	
6 weeks to 6 months	14 (14.4)	9 (14.1)	14 (15.4)	7 (19.4)	
6 months to 1 year	21 (21.6)	21 (32.8)	20 (22.0)	10 (27.8)	
Over 1 year	60 (61.9)	32 (50.0)	52 (57.1)	18 (50.0)	
Analgesic drug use, n (%)					0.499
Not regularly	14 (14.4)	11 (17.2)	14 (15.4)	3 (8.3)	
Weekly	7 (7.2)	3 (4.7)	12 (13.2)	4 (11.1)	
Daily	76 (78.4)	50 (78.1)	65 (71.4)	29 (80.6)	
Ability to work, n (%)					0.012*
Full	35 (36.1)	18 (28.1)	17 (18.7)	6 (16.7)	
Limited	27 (27.8)	11 (17.2)	18 (19.8)	8 (22.2)	
Unable	35 (36.1)	35 (54.7)	56 (61.5)	22 (61.1)	
NRS back pain severity, mean ± SD	5.04 (2.72)	6.09 (2.74)	6.26 (2.31)	7.22 (2.27)	<0.001*
NRS leg pain severity, mean ± SD	7.19 (1.88)	7.80 (1.32)	7.44 (1.98)	7.75 (2.35)	0.173
Oswestry Disability Index, mean ± SD	38.43 (16.31)	46.41 (16.00)	46.35 (15.59)	59.00 (14.60)	<0.001*
Roland-Morris Disability Questionnaire, mean ± SD	9.48 (4.97)	12.66 (5.22)	12.78 (4.93)	16.50 (3.65)	<0.001*
Health-related quality of life					
Extreme anxiety & depression symptoms, n (%)	1 (1.0)	3 (4.7)	7 (7.7)	5 (13.9)	0.003*
Bedriddenness, n (%)	4 (4.1)	3 (4.7)	6 (6.6)	12 (33.3)	<0.001*
Extreme pain or discomfort, n (%)	43 (44.3)	38 (59.4)	56 (61.5)	34 (94.4)	<0.001*
Unable to carry out activities of daily life (ADL), n (%)	17 (17.5)	16 (25.0)	24 (26.4)	21 (58.3)	<0.001*
Unable to care for oneself, n (%)	1 (1.0)	0 (0.0)	0 (0.0)	2 (5.6)	0.002*
EQ-5D index, mean ± SD	0.49 (0.28)	0.35 (0.28)	0.37 (0.30)	0.13 (0.23)	<0.001*
EQ-5D thermometer, mean ± SD	54.04 (16.47)	45.27 (16.64)	50.29 (18.52)	42.08 (18.41)	0.001*

Data are presented after imputation for missing data. Continuous variables are presented as mean (standard deviation), and categorical variables as frequency (percentage).

SD, standard deviation; DDD, degenerative disc disease; ODI, Oswestry disability index; RMDQ, Roland-Morris disability questionnaire; VAS, visual analogue scale;

*p ≤ 0.05

Table 4. Exemplary cases from our cohort. The 5R-STS Web App was applied to five patients, and illustrative information is given on demographics, clinical characteristics, test performance, and health-related quality of life.

Parameter	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Principal complaint	Neurogenic claudication	Neurogenic claudication	Radiating leg pain	Radiating leg pain	Chronic low back pain
Age [yrs.]	48	69	56	35	32
Gender	M	F	M	F	M
Body height [cm]	185	168	180	185	188
Body weight [kg]	78	68	91	115	88
BMI [kg/m²]	22.8	24.1	28.1	33.6	24.9
Smoking Status	Never smoked	Ceased Smoking	Ceased smoking	Active smoker	Ceased smoking
Actual 5R-STS Test Time [sec.]	**4.98**	**12.6**	**16.30**	**17.08**	**Unable to complete test = 30 sec.**
Predicted Test Time [sec.](95%CI)	6.99 (4.21 – 10.09)	7.85 (5.01 – 12.01)	7.30 (4.87 – 10.61)	7.13 (6.80 – 7.98)	6.37 (4.18 – 8.32)
Personalized ULN [sec.]	10.24	12.06	10.86	8.33	8.53
Objective Functional Impairment	**No**	**Yes**	**Yes**	**Yes**	**Yes**
Unsupervised Cluster Assignment	**-**	**Type 1**	**Type 2**	**Type 2**	**Type 3**
Extreme anxiety & depression	No	No	No	No	**Yes**
Bedridden	No	No	No	No	**Yes**
Unable to care for oneself	No	No	No	No	No
Unable to carry out ADL	No	No	No	**Yes**	**Yes**
Unable to work	No	No	No	**Yes**	**Yes**

BMI, body mass index; 5R-STS, five-repetition sit-to-stand test; CI, confidence interval; ULN, upper limit of normal; ADL, activities of daily living

Acknowledgements

The authors are grateful to all participating volunteers, and to Femke Beusekamp, BSc and Nathalie Schouman for study coordination and data collection. We also thank Marlies P. de Wispelaere, PDEng for her efforts in clinical informatics.

Disclosures

Conflict of Interest: The authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Grants and Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Falavigna A, Dozza DC, Teles AR, et al. Current Status of Worldwide Use of Patient-Reported Outcome Measures (PROMs) in Spine Care. *World Neurosurg*. 2017;108:328-335. doi:10.1016/j.wneu.2017.09.002
2. Theodosopoulos PV, Ringer AJ, McPherson CM, et al. Measuring surgical outcomes in neurosurgery: implementation, analysis, and auditing a prospective series of more than 5000 procedures: Clinical article. *J Neurosurg*. 2012;117(5):947-954. doi:10.3171/2012.7.JNS111622
3. Theodosopoulos PV, Ringer AJ. Measuring outcomes for neurosurgical procedures. *Neurosurg Clin N Am*. 2015;26(2):265-269, x. doi:10.1016/j.nec.2014.11.013
4. Fernández-Méndez R, Rastall RJ, Sage WA, et al. Quality improvement of neuro-oncology services: integrating the routine collection of patient-reported, health-related quality-of-life measures. *Neuro-Oncol Pract*. 2019;6(3):226-236. doi:10.1093/nop/npy040
5. Asher AL, McCormick PC, Selden NR, Ghogawala Z, McGirt MJ. The National Neurosurgery Quality and Outcomes Database and NeuroPoint Alliance: rationale, development, and implementation. *Neurosurg Focus*. 2013;34(1):E2. doi:10.3171/2012.10.FOCUS12311
6. Rock AK, Opalak CF, Workman KG, Broaddus WC. Safety Outcomes Following Spine and Cranial Neurosurgery: Evidence From the National Surgical Quality Improvement Program. *J Neurosurg Anesthesiol*. 2018;30(4):328-336. doi:10.1097/ANA.0000000000000474
7. Stienen MN, Ho AL, Staartjes VE, et al. Objective measures of functional impairment for degenerative diseases of the lumbar spine: a systematic review of the literature. *Spine J*. 2019;19(7):1276-1293. doi:10.1016/j.spinee.2019.02.014
8. Staartjes VE, Schröder ML. The five-repetition sit-to-stand test: evaluation of a simple and objective tool for the assessment of degenerative pathologies of the lumbar spine. *J Neurosurg Spine*. 2018;29(4):380-387. doi:10.3171/2018.2.SPINE171416
9. Staartjes VE, Beusekamp F, Schröder ML. Can objective functional impairment in lumbar degenerative disease be reliably assessed at home using the five-repetition sit-to-stand test? A prospective study. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc*. Published online January 24, 2019. doi:10.1007/s00586-019-05897-3
10. Simmonds MJ, Olson SL, Jones S, et al. Psychometric Characteristics and Clinical Usefulness of Physical Performance Tests in Patients With Low Back Pain. *Spine*. 1998;23(22):2412-2421.
11. Teixeira da Cunha-Filho I, Lima FC, Guimarães FR, Leite HR. Use of physical performance tests in a group of Brazilian Portuguese-speaking individuals with low back pain. *Physiother Theory Pract*. 2010;26(1):49-55. doi:10.3109/09593980802602844
12. Klukowska AM, Schröder ML, Stienen MN, Staartjes VE. Objective functional impairment in lumbar degenerative disease: concurrent validity of the baseline severity stratification for the five-repetition sit-to-stand test. *J Neurosurg Spine*. 2020;33(1):4-11. doi:10.3171/2019.12.SPINE191124
13. Ng SSM, Cheung SY, Lai LSW, Liu ASL, Ieong SHI, Fong SSM. Association of seat height and arm position on the five times sit-to-stand test times of stroke survivors. *BioMed Res Int*. 2013;2013:642362. doi:10.1155/2013/642362
14. Ng SSM, Cheung SY, Lai LSW, Liu ASL, Ieong SHI, Fong SSM. Five Times Sit-To-Stand test completion times among older women: Influence of seat height and arm position. *J Rehabil Med*. 2015;47(3):262-266. doi:10.2340/16501977-1915
15. Stienen MN, Smoll NR, Joswig H, et al. Validation of the baseline severity stratification of objective functional impairment in lumbar degenerative disc disease. *J Neurosurg Spine*. 2017;26(5):598-604. doi:10.3171/2016.11.SPINE16683

16. Gautschi OP, Smoll NR, Corniola MV, et al. Validity and Reliability of a Measurement of Objective Functional Impairment in Lumbar Degenerative Disc Disease: The Timed Up and Go (TUG) Test. *Neurosurgery*. 2016;79(2):270-278. doi:10.1227/NEU.0000000000001195
17. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181
18. Von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335(7624):806-808. doi:10.1136/bmj.39335.541782.AD
19. Jones SE, Kon SSC, Canavan JL, et al. The five-repetition sit-to-stand test as a functional outcome measure in COPD. *Thorax*. 2013;68(11):1015-1020. doi:10.1136/thoraxjnl-2013-203576
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. https://www.R-project.org/
21. Kowarik A, Templ M. Imputation with the R Package VIM. *J Stat Softw*. 2016;74(1):1-16. doi:10.18637/jss.v074.i07
22. Koenker R, Chernozhukov V, He X, Peng L. *Handbook of Quantile Regression*. CRC Press; 2017.
23. Koenker R, code) SP (Contributions to CQ, code) PTN (Contributions to SQ, et al. *Quantreg: Quantile Regression*.; 2021. Accessed April 30, 2021. https://CRAN.R-project.org/package=quantreg
24. Martin Bland J, Altman DouglasG. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. *The Lancet*. 1986;327(8476):307-310. doi:10.1016/S0140-6736(86)90837-8
25. Pagana KD, Pagana TJ, Pagana TN. *Mosby's Diagnostic and Laboratory Test Reference - E-Book*. Elsevier Health Sciences; 2018.
26. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63(7):737-745. doi:10.1016/j.jclinepi.2010.02.006
27. Staartjes VE, Klukowska AM, Schröder ML. Association of maximum back and leg pain severity with objective functional impairment as assessed by five-repetition sit-to-stand testing: analysis of two prospective studies. *Neurosurg Rev*. 2020;43(5):1331-1338. doi:10.1007/s10143-019-01168-3
28. Kothe R, Kohlmann Th, Klink T, Rüther W, Klinger R. Impact of low back pain on functional limitations, depressed mood and quality of life in patients with rheumatoid arthritis. *Pain*. 2007;127(1):103-108. doi:10.1016/j.pain.2006.08.011
29. Andersson GB. Epidemiological features of chronic low-back pain. *The Lancet*. 1999;354(9178):581-585. doi:10.1016/S0140-6736(99)01312-4
30. Leveille SG, Guralnik JM, Hochberg M, et al. Low back pain and disability in older women: independent association with difficulty but not inability to perform daily activities. *J Gerontol A Biol Sci Med Sci*. 1999;54(10):M487-493. doi:10.1093/gerona/54.10.m487
31. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593
32. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med*. 2019;380(14):1347-1358. doi:10.1056/NEJMra1814259
33. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
34. Righini M, Van Es J, Den Exter PL, et al. Age-Adjusted D-Dimer Cut-off Levels to Rule Out Pulmonary Embolism: The ADJUST-PE Study. *JAMA*. 2014;311(11):1117-1124. doi:10.1001/jama.2014.2135
35. Ejupi A, Brodie M, Gschwind YJ, Lord SR, Zagler WL, Delbaere K. Kinect-Based Five-Times-Sit-to-Stand Test for Clinical and In-Home Assessment of Fall Risk in Older People. *Gerontology*. 2015;62(1):118-124. doi:10.1159/000381804

36. Staartjes VE, Stienen MN. Data Mining in Spine Surgery: Leveraging Electronic Health Records for Machine Learning and Clinical Research. *Neurospine*. 2019;16(4):654-656. doi:10.14245/ns.1938434.217

[Part II]

Machine Learning-Augmented Operative Imaging

[Chapter 5]

Magnetic resonance imaging-based synthetic computed tomography of the lumbar spine for surgical planning: A clinical proof-of-concept

Victor E. Staartjes
Peter R. Seevinck
W. Peter Vandertop
Marijn van Stralen
Marc L. Schröder

Published in: *Neurosurg Focus. 2021 Jan;50(1):E13.*

[Abstract]

Background

Computed tomography (CT) of the lumbar spine incurs a radiation dose ranging from 3.5 mSv to 19.5 mSv as well as relevant costs and is commonly necessary for spinal neuronavigation. Mitigation of the need for treatment planning CTs in the presence of magnetic resonance imaging (MRI) facilitated by MRI-based synthetic CT would revolutionize navigated lumbar spine surgery. We aim to demonstrate – as a proof of concept – the capability of deep learning-based generation of synthetic CTs from MRI of the lumbar spine in three cases, and to evaluate the potential of synthetic CT for surgical planning.

Methods

Synthetic CT reconstructions were made using a prototype version of the "BoneMRI" software. This deep learning-based image synthesis method relies on a convolutional neural network (CNN) trained on paired MRI-CT data. A specific but generally available 4-minute 3D rf-spoiled T1-weighted multiple gradient echo MRI sequence was supplemented to a 1.5 T lumbar spine MRI acquisition protocol.

Results

In the three presented cases, the prototype synthetic CT method allowed voxel-wise radiodensity estimation from MRI, resulting in qualitatively adequate CT images of the lumbar spine based on visual inspection. Normal as well as pathological structures were reliably visualized. In the first case, in which a spiral CT was available as a control, a volume computed tomography dose index (CTDIvol) of 12.9 mGy could thus have been avoided. Pedicle screw trajectories and screw thickness were estimable based on synthetic CT.

Conclusion

The evaluated prototype BoneMRI method enables generation of synthetic CTs from MRIs with only minor changes in the acquisition protocol, with a potential to reduce workflow complexity, radiation exposure and costs. The quality of the generated CTs was adequate based on visual inspection, and could potentially be used for surgical planning, intraoperative neuronavigation, or for diagnostic purposes in an adjunctive manner.

Introduction

In complex lumbar spinal conditions such as spondylolisthesis, kyphoscoliosis or spinal tumors, the combination of magnetic resonance imaging (MRI) and computed tomography (CT) have long proven complementary. Whereas CT imaging perfectly visualizes osseous structures and allows for assessment of spinal integrity and stability, MRI excels at delineating soft tissue and nervous system structures with high contrast. CT images are often required for neuronavigation and treatment planning, such as for image-guided navigated biopsies or pedicle screw placement.[1]

However, CT is inherently coupled to radiation exposure – an average lumbar spinal CT equals an effective dose of around 3.5 mSv up to 19.5 mSv.[2–6] In addition, multiple imaging sessions increase the patient burden, overall costs, and introduce complex workflows, with potential for intermodality registration errors. For these reasons, MRI-only workflows have gained attention in recent years.[7,8]

In this brief report, we evaluate a prototype of the "BoneMRI" method, for the first time applied to the lumbar spine, that allows generation of synthetic CT (sCT) images from generally available MRI based on deep learning. In this proof-of-concept study we hypothesize that MRI-based sCT images can be used for surgical planning, potentially rendering planning CT scans superfluous.

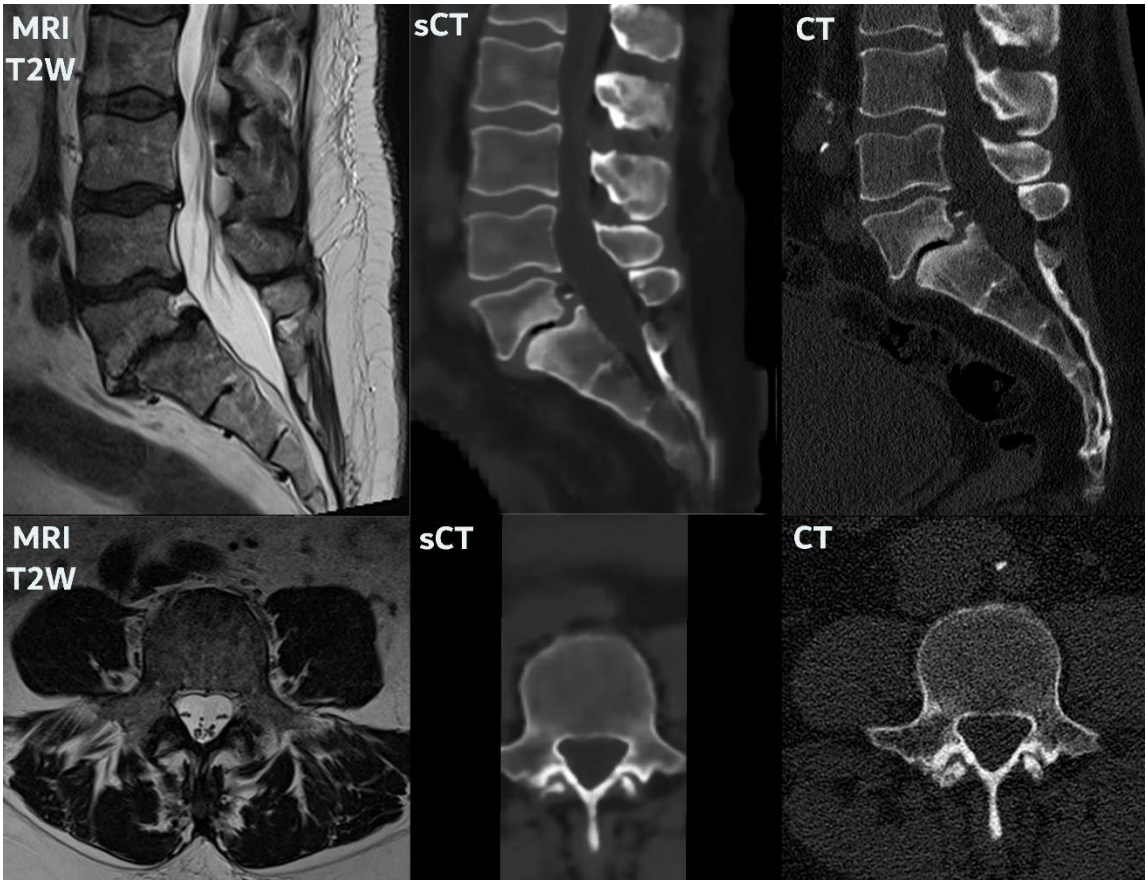


Figure 1. Case 1 (Test Dataset) is depicted. On the far left, conventional T2-weighted MRI images acquired on a 1.5T scanner are provided. In the middle, the synthetic CT (sCT) images generated from the BoneMRI sequence are depicted, along with the "ground truth" spiral CT on the far right. In the top and bottom panel, mid-sagittal and axial (L5 pedicles) cuts are displayed, respectively. This patient's BoneMRI sequence was acquired with a field of view width of 7.2 cm, with the transverse processes being cut off consequently.

Methods

Overview

We utilized a research prototype of the BoneMRI synthetic CT generation method (BoneMRI, MRIguidance B.V., Utrecht, The Netherlands), which is based on preliminary work.[9–11] The deep learning model was trained using paired MRI and CT data, partly obtained in the context of this study and partly in other studies.[9–11] We included nine patients who were scheduled for robotic lumbar fusion surgery, of which eight were used in the training set, independent from the test set. The test set consisted of one of the nine included patients scheduled for lumbar fusion surgery, and two volunteers.

Image Acquisition

CT images were acquired in supine position using a Philips ICT 256 scanner [318 slices / 1mm thickness] and a lumbar spine protocol with iDose reconstruction. MRI images were acquired in a fixed supine position using a Siemens Magnetom Essenza (1.5 T field strength). A standard lumbar spine protocol including conventional T_1W and T_2W sequences (acquisition time = 15 minutes and 4 seconds), was complemented with a sagittal 3D rf-spoiled T_1-weighted multiple gradient echo (T_1w-MGE) sequence for BoneMRI reconstruction (2 echoes; TR/TE1/TE2 = 7ms/2.1ms/4.2ms; FOV = 250x250x90mm; reconstructed voxel size = 0.74x0.74x0.9mm, acquisition time = 3 minutes and 53 seconds). This dedicated sequence utilized a high-frequency encode bandwidth (BW > 500Hz/pix) to minimize potential geometrical distortions.

Model Development

Synthetic CTs were generated from MRI inputs using a patch-based convolutional neural network, similar to U-Net, with CT as the ground truth.[9,12] The model inputs consisted of 4D MRI scans with 3 spatial dimensions and one channel dimension.[9] The network was implemented in Keras[13] with a TensorFlow backend (Google Brain Team, Google LLC, Mountainview, CA, USA).

Proof of Concept Study

The prototype algorithm was applied to three subjects: one patient scheduled for lumbar fusion surgery (Case 1) and two healthy volunteers (Cases 2 & 3): In the first case (Case 1), conventional T1W and T2W MRI sequences as well as the BoneMRI sequence and the generated synthetic CT were available, along with a conventional CT of the lumbar spine. For cases 2 and 3, we test the algorithm in its intended use-case: Only a BoneMRI sequence of the lumbar spine was available, and from it a synthetic CT was generated. These three individuals were never before encountered by the algorithm, thus representing a valid test object. Multiplanar reconstructions (MPR) and 3D volume renders were generated in RadiAnt Version 2020.1, and manual as well as semi-automated measurements and pedicle screw trajectory plannings were carried out in Surgimap Version 2.3.2.1.

Ethical Considerations

The development and proof-of-concept testing of the model, and the associated use of patient data was approved by the local ethical review board (Medical Ethics Committees United (MEC-U), Registration Number: W18.157). All patients signed informed consent forms that allow for use of their data for research and publication purposes.

Results

Synthetic CT images of the lumbar spines of all three cases were successfully generated from BoneMRI sequences. Based on visual inspection, the quality of the synthetic CTs was adequate.

Figure 1 illustrates a comparison of T2-weighed MRI, spiral CT, and synthetic CT images in Case 1. In this case, a spiral CT was available as a control, with a volume computed tomography dose index ($CTDI_{vol}$) of 12.9 mGy that could thus have been avoided. 3D volume renders of the lumbar spine were also calculated from both spiral CT and synthetic CT for comparison (**Figure 2**). In addition, exemplary comparative measurements of anterior and posterior vertebral body height and spinal central canal diameter were performed on Case 1 (**Table 1**), with a mean absolute difference of 0.26 ± 0.24 millimeters.

In order to evaluate the proof-of-concept MRI-only surgical planning workflow the algorithm was validated in two cases in which no CT scan was present (Cases 2 and 3). **Figure 3** illustrates the BoneMRI sequence and the synthetic CTs generated from it for both test cases. Normal as well as pathological structures were reliably visualized, e.g. the relevant spondylolisthesis of one of the volunteers. Figure 4 illustrates that conventional measurements such as spinal canal diameter, lumbar lordosis, and spondylolisthesis grading, as well as semi-automated measurements such as vertebral body segmentation could be carried out reliably on synthetic CT. In addition, we were able to plan pedicle screw trajectories and screw thicknesses based on synthetic CT.

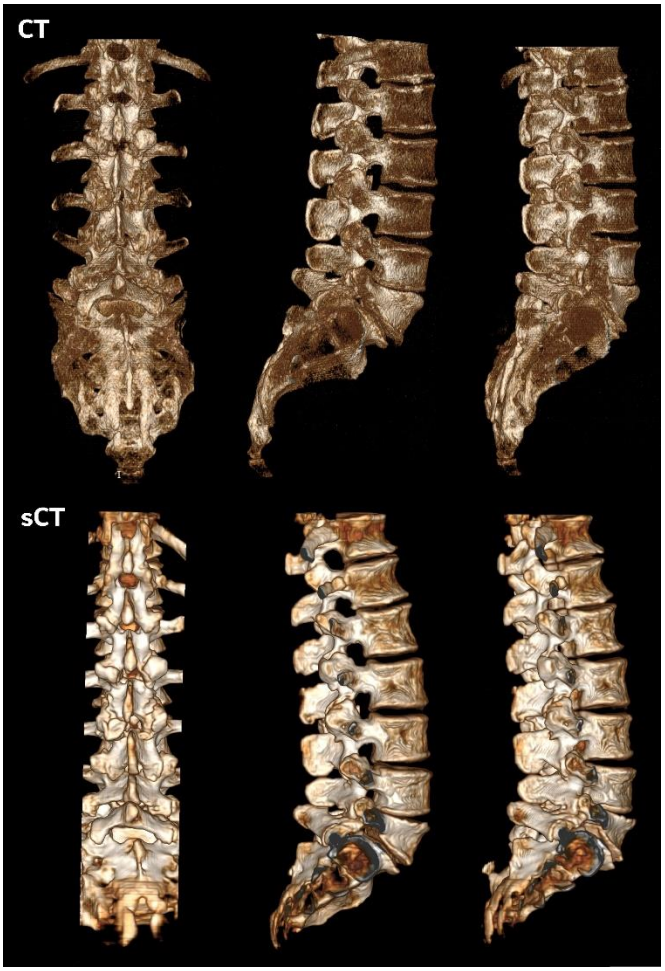


Figure 2. Comparison of 3D volume renders of the lumbar spine calculated from conventional CT (top panel) and from a synthetic CT (sCT) generated from the BoneMRI sequence in the same patient (Case 1, Test Dataset). From left to right, posterior, lateral, and oblique views are provided. This patient's BoneMRI sequence was acquired with a field of view width of 7.2 cm, with the transverse processes being cut off consequently.

Discussion

We show that generation of synthetic CT images of the lumbar spine from MRI is feasible. The ability to visualize the osseous structures in 3D in a similar fashion as traditionally done using CT imaging without radiation and without the need for a separate second examination will be useful in the neurosurgical treatment of spinal disorders, both for diagnostic and therapeutic purposes such as in neuronavigation.

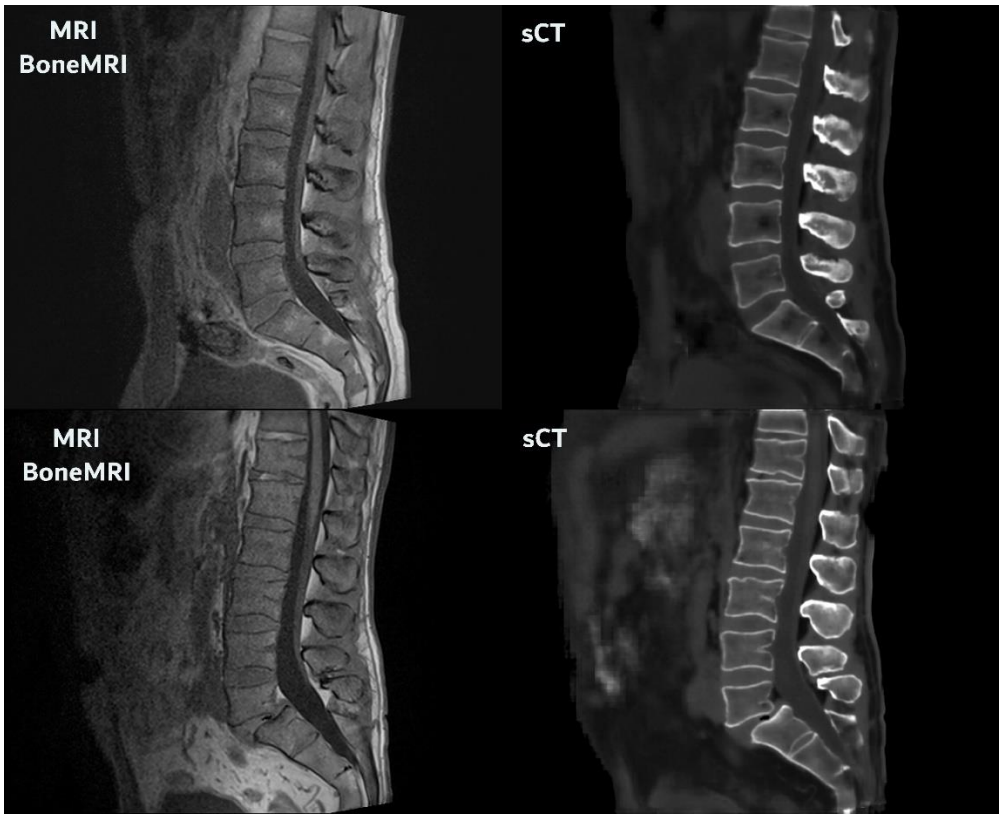


Figure 3. Depicted are the BoneMRI sequences (Left) acquired in two volunteers (Case 2 [top panel] and 3 [bottom panel]) not represented in the training dataset, along with the corresponding synthetic CT (sCT) images generated (Right). Mid-sagittal cuts of the lumbar spine are shown.

The use of image translation algorithms in medicine has previously gained interested in other applications – notably concomitant with an increase of combined use of MRI and CT in the field of radiotherapy.[14,15] For example, generation of synthetic CTs from MRI has been described for radiotherapy purposes in the head and neck, pelvis, prostate, torso, and brain, again mostly for radiotherapy planning.[16–21] A variety of atlas-based or voxel-based methods have been described, using different input sequences, as summarized by Florkow et al.[9] Concomitantly, improvements in image processing techniques such as statistical or machine learning models have also helped the field leap forward.[21] However, most applications have created substitute CTs for radiation treatment planning purposes, often not achieving an image quality that would also be sufficient for i.e. diagnostic imaging or detailed neurosurgical planning. Apart from these aspects, generation of synthetic CTs of the lumbar spine has not previously been demonstrated, although preliminary work has been carried out focusing on the cervical spine.[10,11]

The BoneMRI technique evaluated in this brief report is a deep learning-based method which requires dedicated input data obtained using a generally available sagittal 3D rf-spoiled T_1-weighted multiple gradient sequences, with its parameters carefully chosen in order to sensitize for specific tissue

properties. Due to the dual-echo approach, information about proton density, water and fat fractions, relaxation constants, and susceptibility was intrinsically provided to the deep learning model.[9] Indeed, the results of the BoneMRI technique appear promising in providing high-fidelity synthetic CT images from MRI, with relevant improvements in elimination of radiation exposure, total examination time, and overall logistic efficacy.

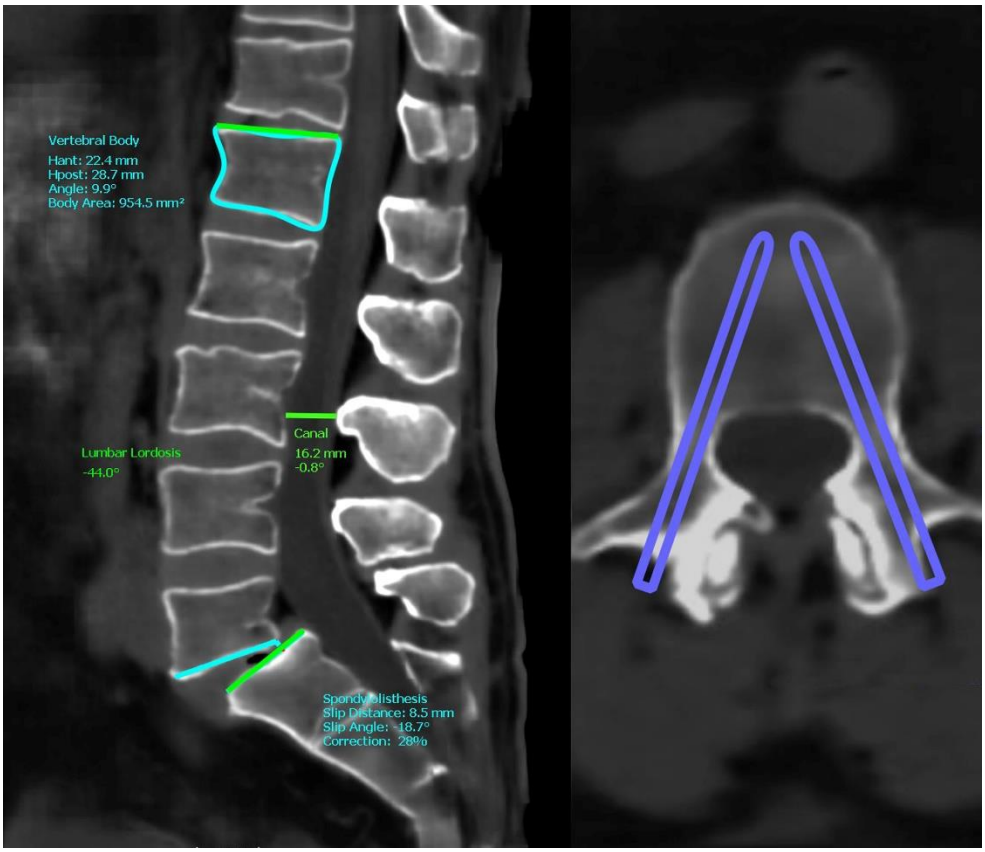


Figure 4. A mid-sagittal synthetic CT cut (Left) of case 3 along with an axial reconstruction (Right) are illustrated. Conventional measurements such as spinal canal diameter, lumbar lordosis, and spondylolisthesis grading (Meyerding Grade II) as well as semi-automated measurements such as vertebral body segmentation were carried out. In addition, pedicle screw trajectories and screw thickness for both L5 pedicles were estimated on synthetic CT imaging.

Recent years have seen an increase in MRI-only workflows.[7,8] However, this would mean losing the radiodensitometric information provided by the CT, which is problematic because spine surgeons often require CT images to judge osseous anatomy such as dysplastic or twisted pedicles, and because MRI-based neuronavigation is still poorly implemented. In spinal neurosurgery, CT imaging has particular importance in surgical navigation, for example for the computer-assisted insertion of pedicle screws.[22] Thus, high-fidelity synthetic CT imaging combined with navigation systems or surgical robotics[22–25] could enable the concept of "radiationless navigated surgery" (RANAS), enabling the use of computer assistance based on preoperative CT imaging without the need for additional radiation. Still, intraoperative fluoroscopy may be necessary for registration and instrumentation control, but these fluoroscopic doses are minor compared to those experienced by the patient during CT scanning.[2,3,26,27] Of course, intraoperative CT imaging can be particularly relevant in spine surgery too, for acquisition of images for spinal neuronavigation based on the actual position of the patient on the operating table. Even under these circumstances, one could imagine the use of intraoperative MRI with generation of synthetic CTs instead of intraoperative CT, especially when considering the increased adoption of intraoperative MRI in

neurosurgical departments[28,29], and the possibility for rapid MRI sequences.[30] In the near future, it is however more likely that clinical applications of synthetic CT will focus around preoperative surgical planning for complex cases, implant sizing, and radiationless intraoperative navigation.

Limitations

The evaluated prototype algorithm is validated on limited data. Thus, although we show the results applied to three cases previously unseen by the algorithm, it was not yet validated in patients with e.g. highly dysplastic or scoliotic pedicles, bone metastases, Modic type endplate changes, and so forth. The limited amount of data prevents us from performing a thorough statistical analysis of geometrical accuracy. However, we applied a high-frequency encode bandwidth to minimize potential geometric distortion. In addition, previous work has demonstrated the robustness of synthetic CT generation to specific degenerative spine diseases, also confirming geometrical accuracy of synthetic CT in comparison to spiral CT.[10,11] Furthermore, validation in patients with implants such as pedicle screws and rods, intervertebral cages, artificial intervertebral discs, neurostimulators, and interspinous process devices is required. Further development and validation on more patients are thus warranted, and the quality of the image might improve further with extended training. Similarly, to demonstrate the feasibility of using the generated synthetic CTs for accurate, near radiationless robotic spine surgery, a case series is currently underway for clinical validation.

Conclusions

We show for the first time in the lumbar spine that, through the use of the BoneMRI acquisition sequence and convolutional neural networks, generation of synthetic CTs from MRIs is feasible within minutes and with visually adequate synthetic CT image fidelity. This novel method has the potential to reduce workflow complexity, radiation exposure and costs associated with adjunctive CT scanning in the lumbar spine. The quality of the generated synthetic CTs – based on visual inspection – is sufficient for surgical planning, neuronavigation, and may even suffice for diagnostics. Further validation of the method is warranted in patients with implants and other artefactants. Likewise, further development of the algorithm based on larger patient cohorts will likely improve image fidelity, and consequently allow early clinical studies focusing on evaluating utility in surgical planning and intraoperative neuronavigation, as well as eventually radiationless robotic pedicle screw insertion.

Table 1 Exemplary measurements performed comparatively on synthetic CT and spiral CT (Case 1). Measurements are provided in millimeters.

Measurement [mm]	Synthetic CT	Spiral CT	Difference
L3			
Anterior VBH	26.5	26.5	0.0
Posterior VBH	32.0	31.8	0.2
Spinal Canal Diameter	14.9	15.0	-0.1
L4			
Anterior VBH	27.2	26.8	0.4
Posterior VBH	29.3	28.9	0.4
Spinal Canal Diameter	19.1	18.9	0.2
L5			
Anterior VBH	29.1	29.1	0.0
Posterior VBH	11.8	12.0	-0.2
Spinal Canal Diameter	24.8	25.6	-0.8
Total (MAD ± SD)			0.26 ± 0.24

CT, computed tomography; VBH, vertebral body height; MAD, mean absolute difference; SD, standard deviation;

Acknowledgements

We thank our radiation technologists Mirjam Visscher and Hester Knottnerus for their kind professional assistance.

Disclosures

Conflict of Interest: MvS and PRS are co-founders and co-owners of MRIguidance B.V. The other authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Grants and Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Härtl R, Lam KS, Wang J, Korge A, Kandziora F, Audigé L. Worldwide Survey on the Use of Navigation in Spine Surgery. *World Neurosurg*. 2013;79(1):162-172. doi:10.1016/j.wneu.2012.03.011
2. Bohl DD, Hijji FY, Massel DH, et al. Patient knowledge regarding radiation exposure from spinal imaging. *Spine J Off J North Am Spine Soc*. 2017;17(3):305-312. doi:10.1016/j.spinee.2016.09.017
3. Biswas D, Bible JE, Bohan M, Simpson AK, Whang PG, Grauer JN. Radiation exposure from musculoskeletal computerized tomographic scans. *J Bone Jt Surg - Ser A*. 2009;91(8):1882-1889. doi:10.2106/JBJS.H.01199
4. Lin EC. Radiation Risk From Medical Imaging. *Mayo Clin Proc*. 2010;85(12):1142-1146. doi:10.4065/mcp.2010.0260
5. Papachristodoulou A, Pliamis N, Volford G, et al. Radiation dose of lumbar spine CT: analysis and comparison between different modes of acquisition in two European imaging centers. ECR 2016 EPOS. Published March 2, 2016. Accessed August 22, 2020. https://epos.myesr.org/poster/esr/ecr2016/C-2386
6. Richards PJ, George J, Metelko M, Brown M. Spine computed tomography doses and cancer induction. *Spine*. 2010;35(4):430-433. doi:10.1097/BRS.0b013e3181cdde47
7. De Silva T, Uneri A, Ketcha MD, et al. Registration of MRI to Intraoperative Radiographs for Target Localization in Spinal Interventions. *Phys Med Biol*. 2017;62(2):684-701. doi:10.1088/1361-6560/62/2/684
8. Owrangi AM, Greer PB, Glide-Hurst CK. MRI-only treatment planning: benefits and challenges. *Phys Med Biol*. 2018;63(5):05TR01. doi:10.1088/1361-6560/aaaca4
9. Florkow MC, Zijlstra F, Willemsen K, et al. Deep learning-based MR-to-CT synthesis: The influence of varying gradient echo-based MR images as input channels. *Magn Reson Med*. Published online October 8, 2019. doi:10.1002/mrm.28008
10. van der Kolk BYM, van Stralen M, Podlogar M, et al. Reconstruction of Osseous Structures in MRI scans of the cervical Spine with BoneMRI: a Quantitative Analysis. In: *ASNR Meeting*. ; 2018:2.
11. van Stralen M, Podlogar M, Hendrikse J, et al. BoneMRI of the cervical spine: Deep learning-based radiodensity contrast generation for selective visualization of osseous structures. In: *ISMRM Meeting*. ; 2019:3.
12. Cui Z, Yang J, Qiao Y. Brain MRI segmentation with patch-based CNN approach. In: *2016 35th Chinese Control Conference (CCC)*. ; 2016:7026-7031. doi:10.1109/ChiCC.2016.7554465
13. Chollet F. Keras: Deep learning library for Theano and TensorFlow. *URL Httpskeras Iok*. 2015;7:8.
14. Pollard JM, Wen Z, Sadagopan R, Wang J, Ibbott GS. The future of image-guided radiotherapy will be MR guided. *Br J Radiol*. 2017;90(1073):20160667. doi:10.1259/bjr.20160667
15. Dirix P, Haustermans K, Vandecaveye V. The value of magnetic resonance imaging for radiotherapy planning. *Semin Radiat Oncol*. 2014;24(3):151-159. doi:10.1016/j.semradonc.2014.02.003
16. Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiat Oncol Lond Engl*. 2017;12(1):28. doi:10.1186/s13014-016-0747-y
17. Maspero M, Savenije MHF, Dinkla AM, et al. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys Med Biol*. 2018;63(18):185001. doi:10.1088/1361-6560/aada6d
18. Dinkla AM, Wolterink JM, Maspero M, et al. MR-Only Brain Radiation Therapy: Dosimetric Evaluation of Synthetic CTs Generated by a Dilated Convolutional Neural Network. *Int J Radiat Oncol Biol Phys*. 2018;102(4):801-812. doi:10.1016/j.ijrobp.2018.05.058
19. Dinkla AM, Florkow MC, Maspero M, et al. Dosimetric evaluation of synthetic CT for head and neck radiotherapy generated by a patch-based three-dimensional convolutional neural network. *Med Phys*. 2019;46(9):4095-4104. doi:10.1002/mp.13663

20. Siversson C, Nordström F, Nilsson T, et al. Technical Note: MRI only prostate radiotherapy planning using the statistical decomposition algorithm. *Med Phys*. 2015;42(10):6090-6097. doi:10.1118/1.4931417
21. Edmund JM, Kjer HM, Van Leemput K, Hansen RH, Andersen JAL, Andreasen D. A voxel-based investigation for MRI-only radiotherapy of the brain using ultra short echo times. *Phys Med Biol*. 2014;59(23):7501-7519. doi:10.1088/0031-9155/59/23/7501
22. Staartjes VE, Klukowska AM, Schröder ML. Pedicle Screw Revision in Robot-Guided, Navigated, and Freehand Thoracolumbar Instrumentation: A Systematic Review and Meta-Analysis. *World Neurosurg*. 2018;116:433-443.e8. doi:10.1016/j.wneu.2018.05.159
23. Schröder ML, Staartjes VE. Revisions for screw malposition and clinical outcomes after robot-guided lumbar fusion for spondylolisthesis. *Neurosurg Focus*. 2017;42(5):E12. doi:10.3171/2017.3.FOCUS16534
24. Staartjes VE, Molliqaj G, Kampen PM van, et al. The European Robotic Spinal Instrumentation (EUROSPIN) study: protocol for a multicentre prospective observational study of pedicle screw revision surgery after robot-guided, navigated and freehand thoracolumbar spinal fusion. *BMJ Open*. 2019;9(9):e030389. doi:10.1136/bmjopen-2019-030389
25. Siccoli A, Klukowska AM, Schröder ML, Staartjes VE. A Systematic Review and Meta-Analysis of Perioperative Parameters in Robot-Guided, Navigated, and Freehand Thoracolumbar Pedicle Screw Instrumentation. *World Neurosurg*. 2019;127:576-587.e5. doi:10.1016/j.wneu.2019.03.196
26. Mendelsohn D, Strelzow J, Dea N, et al. Patient and surgeon radiation exposure during spinal instrumentation using intraoperative computed tomography-based navigation. *Spine J Off J North Am Spine Soc*. 2016;16(3):343-354. doi:10.1016/j.spinee.2015.11.020
27. Villard J, Ryang Y, Demetriades A, et al. Radiation exposure to the surgeon and the patient during posterior lumbar spinal instrumentation: a prospective randomized comparison of navigated versus non-navigated freehand techniques. *Spine*. 2014;39(13):1004-1009. doi:10.1097/BRS.0000000000000351
28. Stienen MN, Fierstra J, Pangalu A, Regli L, Bozinov O. The Zurich Checklist for Safety in the Intraoperative Magnetic Resonance Imaging Suite: Technical Note. *Oper Neurosurg Hagerstown Md*. Published online August 7, 2018. doi:10.1093/ons/opy205
29. Staartjes VE, Serra C, Maldaner N, et al. The Zurich Pituitary Score predicts utility of intraoperative high-field magnetic resonance imaging in transsphenoidal pituitary adenoma surgery. *Acta Neurochir (Wien)*. Published online August 7, 2019. doi:10.1007/s00701-019-04018-9
30. Sayah A, Jay AK, Toaff JS, Makariou EV, Berkowitz F. Effectiveness of a Rapid Lumbar Spine MRI Protocol Using 3D T2-Weighted SPACE Imaging Versus a Standard Protocol for Evaluation of Degenerative Changes of the Lumbar Spine. *AJR Am J Roentgenol*. 2016;207(3):614-620. doi:10.2214/AJR.15.15764

[Chapter 6]

Machine vision for real-time intraoperative anatomical guidance: A proof-of-concept study in endoscopic pituitary surgery

Victor E. Staartjes
Anna Volokitin
Luca Regli
Ender Konukoglu
Carlo Serra

Published in: *Oper Neurosurg (Hagerstown). 2021 Jun 15:opab187. [online ahead of print]*

[Abstract]

Background

Current intraoperative orientation methods either rely on preoperative imaging, are resource-intensive to implement, or difficult to interpret. Real-time, reliable anatomical recognition would constitute another strong pillar on which neurosurgeons could rest for intraoperative orientation.

Objective

We aimed to assess the feasibility of machine vision algorithms to identify anatomical structures using only the endoscopic camera without prior explicit anatomo-topographic knowledge in a proof-of-concept study.

Methods

We developed and validated a deep learning algorithm to detect the nasal septum, the middle turbinate, and the inferior turbinate during endoscopic endonasal approaches based on endoscopy videos from 23 different patients. The model was trained in a weakly supervised manner on 18 and validated on 5 patients. Performance was compared against a baseline consisting of the average positions of the training ground truth labels using a semi-quantitative three-tiered system.

Results

We used 367 images extracted from the videos of 18 patients for training, as well as 182 test images extracted from the videos of another 5 patients for testing the fully developed model. The prototype machine vision algorithm was able to identify the three endonasal structures qualitatively well. Compared to the baseline model based on location priors, the algorithm demonstrated slightly but statistically significantly ($p < 0.001$) improved annotation performance.

Conclusion

Automated recognition of anatomical structures in endoscopic videos by means of a machine vision model using only the endoscopic camera without prior explicit anatomo-topographic knowledge is feasible. This proof-of-concept encourages further development of fully automated software for real-time intraoperative anatomical guidance during surgery.

Introduction

Patient safety and risk management are paramount in surgery, especially in delicate anatomical areas where function is at stake. Successful and safe surgery depends strictly on anatomical orientation – the surgeon needs to be aware of the current anatomy as well as anticipating the following steps of the procedure with the relevant anatomy in his "mind's eye". Appropriate anatomical orientation also reduces surgical risk due to unnecessary manipulation of otherwise healthy and functional neural structures.

Over the years, several methods have been developed to assist neurosurgeons in orienting themselves and assessing the anatomy. Recent years have seen the rise of computer-assisted neuronavigation[1,2], which is however based on preoperative imaging and thus unreliable once the arachnoidal cisterns are opened and brain shift occurs.[3] Intraoperative ultrasound[4–6] and magnetic resonance imaging[7–10] (MRI) can offer more real-time anatomical guidance. The use of fluorescent agents such as 5-ALA have also significantly improved orientation and outcomes.[11–13] Electrophysiological neuromonitoring[14,15] and awake surgery[16] can also assist in navigating around eloquent brain tissue. These methods are effective and rely on exploiting physical characteristics other than light reflection, but they require expensive infrastructure and the operating surgeon to learn to interpret a new imaging modality. Real-time, reliable anatomical recognition would constitute another strong pillar on which neurosurgeons could rest for intraoperative orientation.

Machine vision algorithms constitute a subset of machine learning, and have already proven useful in several domains, notably in self-driving technology, automated environmental recognition, and automated radiological diagnosis. The principles of machine vision could also be applied in the neurosurgical operating room by interpreting the digital image captured by the micro- or endoscope and automatically identifying the visible anatomical structures as well as anticipating other anatomical structures. This approach would allow anatomical navigation in real-time based only on micro- or endoscopic video input, without requiring additional infrastructure. Our aim is to assess the feasibility of machine vision algorithms identifying anatomical structures based solely on endoscopy without prior explicit anatomo-topographic knowledge in a proof-of-concept study. Another specific goal is to evaluate whether a weakly-supervised machine learning approach can generate "heatmap" segmentations of anatomical structures automatically from only single-pixel annotations, enabling far more efficient anatomical annotation by experts.

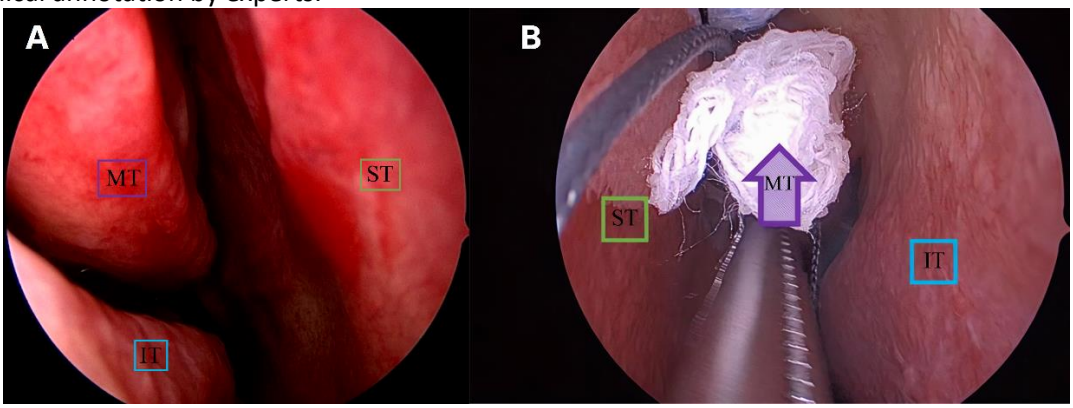


Figure 1. Panel A demonstrates an endoscopic view of the endonasal anatomy, specifically the middle and inferior turbinates as well as the nasal septum. An experienced endoscopic pituitary surgeon labelled these three structures once each second with a single pixel. Panel B demonstrates the intended output of the machine vision model: Anatomical structures in sight such as the septum and inferior turbinate are identified and marked by the user interface. In the future, the model is intended to learn to anticipate structures, for example the middle turbinate hidden behind the cotton patty.
MT, middle turbinate; IT inferior turbinate; ST, nasal septum;

Materials and Methods

Overview

We developed and validated a deep learning algorithm to detect the nasal septum, the middle turbinate, and the inferior turbinate during endoscopic endonasal approaches for pituitary adenoma surgery based on endoscopy videos from 23 different patients.

Ethical Considerations

The use of patient data from the pituitary registry was approved by the local ethical review board (KEK St-V-Nr 2015-0142). All patients signed informed consent forms that allow for use of their data for research and publication purposes.

Data Acquisition and Labelling

Surgical videos were acquired during pituitary surgery performed by two senior neurosurgeons (L.R., C.S.) at the Department of Neurosurgery of the University Hospital Zurich using an endoscopic mono-nostril endonasal technique (Karl Storz GmbH, Tuttlingen, Germany). Both right- and left-nostril cases were included. We included data from 23 different patients, acquired at 30 frames per second. An experienced pituitary surgeon [(C.S.) subsequently labelled the videos at a rate of 1 frame per second (**Figure 1**), marking the nasal septum as well as the middle and inferior turbinates as ground truth points with 1-pixel thick marks around the center for each structure. These ground truth landmarks were blurred with a Gaussian kernel with a 30-pixel standard deviation to create the training set for the learning algorithm. All images were downsampled to a 256-by-256-pixel size for computational reasons.

Model Development and Evaluation

We trained a deep learning model to predict the locations of the ground truth labels in a weakly supervised manner. Videos from 18 patients were randomly selected as training data, while the videos from the remaining 5 patients were used as the holdout (test) set. From the 18 patient videos in the training set, 367 images were extracted for training a network. We use a U-Net[17] neural network to perform heatmap regression, based on the approach described by Payer et al.[18,19] We trained the network for 500 epochs using adaptive moment estimation (ADAM)[20] with a learning rate of 0.001.

For comparison, we also constructed a baseline model. We computed average heatmaps for each structure in the training set as a location prior and we used these average heatmaps as baseline prediction for each test image. Note that predictions of this baseline method are indeed independent of input images, only implementing the location prior.

Due to the weak labelling strategy, quantitative measurements of model performance are not feasible. For semi-quantitative evaluation of model performance, we implemented a three-tiered grading system for the accuracy of anatomical structure segmentation and evaluated predicted heatmaps of both the deep learning as well as the "benchmark" baseline model in this way. All holdout frames were graded. The grading system (**Supplementary Content 1**) was defined in the following way:

> **Grade 1**: The predicted heatmap is completely contained within a single, correctly recognized anatomical structure.
> **Grade 2**: The predicted heatmap contains only a single, correctly recognized anatomical structure, but "overshoots" without involving another anatomical structure.
> **Grade 3**: The predicted heatmap does not contain the correct anatomical structure, or contains more than one anatomical structure

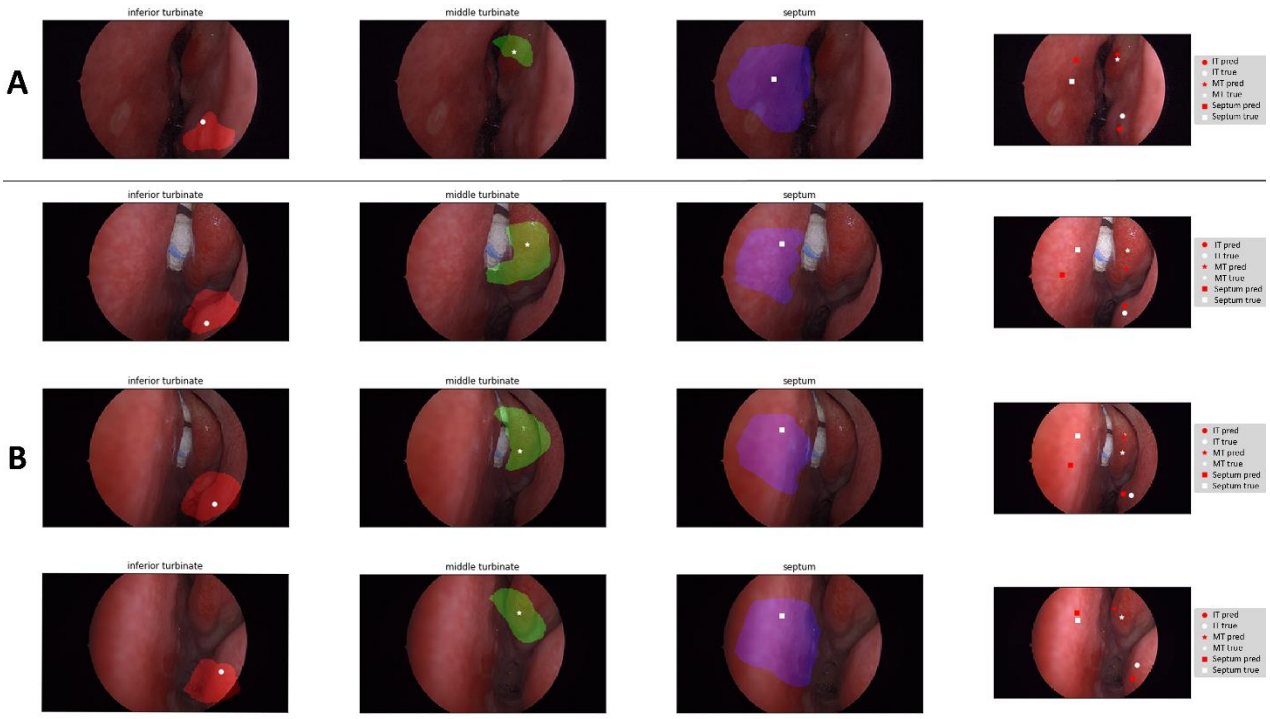


Figure 2. Demonstration of the machine vision model's performance on two patients (Panel A and Panel B) from the test set. The three anatomical structures are correctly identified by the predicted heatmaps, and the ground truth labels (white markers) are contained within these predicted heatmaps. On the far right, the white ground truth labels are contrasted directly with the red markers with the highest model confidence.
MT, middle turbinate; IT inferior turbinate;

A McNemar-Bowker test was used in R to compare the performance of the "benchmark" baseline and the deep learning models directly in the same endoscopy frames.[21] Statistical significance was set at $p \leq 0.05$.

Results

A final 367 images (1101 structures) from 18 patients were used for training, as well as 182 test images (546 structures) from another 5 patients – who were not included in the training data – for testing of the fully developed model. The prototype machine vision algorithm was able to locate and generate heatmaps for the three endonasal structures relatively accurately. **Figure 2** demonstrates the predicted heatmaps for two patients from the test set.

Figure 3 demonstrates the predictions of the baseline model based on location priors, for the same two patients mentioned above. It is visible that this approach – which predicts the average training heatmaps independent of the input image – produces qualitatively inferior results and does not locate the three endonasal anatomical structures as accurately as our machine vision model.

Table 1 demonstrates the results of the semi-quantitative comparative performance grading. The deep learning model performed statistically significantly better ($p < 0.001$) compared to a baseline model based on location priors, increasing the rate of perfect segmentations (Grade 1) from 27.1% to 36.1%, and increasing the rate of correct anatomical structure labelling (Grade 1 + Grade 2) from 45.4%. to 55.3%.

Discussion

In a proof-of-concept study using endoscopic footage from 23 patients undergoing transnasal pituitary surgery, a machine vision approach without prior anatomical knowledge was able to locate three anatomical structures on new samples, using only information coming from an endoscopic camera. Compared to a baseline model based on location priors, our deep learning algorithm demonstrated slightly but statistically significantly improved annotation performance. These preliminary results lay the groundwork for further development of an automated anatomical guidance system for real-time intraoperative surgical guidance.

Table 1. Comparative performance grading of the deep learning model and the baseline model based on location priors as a benchmark. All 182 frames from the test set as well as all three anatomical structures were pooled and evaluated using a three-tiered semi-quantitative grading system **(Supplementary Content 1)**. A McNemar-Bowker test was used to assess the difference in performance among the two methods on the same endoscopy frames.

	Deep Learning Model	Baseline Model (Benchmark)	*P* Value
Grade 1	197 (36.1%)	148 (27.1%)	< 0.001*
Grade 2	105 (19.2%)	100 (18.3%)	
Grade 3	244 (44.7%)	298 (54.6%)	

***Grade 1**: The predicted heatmap is completely contained within a single, correctly recognized anatomical structure.*
***Grade 2**: The predicted heatmap contains only a single, correctly recognized anatomical structure, but "overshoots" without involving another anatomical structure.*
***Grade 3**: The predicted heatmap does not contain the correct anatomical structure, or contains more than one anatomical structure*
** $p \leq 0.05$*

Reliable real-time anatomical recognition would constitute a breakthrough in intraoperative surgical guidance and would likely improve patient safety and potentially clinical outcomes. Especially if machine vision implementations could help identify and anticipate critical eloquent structures, it is conceivable that some complications could be avoided through greater anatomical orientation, especially when the neuroanatomy is distorted e.g. in complex skull base tumors. Moreover, by clarifying the steps of complex surgical procedures, real-time anatomical guidance could result in shorter surgical times and thus consequently lead to improvements in logistics and costs. State-of-the-art machine vision techniques could enable the concept of real-time anatomical guidance without the need for additional infrastructure. The presented pilot study is a step towards this goal.

The past contributions of machine vision in the field of clinical neurosurgery have been limited. This is especially true when considering real-time surgical field recognition that is not based on preoperative imaging or anatomical atlases. The major applications of deep learning in neurosurgery have focused on the interpretation of data in neuroradiology and neuropathology: For example, Titano et al.[22] have used convolutional neural networks (CNNs) to automatically detect acute neurological events on cranial imaging. Similarly, Hollon et al.[23] used CNNs to interpret intraoperative fresh-frozen sections in near real-time using stimulated Raman spectroscopy data.

Indeed, when focusing specifically on machine vision and its intraoperative applications, most applications in surgery have been based on preoperative imaging. For example, Gong et al.[24] reported on a machine vision algorithm that recognizes parts of a model of a debulked brain tumor cavity based on endoscopic cameras and a pre-existing three-dimensional scan of the field.
This then helped guide a robotic arm with a mock surgical tool move to a target – namely the tumor margin – using fluorescence. Similarly, applications of machine vision exist to guide the placement of augmented reality objects into the surgeon's view, based on preoperative delineation of structures.[25]

Mohri et al.[26] have developed a machine vision based method of position-recognition in CyberKnife radiosurgery, that allows positional adjustments based on device cameras. Heilbrun et al.[27] have succeeded in stereotactic localization and guidance of surgical instruments within the cranial vault based on machine vision alone, in relation to a preoperative gold standard.

To the best of our knowledge, no applications of real-time machine learning – i.e. not on still images, and not based on preoperative imaging or anatomical atlases – in neurosurgery have been reported yet. In other fields, preliminary results in intraoperative anatomical recognition have been obtained. Twinanda et al.[28] have applied machine vision to recognize the type of laparoscopic surgical procedure carried out using surgical videos alone. Petscharnig et al.[29] have been able to classify still images from laparoscopic videos to recognize anatomical structures in gynecological surgery, for example differentiating among colon, liver, ovary, and uterus, as well as among surgical tasks such as cutting, coagulation, and suturing. Similarly, Takiyama et al.[30] used CNNs to classify still images from esophagoduodenoscopy procedures automatically and with high accuracy, and found that their algorithm can recognize specific anatomical locations reliably.

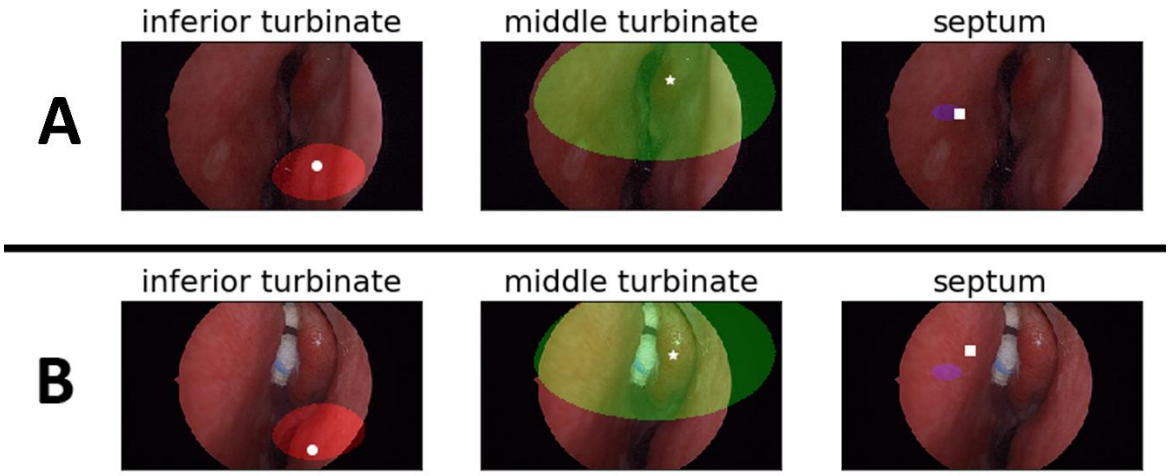


Figure 3. Demonstration of the "benchmark" baseline model based on location priors on the same two patients (Panel A and Panel B) from the test set. While the ground truth labels (white markers) of the three anatomical structures are sometimes contained in the predicted heatmaps, the fact that the predicted heatmaps are independent of the input images and solely based on the average locations in the training data, the predicted heatmaps only outline the three structures poorly compared to our machine vision model.

To minimize the annotation burden of the expert neurosurgeon, in this pilot study the supervision of the algorithm is given only as single points on each structure of interest, even though structures covered larger areas in the images ("weakly supervised"). As is standard in the landmark detection framework, our network predicts the locations of structures in the form of heatmaps. Using an appropriate value, the heatmaps can be thresholded to yield an image region that corresponds to the real anatomical structure. This feature however remains to be verified. The model appeared to have learned to identify the rough extents of the three anatomical structures without explicit supervision. It is important to mention that expert annotations – especially concerning microneurosurgical anatomy – are time-intensive and difficult to obtain. This feature leads to a relevantly reduced annotation burden for experts. These heatmaps demonstrated qualitatively reasonable coverage of the anatomical structures, considering the small proof-of-concept sample.

The qualitatively rational behavior of the machine vision method indicate that it may be a promising component to creating a fully automated anatomical recognition software, while reducing the annotation burden on expert neurosurgeons. Further development – first on relatively standardized surgical approaches such as the pterional craniotomy and sylvian fissure split – with larger training samples and many more anatomical structures will be the next step towards real-time intraoperative anatomical guidance.

Limitations

Our method performed relatively well when applied to new cases, i.e. those in the test set which were never previously encountered by the algorithm. Nonetheless, the machine vision model was trained on a limited amount of endoscopy video data and has only encountered a limited amount of anatomical variability. For these reasons, we cannot judge whether the algorithm would generalize well to cases with anatomical variants such as a perforated septum or when the mucosa is covered by more blood than in the current dataset. However, the fact that this machine vision approach achieved satisfactory performance with limited training data is encouraging towards the likely improved generalization and performance with relevantly more numerous and varying training samples, although the present samples were not specifically selected to be anatomically similar. Further training and in-silico validation on more patients are thus warranted, and the quality of the predictions generated will undoubtedly increase with extended training and further technical innovation.

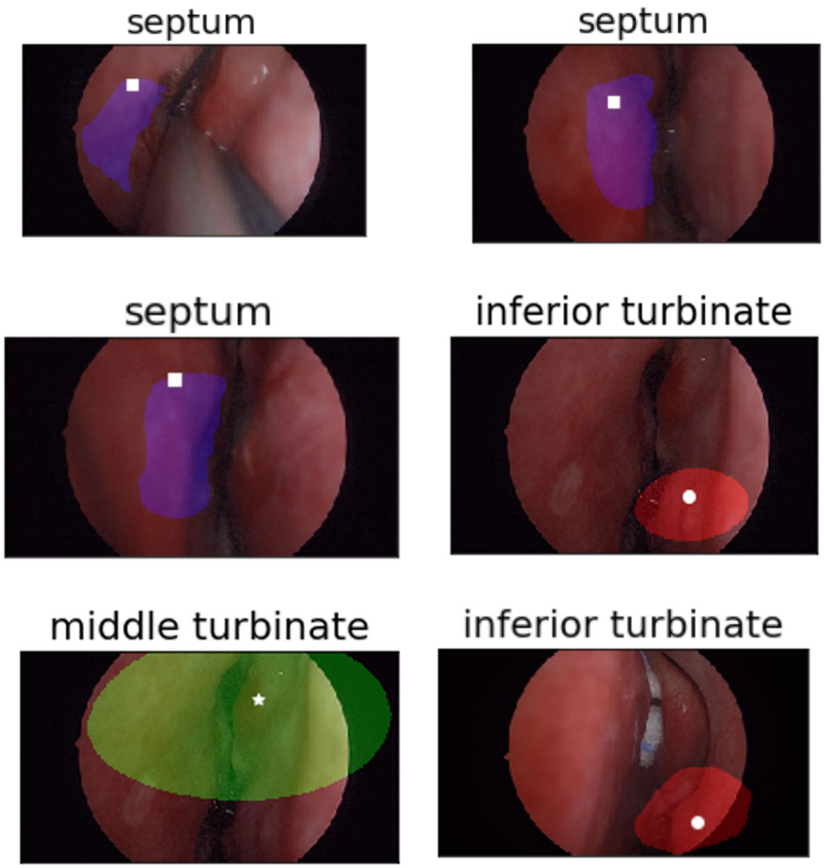
Given our limited annotations and the weakly supervised nature of our learning approach, evaluating our method quantitatively is inherently difficult. Since there are no negative examples, it is unclear where to set the threshold value for predictions. Additionally, even when a threshold is picked heuristically, there is no way of distinguishing whether predicted points actually belong to the structure or not. It is only possible to quantify whether the ground truth point is included in a given prediction, but this does not give very much information about specificity and sensitivity of the proposed machine vision model. In this case, a qualitative evaluation – which we carried out using a three-tiered grading – may be more appropriate, since it is easy to check visually whether a predicted region matches a structure well, or whether it does not. Furthermore, given the visual results, the presence of false-positive predictions – i.e. an anatomical structure identified where there is no such structure – cannot be excluded. Because our ground truth samples did not include negative examples, we were unable to assess the false-positive rate.

The use-case of the model – predicting only three endonasal anatomical structures – is still limited and serves only as a proof-of-concept. This technique will become of particular relevance once it is able to identify abnormal anatomical structures such as a tortuous carotid or in revision surgery, and also once the technique is able to recognize a far wider array of anatomical structures than only the superficial ones that were initially chosen to assess our method in this proof-of-concept. Lastly, even when further development with larger sample sizes will allow for the desired levels of accuracy, machine vision for real-time intraoperative anatomical guidance should only be seen as an adjunct to other surgical tools such as neuronavigation.

Conclusions

In a proof-of-concept study, we demonstrate that automated recognition of anatomical structures by means of a machine vision model without prior anatomical knowledge is feasible. Our weakly supervised learning method is able to learn heatmap representations from single-pixel annotations, which enables efficient labelling by experts, which in turn is the key to developing further applications in complex cranial surgery. Our algorithm recognized endonasal structures with qualitatively rational behavior after training

on a relatively small amount of endoscopy video data and performed slightly but statistically significantly better than a "benchmark" baseline model based on location priors. This proof-of-concept study encourages further development of fully automated software for real-time intraoperative anatomical guidance during surgery.



Supplementary Content 1. Semi-quantitative, three-tiered grading system to assess weakly-supervised anatomical structure recognition performance. Grade 1: The predicted heatmap is completely contained within a single, correctly recognized anatomical structure; Grade 2: The predicted heatmap contains only a single, correctly recognized anatomical structure, but "overshoots" without involving another anatomical structure; Grade 3: The predicted heatmap does not contain the correct anatomical structure, or contains more than one anatomical structure.

Disclosures

Conflict of Interest: The authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
Grants and Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Härtl R, Lam KS, Wang J, Korge A, Kandziora F, Audigé L. Worldwide Survey on the Use of Navigation in Spine Surgery. *World Neurosurgery*. 2013;79(1):162-172. doi:10.1016/j.wneu.2012.03.011
2. Orringer DA, Golby A, Jolesz F. Neuronavigation in the surgical management of brain tumors: current and future trends. *Expert Rev Med Devices*. 2012;9(5):491-500. doi:10.1586/erd.12.42
3. Iversen DH, Wein W, Lindseth F, Unsgård G, Reinertsen I. Automatic Intraoperative Correction of Brain Shift for Accurate Neuronavigation. *World Neurosurgery*. 2018;120:e1071-e1078. doi:10.1016/j.wneu.2018.09.012
4. Ulrich NH, Burkhardt J-K, Serra C, Bernays R-L, Bozinov O. Resection of pediatric intracerebral tumors with the aid of intraoperative real-time 3-D ultrasound. *Childs Nerv Syst*. 2012;28(1):101-109. doi:10.1007/s00381-011-1571-1
5. Burkhardt J-K, Serra C, Neidert MC, et al. High-frequency intra-operative ultrasound-guided surgery of superficial intra-cerebral lesions via a single-burr-hole approach. *Ultrasound Med Biol*. 2014;40(7):1469-1475. doi:10.1016/j.ultrasmedbio.2014.01.024
6. Hammoud MA, Ligon BL, Elsouki R, Shi WM, Schomer DF, Sawaya R. Use of intraoperative ultrasound for localizing tumors and determining the extent of resection: a comparative study with magnetic resonance imaging. *Journal of Neurosurgery*. 1996;84(5):737-741. doi:10.3171/jns.1996.84.5.0737
7. Berkmann S, Schlaffer S, Nimsky C, Fahlbusch R, Buchfelder M. Intraoperative high-field MRI for transsphenoidal reoperations of nonfunctioning pituitary adenoma. *J Neurosurg*. 2014;121(5):1166-1175. doi:10.3171/2014.6.JNS131994
8. Senft C, Bink A, Franz K, Vatter H, Gasser T, Seifert V. Intraoperative MRI guidance and extent of resection in glioma surgery: a randomised, controlled trial. *Lancet Oncol*. 2011;12(11):997-1003. doi:10.1016/S1470-2045(11)70196-6
9. Staartjes VE, Serra C, Maldaner N, et al. The Zurich Pituitary Score predicts utility of intraoperative high-field magnetic resonance imaging in transsphenoidal pituitary adenoma surgery. *Acta Neurochir*. Published online August 7, 2019. doi:10.1007/s00701-019-04018-9
10. Stienen MN, Fierstra J, Pangalu A, Regli L, Bozinov O. The Zurich Checklist for Safety in the Intraoperative Magnetic Resonance Imaging Suite: Technical Note. *Oper Neurosurg (Hagerstown)*. Published online August 7, 2018. doi:10.1093/ons/opy205
11. Stummer W, Stepp H, Wiestler OD, Pichlmeier U. Randomized, Prospective Double-Blinded Study Comparing 3 Different Doses of 5-Aminolevulinic Acid for Fluorescence-Guided Resections of Malignant Gliomas. *Neurosurgery*. 2017;81(2):230-239. doi:10.1093/neuros/nyx074
12. Stummer W, Pichlmeier U, Meinel T, et al. Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase III trial. *Lancet Oncol*. 2006;7(5):392-401. doi:10.1016/S1470-2045(06)70665-9
13. Hadjipanayis CG, Widhalm G, Stummer W. What is the Surgical Benefit of Utilizing 5-ALA for Fluorescence-Guided Surgery of Malignant Gliomas? *Neurosurgery*. 2015;77(5):663-673. doi:10.1227/NEU.0000000000000929
14. De Witt Hamer PC, Robles SG, Zwinderman AH, Duffau H, Berger MS. Impact of intraoperative stimulation brain mapping on glioma surgery outcome: a meta-analysis. *J Clin Oncol*. 2012;30(20):2559-2565. doi:10.1200/JCO.2011.38.4818
15. Sanai N, Mirzadeh Z, Berger MS. Functional Outcome after Language Mapping for Glioma Resection. *New England Journal of Medicine*. 2008;358(1):18-27. doi:10.1056/NEJMoa067819
16. Hervey-Jumper SL, Li J, Lau D, et al. Awake craniotomy to maximize glioma resection: methods and technical nuances over a 27-year period. *J Neurosurg*. 2015;123(2):325-339. doi:10.3171/2014.10.JNS141520

17. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:150504597 [cs]*. Published online May 18, 2015. Accessed August 27, 2020. http://arxiv.org/abs/1505.04597
18. Payer C, Štern D, Bischof H, Urschler M. Multi-label Whole Heart Segmentation Using CNNs and Anatomical Label Configurations. In: Pop M, Sermesant M, Jodoin P-M, et al., eds. *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*. Lecture Notes in Computer Science. Springer International Publishing; 2018:190-198. doi:10.1007/978-3-319-75541-0_20
19. Payer C, Stern D, Bischof H, Urschler M. Regressing Heatmaps for Multiple Landmark Localization Using CNNs. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*. Springer International Publishing AG; 2016:230-238. doi:10.1007/978-3-319-46723-8_27
20. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv:14126980 [cs]*. Published online January 29, 2017. Accessed August 27, 2020. http://arxiv.org/abs/1412.6980
21. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020. https://www.R-project.org/
22. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med*. 2018;24(9):1337-1341. doi:10.1038/s41591-018-0147-y
23. Hollon TC, Pandian B, Adapa AR, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nature Medicine*. 2020;26(1):52-58. doi:10.1038/s41591-019-0715-9
24. Gong Y, Hu D, Hannaford B, Seibel EJ. Toward real-time endoscopically-guided robotic navigation based on a 3D virtual surgical field model. *Proc SPIE Int Soc Opt Eng*. 2015;9415:94150C. doi:10.1117/12.2082872
25. Maruyama K, Watanabe E, Kin T, et al. Smart Glasses for Neurosurgical Navigation by Augmented Reality. *Oper Neurosurg (Hagerstown)*. 2018;15(5):551-556. doi:10.1093/ons/opx279
26. Mohri I, Umezu Y, Fukunaga J, et al. [Development of a new position-recognition system for robotic radiosurgery systems using machine vision]. *Nihon Hoshasen Gijutsu Gakkai Zasshi*. 2014;70(8):751-756. doi:10.6009/jjrt.2014_jsrt_70.8.751
27. Heilbrun MP, McDonald P, Wiker C, Koehler S, Peters W. Stereotactic localization and guidance using a machine vision technique. *Stereotact Funct Neurosurg*. 1992;58(1-4):94-98. doi:10.1159/000098979
28. Twinanda AP, Marescaux J, de Mathelin M, Padoy N. Classification approach for automatic laparoscopic video database organization. *Int J Comput Assist Radiol Surg*. 2015;10(9):1449-1460. doi:10.1007/s11548-015-1183-4
29. Petscharnig S, Schöffmann K. Learning laparoscopic video shot classification for gynecological surgery. *Multimed Tools Appl*. 2018;77(7):8061-8079. doi:10.1007/s11042-017-4699-5
30. Takiyama H, Ozawa T, Ishihara S, et al. Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. *Scientific Reports*. 2018;8(1):1-8. doi:10.1038/s41598-018-25842-6

[Part III]

Clinical Prediction Modelling Using Machine Learning

[Chapter 7]

Development and external validation of a clinical prediction model for functional impairment after intracranial tumor surgery

Victor E. Staartjes
Morgan Broggi
Costanza Maria Zattra
Flavio Vasella
Julia Velz
Silvia Schiavolin
Carlo Serra
Jiri Bartek Jr
Alexander Fletcher-Sandersjöö
Petter Förander
Darius Kalasauskas
Mirjam Renovanz
Florian Ringel
Konstantin R. Brawanski
Johannes Kerschbaumer
Christian F. Freyschlag
Asgeir S. Jakola
Kristin Sjåvik
Ole Solheim
Bawarjan Schatlo
Alexandra Sachkova
Hans Christoph Bock
Abdelhalim Hussein
Veit Rohde
Marike L. D. Broekman
Claudine O. Nogarede
Cynthia M.C. Lemmens
Julius M. Kernbach
Georg Neuloh
Oliver Bozinov
Niklaus Krayenbühl
Johannes Sarnthein
Paolo Ferroli
Luca Regli
Martin N. Stienen

Published in: *J Neurosurg. 2020 Jun 12:1-8. [online ahead of print]*

[Abstract]

Background

Decision-making for intracranial tumor surgery requires balancing the oncological benefit against the risk for resection-related impairment. Risk estimates are commonly based on subjective experience and generalized numbers from the literature, but even experienced surgeons overestimate functional outcome after surgery. Today, there is no reliable and objective way to preoperatively predict an individual patient's risk of experiencing any functional impairment.

Methods

We developed a prediction model for functional impairment at 3 to 6 months after microsurgical resection, defined as a decrease in Karnofsky Performance Status of ≥10 points. Two prospective registries in Switzerland and Italy were used for development. External validation was performed in seven cohorts from Sweden, Norway, Germany, Austria and the Netherlands. Age, gender, prior surgery, tumor histology and maximum diameter, expected major brain vessel or cranial nerve manipulation, resection in eloquent areas and the posterior fossa, and surgical approach were recorded. Discrimination and calibration metrics were evaluated.

Results

In the development (2437 patients; 48.2% male; mean [SD] age: 55 [15] years) and external validation (2427 patients; 42.4% male; mean [SD] age: 58 [13] years) cohorts, functional impairment rates were 21.5% and 28.5%, respectively. In the development cohort, area-under-the-curve (AUC) values of 0.72 (95% CI: 0.69 to 0.74) were observed. In the pooled external validation cohort, the AUC was 0.72 (95% CI: 0.69 to 0.74), confirming generalizability. Calibration plots indicate fair calibration in both cohorts. The tool has been incorporated into a web app available at https://neurosurgery.shinyapps.io/impairment/.

Conclusions

Functional impairment after intracranial tumor surgery remains extraordinarily difficult to predict, although machine learning can help quantify risk. This externally validated prediction tool can serve as the basis for case-by-case discussions and risk-to-benefit estimation of surgical treatment in the individual patient.

Introduction

Patients frequently ask whether they will "stay the same" after the resection of an intracranial tumor—an intricate question often challenging to answer satisfactorily. Clinicians cautiously estimate the likelihood of functional impairment after microsurgical resection by integrating radiological information, anatomo-topographical features, the expected histopathological tumor type, and the complexity of the required surgical approach in view of patient-intrinsic characteristics, generalized numbers from the literature, and the surgeon's own expertise and experience. The answer to this question plays a critical role in the shared decision-making process.

Among multiple centers and surgeons, considerable diversity exists in treatment protocols, surgical techniques, experience, and equipment, which relate to the achieved extent of resection (EOR), survival, functional and patient-reported outcome measures (PROMs).[1–7] Today, evidence is accumulating regarding the lower oncological benefit of complete resection in cases of postoperative neurological and/or functional worsening[8,9], emphasizing the importance of periprocedural safety and the regimen of "maximum safe resection", meaning aiming for the greatest EOR that allows for preservation of neurological function.[5]

Functional impairment after intracranial tumor surgery is an extraordinarily difficult outcome to predict, and neurooncological surgeons often overestimate postoperative functional outcome.[2,10] Currently, risk estimation is based on prior experiences and generalizable rates from the literature, but outcome prediction tailored to a patient's specific features is increasingly becoming a part of modern precision "personalized medicine".[11–13] Recently, machine learning (ML) methods have been applied to generate patient-specific predictive analytics for outcomes in neurosurgery, and these often outperform classification schemes as well as conventional modelling techniques such as logistic regression.[11–16] The present study aimed to develop and externally validate a novel prediction model that forecasts individualized postoperative functional impairment from a set of variables usually available at the time of preoperative informed patient consent.

Methods

Overview

From a large bi-centric sample of patients who underwent microsurgical resection of intracranial tumors, we developed an ML-based prediction tool for new postoperative functional impairment. The prediction tool was externally validated with data from seven European centers. This study was compiled according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.

Ethical Considerations

The scientific workup of registry data was approved by the institutional review boards (IRBs) of all informed institutions. The study was registered at the University Hospital Zurich (ClinicalTrials.gov Identifier: NCT01628406). Patients provided informed consent or informed consent was waived, depending on the demands of the local IRB.

Data Sources

Prospective institutional databases from two centers were retrospectively analyzed. Consecutive patients undergoing microsurgical resection of intracranial tumors via microscopic craniotomy or transsphenoidal surgery were included. Diagnostic biopsies were excluded. We pooled data from patients undergoing surgery between January 2013 and December 2017 at the Department of Neurosurgery, University

Hospital of Zurich, Switzerland, and between January 2014 and December 2017 at the Department of Neurosurgery, Fondazione IRCCS Istituto Neurologico Carlo Besta in Milan, Italy. The methodological details of these two patient registries were described previously.[2,6,17] Physicians who collected the registry and outcome data in these registries were specifically trained; internal standard operating procedures additionally helped with harmonizing the data collection. Data quality in the registries was regularly reviewed and improved as required. All patients in the derivation cohort had the required variables recorded; there was no need to delete cases or impute missing data.
The use of intraoperative technology to increase EOR while monitoring neurological function, e.g. intraoperative imaging (ultrasound, magnetic resonance imaging, navigation, fluorescence-guidance, etc.), electrophysiological monitoring, or awake surgery, is routinely applied in addition to the use of surgical tools (e.g. intraoperative microscope, ultrasonic aspirators).[3,5,18–21]
The model was evaluated in seven centers from five countries. Göttingen (2014–2017), Innsbruck (2015–2018), and Leiden and the Hague (2015–2018) data were derived from prospective registries. Trondheim data (2007–2015) were based on a prospective registry supplemented with retrospectively collected data. Stockholm (2007–2015), Mainz (2007–2018), and Aachen (2018) data were retrospectively collected. To improve the realistic representation of external validation model performance, neurosurgeons who collected data for the external validation cohort were not specifically trained, apart from receiving the same detailed variable definitions as described in this methods section and as listed in the web-app. All participating centers pursue a "maximum safe resection" philosophy.[5]

Outcome Measures

The primary outcome measure was "new postoperative functional impairment", defined as a 10-point or greater decrease in Karnofsky Performance Status (KPS) at 3 to 6 months postoperatively, compared to preoperative functional status.[2] There is no established minimum clinically important difference (MCID) for KPS after intracranial tumor surgery. We deliberately chose the 10-point cut-off [2], as opposed to a dynamic cut-off with different steps depending on baseline status [22], in order not to overlook subtle differences in performance, since even minor decreases in performance as judged by clinical scales can be perceived as devastating by patients.[7]

Recorded variables included KPS at admission and at 3 to 6 months, age, gender, prior surgery, tumor type and maximum diameter, expected major vessel or cranial nerve manipulation, surgery in the posterior fossa, resection in an eloquent area, and whether a transsphenoidal or transcranial resection was performed. We defined major brain vessel manipulation as the expected manipulation of major vessels encased by or in proximity of the tumor. Major vessels included the internal carotid, the anterior, middle, and posterior cerebral arteries, basilar and vertebral arteries, as well as the large venous sinuses and internal, Trolard, and Labbé veins. Eloquent areas were defined as motor, sensory, language, or visual areas, as well as the hypothalamus, thalamus, internal capsule, brainstem, and pineal region.[2] These variables were chosen as inputs for the model owing to their demonstrated relationships to functional impairment, and their number was limited to ensure the practical applicability of the prediction model.[2]

Model Development and Validation

Continuous data are reported as mean ± standard deviation (SD) or median (interquartile range, IQR), and categorical data as numbers (percentages). Non-dichotomous categorical input variables were one-hot encoded. Numerical input variables were standardized using centering and scaling.
A logistic generalized additive model (GAM) based on locally estimated scatterplot smoothing (LOESS) was developed on the derivation cohort to predict any functional impairment, using the "caret" and "gam" packages.[23–26] The model parameters were fitted in 50 bootstrap resamples with replacement, hyperparameters were tuned, and the final model was selected based on area-under-the-curve (AUC).

The final model had a span of 0.5. A *k*-nearest neighbors (KNN) algorithm was trained on the derivation set to impute any potential missing data during prediction on new data.[27] The threshold for binary classification was selected on the derivation cohort based on the "closest-to-(0,1)-criterion".[28]
The prediction model was subsequently externally validated. No recalibration was carried out.[29] When predicting on the external validation cohort, the co-trained KNN algorithm was applied to impute missing data.[27] Calibration was visually assessed using calibration plots. Quantile-based 95% confidence intervals (CIs) of the discrimination and calibration metrics were obtained in 1000 bootstrap resamples.
All analyses were carried out in *R* version 3.5.2 (The R Foundation for Statistical Computing, Vienna, Austria). The *Supplementary Methods* contain the statistical code.

Results

Derivation Cohort

A total of 2437 patients were available in the two prospective registries. There was no missing data. Mean patient age was 55 ± 15 years, and 1175 patients (48.2%) were male. The median KPS at admission was 90 (IQR: 80–90), and 440 patients (18.1%) had undergone prior surgery. The majority of patients (n = 2148, 88.1%) underwent open craniotomy, while 289 patients (11.9%) underwent transsphenoidal surgery. New functional impairment was observed in 525 patients (21.5%). Early mortality occurred in 85 patients (3.5%). Detailed patient characteristics are provided in **Table 1**.

External Validation Cohort

Seven centers in five countries provided data for external validation. The external validation cohort was made up of 2427 patients. Patient characteristics per center are provided in **Supplementary Table S1**. Overall, 392 of 26,697 baseline data fields (1.5%) were incomplete, and the primary outcome was available for all patients. Mean patient age was 58 ± 13 years, and 1023 patients (42.4%) were male. Median admission KPS was 80 (IQR: 70–90). Three hundred and six patients (12.6%) had undergone prior surgery. Open craniotomy was carried out in 2326 patients (95.8%), while 101 patients (4.2%) underwent transsphenoidal surgery. In the external validation cohort, the rate of functional impairment was 28.5% (n = 692). Early mortality occurred in 74 cases (3.1%).

Model Performance

The prediction model resulted in an AUC of 0.72 (95% CI: 0.69–0.74) on the derivation cohort (**Figure 1**). A threshold of 0.205 for binary classification of functional impairment was determined based on AUC. Sensitivity and specificity of 0.73 (95% CI: 0.69–0.77) and 0.59 (0.57–0.62) were observed, respectively (**Table 2**). The prediction model was well-calibrated on the development cohort, with a calibration slope of 1.01 (95% CI: 0.87–1.15) and intercept of –0.00 (–0.10–0.10) (**Figure 2**).

In the external validation cohort, a pooled AUC of 0.72 (95% CI: 0.69–0.74) was observed. Sensitivity and specificity amounted to 0.62 (95% CI: 0.59–0.66) and 0.70 (95% CI: 0.67–0.72), respectively. Among the external validation centers, AUC values ranged from 0.54 (95% CI: 0.47–0.61) to 0.78 (95% CI: 0.73–0.82). In terms of calibration, a slope of 0.88 (95% CI: 0.77–0.99) and intercept of 0.58 (95% CI: 0.48–0.67) were observed. Location in an eloquent area, surgical approach, tumor histology, KPS at admission, and gender demonstrated the highest variable importance in the prediction model (**Supplementary Table S2**). Partial dependence plots for each variable are provided in **Supplementary Figure S1**.

Model Deployment

The prediction model was integrated into a free, user-friendly, web-based application accessible at https://neurosurgery.shinyapps.io/impairment.

Table 1. Patient characteristics and incidence of functional impairment.

Variable	Cohort	
	Development (n = 2437)	**External Validation (n = 2427)**
Male gender, n (%)	1175 (48.2%)	1023 (42.4%)
No. missing	*0 (0.0%)*	*12 (0.5%)*
Age [yrs.]		
Mean ± SD	54.6 ± 15.3	58.2 ± 13.3
Median (IQR)	55 (44 - 67)	59 (49 - 68)
Range	18 - 92	18 - 91
No. missing	*0 (0.0%)*	*2 (0.1%)*
Max. tumor diameter [cm]		
Mean ± SD	3.5 ± 1.6	3.7 ± 1.7
Median (IQR)	3.2 (2.3 - 4.5)	3.5 (2.5 - 4.9)
Range	0.1 - 10.0	0.3 - 10.2
No. missing	*0 (0.0%)*	*3 (0.1%)*
Histology, n (%)		
Meningioma	636 (26.1%)	1348 (55.5%)
Glioblastoma	514 (21.1%)	554 (22.8%)
Metastasis	324 (13.3%)	259 (10.7%)
Adenoma	243 (10.0%)	103 (4.2%)
Low-grade glioma	121 (5.0%)	44 (1.8%)
Schwannoma	120 (4.9%)	35 (1.4%)
Anaplastic astrocytoma	112 (4.6%)	48 (2.0%)
Craniopharyngioma	39 (1.6%)	2 (0.1%)
(Epi-)Dermoid cyst	30 (1.2%)	6 (0.2%)
Chordoma	25 (1.0%)	0 (0.0%)
Other	273 (11.2%)	28 (1.2%)
No. missing	*0 (0.0%)*	*0 (0.0%)*
Prior surgery, n (%)	440 (18.1%)	306 (12.6%)
No. missing	*0 (0.0%)*	*2 (0.1%)*
Open craniotomy, n (%)	2148 (88.1%)	2326 (95.8%)
No. missing	*0 (0.0%)*	*0 (0.0%)*
Surgery in eloquent area, n (%)	1197 (49.1%)	879 (36.2%)
No. missing	*0 (0.0%)*	*1 (0.0%)*
Brain vessel manipulation, n (%)	898 (36.8%)	995 (41.0%)
No. missing	*0 (0.0%)*	*185 (7.6%)*
Cranial nerve manipulation, n (%)	715 (29.3%)	487 (20.1%)
No. missing	*0 (0.0%)*	*185 (7.6%)*
Surgery in posterior fossa, n (%)	413 (16.9%)	361 (14.9%)
No. missing	*0 (0.0%)*	*1 (0.0%)*
KPS at admission		
Mean ± SD	84.3 ± 13.9	82.0 ± 13.9
Median (IQR)	90 (80 - 90)	80 (70 - 90)
Range	20 - 100	10 - 100
No. missing	*0 (0.0%)*	*1 (0.0%)*
New functional impairment[a], n (%)	525 (21.5%)	692 (28.5%)

SD, standard deviation; KPS, Karnofsky Performance Status; IQR, interquartile range.

[a] New functional impairment was defined as a ≥ 10 point decrease in KPS from baseline to the 3-month follow-up.

Table 2. Discrimination and calibration metrics of the machine learning-based prediction model.

Metric	Cohort	
	Development (n = 2437)	**External Validation (n = 2427)**
Discrimination		
AUC	0.72 (0.69 - 0.74)	0.72 (0.69 - 0.74)
Accuracy	0.62 (0.60 - 0.64)	0.68 (0.66 - 0.69)
Sensitivity	0.73 (0.69 - 0.77)	0.62 (0.59 - 0.66)
Specificity	0.59 (0.57 - 0.62)	0.70 (0.67 - 0.72)
PPV	0.33 (0.30 - 0.36)	0.45 (0.42 - 0.48)
NPV	0.89 (0.87 - 0.90)	0.82 (0.80 - 0.84)
Calibration		
Intercept	-0.00 (-0.10 - 0.10)	0.58 (0.48 - 0.67)
Slope	1.01 (0.87 - 1.15)	0.88 (0.77 - 0.99)

Metrics are provided with bootstrapped 95% confidence intervals.
AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value.

Discussion

Prediction tools can assist in the shared surgical decision-making process.[11–14] Compared to other pathologies, where scoring systems are broadly applied to estimate postoperative outcome (e.g. for arteriovenous malformations[30] or intracranial aneurysms[15]), there is little research on classification or prediction tools for postoperative functional impairment after resection of intracranial tumors. In addition, what is known about postoperative functional impairment usually focuses on a particular histopathological entity instead of principles that apply to various kinds of intracranial neoplastic lesions. The Milan Complexity Scale is a classification system based on objective surgical complexity, which correlates with the risk of functional impairment.[2] The scale can help judge case complexity and thus provides benchmarks for complication risk, resident training, and health system management.[31] We expanded on this concept by applying ML techniques to multicentric data and incorporating additional variables in a nonlinear fashion. Learning of non-linear structures in the data may reveal patterns that linear models are blind to, potentially leading to better predictions.[14]

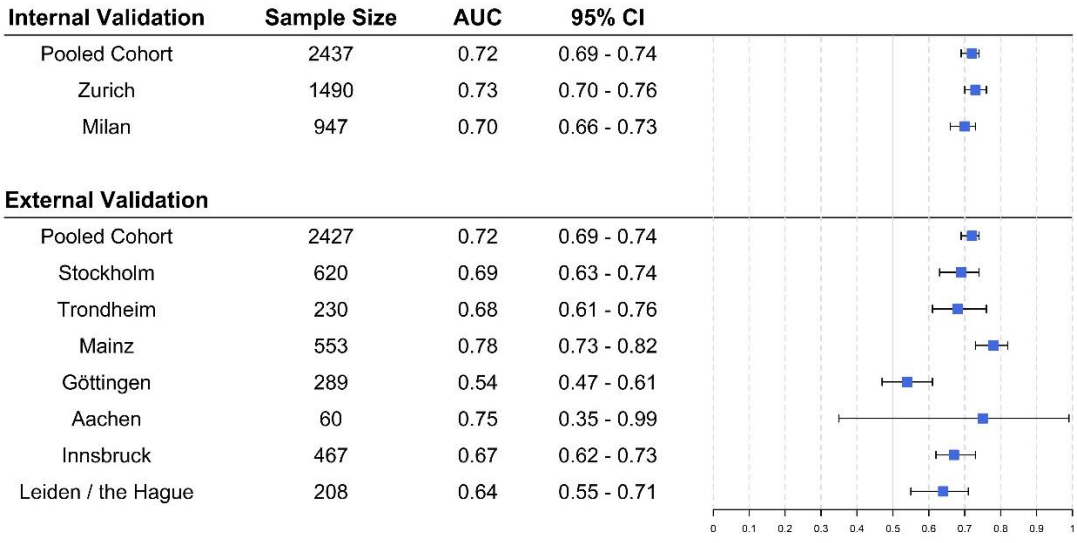


Figure 1. Area under the receiver operating characteristic (AUC) values of the prediction model among the different centers. AUC values are provided with bootstrapped 95% confidence intervals.

No tools exist to enable the prediction of an individual patient's risk of functional impairment after intracranial tumor surgery. Experienced clinicians are proficient at judging this risk by integrating clinical and imaging findings and the proposed procedure into their personal pool of experience. However, studies assessing the accuracy of these subjective predictions have raised concern about the accuracy of the information available to patients at preoperative informed consent. It appears that neurosurgeons tend to overestimate patients' postoperative functional status.[10] Our study provides a first objective benchmark of this accuracy and the functional result that can be expected by patients. The free web-based application can be used by physicians and patients alike as a basis for individual case-by-case discussions of the risk-to-benefit estimation of surgical treatment.

From specific pathologies such as pituitary adenomas, we know that classification systems and experienced clinicians are usually adept at identifying patients who are either at very high or low risk of a certain endpoint.[2,16] Thus, they excel at identifying extreme cases, such as large glioblastomas in eloquent areas, but are less successful in differentiating between good and bad outcomes in cases with moderate risk, such as diffuse low-grade gliomas in non-eloquent areas but adjacent to critical structures. The hope is that ML enables better differentiation in these moderate cases, leading to more accurate predictions.[16] This notion is corroborated by a systematic review demonstrating that artificial intelligence, including ML, is often superior to experienced raters (coined "natural intelligence") in terms of neurosurgical decision making.[32] Notably, in studies where clinical experts assisted by ML models were compared to clinical experts alone, the ML-assisted group consistently performed better.[32]

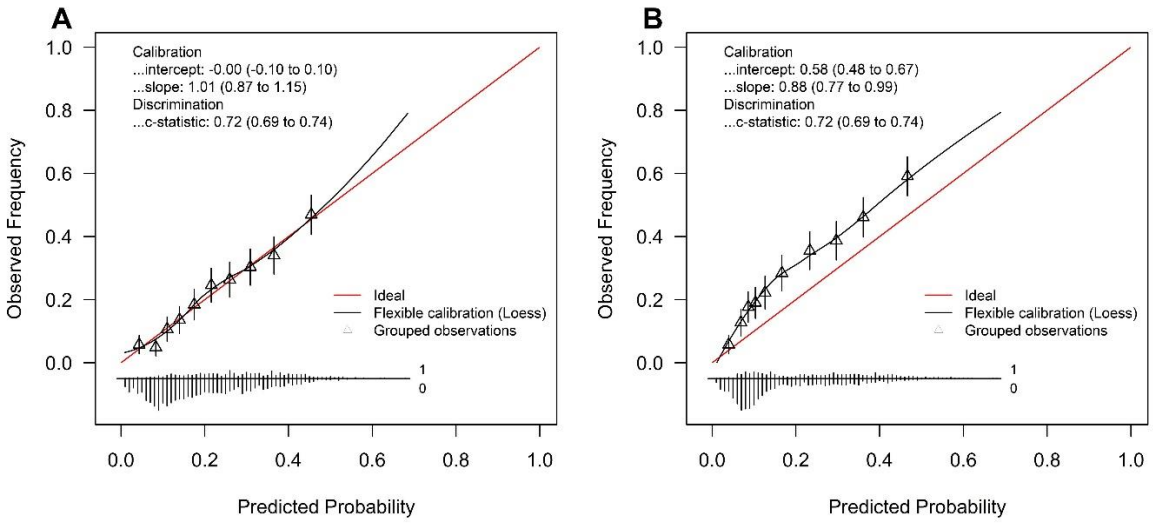


Figure 2. Calibration curves of the prediction model on the internal (Panel A) and external (Panel B) validation cohorts. The predicted probabilities for functional impairment are distributed into ten equally sized groups, and contrasted with the actually observed frequencies of functional impairment. Calibration intercept and slope are calculated. A perfectly calibrated model has a calibration intercept of 0 and slope of 1. The calibration intercept is influenced by the frequency of the outcome of interest in a certain population. Metrics are provided with bootstrapped 95% confidence intervals.

This underlines that prediction models such as ours are not meant to be used as absolute red or green lights, but rather as a supplement to neurosurgeons' clinical expertise. The current model mainly provides the ability to rule out functional impairment at 3 to 6 months postoperatively, due to its relatively high negative predictive value (NPV). However, the objective risk estimates produced by the model are more informative than the derived binary classifications. For example, a predicted risk of functional impairment of 55% may not accurately classify patients in a binary fashion but may be useful to communicate a

relatively high risk of impairment to a patient. The risk estimates our model calculates appear well-calibrated. In the external validation cohort, major heterogeneities were observed, including a higher rate of new functional impairment, which explains the calibration intercept of 0.58 observed at external validation. This would mean that – because the incidence of functional impairment was 33% higher in the external validation cohort – the model slightly underestimates functional impairment in this new cohort. For example, in a different cohort with a massively increased incidence of functional impairment of 42%, the model would predict an impairment risk of 10%, while the actual risk would be around 20%. This phenomenon is frequently observed, and in fact unavoidable unless the variables that explain the increased rate of functional impairment, such as potentially center caseload or surgeon experience, et cetera are included in the model.[29,33] The calibration intercept at external validation can be artificially improved by recalibrating onto the new population by changing mode intercepts. We chose not to recalibrate our model to the external validation data in order to evaluate its external validity in a more realistic setup. Still, the calibration of our model appears to generalize well in terms of slope, and when applying the prediction model to different demographics with different rates of new functional impairment, the model can be recalibrated by updating its intercept accordingly or by other rescaling techniques.[29,33]

Even with a large amount of development data and the application of machine learning techniques, functional impairment after intracranial tumor surgery remains difficult to predict with high reliability. One likely cause is the lack of functional anatomo-topographical data as inputs for our model, which was designed to include only a few simple, preoperatively and easily available variables. This was intended to keep it applicable to primary care and other non-neurosurgery physicians, who are typically the first and most important contact for patients facing the new diagnosis of an intracranial tumor. The introduction of anatomical features and the ability to account for intraoperative parameters and complications in a second postoperative model would surely improve performance to some extent.

In the case of intracranial tumor surgery, a key factor for variability is the use of different treatment protocols. Different surgical approaches, availability of intraoperative imaging, functional mapping, and fluorescents, as well as varying “aggressiveness” in terms of resection but also handling of critical structures introduces biases that are difficult to statistically account or adjust for.[3–5,18–21] Depending on case complexity, surgical experience may also influence outcome.[31] Even an externally validated prediction model lacks generalizability to cohorts with radically different treatment protocols.

An often-cited drawback of ML models is the inability to understand why a certain prediction has been generated. Whereas logistic regression models provide interpretable odds ratios, ML models are often considered “black boxes”—that is, inputs and outputs are known, but the internal decision-making process is not necessarily interpretable. Some insight can be gained by assessing overall variable importance (*eTable 2*). Additionally, GAMs are somewhat of an exception, since one can exploit their inherent additivity to examine each variable for the purpose of inference (see *eFigure 1*).[23,24] Surgery in eloquent areas may double the rate of postoperative functional impairment, as high-grade tumors do,[2,7,34] and preoperative status has been demonstrated to relate to complications and outcome.[2,6,35] It is not always feasible for clinicians to integrate these many independent risk factors into a single communicable risk for outcomes such as impairment. Prediction tools represent an interface between these patient factors with complex interactions and output a risk that is interpretable and clinically useful to clinicians and patients alike.[12,13]

Decision-making for intracranial tumor surgery requires balancing oncological benefit against the risk of resection-related impairment. Our study demonstrates that ML-based prediction of functional impairment is feasible and externally valid with simple inputs. Integrating artificial intelligence as

supportive means into the clinical routine is likely to provide valuable improvements in patient information, objective risk assessment, and shared surgical decision making.

Strengths and Limitations

Our study used datasets from nine large institutional registries of national referral centers, encompassing several different cultural and linguistic regions. Variable definitions were unified in all centers, allowing us to generate results with fair external validity and generalizability. The primary outcome of our study was based on a clearly defined and well-established outcome measure that correlates with PROMs.[6,7,36] The final model is accessible as a free web-based tool, allowing clinicians and patients to access the objective risk estimates.

A range of tumor types were analyzed, which may bias our prediction model towards more common tumor types, whereas performance may be limited for the less frequently included tumor types. However, the resulting model enables outcome prediction for most major classes of intracranial tumors. In addition, one might expect especially pituitary adenomas and recurrent craniotomies to exhibit an inherently different risk profile, potentially limiting performance of the model – However, we found that their inclusion did not alter overall model performance. In addition, the local regression algorithm on which our model relies is limited in terms of extrapolation to unseen, extreme input variable values.[23,24] For this reason, predictions made from inputs not available in the derivation data, such as ages over 92 and tumor sizes over 10 cm, should be cautiously interpreted.

Although external validation was successful, no conclusions can be drawn regarding performance in centers with radically different resection protocols and vastly different rates of new functional impairment. The high negative predictive value can be seen as one of the model's strengths. However, predictive values are inherently dependent on the prevalence of the outcome and as such, the setting in which the prognostic model is used.[29] The predictive values should therefore be interpreted with caution, especially when generalizing to other centers.

Although all participating centers followed a "maximum safe resection" philosophy, potential nuances in EOR may persist, which were not accounted for.[5] We only assessed outcomes at 3 to 6 months postoperatively, and the outcome definition did not include further, relevant aspects such as quality of life, cognitive or work status, and PROMs. Additionally, as with most outcome measures, the interrater agreement of the KPS has been debated, with generally better interrater agreement compared ECOG and palliative performance status (PPS).[37] Lastly, the study protocol of this analysis was not prospectively registered.

Conclusions

Functional impairment after intracranial tumor surgery is extraordinarily difficult to predict preoperatively. A machine learning-based approach resulted in a prediction model capable of forecasting individualized risk for any functional impairment at 3 to 6 months postoperatively with fair performance. Extensive external validation demonstrated the high generalizability of the prediction model. To our knowledge, this study is the first externally validated attempt at preoperatively quantifying the "patient-specific" surgical risk for any functional impairment after intracranial tumor surgery. The web-based application can be used by physicians and patients alike, serving as a basis for case-by-case discussions on the risk-to-benefit estimation of surgical treatment.

Supplementary Table S1. Patient characteristics and incidence of functional impairment among centers.

Variable	Internal Validation		External Validation						
	Zurich (n = 1490)	**Milan (n = 947)**	**Stockholm (n = 620)**	**Trondheim (n = 230)**	**Mainz (n = 553)**	**Göttingen (n = 289)**	**Aachen (n = 60)**	**Innsbruck (n = 467)**	**Leiden / the Hague (n = 208)**
Years of data collection	2013 - 2017	2014 - 2017	2007 - 2015	2007 - 2015	2007 - 2018	2014 - 2017	2018	2015 - 2018	2015 - 2018
Data source	PR	PR	RC	PR/RC	RC	PR	RC	PR	PR
Male gender, n (%)	723 (48.5%)	452 (47.7%)	179 (28.9%)	78 (33.9%)	195 (35.3%)	169 (58.5%)	34 (56.7%)	230 (49.3%)	138 (66.3%)
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	1 (0.4%)	0 (0.0%)	11 (2.4%)	0 (0.0%)
Age [yrs.]									
Mean ± SD	55.5 ± 15.4	53.3 ± 14.9	57.3 ± 12.0	56.5 ± 13.8	58.3 ± 13.6	63.4 ± 10.7	57.5 ± 13.4	56.1 ± 15.2	60.3 ± 12.7
Median (IQR)	57 (45 - 67)	54 (42 - 66)	58 (48 - 66)	58 (49 - 65)	59 (49 - 69)	64 (57 - 71)	58 (52 - 64)	57 (46 - 69)	62 (54 - 70)
Range	18 - 92	18 - 85	26 - 85	18 - 91	22 - 85	26 - 88	22 - 87	18 - 88	20 - 82
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*2 (0.4%)*	*0 (0.0%)*
Max. tumor diameter [cm]									
Mean ± SD	3.6 ± 1.7	3.5 ± 1.5	3.6 ± 1.5	3.7 ± 1.7	3.4 ± 1.8	3.7 ± 1.7	2.1 ± 0.9	3.8 ± 1.5	5.2 ± 1.5
Median (IQR)	3.3 (2.2 - 4.6)	3.0 (2.5 - 4.5)	3.3 (2.5 - 4.6)	3.4 (2.5 - 4.8)	3.1 (2.0 - 4.6)	3.5 (2.5 - 4.8)	1.9 (1.5 - 2.6)	3.6 (2.6 - 4.6)	5.3 (4.1 - 6.3)
Range	0.1 - 9.6	0.1 - 10.0	1.0 - 8.5	0.7 - 8.6	0.3 - 10.2	0.8 - 10.0	0.5 - 4.3	1.2 - 8.7	1.0 - 8.9
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*2 (3.3%)*	*1 (0.2%)*	*0 (0.0%)*
Histology, n (%)									
Meningioma	364 (24.4%)	272 (28.7%)	620 (100%)	230 (100%)	364 (65.8%)	0 (0.0%)	0 (0.0%)	134 (28.7%)	0 (0.0%)
Glioblastoma	298 (20.0%)	216 (22.8%)	0 (0.0%)	0 (0.0%)	125 (22.6%)	114 (39.4%)	0 (0.0%)	108 (23.1%)	207 (99.5%)
Metastasis	265 (17.8%)	59 (6.23%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	175 (60.6%)	0 (0.0%)	84 (18.0%)	0 (0.0%)
Adenoma	148 (9.93%)	95 (10.0%)	0 (0.0%)	0 (0.0%)	25 (4.52%)	0 (0.0%)	57 (95.0%)	21 (4.5%)	0 (0.0%)
Low-grade glioma	114 (7.65%)	7 (0.74%)	0 (0.0%)	0 (0.0%)	6 (1.08%)	0 (0.0%)	0 (0.0%)	38 (8.1%)	0 (0.0%)
Schwannoma	72 (4.83%)	48 (5.07%)	0 (0.0%)	0 (0.0%)	18 (3.25%)	0 (0.0%)	0 (0.0%)	17 (3.6%)	0 (0.0%)
Anaplastic astrocytoma	74 (4.97%)	38 (4.01%)	0 (0.0%)	0 (0.0%)	10 (1.81%)	0 (0.0%)	0 (0.0%)	38 (8.1%)	0 (0.0%)
Craniopharyngioma	27 (1.81%)	12 (1.27%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (0.4%)	0 (0.0%)
(Epi-)Dermoid cyst	15 (1.01%)	15 (1.58%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (1.3%)	0 (0.0%)
Chordoma	3 (0.20%)	22 (2.32%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Other	110 (7.38%)	163 (17.2%)	0 (0.0%)	0 (0.0%)	5 (0.90%)	0 (0.0%)	3 (5.0%)	19 (4.1%)	1 (0.5%)
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*
Prior surgery, n (%)	301 (20.2%)	139 (14.7%)	88 (14.2%)	50 (21.7%)	62 (11.2%)	8 (2.8%)	5 (8.3%)	74 (15.8%)	19 (9.1%)
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	1 (0.4%)	0 (0.0%)	1 (0.2%)	0 (0.0%)
Open craniotomy, n (%)	1321 (88.7%)	827 (87.3%)	620 (100%)	230 (100%)	528 (95.5%)	289 (100%)	3 (5.00%)	448 (95.9%)	208 (100%)
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*

Supplementary Table S1 (continued). Patient characteristics and incidence of functional impairment among centers.

Variable	Internal Validation		External Validation						
	Zurich (n = 1490)	**Milan (n = 947)**	**Stockholm (n = 620)**	**Trondheim (n = 230)**	**Mainz (n = 553)**	**Göttingen (n = 289)**	**Aachen (n = 60)**	**Innsbruck (n = 467)**	**Leiden / the Hague (n = 208)**
Surgery in eloquent area, n (%)	776 (52.1%)	421 (44.5%)	158 (25.5%)	86 (37.4%)	191 (34.5%)	125 (43.3%)	0 (0.0%)	194 (41.5%)	125 (60.1%)
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*1 (0.2%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*
Brain vessel manipulation, n (%)	692 (46.4%)	206 (21.8%)	358 (57.7%)	143 (62.2%)	171 (30.9%)	23 (8.0%)	3 (5.0%)	268 (57.4%)	29 (13.9%)
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	184 (63.7%)	0 (0.0%)	0 (0.0%)	1 (0.5%)
Cranial nerve manipulation, n (%)	456 (30.6%)	259 (27.3%)	151 (24.4%)	66 (28.7%)	168 (30.4%)	1 (0.4%)	4 (6.7%)	90 (19.3%)	7 (3.4%)
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	184 (63.7%)	0 (0.0%)	1 (0.21%)	0 (0.0%)
Surgery in posterior fossa, n (%)	241 (16.2%)	172 (18.2%)	68 (11.0%)	29 (12.6%)	96 (17.4%)	74 (25.6%)	0 (0.0%)	90 (19.3%)	4 (1.9%)
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	0 (0.0%)	0 (0.0%)	1 (0.21%)	0 (0.0%)
KPS at admission									
Mean ± SD	81.9 ± 14.5	88.1 ± 11.8	79.3 ± 15.4	75.8 ± 12.4	81.3 ± 11.6	79.0 ± 15.5	91.0 ± 14.1	89.8 ± 10.4	83.0 ± 12.2
Median (IQR)	90 (80 - 90)	90 (80 - 100)	80 (70 - 90)	70 (70 - 90)	80 (70 - 90)	80 (70 - 90)	100 (90-100)	90 (80 - 100)	80 (80 - 90)
Range	20 - 100	30 - 100	10 - 100	40 - 100	40 - 100	20 - 100	40 - 100	20 - 100	10 - 100
No. missing	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*	*1 (1.7%)*[a]	*0 (0.0%)*	*0 (0.0%)*
New functional impairment[b], n (%)	310 (20.8%)	215 (22.7%)	120 (19.4%)	67 (29.1%)	162 (29.3%)	121 (41.9%)	3 (5.0%)	103 (22.1%)	116 (55.8%)

PR, prospective registry; RC, retrospective collection; SD, standard deviation; KPS, Karnofsky Performance Status; IQR, interquartile range.

[a] The 3 to 6 month KPS for this patient with missing admission KPS was 100. Thus, it was concluded that no new functional impairment had occurred.

[b] New functional impairment was defined as a ≥ 10 point decrease in KPS from baseline to the 3-month follow-up.

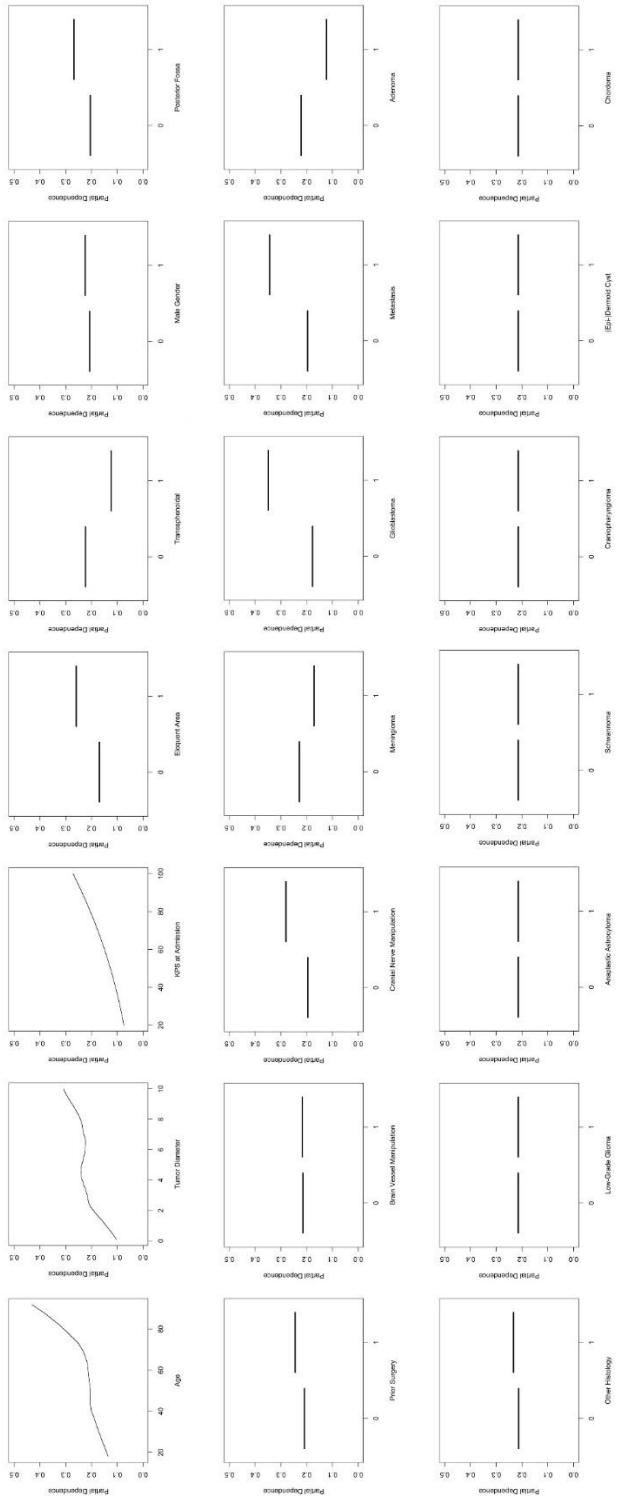
Supplementary Table S2. Variable importance in the fully trained model. Importance measurements were based on AUC. The variables are ordered by importance.

Variable	AUC-Based Importance
Eloquent area	7.94
Surgical approach	7.81
Histology	
Meningioma	6.60
Glioblastoma	5.68
Metastasis	5.16
Adenoma	1.20
Others	0.21
Low-grade glioma	0.00
Anaplastic astrocytoma	0.00
Schwannoma	0.00
Craniopharyngioma	0.00
(Epi-)Dermoid cyst	0.00
Chordoma	0.00
KPS at admission	3.98
Male gender	2.25
Age	1.15
Tumor diameter	1.13
Posterior fossa	1.12
Prior surgery	0.69
Brain vessel	0.32
Cranial nerve	0.16

AUC, area under the curve; KPS, Karnofsky Performance Status.

A series of cut-offs is applied to the prognostic factors to predict the outcome. Sensitivity and specificity are calculated for each cut-off and the corresponding AUC curve is drawn. The trapezoidal rule is used to compute AUC. This AUC is used as the measure of variable importance.

Supplementary Figure S1. Partial dependence plots for each variable contained in the fully trained model.



The partial dependence plots demonstrate the marginal effect that a variable has on the predicted outcome of a prediction model. A partial dependence plot can show whether the association between the predictor variable and the outcome of interest is linear, monotonous or more complex in the trained prediction model. In this way, partial dependence helps explain a model's internal decision-making process, and thus fosters interpretability of prediction models.
KPS, Karnofsky Performance Status;

Acknowledgements

We thank the patients whose anonymized data were used for this research.
The persons listed below contributed in establishing the data collection:

- *Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland:* David Y. Zhang, Dominik Seggewiss.
- *Department of Neurosurgery, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy:* Stefano Villa.
- *Department of Neurosurgery, Haaglanden Medical Center, The Hague, The Netherlands and Department of Neurosurgery, Leiden University Medical Center, Leiden, The Netherlands:* Rishi Nandoe Tewarie, Fred Kloet

Disclosures

Conflict of Interest: The authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
Grants and Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Barker FG, Curry WT, Carter BS. Surgery for primary supratentorial brain tumors in the United States, 1988 to 2000: the effect of provider caseload and centralization of care. *Neuro-Oncol*. 2005;7(1):49-63. doi:10.1215/S1152851704000146
2. Ferroli P, Broggi M, Schiavolin S, et al. Predicting functional impairment in brain tumor surgery: the Big Five and the Milan Complexity Scale. *Neurosurg Focus*. 2015;39(6):E14. doi:10.3171/2015.9.FOCUS15339
3. Yordanova YN, Moritz-Gasser S, Duffau H. Awake surgery for WHO Grade II gliomas within "noneloquent" areas in the left dominant hemisphere: toward a "supratotal" resection. Clinical article. *J Neurosurg*. 2011;115(2):232-239. doi:10.3171/2011.3.JNS101333
4. Sanai N, Berger MS. Glioma extent of resection and its impact on patient outcome. *Neurosurgery*. 2008;62(4):753-766. doi:10.1227/01.neu.0000318159.21731.cf
5. Marko NF, Weil RJ, Schroeder JL, Lang FF, Suki D, Sawaya RE. Extent of resection of glioblastoma revisited: personalized survival modelling facilitates more accurate survival prediction and supports a maximum-safe-resection approach to surgery. *J Clin Oncol*. 2014;32(8):774-782. doi:10.1200/JCO.2013.51.8886
6. Stienen MN, Zhang DY, Broggi M, et al. The influence of preoperative dependency on mortality, functional recovery and complications after microsurgical resection of intracranial tumors. *J Neurooncol*. 2018;139(2):441-448. doi:10.1007/s11060-018-2882-9
7. Schiavolin S, Raggi A, Scaratti C, et al. Patients' reported outcome measures and clinical scales in brain tumor surgery: results from a prospective cohort study. *Acta Neurochir (Wien)*. 2018;160(5):1053-1061. doi:10.1007/s00701-018-3505-0
8. Rahman M, Abbatematteo J, De Leo EK, et al. The effects of new or worsened postoperative neurological deficits on survival of patients with glioblastoma. *J Neurosurg*. 2017;127(1):123-131. doi:10.3171/2016.7.JNS16396
9. Jakola AS, Gulati S, Weber C, Unsgård G, Solheim O. Postoperative deterioration in health related quality of life as predictor for survival in patients with glioblastoma: a prospective study. *PloS One*. 2011;6(12):e28592. doi:10.1371/journal.pone.0028592
10. Sagberg LM, Drewes C, Jakola AS, Solheim O. Accuracy of operating neurosurgeons' prediction of functional levels after intracranial tumor surgery. *J Neurosurg*. 2017;126(4):1173-1180. doi:10.3171/2016.3.JNS152927
11. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181
12. Galovic M, Stauber AJ, Leisi N, et al. Development and Validation of a Prognostic Model of Swallowing Recovery and Enteral Tube Feeding After Ischemic Stroke. *JAMA Neurol*. Published online February 11, 2019. doi:10.1001/jamaneurol.2018.4858
13. Khor S, Lavallee D, Cizik AM, et al. Development and Validation of a Prediction Model for Pain and Functional Outcomes After Lumbar Spine Surgery. *JAMA Surg*. 2018;153(7):634-642. doi:10.1001/jamasurg.2018.0072
14. Senders JT, Staples PC, Karhade AV, et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg*. 2018;109:476-486.e1. doi:10.1016/j.wneu.2017.09.149
15. Jaja BNR, Saposnik G, Lingsma HF, et al. Development and validation of outcome prediction models for aneurysmal subarachnoid haemorrhage: the SAHIT multinational cohort study. *BMJ*. 2018;360:j5745. doi:10.1136/bmj.j5745
16. Staartjes VE, Serra C, Muscas G, et al. Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study. *Neurosurg Focus*. 2018;45(5):E12. doi:10.3171/2018.8.FOCUS18243

17. Sarnthein J, Stieglitz L, Clavien P-A, Regli L. A Patient Registry to Improve Patient Safety: Recording General Neurosurgery Complications. *PloS One*. 2016;11(9):e0163154. doi:10.1371/journal.pone.0163154
18. Stummer W, Stepp H, Wiestler OD, Pichlmeier U. Randomized, Prospective Double-Blinded Study Comparing 3 Different Doses of 5-Aminolevulinic Acid for Fluorescence-Guided Resections of Malignant Gliomas. *Neurosurgery*. 2017;81(2):230-239. doi:10.1093/neuros/nyx074
19. Kubben PL, ter Meulen KJ, Schijns OE, ter Laak-Poort MP, van Overbeeke JJ, Santbrink H van. Intraoperative MRI-guided resection of glioblastoma multiforme: a systematic review. *Lancet Oncol*. 2011;12(11):1062-1070. doi:10.1016/S1470-2045(11)70130-9
20. Gronningsaeter A, Kleven A, Ommedal S, et al. SonoWand, an ultrasound-based neuronavigation system. *Neurosurgery*. 2000;47(6):1373-1379; discussion 1379-1380.
21. Sanai N, Mirzadeh Z, Berger MS. Functional outcome after language mapping for glioma resection. *N Engl J Med*. 2008;358(1):18-27. doi:10.1056/NEJMoa067819
22. Nghiemphu PL, Liu W, Lee Y, et al. Bevacizumab and chemotherapy for recurrent glioblastoma. *Neurology*. 2009;72(14):1217-1222. doi:10.1212/01.wnl.0000345668.03039.90
23. Hastie T, Tibshirani R. *Generalized Additive Models*. 1st ed. Chapman and Hall; 1990.
24. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media; 2013.
25. Hastie T. *Gam: Generalized Additive Models*.; 2019. Accessed January 5, 2020. https://CRAN.R-project.org/package=gam
26. Kuhn M. Building Predictive Models in *R* Using the **caret** Package. *J Stat Softw*. 2008;28(5). doi:10.18637/jss.v028.i05
27. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. *Appl Artif Intell*. 2003;17(5-6):519-533. doi:10.1080/713827181
28. Perkins NJ, Schisterman EF. The Inconsistency of "Optimal" Cut-points Using Two ROC Based Criteria. *Am J Epidemiol*. 2006;163(7):670-675. doi:10.1093/aje/kwj063
29. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76-86. doi:10.1016/j.jclinepi.2007.04.018
30. Spetzler RF, Martin NA. A proposed grading system for arteriovenous malformations. *J Neurosurg*. 1986;65(4):476-483. doi:10.3171/jns.1986.65.4.0476
31. Vasella F, Velz J, Neidert MC, et al. Safety of resident training in the microsurgical resection of intracranial tumors: Data from a prospective registry of complications and outcome. *Sci Rep*. 2019;9(1):954. doi:10.1038/s41598-018-37533-3
32. Senders JT, Arnaout O, Karhade AV, et al. Natural and Artificial Intelligence in Neurosurgery: A Systematic Review. *Neurosurgery*. Published online September 7, 2017. doi:10.1093/neuros/nyx384
33. van Rein EAJ, van der Sluijs R, Voskens FJ, et al. Development and Validation of a Prediction Model for Prehospital Triage of Trauma Patients. *JAMA Surg*. Published online February 6, 2019. doi:10.1001/jamasurg.2018.4752
34. Duffau H, Capelle L, Denvil D, et al. Functional recovery after surgical resection of low grade gliomas in eloquent brain: hypothesis of brain compensation. *J Neurol Neurosurg Psychiatry*. 2003;74(7):901-907. doi:10.1136/jnnp.74.7.901
35. Chang SM, Parney IF, Mcdermott M, et al. Perioperative complications and neurological outcomes of first and second craniotomies among patients enrolled in the Glioma Outcome Project. *J Neurosurg*. 2003;98(6):1175-1181. doi:10.3171/jns.2003.98.6.1175
36. Reponen E, Tuominen H, Korja M. Evidence for the Use of Preoperative Risk Assessment Scores in Elective Cranial Neurosurgery: A Systematic Review of the Literature. *Anesth Analg*. 2014;119(2):420. doi:10.1213/ANE.0000000000000234

37. Chow R, Chiu N, Bruera E, et al. Inter-rater reliability in performance status assessment among health care professionals: a systematic review. *Ann Palliat Med*. 2016;5(2):83-92-92.

[Chapter 8]

FUSE-ML: Development and external validation of a clinical prediction model for mid-term outcomes after lumbar spinal fusion for degenerative disease

Victor E. Staartjes
Vittorio Stumpo
Luca Ricciardi
Nicolai Maldaner
Hubert A.J. Eversdijk
Moira Vieli
Olga Ciobanu-Caraus
Antonino Raco
Massimo Miscusi
Andrea Perna
Luca Proietti
Giorgio Lofrese
Michele Dughiero
Francesco Cultrera
Nicola Nicassio
Seong Bae An
Yoon Ha
Aymeric Amelot
Irene Alcobendas
Jose Viñuela
Maria L. Gandía-González
Pierre-Pascal Girod
Sara Lener
Nikolaus Kögl
Nico Akhavan Safa
Christoph J. Laux
Mazda Farshad
Dave O'Riordan
Markus Loibl
Anne F. Mannion
Alba Scerrati
Granit Molliqaj
Enrico Tessitore
Marc L. Schröder
W. Peter Vandertop
Martin N. Stienen
Carlo Serra
Luca Regli

Eur Spine J 2022 Feb 21. [online ahead of print]

[Abstract]

Background

Indications and outcomes in lumbar spinal fusion for degenerative disease are notoriously heterogenous. Selected subsets of patients show remarkable benefit. However, their objective identification is often difficult. Decision-making may be improved with reliable prediction of long-term outcomes for each individual patient, improving patient selection and avoiding ineffective procedures.

Methods

Clinical prediction models for long-term functional impairment (Oswestry Disability Index [ODI] or Core Outcome Measures Index [COMI]), back pain, and leg pain after lumbar fusion for degenerative disease were developed. Achievement of the minimum clinically important difference (MCID) at 12 months postoperatively was defined as a reduction from baseline of at least 15 points for ODI, 2.2 points for COMI, or 2 points for pain severity.

Results

Models were developed and integrated into a web-app (https://neurosurgery.shinyapps.io/fuseml/) based on a multinational cohort (N=817; 42.7% male; mean [SD] age: 61.19 [12.36] years). At external validation (N=298; 35.6% male; mean [SD] age: 59.73 [12.64] years), areas-under-the-curves for functional impairment (0.67, 95% confidence interval [CI]: 0.59-0.74), back pain (0.72, 95%CI: 0.64-0.79), and leg pain (0.64, 95%CI: 0.54-0.73) demonstrated moderate ability to identify patients who are likely to benefit from surgery. Models demonstrated fair calibration of the predicted probabilities.

Conclusions

Outcomes after lumbar spinal fusion for degenerative disease remain difficult to predict. Although assistive clinical prediction models can help in quantifying potential benefits of surgery and the externally validated FUSE-ML tool may aid in individualized risk-benefit estimation, truly impacting clinical practice in the era of “personalized medicine” necessitates more robust tools in this patient population.

Introduction

Degenerative disease of the lumbar spine, including chronic low back pain (CLBP), lumbar spinal stenosis (LSS), lumbar disc herniation (LDH), and degenerative lumbar spondylolisthesis are part of the top-three causes of disability in Western societies and impose significant direct and indirect socio-economic costs.[1] The standard treatment for these chronic degenerative diseases is conservative therapy including physical therapy, although certain patients who are unresponsive to long-term conservative treatment may benefit from interbody fusion, but this is controversial.[2,3] With some reports showing no benefit compared to conservative treatment in a randomized population, patient selection is vitally important.[4] Various prognostic tests exist to attempt to identify subsets of patients that might truly benefit from surgery as a "last resort", but the validity of these tests is unclear.[5,6] A relevant proportion of patients with intractable, conservative therapy-resistant lumbar degenerative disease do finally profit from lumbar fusion surgery – the difficult question is how to identify these subsets securely and how to avoid unnecessary, unsuccessful surgery.[3]

Clinical prediction models can summarize a large number of factors into a single, potentially more accurate prediction of surgical risk or benefit, tailored to each individual patient.[7–9] The implementation of machine learning (ML) is increasing exponentially, albeit methodological rigor is only seldomly upheld.[8,10] Without thorough methodological foundations, development of clinical prediction models can very easily lead to pseudo-reliable predictions with seemingly high performance measures due to issues such as data leakage, class imbalance, and overfitting.[8,11] If clinical prediction models are not properly externally validated, real-world performance cannot be adequately estimated, and they certainly ought not to be applied in clinical practice.[12,13]

For patients with degenerative disease of the lumbar spine in whom spinal fusion surgery is considered, accurate prediction of long-term outcome in individual patients has been demonstrated to be extraordinarily difficult.[5,14] The aim of the FUSE-ML consortium was to assemble a large multinational dataset of patients undergoing lumbar spinal fusion for degenerative disease. We intended to create robust clinical prediction models that take into account surgical variables and that are thoroughly developed and externally validated in a range of international centers.

Methods

Overview

A substantial multinational (7 countries), multicenter (11 centers) dataset (FUSE-ML) of patients who underwent lumbar spinal fusion for degenerative disease was applied to develop and externally validate a ML-based prediction tool for long-term patient-reported functional impairment, back pain, and leg pain. We then briefly compare the performance to that of the – to our knowledge – only other comparable, externally validated, clinical prediction model that is comparable.[14] This study adheres to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines.[7] The use of patient data for research purposes was approved by each local institutional review board (IRBs), and patients provided informed consent or informed consent was waived, depending on the demands of the local IRB.

Inclusion and Exclusion Criteria

Patients with the following indications for thoracolumbar pedicle screw placement were considered for inclusion: degenerative pathologies (one or multiple of the following: spinal stenosis, spondylolisthesis, degenerative disc disease, recurrent disc herniation, failed back surgery syndrome (FBSS), radiculopathy, pseudarthrosis). Exclusion criteria were: surgery for – as the primary indication – infections, vertebral

tumors, as well as traumatic and osteoporotic fractures or deformity surgery for scoliosis or kyphosis; moderate or severe scoliosis (Coronal Cobb's >30 degrees / Schwab classification sagittal modifier + or ++); surgery at more than 6 vertebral levels; missing endpoint data at 12 months; lack of informed consent; age < 18 years-old.

Data Collection

Each center collected data either retrospectively, from a prospective registry, or from a prospective registry supplemented by retrospectively collected variables, with complete long-term follow-up. The following data were collected: age, gender, surgical indication, index level(s), height, weight, BMI, smoking status, American Society of Anesthesiologists (ASA) Score, preoperative use of opioid pain medication, asthma pulmonale as a comorbidity, prior thoracolumbar spine surgery, race/ethnicity, surgical approach, pedicle screw insertion and minimally invasive technique. In terms of PROMs, we collected preoperative (baseline) and 12-month postoperative Oswestry Disability Index (ODI) [scaled from 0 to 100] or Core Outcome Measures Index (COMI) for subjective functional impairment, numeric rating scale (NRS) for back pain severity, and NRS for leg pain severity.[15,16]

Primary Endpoint Definitions

Clinically relevant improvements in terms of functional impairment (ODI or COMI) and back/leg pain were dichotomized using the minimum clinically important difference (MCID) according to validated thresholds (Improvement from baseline to 12 months postoperatively of ≥15 points for ODI, ≥2.2 points for COMI, and ≥2 points for NRS pain severity).[17–19] Thus, improvements from baseline that are greater than these validated thresholds were counted as achievement of MCID in the respective score. A preoperative or postoperative ODI of ≤ 22[20], COMI of ≤ 3.05[21], or NRS pain severity of ≤ 3[16] was considered as a probable "patient acceptable symptom state" (PASS)[22] based on established cut-offs.

Clinical Prediction Modelling

Numerical input variables were standardized using centering and scaling, and Yeo-Johnson transformation, and highly correlated variables (Pearson correlation coefficient ≥0.8) were filtered. Patients with a preoperative PASS (minimal symptoms) in one of the three outcome dimensions were excluded from training for that respective dimension. Recursive feature elimination (RFE) based on generalized linear models (GLMs) was carried out to identify the optimal, parsimonious set of inputs for each of the three models. Subsequently, GLMs were trained using Elastic Net Regularization using the Caret[23] library. During training, hyperparameters were tuned using 5-fold cross-validation with 10 repeats, maximizing area-under-the-curve (AUC). A *k*-nearest neighbor imputer was trained to impute missing data. The threshold for binary classification was selected based on the "closest-to-(0,1)-criterion" and rounded. The models were then integrated into a web-app and underwent external validation. No recalibration was carried out. Quantile-based 95% confidence intervals (CIs) of the discrimination and calibration metrics were obtained from 1000 bootstrap resamples. Standardized model coefficients are reported to allow for explanation.[23] Finally, the models reported by Khor et al.[14] were reconstructed from the published coefficients and external validation performance was compared. Notably, the Khor et al. model takes insurance status, which was not available within the FUSE-ML consortium. As has been done previously and due to the fact that virtually all inclusions in the FUSE-ML dataset stem from countries with either single-payer healthcare or compulsory health insurance, we adopted "*Medicare/Medicaid*" as the most appropriate choice for the entire cohort.[12] All analyses were carried out in *R* version 4.1.1.

Results

Patient Cohort

Data from 1115 patients were provided by 11 participating centers in total. The development cohort was made up of eight centers (817 patients, 42.7% male, age: 61.19 ± 12.36 years), while the remaining three centers carried out external validation (298 patients, 35.6% male, age: 59.73 ± 12.64 years). Achievement of MCID at 12-months was recorded in 761 (68.3%) patients for functional impairment, 862 (77.3%) patients for back pain severity, and 796 (71.4%) patients for leg pain severity. An overview of patient characteristics is provided in **Table 1**, and detailed patient characteristics including missingness and data per center are shown in **Supplementary Table 1**. Overall, 3074 of 52'405 baseline data fields (5.9%) were incomplete.

Performance Evaluation

Detailed model performance, including resampled development and external validation performance, is summarized in **Table 2**, and standardized model coefficients – enabling judgement of variable importance – are provided in **Table 3**. Calibration plots generated from the external validation cohort are shown in **Figure 1** including resampled training calibration, external validation calibration, and calibration from the Khor et al. model applied to the FUSE-ML external validation cohort. A detailed performance comparison with the Khor et al. model is available in **Supplementary Table 2**.

Prediction of Functional Impairment

At external validation, the FUSE-ML prediction model for clinical success in terms of functional impairment achieved an AUC of 0.67 (95% CI: 0.59 – 0.74), sensitivity of 0.59 (95% CI: 0.52 – 0.66) and specificity of 0.66 (95% CI: 0.55 – 0.77). In terms of calibration, we measured a calibration intercept -0.07 (95% CI: -0.36 – 0.22) of and a calibration slope of 0.63 (95% CI: 0.34 – 0.93). When studying the standardized model coefficients, it becomes clear that predictions were mostly driven by greater baseline impairment, greater age, lower back pain severity preoperatively, and application of a lateral surgical approach. The Khor et al. model achieved an AUC of 0.71 (95% CI: 0.64 – 0.77) on the same external validation cohort.

Prediction of Back Pain Severity

Prediction of clinical success in terms of back pain severity was achieved with an externally validated AUC of 0.72 (95% CI: 0.64 – 0.79), sensitivity of 0.72 (95% CI: 0.65 – 0.77) and specificity of 0.64 (95% CI: 0.51 – 0.78). Calibration intercept -0.38 (95% CI: -0.70 – 0.06) and slope 1.10 (95% CI: 0.62 – 1.57) were evaluated. Higher baseline back pain and a lateral surgical approach were assigned the highest importance by the model. Similarly, the Khor et al. model demonstrated an AUC of 0.73 (95% CI: 0.65 – 0.79) at external validation.

Prediction of Leg Pain Severity

At external validation, we predicted long-term leg pain severity with an AUC of 0.64 (95% CI: 0.54 – 0.73), sensitivity of 0.76 (95% CI: 0.71 – 0.82) and specificity of 0.42 (95% CI: 0.26 – 0.57). In terms of calibration, we measured a calibration intercept 0.14 (95% CI: -0.22 – 0.51) of and a calibration slope of 0.49 (95% CI: -0.12 – 0.86). Looking at model coefficients, it appears that greater baseline leg pain, a posterior surgical approach, and absence of prior thoracolumbar surgery contributed most to the predictions of leg pain. The Khor et al. model performed similarly with an AUC of 0.63 (95% CI: 0.54 – 0.71) on the same data.

Model Deployment

The prediction model was integrated into a freely available, web-based application accessible at https://neurosurgery.shinyapps.io/fuseml/.

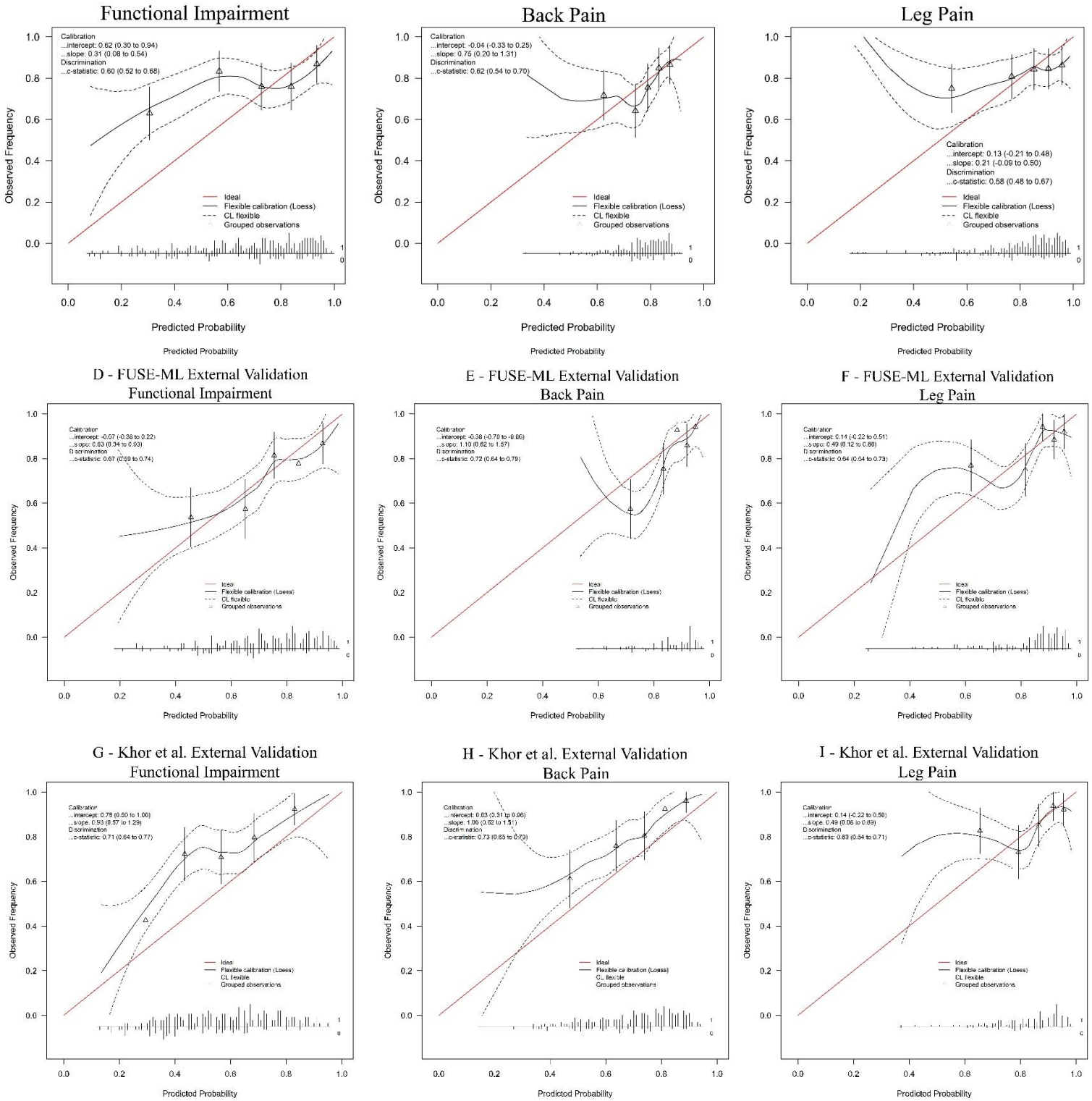


Figure 1. Calibration curves of the three clinical prediction models for function, back pain, and leg pain on the resampled development cohort (**A-C**, cross-validation performance), the external validation cohort (**D-F**, FUSE-ML models at external validation), as well as those generated from the performance of the Khor et al.14 prediction model applied to the FUSE-ML external validation cohort (**G-I**).

The predicted probabilities for functional impairment are distributed into five equally sized groups, and contrasted with the actually observed frequencies of functional impairment. Calibration intercept and slope are calculated. A perfectly calibrated model has a calibration intercept of 0 and slope of 1. Metrics are provided with bootstrapped 95% confidence intervals.

Table 1. Summary of patient characteristics and outcome measures.

Center	Overall (Pooled)	Development Cohort	External Validation Cohort
N	**1115**	**817**	**298**
Male gender, n (%)	455 (40.8)	349 (42.7)	106 (35.6)
Age, mean (SD) [yrs.]	60.80 (12.45)	61.19 (12.36)	59.73 (12.64)
Height, mean (SD) [cm]	166.47 (9.82)	167.57 (9.62)	162.09 (9.43)
Weight, mean (SD) [kg]	73.53 (14.91)	74.74 (14.77)	69.14 (14.63)
Body Mass Index, mean (SD) [kg/m²]	26.58 (4.61)	26.80 (4.86)	26.07 (3.92)
Smoking status, n (%)			
Active smoker	306 (27.4)	236 (29.0)	70 (24.1)
Ceased smoking	192 (17.2)	166 (20.4)	26 (9.0)
Never smoked	607 (54.4)	413 (50.7)	194 (66.9)
ASA Score ≥3, n (%)	324 (29.1)	251 (31.4)	73 (24.5)
Opioid analgetic use, n (%)	364 (32.6)	314 (43.9)	50 (16.8)
Bronchial asthma, n (%)	63 (5.7)	51 (7.1)	12 (4.0)
Race/Ethnicity, n (%)			
White	861 (77.2)	667 (93.0)	194 (65.5)
Black	30 (2.7)	29 (4.0)	1 (0.3)
Asian	106 (9.5)	6 (0.8)	100 (33.8)
Other	16 (1.4)	15 (2.1)	1 (0.3)
Prior thoracolumbar surgery, n (%)	257 (23.0)	204 (25.0)	53 (26.8)
Indication(s) for Surgery, n (%)			
Spondylolisthesis	599 (53.7)	414 (50.7)	185 (62.1)
Lumbar disc herniation	202 (18.1)	139 (17.0)	63 (21.1)
Radiculopathy	323 (29.0)	230 (32.1)	93 (31.2)
Discogenic CLBP / DDD	457 (41.0)	337 (41.2)	120 (40.3)
FBSS	47 (4.2)	31 (4.3)	16 (5.4)
Lumbar spinal stenosis	618 (55.4)	429 (52.5)	189 (63.4)
Pseudarthrosis	56 (5.0)	55 (7.7)	1 (0.3)
Surgical index level(s), n (%)			
T12/L1	39 (3.5)	36 (4.4)	3 (1.0)
L1/L2	24 (2.2)	19 (2.3)	5 (1.7)
L2/L3	126 (11.3)	114 (14.0)	12 (4.0)
L3/L4	305 (27.4)	245 (30.0)	60 (20.1)
L4/L5	657 (58.9)	529 (64.7)	128 (64.6)
L5/S1	401 (36.0)	344 (42.1)	57 (28.8)
Surgical Technique, n (%)			
TLIF	373 (33.5)	199 (27.8)	174 (58.4)
PLIF	449 (40.3)	325 (45.3)	124 (41.6)
ALIF	7 (0.6)	7 (1.0)	0 (0.0)
Lateral	73 (6.5)	73 (10.2)	1 (0.3)
Minimally invasive, n (%)	310 (27.8)	207 (25.3)	103 (34.6)
Pedicle screw insertion, n (%)	1081 (97.0)	783 (95.8)	298 (100.0)
Baseline patient-reported outcome			
Baseline ODI, mean (SD)	50.17 (17.93)	51.45 (17.50)	47.37 (18.55)
Baseline COMI, mean (SD)	7.47 (1.72)	7.47 (1.72)	-
Baseline back pain, mean (SD)	6.81 (2.32)	6.87 (2.29)	6.65 (2.39)
Baseline leg pain, mean (SD)	6.29 (2.77)	6.20 (2.80)	6.53 (2.68)
Baseline PASS[a] for function, n (%)	58 (5.2)	29 (3.8)	29 (9.7)
Baseline PASS[a] for back pain, n (%)	102 (9.1)	68 (8.4)	34 (11.4)
Baseline PASS[a] for leg pain, n (%)	192 (17.2)	152 (19.0)	40 (13.4)
12-month patient-reported outcome			
12-month ODI, mean (SD)	21.59 (16.49)	21.57 (16.65)	21.62 (16.13)
12-month COMI, mean (SD)	3.42 (2.85)	3.42 (2.85)	-
12-month back pain, mean (SD)	3.08 (2.39)	3.05 (2.40)	3.14 (2.36)
12-month leg pain, mean (SD)	2.51 (2.50)	2.52 (2.48)	2.48 (2.56)
12-month MCID[b] for function, n (%)	761 (68.3)	563 (74.4)	198 (66.4)
12-month MCID[b] for back pain, n (%)	862 (77.3)	640 (80.2)	222 (74.5)
12-month MCID[b] for leg pain, n (%)	796 (71.4)	564 (71.2)	232 (77.9)

SD, standard deviation; ASA, American Society of Anesthesiologists; CLBP, chronic low back pain; DDD, degenerative disc disease; FBSS, failed back surgery syndrome; TLIF, transforaminal lumbar interbody fusion; PLIF, posterior lumbar interbody fusion; ALIF, anterior lumbar interbody fusion; ODI, Oswestry Disability Index; COMI, Core Outcome Measures Index; MCID, minimum clinically important difference; PASS, patient-acceptable symptom state;

[a]PASS (patient acceptable symptom state) was defined as a ODI of ≤ 22, COMI of ≤ 3.05, or a NRS of ≤ 3 for back and leg pain.

[b]MCID (minimum clinically important difference) was defined as a 15-point or greater improvement in ODI or a 2.2-point or greater improvement in COMI (function), or as a 2-point or greater improvement in NRS pain scores at 12 months compared to baseline, respectively.

Table 2. Discrimination and calibration metrics of the machine learning-based prediction models for clinically relevant improvement.

Metric	Models for Improvement					
	Functional Impairment (MCID)		Back Pain (MCID)		Leg Pain (MCID)	
	Development	External Validation	Development	External Validation	Development	External Validation
Model	Elastic Net-Regularized GLM		Elastic Net-Regularized GLM		Elastic Net-Regularized GLM	
Dichotomization Cut-off	0.75		0.85		0.80	
No. Observations	730	269	724	264	640	258
No. Input Variables	10		8		8	
Sampling	-	-	-	-	-	-
Discrimination						
AUC	0.75 (0.73 – 0.76)	0.67 (0.59 – 0.74)	0.71 (0.69 – 0.73)	0.72 (0.64 – 0.79)	0.72 (0.71 – 0.73)	0.64 (0.54 – 0.73)
Accuracy	0.70 (0.69 – 0.71)	0.61 (0.55 – 0.67)	0.68 (0.66 – 0.69)	0.70 (0.64 – 0.75)	0.74 (0.73 – 0.74)	0.71 (0.65 – 0.77)
Sensitivity	0.70 (0.68 - 0.72)	0.59 (0.52 – 0.66)	0.68 (0.67 – 0.69)	0.72 (0.65 – 0.77)	0.77 (0.76 – 0.78)	0.76 (0.71 – 0.82)
Specificity	0.70 (0.68 – 0.72)	0.66 (0.55 – 0.77)	0.63 (0.60 – 0.66)	0.64 (0.51 – 0.78)	0.58 (0.56 – 0.60)	0.42 (0.26 – 0.57)
PPV	0.88 (0.87 – 0.89)	0.81 (0.74 – 0.88)	0.91 (0.91 – 0.92)	0.90 (0.85 – 0.94)	0.90 (0.89 – 0.91)	0.88 (0.83 – 0.92)
NPV	0.43 (0.41 – 0.45)	0.39 (0.31 – 0.48)	0.26 (0.24 – 0.27)	0.34 (0.24 – 0.44)	0.34 (0.33 – 0.36)	0.23 (0.14 – 0.33)
F1 Score	0.54 (0.52 – 0.55)	0.49 (0.41 – 0.58)	0.37 (0.34 – 0.39)	0.45 (0.34 – 0.54)	0.43 (0.42 – 0.45)	0.30 (0.19 – 0.41)
Calibration						
Intercept	0.00 (-0.05 – 0.06)	-0.07 (-0.36 – 0.22)	-0.00 (-0.07 – 0.07)	-0.38 (-0.70 – 0.06)	0.00 (-0.04 – 0.05)	0.14 (-0.22 – 0.51)
Slope	0.89 (0.84 – 0.95)	0.63 (0.34 – 0.93)	0.86 (0.77 – 0.94)	1.10 (0.62 – 1.57)	0.84 (0.79 – 0.89)	0.49 (0.12 – 0.86)

MCID, minimum clinically important difference; GLM, generalized linear model; AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value;
Metrics are provided with bootstrapped 95% confidence intervals based on 1000 samples with replacement. Reported development performance is the resampled cross-validation performance.

Table 3. Model coefficients of the fully trained models. Since centering and scaling were applied to the training data, the magnitude of the coefficients corresponds to variable importance.

	Model Coefficients (MCID)		
Variable	**Function**	**Back Pain**	**Leg Pain**
Model Intercept	*1.399*	*2.021*	*1.828*
Male gender			0.214
Age	0.291		
Height	0.190		
ASA Score ≥3	-0.188		
Opioid analgetic use	-0.156		
Prior thoracolumbar surgery		-0.206	-0.293
Indication(s) for Surgery			
Lumbar disc herniation		0.157	
Radiculopathy	-0.131	-0.126	
Discogenic CLBP / DDD			-0.238
Surgical index level(s)			
L4/L5			-0.160
L5/S1		-0.211	
Surgical Technique			
TLIF	-0.139	0.284	0.135
PLIF	0.169	0.271	0.299
Lateral	0.347	0.666	
Baseline patient-reported outcome			
Baseline ODI/COMI	1.026		
Baseline back pain	-0.340	0.725	-0.187
Baseline leg pain			0.812

MCID, Minimum Clinically Important Difference; ASA, American Society of Anesthesiologists; CLBP, chronic low back pain; DDD, degenerative disc disease; TLIF, transforaminal lumbar interbody fusion; PLIF, posterior lumbar interbody fusion; ODI, Oswestry Disability Index; COMI, Core Outcome Measures Index;

Discussion

The rationale of the FUSE-ML study was to develop and thoroughly externally validate clinical prediction models for 12-month MCID in function, back pain, and leg pain in patients undergoing lumbar fusion for degenerative disease of the lumbar spine. Using data from 11 centers in 7 countries, a web-app was generated. After thorough external validation, we found that the fully trained clinical prediction models demonstrated only moderate ability to dichotomize patients who will and those who will not benefit from lumbar fusion surgery (discrimination performance). Calibration performance – the reliability of the predicted probabilities – was fair. Generally, our models performed comparably well to those published previously by Khor et al., although our models appeared to often require only half of the inputs to achieve the same performance, which streamlines implementation.

Our findings demonstrate that accurate prediction of long-term postoperative PROMs in this patient population remains remarkably difficult, and that clinical decision-making based partially on clinical prediction models should only have a minor role, considering the current state of clinical prediction models in patients with degenerative lumbar spine disease. It is well-known that even expert surgeons may often overestimate the benefits and underestimate complications of certain procedures.[24] Clinical outcomes in degenerative disease of the lumbar spine and spinal fusion – and in particular CLBP, FBSS, and low-grade spondylolisthesis – are known as distinctly difficult to anticipate, and few independent predictors with a sufficiently large effect size are known.[5,14,25] Taking the example of discogenic CLBP, all recent randomized studies show that fusion surgery – overall – does not produce significantly better results than conservative treatment.[4] While surgery may not provide a benefit compared to conservative treatment for CLBP in the general patient population, there are subsets of patients that will truly benefit.[5,6] Rigorous patient selection is key to success in degenerative spine surgery.

In theory, clinical prediction models can provide valuable insights, since they enable calculation of individualized likelihoods of improvements or complications for each patient - as opposed to informing patients about a generalized treatment success rate that is based on historical data in the literature.[26] The hopes of being able to predict the effects of fusion surgery more robustly by generating “objective” risk-benefit profiles for each individual patient have not been fulfilled to date.[26] To our best knowledge, there is only one other externally validated prediction tool for this population: The prediction models for functional impairment and back/leg pain generated by Khor et al.[14] have been developed on 1965 adult lumbar fusion surgery patients collected from a registry of fifteen Washington state hospitals. This model has recently been externally validated at a single Dutch center, demonstrating AUCs of 0.71 to 0.83, sensitivities of 0.64 to 1.00, and specificities of 0.38 to 0.65, with fair calibration.[12] This analysis demonstrated that the discrimination and calibration performance generalized relatively well to a new population, although this level of performance unfortunately still would not allow any reliable decision support in actual clinical practice. The Khor et al.[14] tool is largely based on the same inputs as FUSE-ML, albeit we attempted to improve upon these predictions by introducing surgical variables. In our extensive, multinational external validation study, the FUSE-ML models demonstrated only moderate discrimination and calibration, both of which appeared similar to the performance of the Khor et al. models when applied to our external validation dataset. Still, judging by these performance measures, these models would likely not be very helpful in clinical practice. The discrimination and calibration performance of expert surgeons has not been established as of yet for lumbar fusion in degenerative disease. As long as these metrics remain unknown and as long as comparative or randomized studies do not demonstrate superiority of a decision-making approach integrating machine learning, these supportive tools ought to be used only adjunctively and with great caution in this patient population.

Even with the considerable amount of development data available to us for FUSE-ML, and the application of e.g. regularization techniques, outcomes after lumbar spinal fusion remain difficult to predict with high reliability. One likely contributing factor is the input data: While we included a wide range of relevant socio-demographic, disease-specific, and surgical variables, the addition of imaging data for radiomic analysis and the inclusion of psychological factors could potentially improve predictions. The rationale behind the current approach was to only include few simple, preoperatively and easily available variables, with the intention to keep prediction tools simple, accessible, and quick. This goal was also achieved: We demonstrate that our models generalize to an external validation dataset approximately equally well as previously published, robust models do – although the FUSE-ML models appear to enable the same level of performance with only around half of the inputs required.[14] More parsimonious models, rather than more complex models that require hard-to-collect inputs, are more prone to overfitting, and may not be interpretable at all ("black box")[27,28], may in the end improve accessibility and adoption of models into clinical practice.

Still, even generally – in other patient populations – there is little to no high-quality evidence that clinical prediction models have any measurable clinical impact in their current state. A simulation analysis by Joshi et al.[29] found that, only if applied on a population scale, prediction models in adult spinal deformity may overall decrease healthcare costs by better redirection of resources. Prospective clinical studies evaluating the real-world impact of integrating decision support tools into practice are currently not available. All of the above indicates a need for improving the methods, performance, and in-silico/in-vivo validation of clinical prediction models. However, caution must be taken: The increase in publications of clinical prediction models has increased exponentially over the past few years, as a result of equally exponential access to computing power and "big data".[8] Exactly because it has become relatively easy to generate prediction models, many of these publications fall into common methodological ML "traps", which reviewers of expert medical journals cannot always be generally expected to catch outright. An important notion is the fact that it is relatively easy to generate prediction models with seemingly high performance measures if certain concepts are disregarded – such as class imbalance, data leakage, adequate resampling, and proper validation, among others.[8,11] Furthermore, the vast majority of published models have not undergone external validation and would very likely perform considerably worse in external validation studies.[10,13] A recent review by Lubelski et al. demonstrated the vast methodological deficits in the spinal prediction modelling literature.[10] Lastly, the hopes that ML may help improve predictive performance compared to "traditional statistical modelling" have not been fulfilled, as a systematic analysis by Christodoulou et al. concludes.[30] ML certainly has advantages when analysing highly dimensional data, imaging data, or in natural language processing and time series analysis, but for "simple" tabulated clinical data as is the case with most prediction models, the advantages of ML over e.g. "traditional" generalized linear models likely do not outweigh their drawbacks.[8,30]

We do not recommend the use of clinical prediction models – even those with very high performance metrics – as absolute "red light" or "green light" indicators, but advocate carefully balancing all available clinical data against patient wishes and expectations as well as clinical expertise. There is a need for improved clinical prediction models in spinal fusion for degenerative disease of the lumbar spine, and development will require major international collaborative efforts to collect larger amounts of data and to enable thorough validation of developed models. The FUSE-ML collaborators will continue investigating approaches to improving patient selection in this population.

Strengths and Limitations

Our study used data from eleven centers in different countries, with unified variable definitions. The models have been made available as a web-based tool. Different degenerative spinal diseases were

included. Consequently, our models may perform better for more common pathologies, whereas performance may be limited for the less prevalent ones. Conversely, this heterogeneity in training data may equip the models for the heterogenous presentations of spinal degenerative disease. We also directly compare the performance of our models to the current "benchmark" model in spinal fusion surgery, and demonstrate approximate equivalence of our performance at external validation, as well as fair calibration of our models.

Our data consisted of a mix of retrospectively and prospectively collected data from institutional registries. Many definitions of MCID – and, in the same vein, of PASS – exist, and their choice determines the interpretation of generated predictions.[15] We chose a MCID based on robust MCID studies[17–19], and we excluded patients unlikely to improve by determining a minimally symptomatic state (PASS) based on thresholds from analyses that were anchored to patient-rated symptom satisfaction.[16,20,21] Our prediction tool does not include measures of quality of life and psychological factors, which may improve performance. Learning techniques rely on large amounts of development data, and often improve their performance linearly with an increasing number of training samples. Thus, although we included a relatively large cohort of patients, further training with a larger sample is likely to improve the performance and generalization of the models. We excluded patients under the age of 18 and those with spinal deformity. Our models may not necessarily generalize when extrapolating to these patients.

Conclusions

With the great heterogeneity of outcomes after lumbar spinal fusion for degenerative disease and the countless physical and psychological factors that may modulate the effects of procedures, identifying those patients most likely to benefit from surgical treatment in an objective fashion remains difficult. Although assistive clinical prediction models can help in quantifying potential benefits of surgery and the externally validated FUSE-ML tool (https://neurosurgery.shinyapps.io/fuseml) may aid in individualized risk-benefit estimation, truly impacting clinical practice in the era of "personalized medicine" will necessitate improvements in reliability of clinical prediction models in this patient population. When thoroughly externally validated, current approaches based on tabulated clinical data fail to break the performance barrier required to prevent ineffective surgery or to allow meaningful decisions that are at least partially informed by such clinical prediction models.

Supplementary Table 1. Detailed patient data per center, including missingness.

Variable	Overall	Development Cohort								External Validation Cohort		
Center	Pooled	Balgrist	Paris	St. Andrea	Cesena	Schulthess	Ferrara	Gemelli	Madrid	Amsterdam	Innsbruck	Seoul
N	1115	100	185	76	111	100	61	84	100	100	99	99
Male gender, n (%)	455 (40.8)	52 (52.0)	74 (40.0)	33 (43.4)	61 (55.0)	35 (35.0)	30 (49.2)	30 (35.7)	34 (34.0)	51 (51.0)	34 (34.3)	21 (21.2)
Age, mean (SD) [yrs.]	60.80 (12.45)	64.30 (11.99)	58.62 (12.60)	63.58 (7.18)	57.80 (12.01)	66.04 (11.56)	63.90 (10.58)	54.55 (13.50)	63.87 (12.46)	50.41 (11.39)	62.61 (12.61)	66.27 (7.32)
Height, mean (SD) [cm]	166.47 (9.82)	167.10 (9.14)	168.14 (9.69)	172.87 (7.80)	170.26 (9.90)	167.18 (9.30)	167.40 (7.67)	164.87(10.64)	162.71 (7.94)	-	166.70 (8.49)	157.48 (7.98)
Weight, mean (SD) [kg]	73.53 (14.91)	77.92 (15.16)	73.61 (13.88)	-	75.59 (14.67)	74.70 (16.63)	78.23 (15.16)	72.51 (14.83)	73.42 (13.50)	-	77.26 (13.99)	61.03 (10.08)
Body Mass Index, mean (SD) [kg/m^2]	26.58 (4.61)	27.81 (4.57)	26.14 (5.15)	-	26.05 (4.55)	26.65 (5.05)	27.85 (4.63)	26.59 (4.41)	27.76 (4.90)	25.86 (3.42)	27.77 (4.32)	24.59 (3.29)
Smoking status, n (%)												
Active smoker	306 (27.4)	20 (20.0)	66 (35.7)	45 (59.2)	41 (36.9)	23 (23.0)	9 (14.8)	14 (16.7)	18 (18.0)	30 (30.0)	35 (35.4)	5 (5.1)
Ceased smoking	192 (17.2)	11 (11.0)	38 (20.5)	28 (36.8)	34 (30.6)	0 (0.0)	7 (11.5)	38 (45.2)	10 (10.0)	18 (18.0)	2 (2.0)	6 (6.1)
Never smoked	607 (54.4)	67 (67.0)	81 (43.8)	3 (3.9)	36 (32.4)	77 (77.0)	45 (73.8)	32 (38.1)	72 (72.0)	52 (52.0)	54 (54.5)	88 (88.9)
No. missing, n (%)	10 (0.9)	2 (2.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	8 (8.1)	0 (0.0)
ASA Score ≥3, n (%)	324 (29.1)	39 (39.0)	40 (21.6)	19 (25.0)	29 (26.1)	39 (39.0)	28 (45.9)	19 (22.6)	38 (38.0)	2 (2.0)	25 (25.3)	46 (46.5)
No. missing, n (%)	17 (1.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	17 (27.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Opioid analgetic use, n (%)	364 (32.6)	22 (22.0)	100 (54.1)	1 (1.3)	62 (55.9)	0 (0.0)	5 (8.2)	84 (100.0)	40 (40.0)	25 (25.0)	23 (23.2)	2 (2.0)
No. missing, n (%)	102 (9.1)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.9)	100 (100.0)	1 (1.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Bronchial asthma, n (%)	63 (5.7)	11 (11.0)	11 (5.9)	3 (3.9)	4 (3.6)	0 (0.0)	3 (4.9)	11 (13.1)	8 (8.0)	1 (1.0)	8 (8.1)	3 (3.0)
No. missing, n (%)	101 (9.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	100 (100.0)	1 (1.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Ethnicity, n (%)												
White	861 (77.2)	99 (99.0)	157 (84.9)	76 (100.0)	108 (97.3)	0 (0.0)	61 (100.0)	72 (85.7)	94 (94.0)	95 (95.0)	99 (100.0)	0 (0.0)
Black	30 (2.7)	0 (0.0)	19 (10.3)	0 (0.0)	1 (0.9)	0 (0.0)	0 (0.0)	9 (10.7)	0 (0.0)	1 (1.0)	0 (0.0)	0 (0.0)
Asian	106 (9.5)	1 (1.0)	3 (1.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (2.4)	0 (0.0)	3 (3.0)	0 (0.0)	97 (98.0)
Other	16 (1.4)	0 (0.0)	6 (3.2)	0 (0.0)	2 (1.8)	0 (0.0)	0 (0.0)	1 (1.2)	6 (6.0)	1 (1.0)	0 (0.0)	0 (0.0)
No. missing, n (%)	102 (9.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	100 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (2.0)
Prior thoracolumbar surgery, n (%)	257 (23.0)	30 (30.0)	54 (29.2)	6 (7.9)	36 (32.4)	50 (50.0)	8 (13.1)	0 (0.0)	20 (20.0)	0 (0.0)	29 (29.3)	24 (24.2)
No. missing, n (%)	100 (9.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	100 (100.0)	0 (0.0)	0 (0.0)
Indication(s) for Surgery, n (%)												
Spondylolisthesis	599 (53.7)	41 (41.0)	114 (61.6)	23 (30.3)	35 (31.5)	39 (39.0)	26 (42.6)	84 (100.0)	52 (52.0)	79 (79.0)	55 (55.6)	51 (51.5)
Lumbar disc herniation	202 (18.1)	24 (24.0)	1 (0.5)	23 (30.3)	28 (25.2)	16 (16.0)	11 (18.0)	2 (2.4)	34 (34.0)	8 (8.0)	14 (14.1)	41 (41.4)
Radiculopathy	323 (29.0)	34 (34.0)	14 (7.6)	17 (22.4)	92 (82.9)	0 (0.0)	21 (34.4)	12 (14.3)	40 (40.0)	5 (5.0)	23 (23.2)	65 (65.7)
No. missing, n (%)	101 (9.1)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.9)	100 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Discogenic CLBP / DDD	457 (41.0)	61 (61.0)	59 (31.9)	47 (61.8)	85 (76.6)	35 (35.0)	9 (14.8)	19 (22.6)	22 (22.0)	35 (35.0)	46 (46.5)	39 (39.4)
FBSS	47 (4.2)	21 (21.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	4 (6.6)	0 (0.0)	6 (6.0)	14 (14.0)	0 (0.0)	2 (2.0)
No. missing, n (%)	100 (9.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	100 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Lumbar spinal stenosis	618 (55.4)	84 (84.0)	67 (36.2)	40 (52.6)	45 (40.5)	64 (64.0)	33 (54.1)	18 (21.4)	78 (78.0)	46 (46.0)	48 (48.5)	95 (96.0)
Pseudarthrosis	56 (5.0)	2 (2.0)	34 (18.4)	0 (0.0)	0 (0.0)	0 (0.0)	5 (8.2)	0 (0.0)	14 (14.0)	0 (0.0)	1 (1.0)	0 (0.0)
No. missing, n (%)	100 (9.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	100 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Surgical index level(s), n (%)												
T12/L1	39 (3.5)	4 (4.0)	30 (16.2)	2 (2.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (3.0)	0 (0.0)
L1/L2	24 (2.2)	8 (8.0)	0 (0.0)	3 (3.9)	1 (0.9)	6 (6.0)	0 (0.0)	0 (0.0)	1 (1.0)	0 (0.0)	4 (4.0)	1 (1.0)
L2/L3	126 (11.3)	18 (18.0)	48 (25.9)	5 (6.6)	14 (12.6)	15 (15.0)	3 (4.9)	7 (8.3)	4 (4.0)	0 (0.0)	6 (6.1)	6 (6.1)
L3/L4	305 (27.4)	33 (33.0)	78 (42.2)	14 (18.4)	27 (24.3)	35 (35.0)	21 (34.4)	10 (11.9)	27 (27.0)	0 (0.0)	28 (28.3)	32 (32.3)
L4/L5	657 (58.9)	63 (63.0)	130 (70.3)	55 (72.4)	77 (69.4)	65 (65.0)	41 (67.2)	30 (35.7)	68 (68.0)	0 (0.0)	55 (55.6)	73 (73.7)
L5/S1	401 (36.0)	50 (50.0)	87 (47.0)	37 (48.7)	50 (45.0)	42 (42.0)	17 (27.9)	31 (36.9)	30 (30.0)	0 (0.0)	35 (35.4)	22 (22.2)
No. missing, n (%)	100 (9.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	100 (100.0)	0 (0.0)	0 (0.0)
Surgical Technique, n (%)												
TLIF	373 (33.5)	73 (73.0)	0 (0.0)	45 (59.2)	32 (28.8)	0 (0.0)	1 (1.6)	47 (56.0)	1 (1.0)	62 (62.0)	99 (100.0)	13 (13.1)
PLIF	449 (40.3)	33 (33.0)	171 (92.4)	0 (0.0)	58 (52.3)	0 (0.0)	60 (98.4)	3 (3.6)	0 (0.0)	38 (38.0)	0 (0.0)	86 (86.9)
ALIF	7 (0.6)	0 (0.0)	5 (2.7)	0 (0.0)	2 (1.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Lateral	73 (6.5)	0 (0.0)	5 (2.7)	32 (42.1)	1 (0.9)	0 (0.0)	0 (0.0)	35 (41.7)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
No. missing, n (%)	101 (9.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	100 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.0)	0 (0.0)
Minimally invasive, n (%)	310 (27.8)	26 (26.0)	0 (0.0)	76 (100.0)	6 (5.4)	12 (12.0)	41 (67.2)	38 (45.2)	8 (8.0)	100 (100.0)	3 (3.0)	0 (0.0)
Pedicle screw insertion, n (%)	1081 (97.0)	97 (97.0)	173 (93.5)	59 (77.6)	110 (99.1)	100 (100.0)	60 (98.4)	84 (100.0)	100 (100.0)	100 (100.0)	99 (100.0)	99 (100.0)
Baseline ODI, mean (SD)	50.17 (17.93)	39.52 (16.25)	53.39 (19.88)	52.95 (17.33)	61.33 (12.06)	-	42.00 (NA)	41.26 (11.23)	56.30 (12.70)	45.56 (17.40)	46.55 (20.73)	50.02 (17.21)
Baseline COMI, mean (SD)	7.47 (1.72)	-	-	-	-	7.47 (1.72)	-	-	-	-	-	-
Baseline back pain, mean (SD)	6.81 (2.32)	6.50 (2.00)	6.68 (2.40)	5.82 (1.68)	7.41 (2.41)	5.90 (2.71)	7.23 (1.01)	7.60 (1.53)	7.90 (2.44)	6.77 (2.42)	6.59 (2.27)	6.61 (2.48)
Baseline leg pain, mean (SD)	6.29 (2.77)	5.73 (2.64)	5.98 (2.71)	2.83 (2.52)	7.52 (2.80)	5.88 (2.74)	7.16 (1.01)	6.20 (1.50)	7.92 (2.29)	6.54 (2.72)	5.60 (2.71)	7.45 (2.30)
Baseline PASS[a] for function, n (%)	58 (5.2)	14 (14.0)	11 (5.9)	1 (1.3)	0 (0.0)	2 (2.0)	0 (0.0)	0 (0.0)	1 (1.0)	12 (12.0)	11 (11.1)	6 (6.1)
No. missing, n (%)	60 (5.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	60 (98.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Baseline PASS[a] for back pain, n (%)	102 (9.1)	8 (8.0)	16 (8.6)	7 (9.2)	7 (6.3)	20 (20.0)	1 (1.6)	0 (0.0)	9 (9.0)	11 (11.0)	12 (12.1)	11 (11.1)
No. missing, n (%)	10 (0.9)	10 (10.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Baseline PASS[a] for leg pain, n (%)	192 (17.2)	19 (19.0)	35 (18.9)	55 (72.4)	14 (12.6)	19 (19.0)	1 (1.6)	3 (3.6)	6 (6.0)	12 (12.0)	21 (21.2)	7 (7.1)
No. missing, n (%)	16 (1.4)	11 (11.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	5 (8.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
12-month ODI, mean (SD)	21.59 (16.49)	21.76 (16.46)	16.72 (15.12)	11.74 (5.52)	25.68 (19.40)	-	21.02 (17.74)	24.64 (8.10)	31.00 (19.72)	16.24 (14.81)	21.16 (15.43)	27.53 (16.26)
12-month COMI, mean (SD)	3.42 (2.85)	-	-	-	-	3.42 (2.85)	-	-	-	-	-	-
12-month back pain, mean (SD)	3.08 (2.39)	3.62 (2.17)	2.98 (2.29)	1.57 (0.75)	2.82 (2.59)	2.96 (2.50)	2.93 (2.48)	3.19 (1.69)	4.14 (3.06)	3.11 (2.71)	2.93 (2.13)	3.38 (2.20)
12-month leg pain, mean (SD)	2.51 (2.50)	2.76 (2.76)	1.69 (2.04)	1.20 (0.65)	2.69 (2.90)	2.48 (2.61)	2.98 (2.45)	2.98 (1.39)	4.02 (2.91)	2.13 (2.56)	2.09 (2.26)	3.24 (2.71)
12-month MCID[b] for function, n (%)	761 (68.3)	54 (54.0)	155 (83.8)	75 (98.7)	88 (79.3)	70 (70.0)	0 (0.0)	47 (56.0)	74 (74.0)	73 (73.0)	67 (67.7)	58 (58.6)
No. missing, n (%)	60 (5.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	60 (98.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
12-month MCID[b] for back pain, n (%)	862 (77.3)	56 (56.0)	149 (80.5)	74 (97.4)	91 (82.0)	65 (65.0)	50 (82.0)	81 (96.4)	74 (74.0)	77 (77.0)	75 (75.8)	70 (70.7)
No. missing, n (%)	19 (1.7)	19 (19.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
12-month MCID[b] for leg pain, n (%)	796 (71.4)	49 (49.0)	145 (78.4)	28 (36.8)	82 (73.9)	67 (67.0)	46 (75.4)	68 (81.0)	79 (79.0)	76 (76.0)	73 (73.7)	83 (83.8)
No. missing, n (%)	25 (2.2)	20 (20.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	5 (8.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

PR, prospective registry; RC, retrospective collection; SD, standard deviation; ASA, American Society of Anesthesiologists; CLBP, chronic low back pain; DDD, degenerative disc disease; FBSS, failed back surgery syndrome; TLIF, transforaminal lumbar interbody fusion; PLIF, posterior lumbar interbody fusion; ALIF, anterior lumbar interbody fusion; ODI, Oswestry Disability Index; COMI, Core Outcome Measures Index; MCID, minimum clinically important difference; PASS, patient-acceptable symptom state;

[a]PASS (patient acceptable symptom state) was defined as a ODI of ≤ 22, COMI of ≤ 3.05, or a NRS of ≤ 3 for back and leg pain.

[b]MCID (minimum clinically important difference) was defined as a 15-point or greater improvement in ODI or a 2.2-point or greater improvement in COMI (function), or as a 2-point or greater improvement in NRS pain scores at 12 months compared to baseline, respectively.

Supplementary Table 2. External validation discrimination and calibration metrics of the machine learning-based prediction models for the minimum clinically important difference (MCID) generated in FUSE-ML in comparison with the models published by Khor et al.[14]

Metric	External Validation Performance					
	Functional Impairment		Back Pain		Leg Pain	
	FUSE-ML Model	Khor et al. Model	FUSE-ML Model	Khor et al. Model	FUSE-ML Model	Khor et al. Model
Cut-off	0.75	0.5	0.85	0.5	0.80	0.5
Model	Elastic Net GLM	GLM	Elastic Net GLM	GLM	Elastic Net GLM	GLM
No. Inputs	10	16	8	16	8	16
Discrimination						
AUC	0.67 (0.59 – 0.74)	0.71 (0.64 – 0.77)	0.72 (0.64 – 0.79)	0.73 (0.65 – 0.79)	0.64 (0.54 – 0.73)	0.63 (0.54 – 0.71)
Accuracy	0.61 (0.55 – 0.67)	0.65 (0.59 – 0.71)	0.70 (0.64 – 0.75)	0.78 (0.73 – 0.83)	0.71 (0.65 – 0.77)	0.84 (0.80 – 0.88)
Sensitivity	0.59 (0.52 – 0.66)	0.68 (0.61 – 0.74)	0.72 (0.65 – 0.77)	0.92 (0.88 – 0.95)	0.76 (0.71 – 0.82)	0.98 (0.96 – 1.00)
Specificity	0.66 (0.55 – 0.77)	0.58 (0.47 – 0.69)	0.64 (0.51 – 0.78)	0.22 (0.11 – 0.33)	0.42 (0.26 – 0.57)	0.03 (0.00 – 0.09)
PPV	0.81 (0.74 – 0.88)	0.80 (0.80 – 0.86)	0.90 (0.85 – 0.94)	0.84 (0.79 – 0.88)	0.88 (0.83 – 0.92)	0.85 (0.81 – 0.90)
NPV	0.39 (0.31 – 0.48)	0.42 (0.32 – 0.52)	0.34 (0.24 – 0.44)	0.38 (0.19 – 0.56)	0.23 (0.14 – 0.33)	0.20 (0.00 – 0.67)
F1 Score	0.49 (0.41 – 0.58)	0.49 (0.40 – 0.58)	0.45 (0.34 – 0.54)	0.27 (0.14 – 0.40)	0.30 (0.19 – 0.41)	0.91 (0.87 – 0.94)
Calibration						
Intercept	-0.07 (-0.36 – 0.22)	0.78 (0.50 – 1.06)	-0.38 (-0.70 – 0.06)	0.63 (0.31 – 0.96)	0.14 (-0.22 – 0.51)	0.14 (-0.22 – 0.50)
Slope	0.63 (0.34 – 0.93)	0.93 (0.57 – 1.29)	1.10 (0.62 – 1.57)	1.06 (0.62 – 1.51)	0.49 (0.12 – 0.86)	0.49 (0.08 – 0.89)

GLM, generalized linear model; AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value; Metrics are provided with bootstrapped 95% confidence intervals based on 1000 samples with replacement.

Acknowledgements

We thank the patients whose anonymized data were used for this research.

Disclosures

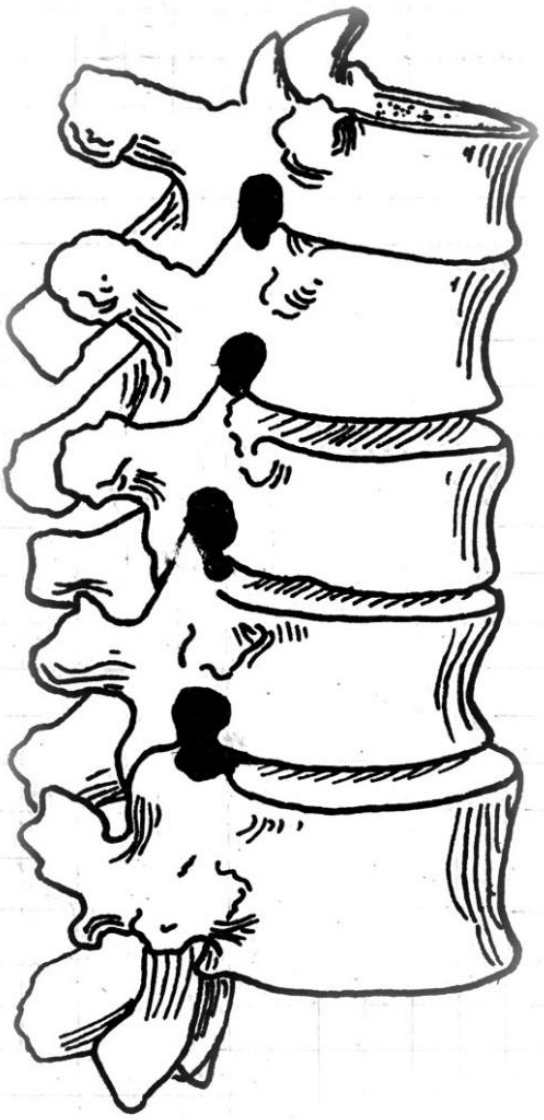
Conflict of Interest: The authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Grants and Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Ravindra VM, Senglaub SS, Rattani A, et al. Degenerative Lumbar Spine Disease: Estimating Global Incidence and Worldwide Volume. *Glob Spine J*. 2018;8(8):784-794. doi:10.1177/2192568218770769
2. Manchikanti L, Abdi S, Atluri S, et al. An update of comprehensive evidence-based guidelines for interventional techniques in chronic spinal pain. Part II: guidance and recommendations. *Pain Physician*. 2013;16(2 Suppl):S49-283.
3. Bono CM, Lee CK. Critical analysis of trends in fusion for degenerative disc disease over the past 20 years: influence of technique on fusion rate and clinical outcome. *Spine*. 2004;29(4):455-463; discussion Z5. doi:10.1097/01.brs.0000090825.94611.28
4. Mannion AF, Brox J-I, Fairbank JC. Consensus at last! Long-term results of all randomized controlled trials show that fusion is no better than non-operative care in improving pain and disability in chronic low back pain. *Spine J Off J North Am Spine Soc*. 2016;16(5):588-590. doi:10.1016/j.spinee.2015.12.001
5. Staartjes VE, Vergroesen P-PA, Zeilstra DJ, Schröder ML. Identifying subsets of patients with single-level degenerative disc disease for lumbar fusion: the value of prognostic tests in surgical decision making. *Spine J*. 2018;18(4):558-566. doi:10.1016/j.spinee.2017.08.242
6. Willems P. Decision making in surgical treatment of chronic low back pain: the performance of prognostic tests to select patients for lumbar spinal fusion. *Acta Orthop Suppl*. 2013;84(349):1-35. doi:10.3109/17453674.2012.753565
7. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
8. Kernbach JM, Staartjes VE. Machine learning-based clinical prediction modelling -- A practical guide for clinicians. *ArXiv200615069 Cs Stat*. Published online June 23, 2020. Accessed March 13, 2021. http://arxiv.org/abs/2006.15069
9. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media; 2008.
10. Lubelski D, Hersh A, Azad TD, et al. Prediction Models in Degenerative Spine Surgery: A Systematic Review. *Glob Spine J*. 2021;11(1_suppl):79S-88S. doi:10.1177/2192568220959037
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953
12. Quddusi A, Eversdijk HAJ, Klukowska AM, et al. External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc*. Published online October 22, 2019. doi:10.1007/s00586-019-06189-6
13. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40. doi:10.1186/1471-2288-14-40
14. Khor S, Lavallee D, Cizik AM, et al. Development and Validation of a Prediction Model for Pain and Functional Outcomes After Lumbar Spine Surgery. *JAMA Surg*. 2018;153(7):634-642. doi:10.1001/jamasurg.2018.0072
15. Ostelo RWJG, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine*. 2008;33(1):90-94. doi:10.1097/BRS.0b013e31815e3a10
16. Fekete TF, Haschtmann D, Kleinstück FS, Porchet F, Jeszenszky D, Mannion AF. What level of pain are patients happy to live with after surgery for lumbar degenerative disorders? *Spine J Off J North Am Spine Soc*. 2016;16(4 Suppl):S12-18. doi:10.1016/j.spinee.2016.01.180

17. Mannion AF, Porchet F, Kleinstück FS, et al. The quality of spine surgery from the patient's perspective: Part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index. *Eur Spine J*. 2009;18(Suppl 3):374-379. doi:10.1007/s00586-009-0931-y
18. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole MR. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*. 2001;94(2):149-158. doi:10.1016/S0304-3959(01)00349-9
19. Parker SL, Adogwa O, Paul AR, et al. Utility of minimum clinically important difference in assessing pain, disability, and health state after transforaminal lumbar interbody fusion for degenerative lumbar spondylolisthesis. *J Neurosurg Spine*. 2011;14(5):598-604. doi:10.3171/2010.12.SPINE10472
20. van Hooff ML, Mannion AF, Staub LP, Ostelo RWJG, Fairbank JCT. Determination of the Oswestry Disability Index score equivalent to a "satisfactory symptom state" in patients undergoing surgery for degenerative disorders of the lumbar spine—a Spine Tango registry-based study. *Spine J*. 2016;16(10):1221-1230. doi:10.1016/j.spinee.2016.06.010
21. Genevay S, Marty M, Courvoisier DS, et al. Validity of the French version of the Core Outcome Measures Index for low back pain patients: a prospective cohort study. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc*. 2014;23(10):2097-2104. doi:10.1007/s00586-014-3325-8
22. Tubach F, Dougados M, Falissard B, Baron G, Logeart I, Ravaud P. Feeling good rather than feeling better matters more to patients. *Arthritis Care Res*. 2006;55(4):526-530. doi:10.1002/art.22110
23. Kuhn M. Building Predictive Models in *R* Using the **caret** Package. *J Stat Softw*. 2008;28(5). doi:10.18637/jss.v028.i05
24. Sacks GD, Dawes AJ, Ettner SL, et al. Surgeon Perception of Risk and Benefit in the Decision to Operate. *Ann Surg*. 2016;264(6):896-903. doi:10.1097/SLA.0000000000001784
25. Alentado VJ, Caldwell S, Gould HP, Steinmetz MP, Benzel EC, Mroz TE. Independent predictors of a clinically significant improvement after lumbar fusion surgery. *Spine J Off J North Am Spine Soc*. 2017;17(2):236-243. doi:10.1016/j.spinee.2016.09.011
26. Steinmetz MP, Mroz T. Value of Adding Predictive Clinical Decision Tools to Spine Surgery. *JAMA Surg*. Published online March 7, 2018. doi:10.1001/jamasurg.2018.0078
27. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206. doi:10.1038/s42256-019-0048-x
28. Ariew R. Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony. Published online 1976.
29. Joshi RS, Serra-Burriel M, Pellise F, et al. 15. Use of predictive machine learning models at the population level has the potential to save cost by directing economic resources to those likely to improve most: a simulation analysis stratified by risk in largest combined US/European ASD registry. *Spine J*. 2020;20(9, Supplement):S8. doi:10.1016/j.spinee.2020.05.118
30. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004



[Chapter 9]

General Discussion and Future Directions

General Discussion

Applications of ML in clinical neuroscience have come a long way since the first ML-based clinical prediction models were published in the late 1980s[1,2]. Contemporary applications range from clinical prediction modelling to detecting genetic mutations from a brain tumor MRI with relatively high precision.[3] Machine vision has also been applied to structure neuroradiological worklists by prioritizing scans that are likely to be more urgent than others.[4] Intraoperatively, histological analyses to distinguish brain tumor from healthy brain tissue can be achieved within minutes using stimulated Raman spectroscopy combined with machine vision.[5] In the neurointensive care unit, the flood of false alarms due to artefacts can be significantly reduced by considering video recordings and other data of patients.[6] Brain shift can be partially corrected intraoperatively using deformation algorithms, and it has even been suggested that intraoperative histological diagnosis can be achieved by analysing electrocautery smoke.[7,8] Similarly to these applications, there are many others that until recently would have been considered impossible. Especially those problems that would ordinarily not be amenable to solving even by expert physicians, such as reading genetic mutations from a brain MRI, are those that are likely to provide the greatest benefit.

In fact, many applications – rudimentary or not – of ML have already long seen introduction into clinical practice and may not even be noticed by most clinicians. It is not uncommon nowadays for deep learning algorithms to be applied to improve MRI, CT, or radiographic imaging or to reduce their radiation dose.[9] Image registration algorithms are applied daily in the operating room when using neuronavigation, at the press of a button.[10] Electrocardiograms usually include an automated analysis.[11] Similarly to these examples, there are many other routine applications of ML that may go unnoticed within the hospitals of the 21st century. Even simple risk scores such as the Wells score for thrombotic risk[12], CHA_2DS_2-VASc score for thromboembolic events in atrial fibrillation[13], or PHASES score for rupture risk in unruptured intracranial aneurysms[14] can be considered decision rule-based or regression algorithms in their own right – although they may, or may not be, machine-generated.

More and more diverse applications are emerging and publication numbers on ML in clinical neuroscience are still increasing exponentially each year. In neurosurgery, possible applications can be roughly divided into four phases: the diagnostic process including imaging, preoperative decision-making, intraoperative applications, and the postoperative period. In this thesis, the first three phases have been touched upon in some way and will be discussed further below. Nonetheless, the final phase – the postoperative period on the intensive care, post-acute care, or step-down unit or on the ward for that matter – is not to be brushed over, since the wealth of data that is collected in these settings is immense. Longitudinal data such as pressure curves or frequent laboratory evaluations are a perfect breeding ground for contemporary, complex algorithms, with all their limitations in interpretability and generalizability, too.

This particular field of ML in clinical neuroscience has certainly undergone evolution throughout the past decade, and would deserve a thesis of its own.[15,16]

Machine Learning-Augmented Testing

What is however reported in detail from the very beginning to the end in this thesis is the application of ML to augment clinical testing. In today's Western medicine and medical research, where terms such as "big data", "personalized medicine" and "precision medicine" are continually heard, most clinical quantitative tests are still not personalized in any way. Some developments have been made, for example age-adjusted cut-offs for D-dimers in suspected pulmonary embolism or estimation of glomerular filtration rate based on demographic data from creatinine.[17,18] Fixed cut-offs based on entire normative populations do work in most cases and are still the standard. In some cases however, specificity of a test becomes extremely low based on one or multiple simple demographic factors, such as age or gender, which could be adjusted for with relative ease.

Instead of fixed cut-offs, dynamic cut-offs that are calculated based on each individual patients' data has the potential to improve patient assessment and to somewhat approach the goal of "personalized medicine" or "precision medicine". Some ideological developments to generate more personalized cut-offs have been made already: Multiple fixed cut-offs that are calculated based on subgroups (e.g. under and over 65 years of age, or male versus female) can be applied and shown for reference in a table.[19,20] However, memorizing a range of fixed cut-offs makes clinical application cumbersome. In addition, the goal of "precision medicine" should be to tailor disease diagnosis, grading, prognosis, and prediction to each individual patient, and not to groups or subgroups, which can hardly be considered individual.

The testing process that we describe is not specific to a population with degenerative disease of the lumbar spine, but could really be applied to any quantitative test: The same normative population is taken as per usual, but quantiles are calculated for each patient individually instead of for the whole entire population. If – based on these quantiles – the patient is deemed to be "abnormal", grading of the level of abnormality can be achieved by unsupervised clustering into – hopefully – somewhat prognostically relevant subgroups.[21,22]

Focusing on spine surgery specifically, why is there a need for objective functional tests at all?[23] The first point is that the success of surgical procedures, especially in patients with degenerative disease of the lumbar spine, is sometimes rated very differently by the operating surgeon and by the patient.[24–26] Second, questionnaires on subjective functional impairment may not necessarily capture deficits such as limping, foot drop, and others, but objective tests often do.[27] Such tests appear to capture a different dimension of the patient's disease state that correlates well with important daily tasks such as getting up from a chair or climbing stairs, and may also have prognostic value, although this has not been well-studied as of yet.[28] Some objective tests may also be more inert regarding psychological confounders.[29] Lastly, for research or follow-up purposes, patients do not enjoy the process of filling in a battery of questionnaires and have been reported to prefer objective functional testing.[30]

We combined objective testing and the concept of a personalized testing pipeline augmented by ML by first validating the 5R-STS for use in patients with degenerative disease of the lumbar spine, and then developing the two necessary algorithms and web-app. In this way, we have demonstrated that testing

patients with expectations/cut-offs set to their specific demographics is not only feasible, but potentially even superior in the sense that the patients who were judged as impaired previously using fixed thresholds were not the exact same group as those who were judged as impaired with this personalized method. We described the methodology in detail and reproduction should be feasible, regardless of what data or what disease domain is used.

Machine Learning-Augmented Operative Imaging

Imaging, in all of its forms, has been one of the most popular and successful subfields of machine intelligence in medicine. The wealth of data and the rising frequency of neuroimaging in general have allowed research in this area to prosper. Still, there are many untouched potential applications and concepts that can be elaborated on even further. Operative imaging is vital in neurosurgery: Many decisions and fates are at least partially decided on the basis of tomographic imaging nowadays, operating rooms are plastered with screens demonstrating preoperative images, intraoperative MRI and ultrasound are being adopted more and more, and functional neuroimaging is applied to locate critical areas.[31–34]

In a worldwide survey among neurosurgeons that included a question on what neurosurgeons use ML for in their daily practice, 50.5% of ML users reported using these techniques to interpret or quantify medical imaging.[35] Again, applications that have already entered routine and are sometimes considered more "mundane" in imaging include image registration, upscaling the resolution of low-dose images, or correction for brain shift.[7,34,36] Newer applications are emerging, for example the use of machine vision in open spine surgery to avoid wrong-level surgery or to place pedicle screws based on recognition of the spinal surface anatomy using cameras in the overhead operating room lighting.[37–39]

Common objectives of ML in neuroimaging today include segmentation of structures such as tumours, including generation of augmented reality models from these segmentations, workflow improvement, and "simple" classification of images for diagnosis or prognosis.[40] Our goal was to test expansions of these current objectives towards novel and clinically advantageous uses.

First, navigated instrumented spine surgery is based on tomographic imaging in most cases. The drawbacks of having to undergo an additional CT scan of the lumbar spine when a MRI is often available already will be clear to the reader and are elucidated in detail in chapter 5. Generating one imaging method from another – keeping in mind that these methods are based on completely different physical processes – seems a bit like attempting alchemy, but efforts in image conversion are not completely novel: Combined use of MRI and CT or their generation from each other is not uncommon in radiation oncology.[41–48] A variety of different methods has been applied to achieve image conversion, however, most of these results do not achieve a visually or quantitatively high enough fidelity for considering their use in diagnostic imaging or neurosurgical operative planning or intraoperative navigation.[49] The necessary extra MRI sequence for the technique that we described only takes a few minutes longer and enables gathering of specific information that enables relatively efficient generation of a synthetic lumbar spine CT from its MRI counterpart. We demonstrated that quantitative measurements performed on the synthetic CT are within the range necessary for intraoperative navigation in spine surgery, and that image quality appears sufficiently good upon visual inspection. Still, this remains a proof-of-concept study, with several limitations that are outlined in chapter 5. In particular, quantitative evaluation by statistically testing each voxel against each ground truth voxel will be important, along with evaluation of the performance of the algorithm in rarer situations where implants or artefacts are present. It is possible, or even likely, that, as long as the algorithm is not trained on enough of such special cases, it will not be able to infer a correct CT instance from an MRI. Clinical validation is also needed: It might of course be possible

to plan pedicle screws in silico on a synthetic CT using a robotic workstation, but this does not necessarily mean that pedicle screws placed with this approach will necessarily exhibit the level of accuracy required to compete with a "real" spiral CT or an intraoperative cone-beam CT. Lastly, although current results based on a small sample size are already visually very similar to their ground truth counterparts, it is unclear whether adjunctive use of synthetic CTs using our technique allows reliable diagnostic neuroradiology, or not. However, taken together this approach may soon allow on-demand synthetic CT imaging coupled with a fraction of the hassle, cost, and time required for a separate, real CT, as long as the correct MRI sequences are adopted as standard for lumbar spine imaging in a particular center.

Second, Intraoperative orientation is vital in neurosurgery. Successful and safe intracranial tumor resection is strictly dependent on surgical orientation and understanding of the anatomy – both the anatomy that is currently visible in the endoscope or microscope (identification), but also the structures that lie one layer deep to the current view (anticipation). Appropriate surgical orientation leads to improved surgical efficacy and, most importantly, may be the most critical step to prevent neurological impairment due to manipulation of otherwise healthy neural tissue. Especially in intracranial tumor surgery, normal anatomy may not be recognizable as clearly and even the mere differentiation of what still is healthy brain tissue and what is tumor becomes both critical and difficult. For these reasons, several assistive tools have been introduced to help guide surgeons, including, but not limited to, intraoperative ultrasound[33] or fluorescence-based methods.[50] These methods however, are highly user-dependent and require the surgeon to acquire knowledge on interpreting a new imaging method that consequently has limited sensitivity or specificity. Neuronavigation[10] – meaning the orientation within the three-dimensional operating room by use of cameras and fiducial arrays, overlaid on preoperative neuroimaging, - has also been applied to enable identification of visible structures as well as to allow more accurate assessment of those structures that are not yet visible in the current field of view. However, neuronavigation relies on preoperative imaging only, with some potential intraoperative adjustments or updating after an intraoperative MRI. This means that it cannot account for any intraoperatively occurring unexpected events. In some sense, this is like driving in a self-driving car that navigates purely based on roadmaps, but without any sensors to spot current, unforeseeable dangers, such as a swiftly braking car to its front or a hastily crossing pedestrian. Lastly, intraoperative MRI and intraoperative functional brain mapping as well as awake surgery have proven highly advantageous, but also require certain infrastructure and expertise that might not be available everywhere.[51,52]

Our aim was to tackle some of these issues by enabling real-time navigation – not based on preoperative imaging – based on only the intraoperative microscopic or endoscopic view, just like a master surgeon. Our study demonstrated promising results in the sense that labelling anatomical structures based only on video data and without providing a priori knowledge is at least feasible with the simple structures that we piloted. If developed further, such systems could provide valuable intraoperative assistance with relatively little resources needed. Especially once integrated into an operating microscope and structures that are in view can be sufficiently accurately labelled upon the press of a button, our method may add another building block to patient safety in cranial surgery.

The specific method of machine vision applied here is able to not only label structures but generate heatmap predictions – similar to segmentations – of each identified structure.[53] The most striking progress of our study is probably the fact that labelling the training data (video frames) by experts is made vastly more time-efficient by use of this method, as it appears to be able to learn heatmaps from single pixel labels that the expert sets. Compared to having to segment each structure on each frame, this is much quicker. As data quality and quantity – including the quality of labelling – become more important with deep learning methods, this progress is significant.[54] Especially labelling by highly-qualified experts

is already time-consuming, but on the other hand, an anatomical recognition algorithm trained on data labelled by medical students can probably not be expected to truly outperform and help neurosurgeons in those critical situations where they would most need this tool. One important issue in our study was that, due to the heatmap regression approach that generates heatmaps from single pixels, quantitative performance assessment is tricky since there are no ground truth labels for the heatmaps to compare to as a "gold standard". We have attempted to at least partially tackle this drawback by using a semi-quantitative grading system based on visual inspection of the generated heatmaps, and by comparing to a model that only always predicts the same heatmap that is based on the average distribution of the location of each anatomical structure in the training data. Performance evaluation is critical in ML, and this is a point that will need to be addressed in the near future.

Clinical Prediction Modelling

The fact that ML algorithms learn purely from historical data and that they are sometimes able to derive generalizable interactions among inputs from these data – without specific programming or instruction – is an advantage and a disadvantage at the same time. "Without specific programming or instruction" refers to the fact that, through optimization, algorithms are able to converge on a set of parameters that enables solving a classification or regression problem without provision of any specific rules. For example, when predicting the likelihood of deep vein thrombosis in hospitalized patients, the parameters are randomly initialized, and we do not usually provide it with any a priori useful information such as "*Patients who have had recent surgery, who are immobilized, or who have a malignant tumour are at higher risk of venous thrombosis*". Instead, the randomly initialized parameters are iteratively and empirically improved until a minimum of the error function is arrived at – without any specific other instructions. However, this feature of most ML algorithms has one drawback that becomes particularly apparent in clinical prediction modelling – or any subfield of ML for that matter: Humans do not have to have seen an event to know how to potentially avoid it, but can be told about the existence of a such event and instructed about how to deal with it. For example, an experienced neurosurgeon can verbally warn a beginning resident about a very rare but serious complication that can be avoided with relative ease, if anticipated. Instead, ML models are only knowledgeable about what they have previously seen in the training dataset. Sometimes, when a trained ML algorithm encounters an event it has never seen, its decisions may appear random and out of line. This effect stresses the importance of using enough, representative training data, to avoid extrapolation, and to anticipate that models may fail if they encounter situations that were not present within the training data.

In clinical prediction modelling, for example when predicting a complication, this leads to a frequently observed effect in which the performance of one and the same clinical prediction model differs markedly among rare versus frequent events, among "easy" and "hard" predictions, or among "typical" patients and "atypical" patients. To illustrate, when predicting gross total resection after endoscopic pituitary surgery, it was found that the model performed admirably on patients with extreme Knosp scores – those patients who were clearly without cavernous sinus invasion and those who clearly had their internal carotid artery encased by the adenoma.[55] In these cases, where predictions are rather simple (A purely intrasellar adenoma is almost always amenable to gross total resection, and a Knosp grade 4 adenoma that completely surrounds the internal carotid artery is only rarely amenable to complete resection) and a prediction model may be of little actual clinical use, the model was highly precise in its prediction of resection status. Conversely, however, in patients with Knosp intermediate-grade tumours, where it is rather difficult to predict on an individual basis if complete resection can be achieved, the model's

precision was markedly decreased. Unfortunately, in exactly these "borderline" cases, a reliable prediction model would be useful.

As a consequence, this drawback of current prediction models significantly limits their clinical usefulness. Overall, when comparing machine learning and human expert performance[56], their performance gain is minimal in most cases. However, even if something can be detected with far greater sensitivity or accuracy or if a rare complication can be predicted very reliably using ML, the question still remains whether medical professionals are then able to actually improve outcomes based on these insights. Very small and hard-to-detect, asymptomatic pneumonia, lumbar disc herniation, or arthrosis may be more frequently described by a deep learning algorithm, but their treatment is probably unlikely to improve the patient's health status. In the end, patient beliefs and wishes along with their by definition subjective perception of their symptoms do influence outcomes greatly, especially in spine surgery.[57,58] The literature on the actual, measurable clinical benefits of applying prediction models in medicine is extremely scarce. Currently, to our best knowledge, no randomized trials have been performed to compare outcomes, adverse events, patient satisfaction, and cost-effectiveness of an approach with, versus without, clinical prediction models. These studies would make, or break, the development of prediction models in a particular patient population: If real-world benefits are not produced by the use of clinical prediction models, their continued development has to be regarded as redundant.

As demonstrated in the two studies included within this thesis, clinical prediction modelling can work and provide potentially useful information, or fail to predict anything at all. We were able to predict outcomes after intracranial tumour surgery – not with extremely high performance, but perhaps slightly more accurate than neurosurgeons.[59–61] Apart from the performance in discrimination, the calibration performance of this model was remarkable. This indicates that, while the binary predictions may be less reliable, the predictions of risk that is provided by the web-app ("*Your risk of experiencing new functional impairment is estimated at 12%")* correlates well with real-world risk. Since patients are not binary, and since medical doctors are experts at balancing risks and benefits, well-calibrated clinical prediction models may portend benefits when counselling patients, although there is no evidence that such information leads to any changes in management, outcomes, or patient satisfaction and stress.[62–65]

Conversely, when we attempted to tackle the issue of which patients are most likely to profit from spinal fusion surgery for chronic degenerative disease, all modelling approaches failed to produce generalizable predictions with high enough discrimination or calibration performance to allow for any meaningful patient counselling. Other "hard-to-predict" ground truths such as IDH mutation status from a brain tumour scan seem to be possible, however – provided that these models were also thoroughly validated on new data.[3] In the end, the parsimonious explanation of the differences in success is that some future events are simply unpredictable, or in other words and assuming determinism, that some future events are governed by so many different and hard-to-capture influencing factors that simple prediction modelling becomes unreliable. In theory, one could collect thousands of data points for each patient to allow for a potentially more accurate prediction. In practice, however, it is unlikely that anyone would be interested in using this type of model, since the effort to collect this wealth of data for a single prediction would be immense. Even a web-app with 20 different input variables already will take up some minutes of time in clinic hours, and time is scarce. Having to input many dozens of variables will therefore likely rarely see entry into clinical routine – Unless automation sets in.

Future Directions

It is likely that a range of applications of ML, and in particular deep learning and reinforcement learning, that would not have been considered possible up to now, will emerge in the near future. In this thesis, we already describe several new developments or evolutions of existing techniques: Generating a totally different imaging modality from another one, teaching a machine vision algorithm to recognize anatomical structures without prior knowledge of anatomy and topography, and a workflow to personalize quantitative testing. All applications of ML to problems that would ordinarily seem impossible for human experts to solve are also those that generate the highest level of excitement. For example: A recent study suggests that application of a ML algorithm to "smell" lung cancer in human exhaled breath with surprisingly high accuracy.[66] Similarly, there is some data that suggests that differentiating tumor from healthy brain tissue by analysing electrocautery smoke or IDH mutations status from routine brain MRI sequences, as mentioned before.[3,8] Using machine vision algorithms, tissue can be analysed and quick histopathological diagnoses can be made within minutes.[5] Such developments are to be pushed further. One aspect about ML that is not shared with other technological developments such as robotics is the relatively low cost of application, in many cases: While development may be cumbersome, many models do not require great amounts of computing power and could thus be relatively easily made available worldwide and for free, enabling application even in rural areas on mobile devices.

In terms of the projects described within this thesis, further developments are planned. Studies are currently ongoing to evaluate the predictive value of ML-augmented personalized testing: Could it be that patients with a lumbar disc herniation who are objectively impaired are more likely to already having undergone permanent motor fiber loss, thus making them less likely to profit from discectomy at that point? Such questions will need to be answered in prospective studies. Similarly, there should be attempts at applying the same rationale of patient-specific cut-offs to other tests – Whether those are other objective functional tests such as the popular timed up-and-go test (TUG)[20], or whether those are laboratory values that are known to change e.g. with age, such as D-dimers or erythrocyte sedimentation rate (ESR).[17,67]

The concept of synthetic CTs is also currently being pursued further: Larger amounts of paired CT-MRI data for training of the models are being collected, and studies applying synthetic CTs clinically to guide robotic surgery are in consideration. Only with extensive clinical validation can synthetically generated medical images be taken seriously in the diagnostic realm, and only if they demonstrate equal or better performance in placing pedicle screws compared to traditional navigated or robotic techniques can they be widely introduced here, too.[68] Similarly, real-time machine vision-assisted anatomical guidance is currently under further development: More data is being collected and models that have temporal and spatial memory and that can not only detect currently visible structures, but also predict the location of currently invisible but critical structures are being developed. Integration into the clinical workflow by means of collaboration with industry leaders in operating microscopes and endoscopes could finally enable anatomical structure labelling at the push of a button – At least, that is the hope for this technique.

Other issues pertain to data quality and quantity – all of this applies to clinical prediction modelling as well. In terms of data collection, many models are still developed on retrospective data, which demonstrably misses out on many complications and other important events that simply have not been captured.[69] This also applies to labelling when it comes to supervised learning: The model can only be as good as the labels, and thus as the expert labelling the data. A frequently cited phrase is "garbage in,

garbage out". Because ML cannot magically learn things correctly if it has been taught to learn them incorrectly, this will probably always hold true. Considerations on data quantity have also become more frequent. It is still difficult to foresee how many training observations will be necessary to arrive at a robust model. Many "rule of thumb" approaches exist, although there is no clear consensus.[70,71] Regardless of this particular discussion, efforts are also being made to learn better from less data by improving modelling techniques. This approach will probably meet a glass ceiling at some point, since the models can only really ever learn from the data they have seen during training. Transfer learning – the use of pre-trained models that are then subsequently fine-tuned to solve a specific problem – is such an approach.[40] Other novel approaches to improve predictions from less data include collaborative "hive learning", in which for example a competition is created and many research groups participate for a prize.[72] Finally the best models generated can even be used as an ensemble – all together – to further increase performance.[73] Lastly, taking into account the ever more strict regulations on data privacy and security, a novel approach termed "federated learning" may help improve predictions thorough collaboration: In federated learning, data is not shared among centers, but models are improved step-by-step by learning on the training data of each center individually, eliminating the need for data transfers.[74]

Another point concerning "big data" in the near future is the potential of applying natural language processing (NLP) to automatically extract far greater amounts of training data, in turn allowing more extensive model training and likely improved performance – Especially highly complex or large models such as deep neural networks profit immensely from larger sample sizes, with their performance sometimes increasing linearly with increases in training data availability. The caveat here, as stated before, will be data quality: Especially supervised models can only predict what the ground truth labels teach them. If a data collection NLP algorithm is "only" 90% accurate in labelling data, a significant amount of falsely classified training samples will be learnt. The question is whether – first of all – human raters are more accurate than such data collection models: A large part of retrospective data collection is done by medical students or residents who may be less experienced in e.g. rating a brain scan or interpreting medical records. If data collection algorithms perform approximately equally well compared to human raters, data quality in the sense discussed above may only be a minor issue. The second question is whether a fully trained algorithm based on a much larger sample of partially falsely labelled data would potentially even perform better overall, compared to a fully trained algorithm based on a smaller sample that has been more reliably collected by expert human raters. In the end, the field of ML – apart from considering the basic biostatistical and epidemiological principles of medical research – has up to now remained an empirical science, with relatively little agreement on standard methods or on what is "best practice" when compared to e.g. prospective clinical trials or biostatistics in general.

This leads into the next problem that data scientists will have to face in the 21st century: Methodological agreement and improving the quality of publications on ML in medicine. The democratization of ML through various means requires a new approach to developing models and more acutely to evaluating ML papers. Most clinicians are not equipped with the necessary knowledge and experience to, say, evaluate a ML paper at a journal club. It is perhaps unrealistic to expect all clinicians – experts in clinical medicine – to have the same methodological foundation as someone with a background in data science. In fact, there are some considerable differences in how learning problems are approached and evaluated by medical researchers applying machine intelligence versus data scientists tackling medical problems. The obvious solution to ensure correct research practices in this field is collaboration among the two, which most definitely yields the "best of both worlds": Evidence-based medicine, epidemiology, and biostatistical principles, as well as the programming prowess and theoretical foundational knowledge possessed by machine learning experts. Still, this does not solve all of the abovementioned problems. Many poor quality ML papers are published in neurosurgical journals every week, because their reviewers – being largely clinical experts – may simply not possess the necessary methodological toolkit to dissect a

machine learning paper. This is much akin to a machine learning expert evaluating a randomized clinical trial: With less experience in clinical studies or less medical domain knowledge, such studies usually appear very promising upon initial assessment. Academic medical doctors however, are often able to dissect these studies upon further evaluation to identify major flaws or limitations in the generalizability of the results, making a large part of the randomized studies that are published relatively poorly applicable to clinical practice. Only a small percentage of randomized studies are conducted with rigorous enough methods and a clinical concept that is both relevant and generalizable enough to allow adoption into clinical practice. ML papers are also not without flaws, which is why some neurosurgical journals have started to designate special reviewers who review all or most of the ML manuscripts that are submitted to said journal. A further aspect in solving the methodological quality issue is education about the topic. Many medical schools have introduced optional courses on the basics of ML, and the number of online resources is also steadily increasing. Many of the high-quality online resources will be relatively hard to grasp for complete beginners and may not convey the exact toolkit with which a clinician can decide whether a certain clinical prediction model should really be trusted or not. Resources that are targeted towards clinicians interested in getting started with ML are therefore of prime importance.[75–77] These resources will hopefully improve the reporting standards of manuscripts including ML techniques submitted to neurosurgical journals and elsewhere.[78,79,79]

The last problem for the future of ML in medicine that is often brought up is the so-called “black box”.[80] Complex models such as deep neural networks usually are poorly calibrated and also have less ways – or no way – to explain why and how the model made a certain prediction.[65] Especially in medicine and in the medicolegal arena, being able to explain decisions is crucial. One approach is to use similarly complex methods such as local interpretable model-agnostic explanation (LIME) to arrive at some overview of how predictions are made and which factors contributed most.[81] However, this still is explaining a black box using another black box, and the inner workings of deep neural networks will still be poorly understood. The more elegant and sustainable way is to simply apply models that are appropriately complex for the complexity of the training data that is being used: Tabulated medical data usually will not profit significantly from using a deep neural network compared to a “simple”, shallow neural network or anything less complex. This aspect is crucial, because simpler models usually have ways (such as odds ratios for logistic regression, partial dependence for generalized additive models, or even visual decision rules for tree-based methods) to explain predictions natively, while more complex models rarely do so and often only provide a minute benefit in discrimination performance at the cost of interpretability.[76,82–84]

References

1. Grigsby J, Kramer RE, Schneiders JL, Gates JR, Smith WB. Predicting Outcome of Anterior Temporal Lobectomy Using Simulated Neural Networks. *Epilepsia*. 1998;39(1):61-66. doi:10.1111/j.1528-1157.1998.tb01275.x
2. Mathew B, Norris D, Mackintosh I, Waddell G. Artificial intelligence in the prediction of operative findings in low back surgery. *Br J Neurosurg*. 1989;3(2):161-170. doi:10.3109/02688698909002791
3. Chang K, Bai HX, Zhou H, et al. Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas from MR Imaging. *Clin Cancer Res*. 2018;24(5):1073-1081. doi:10.1158/1078-0432.CCR-17-2236
4. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med*. 2018;24(9):1337-1341. doi:10.1038/s41591-018-0147-y
5. Hollon TC, Pandian B, Adapa AR, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nature Medicine*. 2020;26(1):52-58. doi:10.1038/s41591-019-0715-9
6. Schwab P, Keller E, Muroi C, Mack DJ, Strässle C, Karlen W. Not to Cry Wolf: Distantly Supervised Multitask Learning in Critical Care. *arXiv:180205027 [cs, stat]*. Published online June 7, 2018. Accessed May 3, 2020. http://arxiv.org/abs/1802.05027
7. Iversen DH, Wein W, Lindseth F, Unsgård G, Reinertsen I. Automatic Intraoperative Correction of Brain Shift for Accurate Neuronavigation. *World Neurosurgery*. 2018;120:e1071-e1078. doi:10.1016/j.wneu.2018.09.012
8. Haapala I, Karjalainen M, Kontunen A, et al. Identifying brain tumors by differential mobility spectrometry analysis of diathermy smoke. *Journal of Neurosurgery*. 2019;133(1):100-106. doi:10.3171/2019.3.JNS19274
9. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*. 2019;29(2):102-127. doi:10.1016/j.zemedi.2018.11.002
10. Härtl R, Lam KS, Wang J, Korge A, Kandziora F, Audigé L. Worldwide Survey on the Use of Navigation in Spine Surgery. *World Neurosurgery*. 2013;79(1):162-172. doi:10.1016/j.wneu.2012.03.011
11. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11(1):1760. doi:10.1038/s41467-020-15432-4
12. Wells JK, Hagino RT, Bargmann KM, et al. Venous morbidity after superficial femoral-popliteal vein harvest. *J Vasc Surg*. 1999;29(2):282-289; discussion 289-291. doi:10.1016/s0741-5214(99)70381-2
13. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137(2):263-272. doi:10.1378/chest.09-1584
14. Greving JP, Wermer MJH, Brown RD, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. *Lancet Neurol*. 2014;13(1):59-66. doi:10.1016/S1474-4422(13)70263-1
15. Al-Mufti F, Kim M, Dodson V, et al. Machine Learning and Artificial Intelligence in Neurocritical Care: a Specialty-Wide Disruptive Transformation or a Strategy for Success. *Curr Neurol Neurosci Rep*. 2019;19(11):89. doi:10.1007/s11910-019-0998-8
16. van Niftrik CHB, van der Wouden F, Staartjes VE, et al. Machine Learning Algorithm Identifies Patients at High Risk for Early Complications After Intracranial Tumor Surgery: Registry-Based Cohort Study. *Neurosurgery*. 2019;85(4):E756-E764. doi:10.1093/neuros/nyz145
17. Righini M, Van Es J, Den Exter PL, et al. Age-Adjusted D-Dimer Cut-off Levels to Rule Out Pulmonary Embolism: The ADJUST-PE Study. *JAMA*. 2014;311(11):1117-1124. doi:10.1001/jama.2014.2135

18. Cristelli MP, Cofán F, Rico N, et al. Estimation of renal function by CKD-EPI versus MDRD in a cohort of HIV-infected patients: a cross-sectional analysis. *BMC Nephrology*. 2017;18(1):58. doi:10.1186/s12882-017-0470-4
19. Stienen MN, Smoll NR, Joswig H, et al. Validation of the baseline severity stratification of objective functional impairment in lumbar degenerative disc disease. *Journal of Neurosurgery: Spine*. 2017;26(5):598-604. doi:10.3171/2016.11.SPINE16683
20. Gautschi OP, Smoll NR, Corniola MV, et al. Validity and Reliability of a Measurement of Objective Functional Impairment in Lumbar Degenerative Disc Disease: The Timed Up and Go (TUG) Test. *Neurosurgery*. 2016;79(2):270-278. doi:10.1227/NEU.0000000000001195
21. Meinshausen N. Quantile Regression Forests. *J Mach Learn Res*. 2006;7:983-999.
22. Koenker RW. *Quantile Regression*. Cambridge University Press; 2005. doi:10.1017/CBO9780511754098
23. Gautschi OP, Corniola MV, Schaller K, Smoll NR, Stienen MN. The need for an objective outcome measurement in spine surgery—the timed-up-and-go test. *The Spine Journal*. 2014;14(10):2521-2522. doi:10.1016/j.spinee.2014.05.004
24. Mazur MD, McEvoy S, Schmidt MH, Bisson EF. High self-assessment of disability and the surgeon's recommendation against surgical intervention may negatively impact satisfaction scores in patients with spinal disorders. *Journal of Neurosurgery: Spine*. 2015;22(6):666-671. doi:10.3171/2014.10.SPINE14264
25. Roitberg BZ, Thaci B, Auffinger B, et al. Comparison between patient and surgeon perception of degenerative spine disease outcomes—a prospective blinded database study. *Acta Neurochir*. 2013;155(5):757-764. doi:10.1007/s00701-013-1664-6
26. Porchet F, Lattig F, Grob D, et al. Comparison of patient and surgeon ratings of outcome 12 months after spine surgery: presented at the 2009 Joint Spine Section Meeting. *J Neurosurg Spine*. 2010;12(5):447-455. doi:10.3171/2009.11.SPINE09526
27. Stienen MN, Maldaner N, Sosnova M, et al. Lower Extremity Motor Deficits Are Underappreciated in Patient-Reported Outcome Measures: Added Value of Objective Outcome Measures. *Neurospine*. 2020;17(1):270-280. doi:10.14245/ns.1938368.184
28. Stienen MN, Ho AL, Staartjes VE, et al. Objective measures of functional impairment for degenerative diseases of the lumbar spine: a systematic review of the literature. *Spine J*. 2019;19(7):1276-1293. doi:10.1016/j.spinee.2019.02.014
29. Stienen MN, Smoll NR, Joswig H, et al. Influence of the mental health status on a new measure of objective functional impairment in lumbar degenerative disc disease. *The Spine Journal*. 2017;17(6):807-813. doi:10.1016/j.spinee.2016.12.004
30. Joswig H, Stienen MN, Smoll NR, et al. Patients' Preference of the Timed Up and Go Test or Patient-Reported Outcome Measures Before and After Surgery for Lumbar Degenerative Disk Disease. *World Neurosurg*. 2017;99:26-30. doi:10.1016/j.wneu.2016.11.039
31. Adamczak SE, Bova FJ, Hoh DJ. Intraoperative 3D Computed Tomography: Spine Surgery. *Neurosurgery Clinics of North America*. 2017;28:585-594. doi:10.1016/j.nec.2017.06.002
32. Bellut D, Hlavica M, Schmid C, Bernays RL. Intraoperative magnetic resonance imaging-assisted transsphenoidal pituitary surgery in patients with acromegaly. *Neurosurg Focus*. 2010;29(4):E9. doi:10.3171/2010.7.FOCUS10164
33. Gronningsaeter A, Kleven A, Ommedal S, et al. SonoWand, an ultrasound-based neuronavigation system. *Neurosurgery*. 2000;47(6):1373-1379; discussion 1379-1380.
34. Archip N, Clatz O, Whalen S, et al. Non-rigid alignment of pre-operative MRI, fMRI, and DT-MRI with intra-operative MRI for enhanced visualization and navigation in image-guided neurosurgery. *NeuroImage*. 2007;35:609-624. doi:10.1016/j.neuroimage.2006.11.060
35. Staartjes VE, Stumpo V, Kernbach JM, et al. Machine learning in neurosurgery: a global survey. *Acta Neurochir (Wien)*. 2020;162(12):3081-3091. doi:10.1007/s00701-020-04532-1

36. Chen Y, Christodoulou AG, Zhou Z, Shi F, Xie Y, Li D. MRI Super-Resolution with GAN and 3D Multi-Level DenseNet: Smaller, Faster, and Better. *arXiv:200301217 [cs, eess]*. Published online March 6, 2020. Accessed September 17, 2021. http://arxiv.org/abs/2003.01217
37. Kalfas IH. Machine Vision Navigation in Spine Surgery. *Frontiers in Surgery*. 2021;8:41. doi:10.3389/fsurg.2021.640554
38. Jakubovic R, Guha D, Gupta S, et al. High Speed, High Density Intraoperative 3D Optical Topographical Imaging with Efficient Registration to MRI and CT for Craniospinal Surgical Navigation. *Sci Rep*. 2018;8(1):14894. doi:10.1038/s41598-018-32424-z
39. Guha D, Yang VXD. Perspective review on applications of optics in spinal surgery. *J Biomed Opt*. 2018;23(6):1-8. doi:10.1117/1.JBO.23.6.060601
40. Stumpo V, Kernbach JM, van Niftrik CHB, et al. Machine Learning Algorithms in Neuroimaging: An Overview. In: *Machine Learning in Clinical Neuroscience: Foundations and Clinical Applications*. Acta Neurochirurgica Supplement. Springer International Publishing [in press]; 2022. https://www.springer.com/gp/book/9783030852917
41. Pollard JM, Wen Z, Sadagopan R, Wang J, Ibbott GS. The future of image-guided radiotherapy will be MR guided. *Br J Radiol*. 2017;90(1073):20160667. doi:10.1259/bjr.20160667
42. Dirix P, Haustermans K, Vandecaveye V. The value of magnetic resonance imaging for radiotherapy planning. *Semin Radiat Oncol*. 2014;24(3):151-159. doi:10.1016/j.semradonc.2014.02.003
43. Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiat Oncol*. 2017;12(1):28. doi:10.1186/s13014-016-0747-y
44. Maspero M, Savenije MHF, Dinkla AM, et al. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys Med Biol*. 2018;63(18):185001. doi:10.1088/1361-6560/aada6d
45. Dinkla AM, Wolterink JM, Maspero M, et al. MR-Only Brain Radiation Therapy: Dosimetric Evaluation of Synthetic CTs Generated by a Dilated Convolutional Neural Network. *Int J Radiat Oncol Biol Phys*. 2018;102(4):801-812. doi:10.1016/j.ijrobp.2018.05.058
46. Dinkla AM, Florkow MC, Maspero M, et al. Dosimetric evaluation of synthetic CT for head and neck radiotherapy generated by a patch-based three-dimensional convolutional neural network. *Med Phys*. 2019;46(9):4095-4104. doi:10.1002/mp.13663
47. Siversson C, Nordström F, Nilsson T, et al. Technical Note: MRI only prostate radiotherapy planning using the statistical decomposition algorithm. *Med Phys*. 2015;42(10):6090-6097. doi:10.1118/1.4931417
48. Edmund JM, Kjer HM, Van Leemput K, Hansen RH, Andersen JAL, Andreasen D. A voxel-based investigation for MRI-only radiotherapy of the brain using ultra short echo times. *Phys Med Biol*. 2014;59(23):7501-7519. doi:10.1088/0031-9155/59/23/7501
49. Florkow MC, Zijlstra F, Willemsen K, et al. Deep learning-based MR-to-CT synthesis: The influence of varying gradient echo-based MR images as input channels. *Magn Reson Med*. Published online October 8, 2019. doi:10.1002/mrm.28008
50. Stummer W, Stepp H, Wiestler OD, Pichlmeier U. Randomized, Prospective Double-Blinded Study Comparing 3 Different Doses of 5-Aminolevulinic Acid for Fluorescence-Guided Resections of Malignant Gliomas. *Neurosurgery*. 2017;81(2):230-239. doi:10.1093/neuros/nyx074
51. De Witt Hamer PC, Robles SG, Zwinderman AH, Duffau H, Berger MS. Impact of intraoperative stimulation brain mapping on glioma surgery outcome: a meta-analysis. *J Clin Oncol*. 2012;30(20):2559-2565. doi:10.1200/JCO.2011.38.4818
52. Senft C, Bink A, Franz K, Vatter H, Gasser T, Seifert V. Intraoperative MRI guidance and extent of resection in glioma surgery: a randomised, controlled trial. *Lancet Oncol*. 2011;12(11):997-1003. doi:10.1016/S1470-2045(11)70196-6

53. Payer C, Stern D, Bischof H, Urschler M. Regressing Heatmaps for Multiple Landmark Localization Using CNNs. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*. Springer International Publishing AG; 2016:230-238. doi:10.1007/978-3-319-46723-8_27
54. Ruzevick JJ, Strickland BA, Zada G. Commentary: Machine Vision for Real-Time Intraoperative Anatomic Guidance: A Proof-of-Concept Study in Endoscopic Pituitary Surgery. *Oper Neurosurg (Hagerstown)*. Published online June 15, 2021:opab203. doi:10.1093/ons/opab203
55. Staartjes VE, Serra C, Muscas G, et al. Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study. *Neurosurgical Focus*. 2018;45(5):E12. doi:10.3171/2018.8.FOCUS18243
56. Senders JT, Arnaout O, Karhade AV, et al. Natural and Artificial Intelligence in Neurosurgery: A Systematic Review. *Neurosurgery*. Published online September 7, 2017. doi:10.1093/neuros/nyx384
57. Wertli MM, Held U, Lis A, Campello M, Weiser S. Both positive and negative beliefs are important in patients with spine pain: findings from the Occupational and Industrial Orthopaedic Center registry. *Spine J*. 2018;18(8):1463-1474. doi:10.1016/j.spinee.2017.07.166
58. Burgstaller JM, Wertli MM, Steurer J, et al. The Influence of Pre- and Postoperative Fear Avoidance Beliefs on Postoperative Pain and Disability in Patients With Lumbar Spinal Stenosis: Analysis of the Lumbar Spinal Outcome Study (LSOS) Data. *Spine*. 2017;42(7):E425-E432. doi:10.1097/BRS.0000000000001845
59. Staartjes VE, Broggi M, Zattra CM, et al. Development and external validation of a clinical prediction model for functional impairment after intracranial tumor surgery. *Journal of Neurosurgery*. 2020;1(aop):1-8. doi:10.3171/2020.4.JNS20643
60. Sagberg LM, Drewes C, Jakola AS, Solheim O. Accuracy of operating neurosurgeons' prediction of functional levels after intracranial tumor surgery. *J Neurosurg*. 2017;126(4):1173-1180. doi:10.3171/2016.3.JNS152927
61. Viken HH, Iversen IA, Jakola A, Sagberg LM, Solheim O. When Are Complications After Brain Tumor Surgery Detected? *World Neurosurg*. 2018;112:e702-e710. doi:10.1016/j.wneu.2018.01.137
62. Niculescu-Mizil A, Caruana R. Predicting Good Probabilities with Supervised Learning. In: *Proceedings of the 22Nd International Conference on Machine Learning*. ICML '05. ACM; 2005:625-632. doi:10.1145/1102351.1102430
63. Calster BV, Nieboer D, Vergouwe Y, Cock BD, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005
64. Staartjes VE, Kernbach JM. Letter to the Editor. Importance of calibration assessment in machine learning–based predictive analytics. *Journal of Neurosurgery: Spine*. 2020;32(6):985-987. doi:10.3171/2019.12.SPINE191503
65. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. *arXiv:170604599 [cs]*. Published online August 3, 2017. Accessed December 11, 2019. http://arxiv.org/abs/1706.04599
66. Meng S, Li Q, Zhou Z, et al. Assessment of an Exhaled Breath Test Using High-Pressure Photon Ionization Time-of-Flight Mass Spectrometry to Detect Lung Cancer. *JAMA Network Open*. 2021;4(3):e213486-e213486. doi:10.1001/jamanetworkopen.2021.3486
67. Ranganath VK, Elashoff DA, Khanna D, Park G, Peter JB, Paulus HE. Age Adjustment Corrects for Apparent Differences in Erythrocyte Sedimentation Rate and C-Reactive Protein Values at the Onset of Seropositive Rheumatoid Arthritis in Younger and Older Patients. *The Journal of Rheumatology*.:3.
68. Staartjes VE, Klukowska AM, Schröder ML. Pedicle Screw Revision in Robot-Guided, Navigated, and Freehand Thoracolumbar Instrumentation: A Systematic Review and Meta-Analysis. *World Neurosurg*. 2018;116:433-443.e8. doi:10.1016/j.wneu.2018.05.159

69. Campbell PG, Malone J, Yadla S, et al. Comparison of ICD-9–based, retrospective, and prospective assessments of perioperative complications: assessment of accuracy in reporting. *Journal of Neurosurgery: Spine*. 2010;14(1):16-22. doi:10.3171/2010.9.SPINE10151
70. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Statistics in Medicine*. 2019;38(7):1262-1275. doi:10.1002/sim.7993
71. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*. 2019;38(7):1276-1296. doi:10.1002/sim.7992
72. Mak RH, Endres MG, Paik JH, et al. Use of Crowd Innovation to Develop an Artificial Intelligence-Based Solution for Radiation Therapy Targeting. *JAMA Oncol*. 2019;5(5):654-661. doi:10.1001/jamaoncol.2019.0159
73. Roadknight C, Suryanarayanan D, Aickelin U, Scholefield J, Durrant L. An ensemble of machine learning and anti-learning methods for predicting tumour patient survival rates. *arXiv:160706190 [cs]*. Published online October 2015:1-8. doi:10.1109/DSAA.2015.7344863
74. Yang Q, Liu Y, Chen T, Tong Y. Federated Machine Learning: Concept and Applications. *arXiv:190204885 [cs]*. Published online February 13, 2019. Accessed February 2, 2021. http://arxiv.org/abs/1902.04885
75. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media; 2008.
76. Kuhn M, Johnson K. *Applied Predictive Modelling*. Springer Science & Business Media; 2013.
77. Staartjes V, Regli L, Serra C, eds. *Machine Learning in Clinical Neuroscience: Foundations and Applications*. Springer International Publishing; 2022. doi:10.1007/978-3-030-85292-4
78. Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open*. 2020;10(9):e041537. doi:10.1136/bmjopen-2020-041537
79. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
80. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1(5):206. doi:10.1038/s42256-019-0048-x
81. Tulio Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv e-prints*. 2016;1602:arXiv:1602.04938.
82. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2nd ed. Springer-Verlag; 2009. Accessed March 16, 2018. //www.springer.com/la/book/9780387848570
83. Buhrmester V, Münch D, Arens M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *arXiv preprint arXiv:191112116*. Published online 2019.
84. Castelvecchi D. Can we open the black box of AI? *Nature News*. 2016;538(7623):20.

[Chapter 10]

Summary

Machine learning (ML) has increasingly seen introduction into medical research and practice over the past decade. Trends in “big data” (increased collection of large datasets), computing power, and the availability of programming libraries and educational resources on ML have enables this exponential rise. These techniques have the ability to learn patterns from previously collected data without any specific instructions, and to then apply the generalizable patterns that have been identified to either diagnose (detect), prognosticate (foresee the natural course of a disease), or predict (estimate the effects of a specific treatment). At least, that has been the current paradigm of what ML can do in medicine. The aim of this thesis was to both illustrate current applications of ML in neurosurgery, but also to expand the aforementioned paradigm by demonstrating an evolution towards new tasks. In **part I** of this thesis, we focus on the assessment of patients with degenerative diseases of the lumbar spine and develop a methodology that enables testing each patient on an individual basis – taking into consideration their specific sociodemographic factors – instead of through comparison to a whole normative population, as is usually the case. **Part II** contains two proof of concept studies that both attempt to expand the limits of current applications of ML to medical imaging: We aim to apply ML to improve operative imaging in both spine and cranial surgery by respectively enabling generation of one imaging modality from another, and by aiding in intraoperative navigation through real-time detection of anatomical structures using machine vision. Finally, **part III** harkens back to the basics of ML and the current paradigm of its applications in medicine. Here, we attempt to predict two endpoints that are highly relevant to patients and that were previously nearly unpredictable, even by seasoned experts: New functional impairment after brain tumour surgery and long-term patient-reported outcomes after lumbar fusion surgery for degenerative disease of the lumbar spine.

Chapter 2 deals with the formal validation of a so-called objective functional test. These tests have become widely applied in studies and in clinical practice throughout several domains in medicine, as they are supposed to deliver an additional dimension of patient assessment, adjunctive to patient questionnaires, physician-rated outcomes, and radiological or biochemical surrogate markers of success. Especially in spine surgery for degenerative disease, where psychological factors play a great role and outcomes described by the surgeon and experienced by the patient have been demonstrated to differ significantly at times, an objective “second opinion” on the level of functional impairment of patients may be useful. The five-repetition sit-to-stand test (5R-STS) is a standardized test that measure objective functional impairment (OFI), and it has been used successfully as an outcome measure in pulmonary disease, movement disorders, and geriatrics. It consists out of standing up fully from a sitting position, five times, as fast as possible, and has demonstrated excellent test properties – although it had not been formally validated in lumbar degenerative disease. By performing the 5R-STS in a group of patients with degenerative disease of the lumbar spine as well as a normative, spine-healthy population, we assessed the convergent validity of this test to see whether it correlates with “gold standard” measurements, namely patient-reported questionnaires. This correlation was moderate, demonstrating that – while patients with a higher level of OFI generally also subjectively feel more impaired according to their answers on the questionnaires – the 5R-STS does differ from a purely subjective assessment. We also

found that the test-retest reliability (how reproducible is the test?) was excellent. Lastly, using the data from the spine-healthy population, we calculated that being able to perform the test in under 10.5 seconds can be considered "normal", and identified other cut-offs that indicate mild, moderate, and severe levels of "abnormality".

Expanding on these findings, **chapter 3** introduces ML to the process of clinical patient assessment using objective tests. Normally, what is considered abnormal and normal is defined according to an entire population of normative data – exactly what is described in chapter 2. This applies not only to functional testing, but also to many other measurements in medicine, such as blood chemistry. However, this approach fails to consider the inherent differences in test properties among different individuals: Some tests are significantly influenced by patient factors, such as body height for the 5R-STS. A larger body height equals a larger distance to travel from sitting down to standing upright. Consequently, a single cut-off for "normality" of 10.5 seconds or higher cut-offs for what is considered mild, moderate, or severe impairment will be unreliable, and such patient factors need to be taken into account. Unsupervised learning algorithms (those that learn to identify patterns in data without a specific goal) can identify clusters of observations that are somehow related, which can aid in discovering new interactions among variables. We trained an unsupervised learning algorithm that divides patients with OFI into three clusters based on a range of demographic factors. Upon further analysis, the three clusters appear to roughly correspond to mild, moderate, and severe impairment based on patient-reported endpoints such as depression, bedriddenness, and subjective functional impairment – without the use of any fixed threshold.

The findings gained in these two chapters are then integrated into a holistic personalized testing process in **chapter 4**. While the previous chapter describes a method to grade the severity of impairment in those patients who are functionally impaired according to sit-to-stand testing, the goal here was to decide whether a patient really is functionally impaired at all or not, taking into account their personalized testing properties. Usually, whether a test result is considered abnormal or not is based on the „upper limit of normal" (ULN) of the values measured in a normative population, which - defining the ULN as the 99th percentile of the normative population - results in a fixed threshold of 10.5 seconds for the 5R-STS. Again, this approach does not consider any inter-individual differences in test properties such as height, age, gender, and body mass. To achieve a personalized testing strategy, it was thus necessary to develop a method to estimate a personalized ULN for each individual patient. We achieved this by training a ML model that is able to estimate not only a single prediction of a patient's expected 5R-STS performance, but instead an entire distribution of predictions. By calculating the 99th percentile of each patient's distribution, it is now possible to decide - for each patient individually, based on their demographics - whether their 5R-STS performance is within expectations, or whether they are objectively impaired. In the latter case, the clustering algorithm from chapter 3 is applied to grade the level of impairment. This entire testing process, from initial measurement to interpretation, has been integrated into a web-app that enables application (https://neurosurgery.shinyapps.io/5RSTS/). This approach could eventually also be applied to any other kind of quantitative test that has significant confounding factors.

In **chapter 5**, we tackle the issue of surgical planning and intraoperative navigation in instrumented spine surgery. If instrumentation is to be achieved using computer assistance (by use of a visual navigation system or a robotic arm), a computed tomography (CT) scan is usually necessary to enable planning of the screws that are then drilled into the pedicles of the spine. Bony structures can be more accurately displayed on a CT scan than on magnetic resonance imaging (MRI) - However, in most cases, an MRI scan is already available once patients are being considered for this type of intervention. A CT scan also carries with it other potential drawbacks such as ionizing radiation, as well as the additional logistic and financial sequalae of having to plan and carry out a second, separate imaging examination. We demonstrate that

generation of a "synthetic CT" from an MRI of the lumbar spine using deep learning is feasible with relatively high fidelity. Although we only present qualitative and semi-quantitative preliminary results from the first algorithm trained with a rather low amount of training data, the image quality appears to be at least sufficient for pedicle screw trajectory planning. Further studies evaluating the accuracy of synthetic CT-based computer assistance as well as the performance of the algorithm for diagnostic purposes are required.

In **chapter 6**, machine vision is applied to the problem of intraoperative orientation in cranial surgery. In endoscopic or microscopic neurosurgery, locating anatomical structures – particularly when they are distorted by tumours or other pathologies – is not always simple. Several tools, including intraoperative MRI, ultrasound, and electrophysiology, neuronavigation based on preoperative imaging, and fluorescent markers have been introduced to help guide neurosurgeons during complex procedures. Still, all of these methods have significant drawbacks related either to the use of preoperative imaging which becomes unreliable early on during surgery, to their limited specificity or sensitivity, or simply through their costs or logistic burden. Master surgeons often end up relying mostly on their "mind's eye" and anatomical expertise. For less experienced surgeons, relying too much on the aforementioned tools may be hazardous. Our rationale was to demonstrate in a proof-of-concept study that machine vision algorithms can learn to recognize anatomical structures based purely on video footage (from endoscopes or microscopes), without providing any a priori anatomical foundation or preoperative imaging to the algorithm. Our study demonstrated that this is at least feasible for simple structures in endonasal pituitary surgery and provides a seed for the further development of real-time anatomical guidance in cranial surgery.

Chapter 7 deals with a common question that is asked by patients with brain tumours who are told that they may profit from surgery: "Will I be the same after the surgery?". Certainly, experienced neurosurgeons are aware of risk factors for bad neurological outcomes and can roughly estimate which surgeries have a high or low risk of new neurological deficits. Still, studies have demonstrated that even experienced neurosurgeons tend to underestimate complications in this type of surgery, and that their ability to predict which patient may experience new functionally relevant neurological deficits is low on an individual patient level. Therefore, we have developed and externally validated a clinical prediction model for new functional impairment after intracranial tumour surgery. This model demonstrated fair predictive ability with high generalizability to the external validation cohorts, and that demonstrated excellent calibration – The predictions of risk for new functional impairment correlated well with the true risk. If calibration of such prediction models is high enough, and if the outcome that is targeted is predictable at all, such decision support tools can help in answering the questions that patients and their families frequently ask with a slightly higher degree of objectivity.

Finally, **chapter 8** applies similar principles to patients with degenerative disease of the lumbar spine, including common diagnoses such as intractable chronic low back pain, lumbar spinal stenosis, and instability. Experienced spine surgeons tend to say that deciding which of these patients may profit from lumbar fusion surgery – the stabilization of one or more segments of the spine with the eventual goal of bony "fusion" of these segments – and adequately informing patients about their individual chances of success is more difficult than the surgery itself. Here, clinical prediction models would again seem like a plausible aid. We attempted to predict improvements in daily physical function as well as back and leg pain severity that patients report one year after surgery. However, our conclusion was that accurately predicting chances of success in the individual patient remains as difficult as previously: The clinical prediction models were only slightly better than random chance at identifying which patients might profit most from surgery. Surgical decision-making in patients with chronic degenerative disease of the lumbar spine is notoriously tricky and appears to remain so until modelling improves to a great extent.

[Chapter 11]

Samenvatting

In de afgelopen tien jaar is machine learning (ML) een steeds grotere rol gaan spelen in het medische onderzoek en de medische praktijk. Deze exponentiële toename was mogelijk door big-data-trends (meer verzameling van grote datasets), rekenkracht en de beschikbaarheid van programmeerbibliotheken en onderwijsmiddelen over ML. Deze technieken kunnen zonder specifieke instructies patronen leren uit eerder verzamelde gegevens. Die generaliseerbare patronen kunnen vervolgens worden toegepast voor diagnose (opsporing), prognose (voorspelling van het natuurlijke verloop van een ziekte), of voorspelling (inschatting van het effect van een specifieke behandeling). Dat is in elk geval zo in het huidige paradigma over de mogelijkheden van ML in de geneeskunde. Het doel van dit proefschrift is de huidige toepassingen van ML in de neurochirurgie illustreren, en bovengenoemd paradigma uitbreiden door een evolutie naar nieuwe taken aan te tonen. In **deel I** van dit proefschrift richten we ons op de beoordeling van patiënten met degeneratieve aandoeningen van de lumbale wervelkolom. We ontwikkelen een methodologie waarmee elke patiënt op individuele basis kan worden getest. Hierbij houden we rekening met hun specifieke sociodemografische factoren, in plaats van hen zoals gewoonlijk te vergelijken met een hele normatieve populatie. In **deel II** komen twee 'proof-of-concept'-onderzoeken aan bod. In beide onderzoeken wordt geprobeerd de grenzen van de huidige toepassingen van ML bij medische beeldvorming te verleggen. Ons doel is ML toe te passen om de operatieve beeldvorming in de wervelkolom- en de schedelchirurgie te verbeteren, respectievelijk door de ene beeldvormingsmodaliteit uit de andere te genereren en door te helpen bij de intraoperatieve navigatie met real-timedetectie van anatomische structuren door middel van machine vision. Tot slot grijpt **deel III** terug op de grondbeginselen van ML en het huidige paradigma van de toepassingen van ML in de geneeskunde. Hier proberen we twee eindpunten te voorspellen die bijzonder relevant zijn voor patiënten en die voorheen zelfs door zeer ervaren deskundigen vrijwel niet te voorspellen waren: nieuwe functionele beperkingen na een hersentumoroperatie, en door de patiënt gemelde langetermijnresultaten na een lumbale fusieoperatie bij degeneratieve aandoeningen van de lumbale wervelkolom.

Hoofdstuk 2 beschrijft de formele validatie van een zogeheten objectieve functionele test. Deze tests worden inmiddels op grote schaal toegepast in het onderzoek en de klinische praktijk op verschillende geneeskundige gebieden. De bedoeling is dat ze een extra dimensie geven aan de beoordeling van de patiënt in aanvulling op vragenlijsten, door de arts beoordeelde resultaten, en radiologische of biochemische surrogaatmarkers voor succes. Bij wervelkolomchirurgie voor degeneratieve ziekten spelen psychische factoren een grote rol en blijken de door de chirurg beschreven en door de patiënt ervaren resultaten soms aanzienlijk te verschillen. Daarom kan vooral daar een objectieve second opinion over het niveau van de functionele beperkingen van patiënten nuttig zijn. De zitten-naar-staan-test met vijf herhalingen (five-repetition sit-to-stand test; 5R-STS) is een gestandaardiseerde test waarmee objectieve functionele beperkingen (objective functional impairment; OFI) gemeten kunnen worden. Deze test wordt met succes toegepast als middel om het resultaat te meten bij longziekten en bewegingsaandoeningen en in de geriatrie. De patiënt wordt gevraagd om vijf keer achter elkaar zo snel mogelijk volledig op te

staan vanuit een zittende positie. De 5R-STS-test bleek uitstekende testeigenschappen te hebben, maar is nog niet formeel gevalideerd bij lumbale degeneratieve aandoeningen. We hebben de 5R-STS uitgevoerd bij een groep patiënten met degeneratieve aandoeningen van de lumbale wervelkolom en bij een normatieve populatie met een gezonde wervelkolom. Zo hebben we de convergente validiteit van deze test beoordeeld om te zien of deze correleert met de 'gouden standaard'-metingen, namelijk patiëntenvragenlijsten. Er was een matige correlatie, waaruit blijkt dat de 5R-STS verschilt van een puur subjectieve beoordeling, alhoewel patiënten met een hoger OFI-niveau zich volgens hun antwoorden op de vragenlijsten over het algemeen ook subjectief slechter voelen. De test-hertestbetrouwbaarheid (hoe reproduceerbaar de test is) bleek uitstekend te zijn. Tenslotte hebben we aan de hand van gegevens uit de populatie met een gezonde wervelkolom berekend dat een patiënt de test in minder dan 10,5 seconden moet kunnen uitvoeren om als 'normaal' beschouwd te worden. Ook hebben we andere grenswaarden vastgesteld die duiden op lichte, matige en ernstige niveaus van 'abnormaliteit'.

Voortbordurend op deze bevindingen wordt ML in **hoofdstuk 3** geïntroduceerd in het proces van klinische patiëntenbeoordeling met behulp van objectieve tests. Gewoonlijk wordt er aan de hand van een hele populatie van normatieve gegevens bepaald wat als abnormaal en normaal beschouwd wordt, zoals beschreven in hoofdstuk 2. Dat geldt niet alleen voor functionele tests, maar ook voor veel andere geneeskundige bepalingen, zoals bloedchemie. Deze benadering gaat echter voorbij aan de inherente verschillen in testeigenschappen tussen verschillende personen. Sommige tests worden aanzienlijk beïnvloed door patiëntfactoren, zoals de lichaamslengte bij de 5R-STS. Hoe langer de patiënt is, hoe groter de afstand is die moet worden afgelegd van zitten tot rechtop staan. Er moet dus rekening gehouden worden met zulke patiëntfactoren, want één specifieke grenswaarde van 10,5 seconden voor 'normaliteit' of specifieke hogere grenswaarden voor lichte, matige of ernstige functionele beperkingen zijn hierdoor onbetrouwbaar. Zonder toezicht lerende algoritmen (algoritmen die patronen in gegevens leren herkennen zonder een bepaald doel) kunnen clusters van waarnemingen identificeren die op de een of andere manier samenhangen. Dat kan helpen om nieuwe interacties tussen variabelen te ontdekken. Wij hebben een zonder toezicht lerend algoritme getraind dat patiënten met OFI verdeelt in drie clusters op basis van een aantal demografische gegevens. Bij verdere analyse blijken de drie clusters ruwweg overeen te komen met lichte, matige en ernstige beperkingen, op basis van door de patiënt gemelde eindpunten zoals depressie, bedlegerigheid, en subjectieve functionele beperkingen, zonder gebruik van een vaste drempelwaarde.

De resultaten uit deze twee hoofdstukken worden vervolgens in **hoofdstuk 4** geïntegreerd in een holistisch gepersonaliseerd testproces. In het vorige hoofdstuk wordt een methode beschreven om de ernst van de beperking te rangschikken bij patiënten die volgens de zitten-naar-staan-test functioneel beperkt zijn. In dit hoofdstuk was het doel echter om te beslissen of een patiënt werkelijk functioneel beperkt is of niet, waarbij rekening wordt gehouden met zijn/haar gepersonaliseerde testeigenschappen. Gewoonlijk wordt een testresultaat al dan niet als abnormaal beschouwd op basis van de bovengrens van de normaalwaarde ('upper limit of normal'; ULN) van de waarden die in een normatieve populatie gemeten worden. Bij definiëring van de ULN als het 99e percentiel van de normatieve populatie resulteert dit in een vaste drempelwaarde van 10,5 seconden voor de 5R-STS. Ook bij deze aanpak wordt geen rekening gehouden met eventuele verschillen in testeigenschappen tussen individuen, zoals lengte, leeftijd, geslacht en lichaamsgewicht. Voor een gepersonaliseerde teststrategie moesten we dus een methode ontwikkelen om voor elke individuele patiënt een gepersonaliseerde ULN te schatten. Dat bereikten we door een ML-model te trainen dat een hele verdeling van voorspellingen kan schatten in plaats van slechts één enkele voorspelling van de verwachte 5R-STS-prestaties van een patiënt. Voor elke patiënt afzonderlijk kan nu worden bepaald of zijn/haar 5R-STS-prestatie binnen de verwachtingen ligt, of dat hij/zij een objectieve beperking heeft. Hiervoor wordt het 99e percentiel van zijn/haar verdeling berekend op basis van individuele demografische gegevens. In het laatste geval wordt het

clusteringalgoritme uit hoofdstuk 3 toegepast om de mate van beperking te rangschikken. Dit hele testproces, van de eerste meting tot de interpretatie, is geïntegreerd in een webapp waarmee de test kan worden toegepast (https://neurosurgery.shinyapps.io/5RSTS/). Deze aanpak zou uiteindelijk ook gebruikt kunnen worden voor elke ander soort kwantitatieve test met significante verstorende factoren.

Hoofdstuk 5 gaat over de kwestie van chirurgische planning en intraoperatieve navigatie bij geïnstrumenteerde wervelkolomchirurgie. Als de implantaten geplaatst moeten worden met behulp van de computer (met een visueel navigatiesysteem of een robotarm), is er meestal een computertomografie-scan (CT) nodig voor operatieve planning van de schroeven die dan in de ruggenwervels worden gedraaid. De botstructuren kunnen op een CT-scan nauwkeuriger worden weergegeven dan met magnetische-resonantiebeeldvorming (MRI). Echter, een CT-scan brengt potentiële nadelen met zich mee. Daarbij valt te denken aan ioniserende straling, en de bijkomende logistieke en financiële consequenties van het plannen en uitvoeren van een tweede afzonderlijk beeldvormend onderzoek. In de meeste gevallen is er echter al een MRI-scan beschikbaar wanneer patiënten in aanmerking komen voor dit soort ingrepen. Wij tonen aan dat het haalbaar is om met deep learning een 'synthetische CT' te genereren uit een MRI van de lendenwervelkolom, met een betrekkelijk hoge betrouwbaarheid. We presenteren alleen kwalitatieve en semi-kwantitatieve voorlopige resultaten van het eerste algoritme, dat getraind is met een vrij geringe hoeveelheid trainingsgegevens (8 pre-operatieve CT-scans), maar de beeldkwaliteit blijkt in elk geval voldoende voor de trajectplanning van de pedikelschroef. Er is verder onderzoek nodig voor het evalueren van de nauwkeurigheid van synthetische op CT gebaseerde computerondersteuning en van de prestaties van het algoritme voor diagnostische en klinische doeleinden.

In **hoofdstuk 6** wordt machine vision toegepast op het probleem van de intraoperatieve oriëntatie in de schedelchirurgie. Bij endoscopische of microscopische neurochirurgie is het lokaliseren van anatomische structuren niet altijd eenvoudig, vooral wanneer die vervormd zijn door tumoren of andere pathologieën. Neurochirurgen gebruiken verschillende hulpmiddelen voor het begeleiden bij complexe ingrepen, zoals intraoperatieve MRI, echografie en elektrofysiologie, neuronavigatie op basis van preoperatieve beeldvorming, en fluorescerende markers. Al deze methoden hebben wel belangrijke nadelen. Zo wordt preoperatieve beeldvorming al vroeg tijdens de operatie onbetrouwbaar, hebben de methoden een beperkte specificiteit of gevoeligheid, of zijn ze gewoonweg duur of logistiek belastend. Ervaren chirurgen vertrouwen uiteindelijk vaak vooral op hun 'geestesoog' en hun anatomische expertise. Voor minder ervaren chirurgen kan het gevaarlijk zijn om te veel te vertrouwen op de bovengenoemde hulpmiddelen. Onze gedachtegang was om in een proof-of-conceptonderzoek aan te tonen dat machine-visionalgoritmen anatomische structuren kunnen leren herkennen puur op basis van (endoscopische of microscopische) videobeelden, zonder het algoritme vooraf te voorzien van een anatomische basis of preoperatieve beeldvorming. Uit ons onderzoek blijkt dat dit in elk geval haalbaar is bij eenvoudige structuren in endonasale hypofysechirurgie. Daarnaast geeft het onderzoek een eerste aanzet voor de verdere ontwikkeling van real-time anatomische begeleiding bij schedelchirurgie.

Hoofdstuk 7 behandelt een veelvoorkomende vraag van patiënten met hersentumoren die te horen krijgen dat ze baat kunnen hebben bij een operatie: 'Ben ik nog wel dezelfde persoon na de operatie?' Ervaren neurochirurgen zijn zich zeker bewust van risicofactoren voor slechte neurologische resultaten en kunnen globaal inschatten welke operaties een hoog of laag risico geven op nieuwe neurologische gebreken. Uit onderzoek is echter gebleken dat zelfs ervaren neurochirurgen complicaties bij dit soort operaties vaak onderschatten, en dat ze op het niveau van de individuele patiënt niet goed kunnen voorspellen welke patiënten nieuwe neurologische gebreken met invloed op het functioneren zullen krijgen. Daarom hebben we een klinisch voorspellingsmodel voor nieuwe functionele beperkingen na intracraniale tumorchirurgie ontwikkeld en extern gevalideerd. Dit model bleek een redelijk voorspellend vermogen te hebben met een hoge generaliseerbaarheid naar de externe validatiecohorten. De kalibratie

was dus uitstekend: de voorspellingen van het risico op nieuwe functionele beperkingen correleerden goed met het werkelijke risico. Met een voldoende hoge kalibratie en als het beoogde resultaat überhaupt voorspelbaar is, kunnen zulke hulpmiddelen voor besluitvorming helpen om iets objectiever antwoord te geven op veelvoorkomende vragen van patiënten en hun familie.

Tot slot worden in **hoofdstuk 8** soortgelijke methodes toegepast op patiënten met degeneratieve aandoeningen van de lumbale wervelkolom, waaronder veelvoorkomende diagnoses zoals hardnekkige chronische lage rugpijn, lumbale spinale stenose, en instabiliteit. Bij een lumbale fusieoperatie worden een of meer segmenten van de wervelkolom gestabiliseerd, met als uiteindelijk doel het aaneengroeien van het bot in die segmenten. Ervaren wervelkolomchirurgen zeggen wel eens dat het uitvoeren van een fusieoperatie makkelijker is dan het besluiten wie van de patiënten baat kunnen hebben bij deze operatie of patiënten goed kunnen informeren over de individuele slagingskans. Ook hier lijken klinische voorspellingsmodellen plausibel als hulpmiddel. We hebben geprobeerd verbeteringen te voorspellen in de dagelijkse lichamelijke functie en de ernst van de rug- en beenpijn die de patiënten een jaar na de operatie melden. Onze conclusie was echter dat het nauwkeurig voorspellen van de slagingskans bij de individuele patiënt even moeilijk is als voorheen. De klinische voorspellingsmodellen konden slechts iets beter dan het toeval vaststellen welke patiënten het meest baat zouden hebben bij een operatie. Het is bekend dat chirurgische besluitvorming bij patiënten met chronische degeneratieve aandoeningen van de lumbale wervelkolom lastig is, en dat lijkt ook zo te blijven totdat de modellering sterk verbetert.

[Appendix 1]

List of Abbreviations

5-ALA	5-Aminolevulinic Acid
5R-STS	Five-Repetition Sit-to-Stand Test
6MWT	Six-Minute Walk Test
ADAM	Adaptive Moment Estimation
ADL	Activities of Daily Living
AI	Artificial Intelligence
ALIF	Anterior Lumbar Interbody Fusion
ANOVA	Analysis of Variance
ASA	American Society of Anesthesiologists
AUC	Area under the Curve
BMI	Body Mass Index
CHA_2DS_2-VASc	Congestive Heart Failure, Hypertension, Age, Diabetes Mellitus, Stroke/TIA, Vascular Disease, Age, and Sex Score
CI	Confidence Interval
CLBP	Chronic Low Back Pain
CNN	Convolutional Neural Network
COMI	Core Outcome Measures Index
CT	Computed Tomography
$CTDI_{vol}$	Volume-Computed Tomography Dose Index
DDD	Degenerative Disc Disease
EOR	Extent of Resection
EQ-5D	EuroQOL-5D-3L Questionnaire
ESR	Erythrocyte Sedimentation Rate
FBSS	Failed Back Surgery Syndrome
FN	False Negative
FP	False Positive
GAM	Generalized Additive Model
GLM	Generalized Linear Model
GTR	Gross Total Resection
HRQOL	Health-Related Quality of Life
ICC	Intraclass Correlation Coefficient
IDH	Isocitrate Dehydrogenase
IQR	Interquartile Range
IRB	Institutional Review Board
KNN	*K*-Nearest Neighbor
KPS	Karnofsky Performance Status
LDH	Lumbar Disc Herniation
LIME	Local Interpretable Model-Agnostic Explanations
LOESS	Locally Estimated Scatterplot Smoothing
LSS	Lumbar Spinal Stenosis
MAE	Mean Absolute Error

MAR	Missing at Random
MCAR	Missing Completely at Random
MCID	Minimum Clinically Important Difference
mGy	Milligray
ML	Machine Learning
MPR	Multiplanar Reconstruction
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
NPV	Negative Predictive Value
NRS	Numeric Rating Scale
ODI	Oswestry Disability Questionnaire
OFI	Objective Functional Impairment
PASS	Patient-Acceptable Symptom State
PHASES	Population, Hypertension, Age, Size, Earlier Subarachnoid Hemorrhage, and Site Score
PLIF	Posterior Lumbar Interbody Fusion
PPS	Palliative Performance Status
PPV	Positive Predictive Value
PROM	Patient-Reported Outcome Measure
RANAS	Radiationless Navigated Surgery
RFE	Recursive Feature Elimination
RMDQ	Roland-Morris Disability Questionnaire
RMSE	Root Mean Squared Error
sCT	Synthetic Computed Tomography
TLIF	Transforaminal Lumbar Interbody Fusion
TN	True Negative
TP	True Positive
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
TUG	Timed Up-and-Go Test
ULN	Upper Limit of Normal
VAS	Visual Analogue Scale
WCSS	Within-Cluster Sum of Squares
ZOI	Zone of Indifference

[Appendix 2]

PhD Portfolio

Courses	Date achieved	EC
External courses		
Verfassen einer wissenschaftlichen Arbeit	12-12-2017	4.00
Klinische Epidemiologie / Evidenzbasierte Medizin	27-11-2018	4.00
AMS - Research Integrity	23-10-2020	2.00
Swiss Good Clinical Practice (GCP) Course Module 1	19-11-2020	0.30
Swiss Good Clinical Practice (GCP) Course Module 2	26-11-2020	0.30
Swiss Good Clinical Practice (GCP) Course Module 3	01-03-2021	0.30
AMS - Writing a Data Management Plan	15-07-2021	1.00
Research related		
EUROSPINE 2018 Barcelona	21-09-2018	2.00
EANS 2018 Brussels	25-10-2018	2.00
NASBS 2019 Orlando	17-02-2019	2.00
EANS 2019 Dublin	28-09-2019	2.00
4th SFCNS Congress Lausanne	23-10-2019	2.00
AMS - Attendance AMS Annual Meeting	18-03-2021	1.50
EANS Young Neurosurgeons Meeting	19-07-2021	2.00
Other academic activities		
AMS - Attendance AMS PhD Day	06-11-2020	0.20
1st Zurich Machine Intelligence in Clinical Neuroscience Symposium	21-01-2021	0.00
AMS - Attendance AMS PhD Day	27-05-2021	0.20
Teaching/Student supervision		
Neurowissenschaftliches Seminar Sommersemester 2021	29-06-2021	0.05
AMS - Received supervision by (daily) supervisor	01-07-2021	4.00
AMS - Received supervision by (daily) supervisor	01-07-2021	4.00
		+

Total number of ECTS credits		33.85

[Appendix 3]

List of Publications

Original Articles (70)

1. **Staartjes** VE, Stumpo V, Ricciardi L, Maldaner N, Eversdijk HAJ, Vieli M, Ciobanu-Caraus O, Raco A, Miscusi M, Perna A, Proietti L, Lofrese G, Dughiero M, Cultrera F, Nicassio N, An SB, Ha Y, Amelot A, Alcobendas I, Viñuela-Prieto JM, Gandía-González ML, Girod PP, Lener S, Kögl N, Abramovic A, Safa NA, Laux CJ, Farshad M, O'Riordan D, Loibl M, Mannion AF, Scerrati A, Molliqaj G, Tessitore E, Schröder ML, Vandertop WP, Stienen MN, Regli L, Serra C. FUSE-ML: development and external validation of a clinical prediction model for mid-term outcomes after lumbar spinal fusion for degenerative disease. Eur Spine J. 2022 Feb 21. doi: 10.1007/s00586-022-07135-9. Epub ahead of print. PMID: 35188587.
2. Mohamed M, Alamri A, Mohamed M, Khalid N, O'Halloran P, **Staartjes** V, Uff C. Prognosticating outcome using magnetic resonance imaging in patients with moderate to severe traumatic brain injury: a machine learning approach. Brain Inj. 2022 Feb 7:1-6. doi: 10.1080/02699052.2022.2034184. Epub ahead of print. PMID: 35129403.
3. Siccoli A, **Staartjes** VE, Klukowska AM, Muizelaar JP, Schröder ML. Overweight and smoking promote recurrent lumbar disk herniation after discectomy. Eur Spine J. 2022 Mar;31(3):604-613. doi: 10.1007/s00586-022-07116-y. Epub 2022 Jan 24. PMID: 35072795.
4. Klukowska AM, **Staartjes** VE, Vandertop WP, Schröder ML. Five-Repetition Sit- to-Stand Test Performance in Healthy Individuals: Reference Values and Predictors From 2 Prospective Cohorts. Neurospine. 2021 Dec;18(4):760-769. doi: 10.14245/ns.2142750.375. Epub 2021 Dec 31. PMID: 35000330; PMCID: PMC8752709.
5. **Staartjes** VE, Klukowska AM, Vieli M, Niftrik CHBV, Stienen MN, Serra C, Regli L, Vandertop WP, Schröder ML. Machine learning-augmented objective functional testing in the degenerative spine: quantifying impairment using patient-specific five-repetition sit-to-stand assessment. Neurosurg Focus. 2021 Nov;51(5):E8. doi: 10.3171/2021.8.FOCUS21386. PMID: 34724641.
6. Zanier O, Zoli M, **Staartjes** VE, Guaraldi F, Asioli S, Rustici A, Picciola VM, Pasquini E, Faustini-Fustini M, Erlic Z, Regli L, Mazzatenta D, Serra C. Machine learning-based clinical outcome prediction in surgery for acromegaly. Endocrine. 2022 Feb;75(2):508-515. doi: 10.1007/s12020-021-02890-z. Epub 2021 Oct 12. PMID: 34642894; PMCID: PMC8816764.
7. Akeret K, Vasella F, **Staartjes** VE, Velz J, Müller T, Neidert MC, Weller M, Regli L, Serra C, Krayenbühl N. Anatomical phenotyping and staging of brain tumours. Brain. 2021 Sep 23:awab352. doi: 10.1093/brain/awab352. Epub ahead of print. PMID: 34554211.
8. Enodien B, Taha-Mehlitz S, Bachmann M, **Staartjes** VE, Gripp M, Staudner T, Taha A, Frey D. Analysis of Factors Relevant to Revenue Enhancement in Hernia Interventions (SwissDRG G09). Healthcare (Basel). 2021 Jul 8;9(7):862. doi: 10.3390/healthcare9070862. PMID: 34356240; PMCID: PMC8306973.
9. **Staartjes** VE, Volokitin A, Regli L, Konukoglu E, Serra C. Machine Vision for Real-Time Intraoperative Anatomic Guidance: A Proof-of-Concept Study in Endoscopic Pituitary Surgery. Oper Neurosurg (Hagerstown). 2021 Sep 15;21(4):242-247. doi: 10.1093/ons/opab187. PMID: 34131753.
10. Stumpo V, **Staartjes** VE, Quddusi A, Corniola MV, Tessitore E, Schröder ML, Anderer EG, Stienen MN, Serra C, Regli L. Enhanced Recovery After Surgery strategies for elective craniotomy: a systematic review. J Neurosurg. 2021 May 7:1-25. doi: 10.3171/2020.10.JNS203160. Epub ahead of print. PMID: 33962374.
11. Taha A, Aniukstyte L, Enodien B, **Staartjes** V, Taha-Mehlitz S. Intestinal Malrotation Associated With Invagination of the Distal Ileum and Cancer of the Cecum: A Case Report and Literature Review. Cureus. 2021 Mar 1;13(3):e13637. doi: 10.7759/cureus.13637. PMID: 33816035; PMCID: PMC8011629.
12. Taha A, Taha-Mehlitz S, **Staartjes** VE, Lunger F, Gloor S, Unger I, Mungo G, Tschuor C, Breitenstein S, Gingert C. Association of a prehabilitation program with anxiety and depression before colorectal surgery: a post hoc analysis of the pERACS randomized controlled trial. Langenbecks Arch Surg. 2021 Mar 29. doi: 10.1007/s00423-021-02158-0. Epub ahead of print. PMID: 33782738.
13. Akeret K, Stumpo V, **Staartjes** VE, Vasella F, Velz J, Marinoni F, Dufour JP, Imbach LL, Regli L, Serra C, Krayenbühl N. Topographic brain tumor anatomy drives seizure risk and enables machine learning based prediction.

Neuroimage Clin. 2020;28:102506. doi: 10.1016/j.nicl.2020.102506. Epub 2020 Nov 19. PMID: 33395995; PMCID: PMC7711280.

14. **Staartjes** VE, Seevinck PR, Vandertop WP, van Stralen M, Schröder ML. Magnetic resonance imaging-based synthetic computed tomography of the lumbar spine for surgical planning: a clinical proof-of-concept. Neurosurg Focus. 2021 Jan;50(1):E13. doi: 10.3171/2020.10.FOCUS20801. PMID: 33386013.
15. **Staartjes** VE, Joswig H, Corniola MV, Schaller K, Gautschi OP, Stienen MN. Association of Medical Comorbidities With Objective Functional Impairment in Lumbar Degenerative Disc Disease. Global Spine J. 2020 Dec 17:2192568220979120. doi: 10.1177/2192568220979120. Epub ahead of print. PMID: 33334183.
16. **Staartjes** VE, Battilana B, Schröder ML. Robot-Guided Transforaminal Versus Robot-Guided Posterior Lumbar Interbody Fusion for Lumbar Degenerative Disease. Neurospine. 2021 Mar;18(1):98-105. doi: 10.14245/ns.2040294.147. Epub 2020 Dec 14. PMID: 33332936; PMCID: PMC8021835.
17. Siccoli A, Schröder ML, **Staartjes** VE. Association of age with incidence and timing of recurrence after microdiscectomy for lumbar disc herniation. Eur Spine J. 2021 Apr;30(4):893-898. doi: 10.1007/s00586-020-06692-1. Epub 2020 Dec 14. PMID: 33315158.
18. Siccoli A, Schröder ML, **Staartjes** VE. Influence of dynamic preoperative body mass index changes on patient-reported outcomes after surgery for degenerative lumbar spine disease. Neurosurg Rev. 2020 Dec 11. doi: 10.1007/s10143-020-01454-5. Epub ahead of print. PMID: 33305336.
19. Stumpo V, **Staartjes** VE, Klukowska AM, Golahmadi AK, Gadjradj PS, Schröder ML, Veeravagu A, Stienen MN, Serra C, Regli L. Global adoption of robotic technology into neurosurgical practice and research. Neurosurg Rev. 2020 Nov 30. doi: 10.1007/s10143-020-01445-6. Epub ahead of print. PMID: 33252717.
20. Azad TD, Ehresman J, Ahmed AK, **Staartjes** VE, Lubelski D, Stienen MN, Veeravagu A, Ratliff JK. Fostering reproducibility and generalizability in machine learning for clinical prediction modeling in spine surgery. Spine J. 2020 Oct 13:S1529-9430(20)31143-8. doi: 10.1016/j.spinee.2020.10.006. Epub ahead of print. PMID: 33065274.
21. Jin MC, Ho AL, Feng AY, Zhang Y, **Staartjes** VE, Stienen MN, Han SS, Veeravagu A, Ratliff JK, Desai AM. Predictive modeling of long-term opioid and benzodiazepine use after intradural tumor resection. Spine J. 2020 Oct 13:S1529-9430(20)31147-5. doi: 10.1016/j.spinee.2020.10.010. Epub ahead of print. PMID: 33065272.
22. Maldaner N, Zeitlberger AM, Sosnova M, Goldberg J, Fung C, Bervini D, May A, Bijlenga P, Schaller K, Roethlisberger M, Rychen J, Zumofen DW, D'Alonzo D, Marbacher S, Fandino J, Daniel RT, Burkhardt JK, Chiappini A, Robert T, Schatlo B, Schmid J, Maduri R, **Staartjes** VE, Seule MA, Weyerbrock A, Serra C, Stienen MN, Bozinov O, Regli L. Development of a Complication- and Treatment-Aware Prediction Model for Favorable Functional Outcome in Aneurysmal Subarachnoid Hemorrhage Based on Machine Learning. Neurosurgery. 2021 Jan 13;88(2):E150-E157. doi: 10.1093/neuros/nyaa401. PMID: 33017031.
23. **Staartjes** VE, Stumpo V, Kernbach JM, Klukowska AM, Gadjradj PS, Schröder ML, Veeravagu A, Stienen MN, van Niftrik CHB, Serra C, Regli L. Machine learning in neurosurgery: a global survey. Acta Neurochir (Wien). 2020 Dec;162(12):3081-3091. doi: 10.1007/s00701-020-04532-1. Epub 2020 Aug 18. PMID: 32812067; PMCID: PMC7593280.
24. Maldaner N, Sosnova M, Zeitlberger AM, Ziga M, Gautschi OP, Regli L, Weyerbrock A, Stienen MN; **International 6WT Study Group.** Evaluation of the 6-minute walking test as a smartphone app-based self-measurement of objective functional impairment in patients with lumbar degenerative disc disease. J Neurosurg Spine. 2020 Aug 7;33(6):779-788. doi: 10.3171/2020.5.SPINE20547. PMID: 32764182.
25. Akeret K, **Staartjes** VE, Vasella F, Serra C, Fierstra J, Neidert MC, Regli L, Krayenbühl N. Distinct topographic-anatomical patterns in primary and secondary brain tumors and their therapeutic potential. J Neurooncol. 2020 Aug;149(1):73-85. doi: 10.1007/s11060-020-03574-w. Epub 2020 Jul 8. PMID: 32643065; PMCID: PMC7452943.
26. Sorba EL, **Staartjes** VE, Voglis S, Tosic L, Brandi G, Tschopp O, Serra C, Regli L. Diabetes insipidus and syndrome of inappropriate antidiuresis (SIADH) after pituitary surgery: incidence and risk factors. Neurosurg Rev. 2020 Jun 24. doi: 10.1007/s10143-020-01340-0. Epub ahead of print. PMID: 32583307.
27. **Staartjes** VE, Broggi M, Zattra CM, Vasella F, Velz J, Schiavolin S, Serra C, Bartek J Jr, Fletcher-Sandersjöö A, Förander P, Kalasauskas D, Renovanz M, Ringel F, Brawanski KR, Kerschbaumer J, Freyschlag CF, Jakola AS, Sjåvik K, Solheim O, Schatlo B, Sachkova A, Bock HC, Hussein A, Rohde V, Broekman MLD, Nogarede CO, Lemmens CMC, Kernbach JM, Neuloh G, Bozinov O, Krayenbühl N, Sarnthein J, Ferroli P, Regli L, Stienen MN; FEBNS. Development and external validation of a clinical prediction model for functional impairment after intracranial tumor surgery. J Neurosurg. 2020 Jun 12:1-8. doi: 10.3171/2020.4.JNS20643. Epub ahead of print. PMID: 32534490.
28. Voglis S, van Niftrik CHB, **Staartjes** VE, Brandi G, Tschopp O, Regli L, Serra C. Feasibility of machine learning based predictive modelling of postoperative hyponatremia after pituitary surgery. Pituitary. 2020 Oct;23(5):543-551. doi: 10.1007/s11102-020-01056-w. PMID: 32488759.
29. Serra C, **Staartjes** VE, Maldaner N, Holzmann D, Soyka MB, Gilone M, Schmid C, Tschopp O, Regli L. Assessing the surgical outcome of the "chopsticks" technique in endoscopic transsphenoidal adenoma surgery. Neurosurg Focus. 2020 Jun;48(6):E15. doi: 10.3171/2020.3.FOCUS2065. PMID: 32480377.

30. Zoli M, **Staartjes** VE, Guaraldi F, Friso F, Rustici A, Asioli S, Sollini G, Pasquini E, Regli L, Serra C, Mazzatenta D. Machine learning-based prediction of outcomes of the endoscopic endonasal approach in Cushing disease: is the future coming? Neurosurg Focus. 2020 Jun;48(6):E5. doi: 10.3171/2020.3.FOCUS2060. PMID: 32480364.
31. Tosic L, Goldberger E, Maldaner N, Sosnova M, Zeitlberger AM, **Staartjes** VE, Gadjradj PS, Eversdijk HAJ, Quddusi A, Gandía-González ML, Sayadi JJ, Desai A, Regli L, Gautschi OP, Stienen MN. Normative data of a smartphone app-based 6-minute walking test, test-retest reliability, and content validity with patient-reported outcome measures. J Neurosurg Spine. 2020 May 29:1-10. doi: 10.3171/2020.3.SPINE2084. Epub ahead of print. PMID: 32470938.
32. **Staartjes** VE, Sebök M, Blum PG, Serra C, Germans MR, Krayenbühl N, Regli L, Esposito G. Development of machine learning-based preoperative predictive analytics for unruptured intracranial aneurysm surgery: a pilot study. Acta Neurochir (Wien). 2020 Nov;162(11):2759-2765. doi: 10.1007/s00701-020-04355-0. Epub 2020 May 1. PMID: 32358656.
33. Siccoli A, Wispelaere MP, Schröder ML, **Staartjes** VE. Timing of Surgery in Tubular Microdiscectomy for Lumbar Disc Herniation and Its Effect on Functional Impairment Outcomes. Neurospine. 2020 Mar;17(1):204-212. doi: 10.14245/ns.1938448.224. Epub 2020 Mar 31. PMID: 32252169; PMCID: PMC7136121.
34. Serra C, Akeret K, **Staartjes** VE, Ramantani G, Grunwald T, Jokeit H, Bauer J, Krayenbühl N. Safety of the paramedian supracerebellar-transtentorial approach for selective amygdalohippocampectomy. Neurosurg Focus. 2020 Apr 1;48(4):E4. doi: 10.3171/2020.1.FOCUS19909. PMID: 32234984.
35. **Staartjes** VE, Serra C, Zoli M, Mazzatenta D, Pozzi F, Locatelli D, D'Avella E, Solari D, Cavallo LM, Regli L. Multicenter external validation of the Zurich Pituitary Score. Acta Neurochir (Wien). 2020 Jun;162(6):1287-1295. doi: 10.1007/s00701-020-04286-w. Epub 2020 Mar 14. PMID: 32172439.
36. Klukowska AM, Schröder ML, Stienen MN, **Staartjes** VE. Objective functional impairment in lumbar degenerative disease: concurrent validity of the baseline severity stratification for the five-repetition sit-to-stand test. J Neurosurg Spine. 2020 Feb 21:1-8. doi: 10.3171/2019.12.SPINE191124. Epub ahead of print. PMID: 32084632.
37. **Staartjes** VE, Quddusi A, Klukowska AM, Schröder ML. Initial classification of low back and leg pain based on objective functional testing: a pilot study of machine learning applied to diagnostics. Eur Spine J. 2020 Jul;29(7):1702-1708. doi: 10.1007/s00586-020-06343-5. Epub 2020 Feb 18. PMID: 32072271.
38. Siccoli A, **Staartjes** VE, de Wispelaere MP, Schröder ML. Association of time to surgery with leg pain after lumbar discectomy: is delayed surgery detrimental? J Neurosurg Spine. 2019 Oct 25;32(2):160-167. doi: 10.3171/2019.8.SPINE19613. PMID: 31653820.
39. Quddusi A, Eversdijk HAJ, Klukowska AM, de Wispelaere MP, Kernbach JM, Schröder ML, **Staartjes** VE. External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. Eur Spine J. 2020 Feb;29(2):374-383. doi: 10.1007/s00586-019-06189-6. Epub 2019 Oct 22. PMID: 31641905.
40. Vieli M, **Staartjes** VE, Eversdjik HAJ, De Wispelaere MP, Oosterhuis JWA, Schröder ML. Safety and Efficacy of Anterior Lumbar Interbody Fusion for Discogenic Chronic Low Back Pain in a Short-stay Setting: Data From a Prospective Registry. Cureus. 2019 Aug 7;11(8):e5332. doi: 10.7759/cureus.5332. PMID: 31598439; PMCID: PMC6777969.
41. Stienen MN, Gautschi OP, **Staartjes** VE, Maldaner N, Sosnova M, Ho AL, Veeravagu A, Desai A, Zygourakis CC, Park J, Regli L, Ratliff JK. Reliability of the 6-minute walking test smartphone application. J Neurosurg Spine. 2019 Sep 13:1-8. doi: 10.3171/2019.6.SPINE19559. Epub ahead of print. PMID: 31518975.
42. **Staartjes** VE, Molliqaj G, van Kampen PM, Eversdijk HAJ, Amelot A, Bettag C, Wolfs JFC, Urbanski S, Hedayat F, Schneekloth CG, Abu Saris M, Lefranc M, Peltier J, Boscherini D, Fiss I, Schatlo B, Rohde V, Ryang YM, Krieg SM, Meyer B, Kögl N, Girod PP, Thomé C, Twisk JWR, Tessitore E, Schröder ML. The European Robotic Spinal Instrumentation (EUROSPIN) study: protocol for a multicentre prospective observational study of pedicle screw revision surgery after robot- guided, navigated and freehand thoracolumbar spinal fusion. BMJ Open. 2019 Sep 8;9(9):e030389. doi: 10.1136/bmjopen-2019-030389. PMID: 31501123; PMCID: PMC6738706.
43. **Staartjes** VE, Klukowska AM, Schröder ML. Association of maximum back and leg pain severity with objective functional impairment as assessed by five- repetition sit-to-stand testing: analysis of two prospective studies. Neurosurg Rev. 2020 Oct;43(5):1331-1338. doi: 10.1007/s10143-019-01168-3. Epub 2019 Aug 26. PMID: 31451936.
44. **Staartjes** VE, Serra C, Maldaner N, Muscas G, Tschopp O, Soyka MB, Holzmann D, Regli L. The Zurich Pituitary Score predicts utility of intraoperative high- field magnetic resonance imaging in transsphenoidal pituitary adenoma surgery. Acta Neurochir (Wien). 2019 Oct;161(10):2107-2115. doi: 10.1007/s00701-019-04018-9. Epub 2019 Aug 7. PMID: 31392567.
45. **Staartjes** VE, Zattra CM, Akeret K, Maldaner N, Muscas G, Bas van Niftrik CH, Fierstra J, Regli L, Serra C. Neural network-based identification of patients at high risk for intraoperative cerebrospinal fluid leaks in endoscopic pituitary surgery. J Neurosurg. 2019 Jun 21:1-7. doi: 10.3171/2019.4.JNS19477. Epub ahead of print. PMID: 31226693.
46. van Niftrik CHB, van der Wouden F, **Staartjes** VE, Fierstra J, Stienen MN, Akeret K, Sebök M, Fedele T, Sarnthein J, Bozinov O, Krayenbühl N, Regli L, Serra C. Machine Learning Algorithm Identifies Patients at High Risk for Early

Complications After Intracranial Tumor Surgery: Registry-Based Cohort Study. Neurosurgery. 2019 Oct 1;85(4):E756-E764. doi: 10.1093/neuros/nyz145. PMID: 31149726.

47. Schröder ML, de Wispelaere MP, **Staartjes** VE. Predictors of loss of follow-up in a prospective registry: which patients drop out 12 months after lumbar spine surgery? Spine J. 2019 Oct;19(10):1672-1679. doi: 10.1016/j.spinee.2019.05.007. Epub 2019 May 21. PMID: 31125698.
48. Siccoli A, de Wispelaere MP, Schröder ML, **Staartjes** VE. Machine learning- based preoperative predictive analytics for lumbar spinal stenosis. Neurosurg Focus. 2019 May 1;46(5):E5. doi: 10.3171/2019.2.FOCUS18723. PMID: 31042660.
49. Siccoli A, **Staartjes** VE, De Wispelaere MP, Vergroesen PA, Schröder ML. Tandem Disc Herniation of the Lumbar and Cervical Spine: Case Series and Review of the Epidemiological, Pathophysiological and Genetic Literature. Cureus. 2019 Feb 16;11(2):e4081. doi: 10.7759/cureus.4081. PMID: 31019859; PMCID: PMC6467429.
50. **Staartjes** VE, de Wispelaere MP, Schröder ML. Improving recovery after elective degenerative spine surgery: 5-year experience with an enhanced recovery after surgery (ERAS) protocol. Neurosurg Focus. 2019 Apr 1;46(4):E7. doi: 10.3171/2019.1.FOCUS18646. PMID: 30933924.
51. Akeret K, Serra C, Rafi O, **Staartjes** VE, Fierstra J, Bellut D, Maldaner N, Imbach LL, Wolpert F, Poryazova R, Regli L, Krayenbühl N. Anatomical features of primary brain tumors affect seizure risk and semiology. Neuroimage Clin. 2019;22:101688. doi: 10.1016/j.nicl.2019.101688. Epub 2019 Jan 25. PMID: 30710869; PMCID: PMC6354289.
52. Siccoli A, **Staartjes** VE, de Wispelaere MP, Schröder ML. Is elective degenerative lumbar spine surgery in older adults safe in a short-stay clinic? Data from an institutional registry. Eur Geriatr Med. 2019 Feb;10(1):79-88. doi: 10.1007/s41999-018-0132-5. Epub 2018 Nov 19. PMID: 32720276.
53. **Staartjes** VE, Beusekamp F, Schröder ML. Can objective functional impairment in lumbar degenerative disease be reliably assessed at home using the five- repetition sit-to-stand test? A prospective study. Eur Spine J. 2019 Apr;28(4):665-673. doi: 10.1007/s00586-019-05897-3. Epub 2019 Jan 24. PMID: 30680635.
54. Serra C, Akeret K, Maldaner N, **Staartjes** VE, Regli L, Baltsavias G, Krayenbühl N. A White Matter Fiber Microdissection Study of the Anterior Perforated Substance and the Basal Forebrain: A Gateway to the Basal Ganglia? Oper Neurosurg (Hagerstown). 2019 Sep 1;17(3):311-320. doi: 10.1093/ons/opy345. PMID: 30476312.
55. **Staartjes** VE, Serra C, Muscas G, Maldaner N, Akeret K, van Niftrik CHB, Fierstra J, Holzmann D, Regli L. Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study. Neurosurg Focus. 2018 Nov 1;45(5):E12. doi: 10.3171/2018.8.FOCUS18243. PMID: 30453454.
56. **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 May;19(5):853-861. doi: 10.1016/j.spinee.2018.11.009. Epub 2018 Nov 16. PMID: 30453080.
57. **Staartjes** VE, Siccoli A, de Wispelaere MP, Schröder ML. Patient-reported outcomes unbiased by length of follow-up after lumbar degenerative spine surgery: Do we need 2 years of follow-up? Spine J. 2019 Apr;19(4):637-644. doi: 10.1016/j.spinee.2018.10.004. Epub 2018 Oct 5. PMID: 30296576.
58. Serra C, **Staartjes** VE, Maldaner N, Muscas G, Akeret K, Holzmann D, Soyka MB, Schmid C, Regli L. Predicting extent of resection in transsphenoidal surgery for pituitary adenoma. Acta Neurochir (Wien). 2018 Nov;160(11):2255-2262. doi: 10.1007/s00701-018-3690-x. Epub 2018 Sep 29. PMID: 30267209.
59. **Staartjes** VE, Stricker S, Muscas G, Maldaner N, Holzmann D, Burkhardt JK, Seifert B, Schmid C, Serra C, Regli L. Intraoperative unfolding and postoperative pruning of the pituitary gland after transsphenoidal surgery for pituitary adenoma: A volumetric and endocrinological evaluation. Endocrine. 2019 Feb;63(2):231-239. doi: 10.1007/s12020-018-1758-2. Epub 2018 Sep 21. PMID: 30242602.
60. Siccoli A, **Staartjes** VE, de Wispelaere MP, Schröder ML. Gender differences in degenerative spine surgery: Do female patients really fare worse? Eur Spine J. 2018 Oct;27(10):2427-2435. doi: 10.1007/s00586-018-5737-3. Epub 2018 Aug 21. PMID: 30132176.
61. **Staartjes** VE, Schröder ML. The five-repetition sit-to-stand test: evaluation of a simple and objective tool for the assessment of degenerative pathologies of the lumbar spine. J Neurosurg Spine. 2018 Oct;29(4):380-387. doi: 10.3171/2018.2.SPINE171416. Epub 2018 Jun 29. PMID: 29957147.
62. **Staartjes** VE, de Wispelaere MP, Schröder ML. Recurrent Laryngeal Nerve Palsy Is More Frequent After Secondary than After Primary Anterior Cervical Discectomy and Fusion: Insights from a Registry of 525 Patients. World Neurosurg. 2018 Aug;116:e1047-e1053. doi: 10.1016/j.wneu.2018.05.162. Epub 2018 Jun 1. PMID: 29864565.
63. Schröder ML, de Wispelaere MP, **Staartjes** VE. Are patient-reported outcome measures biased by method of follow-up? Evaluating paper-based and digital follow-up after lumbar fusion surgery. Spine J. 2019 Jan;19(1):65-70. doi: 10.1016/j.spinee.2018.05.002. Epub 2018 May 3. PMID: 29730459.
64. **Staartjes** VE, Schillevoort SA, Blum PG, van Tintelen JP, Kok WE, Schröder ML. Cardiac Arrest During Spine Surgery in the Prone Position: Case Report and Review of the Literature. World Neurosurg. 2018 Jul;115:460-467.e1. doi: 10.1016/j.wneu.2018.04.116. Epub 2018 Apr 26. PMID: 29704693.

65. **Staartjes** VE, Schröder ML. Effectiveness of a Decision-Making Protocol for the Surgical Treatment of Lumbar Stenosis with Grade 1 Degenerative Spondylolisthesis. World Neurosurg. 2018 Feb;110:e355-e361. doi: 10.1016/j.wneu.2017.11.001. Epub 2017 Nov 10. PMID: 29133000.
66. **Staartjes** VE, Vergroesen PA, Zeilstra DJ, Schröder ML. Identifying subsets of patients with single-level degenerative disc disease for lumbar fusion: the value of prognostic tests in surgical decision making. Spine J. 2018 Apr;18(4):558-566. doi: 10.1016/j.spinee.2017.08.242. Epub 2017 Sep 7. PMID: 28890222.
67. Serra C, Maldaner N, Muscas G, **Staartjes** V, Pangalu A, Holzmann D, Soyka M, Schmid C, Regli L. The changing sella: internal carotid artery shift during transsphenoidal pituitary surgery. Pituitary. 2017 Dec;20(6):654-660. doi: 10.1007/s11102-017-0830-x. PMID: 28828722.
68. **Staartjes** VE, de Wispelaere MP, Miedema J, Schröder ML. Recurrent Lumbar Disc Herniation After Tubular Microdiscectomy: Analysis of Learning Curve Progression. World Neurosurg. 2017 Nov;107:28-34. doi: 10.1016/j.wneu.2017.07.121. Epub 2017 Jul 29. PMID: 28765022.
69. Schröder ML, **Staartjes** VE. Revisions for screw malposition and clinical outcomes after robot-guided lumbar fusion for spondylolisthesis. Neurosurg Focus. 2017 May;42(5):E12. doi: 10.3171/2017.3.FOCUS16534. PMID: 28463610.
70. Zeilstra DJ, **Staartjes** VE, Schröder ML. Minimally invasive transaxial lumbosacral interbody fusion: a ten year single-centre experience. Int Orthop. 2017 Jan;41(1):113-119. doi: 10.1007/s00264-016-3273-5. Epub 2016 Aug 23. PMID: 27553062.

Reviews (7)

1. **Staartjes** VE, Togni-Pogliorini A, Stumpo V, Serra C, Regli L. Impact of intraoperative magnetic resonance imaging on gross total resection, extent of resection, and residual tumor volume in pituitary surgery: systematic review and meta-analysis. Pituitary. 2021 May 4. doi: 10.1007/s11102-021-01147-2. Epub ahead of print. PMID: 33945115.
2. Stumpo V, Latour K, Traylor JI, **Staartjes** VE, Giordano M, Caccavella VM, Olivi A, Ricciardi L, Signorelli F. Medical Student Interest and Recruitment in Neurosurgery. World Neurosurg. 2020 Sep;141:448-454.e6. doi: 10.1016/j.wneu.2020.04.248. Epub 2020 May 11. PMID: 32407916
3. Serra C, Guida L, **Staartjes** VE, Krayenbühl N, Türe U. Historical controversies about the thalamus: from etymology to function. Neurosurg Focus. 2019 Sep 1;47(3):E13. doi: 10.3171/2019.6.FOCUS19331. PMID: 31473672.
4. **Staartjes** VE, Klukowska AM, Sorba EL, Schröder ML. Conflicts of interest in randomized controlled trials reported in neurosurgical journals. J Neurosurg. 2019 Aug 16:1-10. doi: 10.3171/2019.5.JNS183560. Epub ahead of print. PMID: 31419788.
5. Siccoli A, Klukowska AM, Schröder ML, **Staartjes** VE. A Systematic Review and Meta-Analysis of Perioperative Parameters in Robot-Guided, Navigated, and Freehand Thoracolumbar Pedicle Screw Instrumentation. World Neurosurg. 2019 Jul;127:576-587.e5. doi: 10.1016/j.wneu.2019.03.196. Epub 2019 Apr 4. PMID: 30954747.
6. Stienen MN, Ho AL, **Staartjes** VE, Maldaner N, Veeravagu A, Desai A, Gautschi OP, Bellut D, Regli L, Ratliff JK, Park J. Objective measures of functional impairment for degenerative diseases of the lumbar spine: a systematic review of the literature. Spine J. 2019 Jul;19(7):1276-1293. doi: 10.1016/j.spinee.2019.02.014. Epub 2019 Mar 2. PMID: 30831316.
7. **Staartjes** VE, Klukowska AM, Schröder ML. Pedicle Screw Revision in Robot- Guided, Navigated, and Freehand Thoracolumbar Instrumentation: A Systematic Review and Meta-Analysis. World Neurosurg. 2018 Aug;116:433-443.e8. doi: 10.1016/j.wneu.2018.05.159. Epub 2018 May 31. PMID: 29859354.

Invited Articles / Editorials (1)

1. **Staartjes** VE, Stienen MN. Data Mining in Spine Surgery: Leveraging Electronic Health Records for Machine Learning and Clinical Research. Neurospine. 2019 Dec;16(4):654-656. doi: 10.14245/ns.1938434.217. Epub 2019 Dec 31. PMID: 31905453; PMCID: PMC6944992.

Monographs (1)

1. **Staartjes** VE, Regli L, Serra C (Editors): Machine Learning in Clinical Neuroscience: Foundations and Applications Acta Neurochirurgica Supplements (Springer Nature), 2021

Book Chapters (12)

1. Jin MC, Schröder ML, **Staartjes** VE. Artificial Intelligence and Machine Learning in Spine Surgery, In: Veeravagu & Wang: Robotic and Navigated Spine Surgery (Elsevier), 2021 (in press)
2. Stumpo V, Kernbach JM, van Niftrik CHB, Sebök M, Fierstra J, Regli L, Serra C, Staartjes VE. Machine Learning Algorithms in Neuroimaging: An Overview. Acta Neurochir Suppl. 2022;134:125-138. doi: 10.1007/978-3-030-85292-4_17. PMID: 34862537.
3. Stumpo V, Staartjes VE, Regli L, Serra C. Machine Learning in Pituitary Surgery. Acta Neurochir Suppl. 2022;134:291-301. doi: 10.1007/978-3-030-85292-4_33. PMID: 34862553.
4. Staartjes VE, Kernbach JM, Stumpo V, van Niftrik CHB, Serra C, Regli L. Foundations of Feature Selection in Clinical Prediction Modeling. Acta Neurochir Suppl. 2022;134:51-57. doi: 10.1007/978-3-030-85292-4_7. PMID: 34862527.
5. Staartjes VE, Kernbach JM. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part V-A Practical Approach to Regression Problems. Acta Neurochir Suppl. 2022;134:43-50. doi: 10.1007/978-3-030-85292-4_6. PMID: 34862526.
6. Stumpo V, Staartjes VE, Esposito G, Serra C, Regli L, Olivi A, Sturiale CL. Machine Learning and Intracranial Aneurysms: From Detection to Outcome Prediction. Acta Neurochir Suppl. 2022;134:319-331. doi: 10.1007/978-3-030-85292-4_36. PMID: 34862556.
7. Muscas G, Orlandini S, Becattini E, Battista F, Staartjes VE, Serra C, Della Puppa A. Radiomic Features Associated with Extent of Resection in Glioma Surgery. Acta Neurochir Suppl. 2022;134:341-347. doi: 10.1007/978-3-030-85292-4_38. PMID: 34862558.
8. Staartjes VE, Kernbach JM. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part III-Model Evaluation and Other Points of Significance. Acta Neurochir Suppl. 2022;134:23-31. doi: 10.1007/978-3-030-85292-4_4. PMID: 34862524.
9. Staartjes VE, Kernbach JM. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part IV-A Practical Approach to Binary Classification Problems. Acta Neurochir Suppl. 2022;134:33-41. doi: 10.1007/978-3-030-85292-4_5. PMID: 34862525.
10. Kernbach JM, Staartjes VE. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II-Generalization and Overfitting. Acta Neurochir Suppl. 2022;134:15-21. doi: 10.1007/978-3-030-85292-4_3. PMID: 34862523.
11. Kernbach JM, Staartjes VE. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part I-Introduction and General Principles. Acta Neurochir Suppl. 2022;134:7-13. doi: 10.1007/978-3-030-85292-4_2. PMID: 34862522.
12. Staartjes VE, Regli L, Serra C. Machine Intelligence in Clinical Neuroscience: Taming the Unchained Prometheus. Acta Neurochir Suppl. 2022;134:1-4. doi: 10.1007/978-3-030-85292-4_1. PMID: 34862521.

Letters (7)

1. Kernbach JM, **Staartjes** VE. Predicted Prognosis of Pancreatic Cancer Patients by Machine Learning-Letter. Clin Cancer Res. 2020 Jul 15;26(14):3891. doi: 10.1158/1078-0432.CCR-20-0523. PMID: 32669273.
2. **Staartjes** VE, Kernbach JM. Significance of external validation in clinical machine learning: let loose too early? Spine J. 2020 Jul;20(7):1159-1160. doi: 10.1016/j.spinee.2020.02.016. PMID: 32624150.
3. **Staartjes** VE, Kernbach JM. Letter to the Editor Regarding "Investigating Risk Factors and Predicting Complications in Deep Brain Stimulation Surgery with Machine Learning Algorithms". World Neurosurg. 2020 May;137:496. doi: 10.1016/j.wneu.2020.01.189. PMID: 32365450.
4. **Staartjes** VE, Kernbach JM. Letter to the Editor. Importance of calibration assessment in machine learning-based predictive analytics. J Neurosurg Spine. 2020 Feb 21:1-2. doi: 10.3171/2019.12.SPINE191503. Epub ahead of print. PMID: 32084640.
5. **Staartjes** VE, Siccoli A, de Wispelaere MP, Schröder ML. Authors' Response to Letter to Editor: Patient-reported outcomes unbiased by length of follow-up after lumbar degenerative spine surgery: do we need 2 years of follow-up? Spine J. 2019 Sep;19(9):1598. doi: 10.1016/j.spinee.2019.05.594. Epub 2019 Jun 6. PMID: 31175993.
6. Serra C, **Staartjes** VE, Maldaner N, Muscas G, Akeret K, Holzmann D, Soyka MB, Schmid C, Regli L. Response to "Going beyond scoring systems for cavernous sinus involvement in trans-sphenoidal pituitary surgery". Acta Neurochir (Wien). 2019 May;161(5):1035-1036. doi: 10.1007/s00701-019-03891-8. Epub 2019 Apr 5. PMID: 30953155.
7. **Staartjes** VE, Schröder ML. Letter to the Editor. Class imbalance in machine learning for neurosurgical outcome prediction: are our models valid? J Neurosurg Spine. 2018 Nov 1;29(5):611-612. doi: 10.3171/2018.5.SPINE18543. PMID: 30117796.

Book Reviews (2)

1. **Staartjes**, Victor E; Stumpo, Vittorio; Vasilios A. Zerris (2019): Neurosurgical Review: For Daily Clinical Use and Oral Board Preparation
2. Stumpo, Vittorio; **Staartjes**, Victor E; Neurosurgery Oral Board Review by Jonathan S. Citow et al.(2019) 376 pp, 335 illustrationsPaperback/softback, ISBN: 9781684201266 Thieme publishers New York/Stuttgart.

[Appendix 4]

Dankwoord / Acknowledgements

Finally, I have arrived at the section of this thesis that most of you are going to indeed read. There are so many individuals that I am thankful of and that in any way, shape, or form have contributed to the success of this thesis (whether directly – consciously – or not!). I am afraid that this will become a very long acknowledgments section… But here we go:

Dear Marc, it all started with you and you alone when I was 15 and first got to know you. You transmitted your love for neurosurgery immediately – on the first day after you had forgotten to pick me up nota bene! – as well as your general attitude to medicine, science, and life. The hundreds of hours we have spent by now in clinics, on Skype, or observing you in the operating room, but also collecting a metric ton of data, jogging, deliberating, and travelling together are simply infinite and have been infinitely valuable to me. We had our arguments from time to time as we are both very stubborn, but "*was sich liebt, das neckt sich*" – In the end, we always got there and have achieved a lot together. This is not the end either, only the early beginnings of a new chapter of learning from you and developing new thoughts together.

Dear Carlo, it was a pure golden stroke of fate and luck that we met in the first place: I had randomly selected you from the USZ website ("skull base surgery" sounded good, even though I had no clue what it was at the time) to try and start doing neurosurgical research in Zurich, and you unwillingly agreed to meet very briefly to discuss. When we then finally met after a couple of attempts, you sat down and asked me repeatedly "What do you want?" and "No, what do you *really* want?", until you somehow took a liking to me and showed me your collection of neuroanatomical books (which was still a lot more petite then), which – as always – made you light up. From that point onwards, we have been building, blossoming, machinating, and theorizing together. Without your constant support and considerate mentoring, all of this would still be very, very far away. I have learnt to respect you greatly not only as a human, as a mentor, surgeon, and anatomist, but also as an intellectual – these are the conversations I cherish most of all, and you never cease to amaze me with your depth of knowledge on almost invariably any topic. Amazingly and with an almost frustrating consistency, up to this day, whenever I have asked you for advice – whether professional or personal – your council has always been right. You also have a fine sense of how to steer me in the right directions when I tend to veer off (such as when you gifted me "*Wider den Methodenzwang*"…), and your influence on me as a person has been enormous. I will beat you at chess at some point! Thank you for all you have done for me.

Prof. Vandertop, education without supervision is not fruitful – I thank you not only for the opportunity to pursue a PhD with you, but also for the always very swift and very thorough, pragmatic, and critical comments on our manuscript. I promise that I will never write "we believe that…" (because we only do so in church) or use as many abbreviations! This has been a great pleasure and a very smooth process thanks to you.

Prof. Regli, how to express my gratitude for how you have I remember the first time we met (at the end of a pituitary case), after Carlo had desperately tried to introduce us to eachother for weeks. When you

asked where I am from and I said Amsterdam, you replied – with a wink – "Well, it's not Utrecht". I would dare to say that we have come a very long way from there. Your continued and very enabling mentorship and support has brought me to where I am now. The always open ear and role as a sort of "patron" have contributed greatly to not only this thesis, but of course the establishment of the MICN Lab, symposia, and countless other things. The start of my neurosurgical traineeship at your department has been a pleasure: I am hopeful for many further "brainstorming sessions" and ready to soak up anything you are willing to teach me.

I also want to thank the members of the PhD commission (Prof. de Witt Hamer, Prof. Peul, Prof. Majoie, Prof. de Groot, Dr. Marquering, Dr. Stadhouder) for their attentive reading of this manuscript. I am looking forward to the day of my defense, and to exchange thoughts on the content of this thesis then.

Dear Nick and Vittorio, dear Broccolo's, thank you for agreeing to be my paranimfs. I know I can always trust you and that this was the obvious choice, and for very good reasons so! Nick, you are my oldest friend, that much is for sure. After going on some truly amazing trips around the world together, we have always kept in brotherly contact, and I can honestly say that we have still never had a single real quarrel about anything. When you moved in with me, this changed my life a great deal – for the better of course. We like to say that we create an environment together in which we inspire eachother to do better (even if that means that you start stretching exercises for three months). Apart from your Swiss self-control, you never cease to amaze me with your quirks (the wrong way you make omelets, the fact that you make spreadsheets for *everything*, for some brief examples). When Vittorio then decided to join us in Bürschtlibau – after moving to this very *elegant* city – the triumvirate was complete. Vittorio, within a few short years you have grown to be a true brother and sort of *partner in crime* to me ("col favore delle tenebre"). We seem to complement eachother quite well, and even now I think it would be hard to combat our daily lives alone – you are truly the worthiest of replacements for Nick. I would like to think that I have become at least 1/10th Italian by now, through the environment you have created, but I hope to make great strides further in that direction soon! Thank you to both of you for all you have done for me and all we have done together. This is merely the beginning of a new, exhilarating era of this triumvirate.

My dear Nonni, Gert and Els: For years you have been wanting to attend my PhD defense, and it seems like we made it! Most of the papers in this thesis have been written in your old kitchen in Amsterdam in the earliest of morning hours, but that does not even begin to describe the impact you have had on my life as a whole: Any passion I have had, you have always uncompromisingly nurtured and enabled, and my most prized memories always include you. Remember, you lit the spark for all of this! Inspiration, support, refuge, and even sponsor – You have always held those roles, and you will always continue to do so in my heart.

My dear parents and dear Oscar: Knowing that you have always wondered why on earth I would spend so much time carrying out this research ("Do you at least get paid?"), I hope that you can still be at least a little bit proud of this. You have gone through so many adversities with me as a small child, somehow still creating a possibility for me to flourish – somehow. I know that at the time, it must have appeared inconsiderable (and at the time probably rightly so) that your son would complete a doctorate, but you have fought to get me here. You have never ceased to do whatever you can to make things happen for me, I know that I can always count on your unwavering support when it really comes to it. Thank you for everything.

Dear Shreya, your uncompromising love has been the greatest support in the final stages of this PhD, and with that I mean not only the repetitive proofreading that you did. You are a box of pure joy and the only one with the ability to take my mind fully off all this, and this has changed my life in the best of ways (ask Vittorio and Nick!). I know research has sometimes cost more hours than a day should have, hours that I would rather have spent with you.

Anita, we have accompanied eachother – and somehow quite closely so – throughout about one third of our lives. You have not only designed the cover and some of the illustrations of this booklet, but in some way also partially my scientific and clinical curriculum vitae: We have always discussed every major decision with eachother, and I am both glad and amazed that this friendship and peership has kept up for so many years already – I am sure there is much more to come. By now, you have also started your PhD training and are preparing for a stellar clinical career back home. I can only thank you for the countless hours spent together on the phone, for the enjoyable trips we have made together, and for the time you have spent helping me with this research.

Elena, Chris, Anthony, Sebi, Vichy, Victoria, Luna, Vivienne, Sung Ju, Marit, Chris, Florine, and many more: Friendships are invaluable, and success is impossible – at least for me – without the outpouring of support I have received from all of you from the very beginning. But many of you have also endured many hours of me sitting at a computer or cancelling plans, often related to this research. Without you, life would be all tones of gray.

Charles, Anneke, Mohamed, Anas: Sometimes, one is in need of mentorship for everything far, far removed from medicine and research, and this is when I have always found shelter with you. I do have a very exclusive “high council” of dearly appreciated friends who are just “good at life” and have helped me make up my mind on many occasions – thank you.

Chazia and Jasmijn, for as long as I have know Marc, I have also had the pleasure of spending time with you. It actually all started with Chazia getting off her bike to meet Nonna, and the rest is history.

Wouter, Jacopo, Elisa, Alessandro: Apart from the pleasant time in STE G 1, more than ever I appreciate the moments together, whether it is a summer evening at the lake, a “Rheinfelder Bierhalle” night, a “categoria becchi” day, or an AS Roma game – You have always made every second fun and we have always supported eachother when in need, let us keep it up!

Moira, Olivier, Raffaele, Olga, Elisa: All of you have somehow supported this thesis, for which I am grateful. But more importantly: You make up the core of the still very young and very new team of the Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, de facto you are laboratory. I am extremely proud of all of you and how you have all very quickly become highly independent researchers. With you, the future is bright and doing research is fun. Hopefully, we can build something together.

Julius, you are unquestionably the superstar when it comes to anything to do with machine intelligence in neurosurgery, at least in my opinion. Pretty sure that a great part of this thesis would not have been possible without your enabling and always open ear. We have written so much together in such a short time! I cannot wait to see what NAILA and MICN will slowly and hopefully grow into, as our friendship does in parallel. You are always welcome here.

Martin, I do not know how you do it all. An almost frighteningly quick growth, and still all those other commitments without losing the zest for all the things you do outside of your clinical and academic

responsibilities – this is what I try to take from you. Thank you for the time we have spent together and for all you have taught me.

Bernhard, every couple days I get to say: "*I learnt this from a very, very experienced internist…".* In the short two months I spent working with you, I really did learn – no exaggeration at all – most of what I know about general medicine and a lot of what I use every day now: I never thought I would be ordering urea-creatinine ratios, c-peptides, haptoglobins, and 24-hour urine sample with such regularity on a neurosurgical ward, but – it is possible! Apart from that, I appreciate that we are still in such regular contact, as well as your humor and our discussions about our shared love for classical music. Hope to see you, Monika, and Linda soon again!

Peter and Marijn, working with you on BoneMRI was always a pleasure and has been an exciting part of my PhD track – even though there have been many external hindrances to our projects, we will not be stopped in our tracks: per aspera ad astra!

Marlies, Paulien, Fleur, Femke, Nathalie, Johan, Hubert, and everyone else at Bergman Clinics who has contributed greatly to Marc's and my research. It certainly was not within the most enabling circumstances, but somehow we still always "made it happen" with blood, sweat, and tears.

Luzian, Thilo, Anne, Ernst, Klaus: You have all contributed immensely to my development in its most critical and impressionable phase. Although most people who might right this text may not understand why, I will never forget what you have done for me. Luzian, especially without your support I would not even have been admitted to the Gymnasium, and neither would my interest in scientific research have been set alight in this way – My first study was together with you for *Schweizer Jugend Forscht*, and see to what ridiculous extent it has grown now from that seed…! Thilo, Anne, Ernst, Klaus, your passion and patience teaching this at times very stubborn and inert boy was immense, still somehow ending up conveying a much more than just foundational education that has made everything else so much easier.

Anna: Thank you for your help with illustrations, apart from also being a great friend to be around. Alessandro, similarly: You have invested so much time into our publications, and I am glad to see that you have grown into an independent researcher yourself now.

Of course there are countless other individuals – both people who have supported me in life as well as valuable collaborators in science or colleagues in Zurich – that I would like to thank, the list is simply too long, but I am not forgetting about anyone. It is impossible to achieve much of anything at all alone, that is very clear. Please let us keep in touch, continue to develop our ideas, and finally I hope to see many of you soon in Amsterdam (or on Zoom, even!). Thank you!

[Appendix 5]

Curriculum Vitae

Victor Egon Staartjes was born on February 16th 1997 in Amsterdam, the Netherlands. After living in Laren – an arboreous suburb of Amsterdam – he moved to Zurich, Switzerland in 2003 and has lived there ever since. He attended the Freies Gymnasium Zurich for high school and received his medical degree from the University of Zurich in 2021. In January 2022, he started his neurosurgical traineeship under Prof. Luca Regli at the University Hospital Zurich.

In 2015, shortly after the start of medical school, Victor started doing scientific research in the field of neurosurgery under the tutelage of Dr. Marc Schröder in Amsterdam at Bergman Clinics, focusing on degenerative disease of the spine and robotic neurosurgery. In 2017, this developed into the start of a PhD trajectory under Prof. W. Peter Vandertop at the Vrije Universiteit Amsterdam, leading to this thesis.

Soon thereafter, Victor took up a research fellowship at the Department of Neurosurgery of the University Hospital Zurich under Prof. Luca Regli. Together with Dr. Carlo Serra as a valiant mentor, they have established a small research group – the Machine Intelligence in Clinical Neuroscience (MICN) Laboratory – at the University of Zurich, which is led by Victor. Together with his mentors, he has currently published over 90 original articles, reviews, and book chapters with over 1000 citations. Research interests are in applications of machine learning to medical imaging and clinical prediction modeling, as well as robotic neurosurgery and personalized / precision medicine, as well as clinical research in pituitary and degenerative spine surgery.