Machine Learning in Spinal Surgery: Prediction Models for Outcome Prediction and Decision Support

Paul Theodor Ogink

Machine Learning in Spinal Surgery: Prediction Models for Outcome Prediction and Decision Support

Machine Learning in de Wervelkolomchirurgie: Predictiemodellen voor Uitkomstvoorspelling en Besluitvormingsondersteuning

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

vrijdag 21 februari 2025 des middags te 12.15 uur

door

Paul Theodor Ogink

geboren op 4 november 1988
te Roosendaal en Nispen

Promotor:

Prof. dr. J.J. Verlaan

Copromotor:

Dr. J.H. Schwab

Beoordelingscommissie:

Prof. dr. ir. C.A.T. van den Berg
Dr. L.A. Daamen
Prof. dr. J.N. Doornberg
Prof. dr. M.P.W. Intven
Prof. dr. L.W. van Rhijn (voorzitter)

Machine Learning in Spinal Surgery: Prediction Models for Outcome Prediction and Decision Support

ISBN/EAN: 978-90-835161-3-4
The financial support for the publication of this thesis from the Nederlandse Orthopaedische Vereniging and the Dutch Spine Society is sincerely appreciated

Preface

"Do you think you are better at assessing psychiatric patients and predicting their risk of suicide than we are?" the psychiatrist asked the orthopedic surgeon aghast. "Clearly, yes" replied the orthopedic surgeon

Humans have always strived to know the future and what lies ahead for us. Whether it's based on data and science, like the weather report or on superstition, like card readers or the weekly horoscope. Predicting the future enables us to anticipate what is to come and prepare accordingly. For instance, my home country of the Netherlands sits well below sea level and is traversed by a variety of rivers. Accurate predictions on storms and rainfall have always been critical to prepare the country for potential disaster. And that brings us to the initial quote at the top. Preventing disaster is what we aimed to do by asking the psychiatrist in consultation with a patient who had sustained vertebral fractures after a jump from height. We had percutaneously stabilized the fracture but were wondering whether or not there was an acute risk of another attempted suicide. Essentially, we asked the psychiatrist whether she could predict if the patient was going to harm herself again during her hospital stay. The psychiatrist had assessed the patient the next morning and concluded there was no acute risk. Two hours later she sneaked out of the ward and jumped from a parking garage. I was ordered to organize a so-called "incident meeting" two weeks later with everybody involved. What ensued was a discussion which included but not explicitly mentioned virtually every concept involved in prediction-making: accuracy, calibration, decision-curve analysis, false-negatives etc. I noticed how certain elements of prediction-making were misinterpreted; a wider problem in society, because as much as we are flooded by predictions, many people do not really understand their meaning. "The pollsters were wrong saying Hillary Clinton would win" and "They said it wouldn't rain today, but it did, so how can they say anything about the climate 30 years from now?" are two examples.

With this thesis I hope to give something of an instruction manual for orthopedic surgeons on how to construct and interpret prediction models.

And as for the initial quote, weather forecasting and the prevention of disaster: our patient survived the second jump; it had rained incessantly the previous days and she fell in a puddle of mud.

Table of contents

Introduction

History of prediction in medicine

Making predictions is not new in medicine. Hippocrates himself at the start of his book Prognosticon encouraged doctors to engage in “prognosis”, defined by him as “foreseeing and foretelling … the present, the past and the future”. He added: “And he will manage the cure best who has foreseen what is to happen from the present state of matters.” “… and by seeing and announcing beforehand those who will live and those who will die, he will thus escape censure.”[1] The Greek word “πϱόνοια” he used, had a much broader meaning than the current medical meaning of prognosis. It ranged from the aforementioned foreseeing to prediction and even prophecy. With the limited knowledge of anatomy and physiology at hand, he spends the rest of the book explaining which symptoms of the patient’s body are predictive of (short-term) mortality and what value must be given to them allowing doctors to either intervene or actually refrain from treatment to avoid unfair criticism. Galen, the Greco-Roman physician in the 2nd century, wrote a commentary on Hippocrates’ book in which he lists his own accomplishments in medical prediction-making.[2] Unfortunately for Galen, this led to his detractors leveling accusations of magic and soothsaying against him. Due to the importance still given to the Hippocratic writings of antiquity, medieval times similarly showed a great interest in prediction, albeit oft descending into mere soothsaying. Onomancy referred to the practice of converting names and dates to numerical equivalents which are then used to predict the outcomes of not just military battles or financial fortunes but also diseases. Popular as well was lunary, which used the position of the moon at any given time to make assumptions about a patient’s chances of curation.[3] Subsequent centuries saw an abundance of developments in medicine. Vesalius described cardiac anatomy, Lind performed the first clinical trial, Laennec invented the stethoscope, Semmelweis mandated washing hands and Flemming discovered penicillin. So while diagnosis, prevention and treatment were making great strides, prognosis and prediction started lagging behind. Doctors were still using physical symptoms to assess a patient's chances of survival and were granted additional tools for acquiring more data, such as the stethoscope, but the concept remained similar to Hippocrates' time.

A paradigm shift occurred in 1954 when Paul Meehl, professor of Psychology at the University of Minnesota, published a book in which he claimed so-called "mechanical" prediction of behavior outperforms "clinical" prediction.[4] Mechanical prediction meant prediction based on algorithmic approaches, as opposed to clinical prediction which referred to the clinical judgment of a doctor or psychologist. While controversial at the time meta-analyses in the following decades proved him right time and time again.[5] In his book he laid the groundwork for prediction as we know it today in medicine.[6] Currently, data driven algorithmic predictions are ubiquitous in clinical practice: the CHADSVASC score for thromboembolic risk, the Wells score for deep vein thrombosis, or more specific for spine surgery the SINS score for tumor related instability. The renewed interest in artificial intelligence – more specifically machine learning – represents a new phase in medical predictions with the prospect of more accurate, highly personalized predictions guiding decision-making.

Personalized Medicine and Machine-Learning

Personalized Medicine - also known as Individualized or Precision Medicine - is a concept that strives to tailor prevention, diagnostic, and treatment strategies to an individual patient, aiming to improve patient care and lower healthcare costs.[7–10] Many elements critical for advancing personalized medicine, such as genome sequencing and predictive analytics, are developing at a rapid speed.[11,12] An increasing number of studies featuring these personalized models are being published in the top medical journals[13–16], prompting those journals to include Statistical Guides and Perspectives to better understand the new concepts.[12,17,18]

One of the current developments propelling personalized medicine is the aforementioned renewed interest in artificial intelligence. Machine learning (ML) is one of the branches of artificial intelligence which lets algorithms learn and self-improve from experience without explicitly being programmed. Even though these techniques have been around for decades, the combination of increasing amounts of patient data, readily available computational hardware, and improved algorithms has led to an enormous expansion of ML prediction models throughout medicine in recent years.[19–22] Patient data is now entered into electronic health records (EHR's) instead of written down by hand, which makes accessing and subsequently using clinical information as data far easier. The increase in computational power is still abiding by Moore's law. This infamous prediction from 1965 by Gordon Moore, co-founder of Intel, poses that the number of transistors on a chip doubles every 2 years, making them twice as powerful as the previous generation. This has provided an enormous computational power to all researchers and even normal consumers.[23] Without these advancements in chip-making widespread use of ML would be impossible. The theoretical advantage of models based on ML lies in the ability to encapsulate nonlinear relationships between the input variables.[24] More traditional methods of creating prediction models such as logistic regression need an explicit search for potential nonlinear relationships.[25] Theoretically, this gives ML the edge in creating better performing prediction models leading to better decision-making for patients and doctors alike.

Spine Surgery

Spinal surgery is one of the major subspecialties of orthopedic surgery, representing a large number of patients, surgeries and accompanying costs. With rising healthcare costs worldwide, spine surgery has come under scrutiny as well, considering the rise in surgeries and even steeper rise in costs associated with spinal care.[26] Total costs of spinal care in the US rose by 65% from 1997 to 2005.[27] With an older, but still active population in the Western world and East Asia this rise in patients, surgeries and expenditure isn't looking to slow down. Additionally, advances in oncological treatments have added a substantial number of patients with spine metastases to the already crowded spine surgeon's practice, essentially creating a new subspecialty within spine surgery. This development is similarly not projected to slow down with an ageing population and evermore advances in medical treatments for cancer.

Challenging in the field of spine surgery is weighing the potential benefits of treatments with the not to be underestimated risks. While all medical specialties have to weigh risks with benefits this is particularly difficult in spinal care considering the wide range of treatment options and complexity of surgical decision-making, as evidenced by a widespread variation in care not just between geographical areas but between surgeons in the same hospital.[28–30] Potentially, individualized predictions by ML algorithms can aid both surgeons and patients in decision-making across the entire spectrum of patient care from diagnosis and (surgical) decision-making to discharge placement and readmission risk. Pre-operatively obtaining an estimate of the risk of complication, mortality, or discharge to rehabilitation enables surgeons to properly inform patients during a consultation. Furthermore, if the decision is made to perform surgery, adequate prevention strategies may be employed. Despite the numerous advantages, not many of these prediction models are available yet and those available are rarely used in practice. This limited usage can be attributed to factors such as unfamiliarity, insufficient external validation, and perceived lack of interpretability.

Thesis outline

The current thesis aims to provide an extensive comprehension of the many facets involved in ML prediction models in the orthopedic subspecialty of spine surgery. In Part I, the current state of ML models in orthopedic surgery is explored. What do these models focus on as outcome and what methodology do they employ to create these models? And when these models are developed, are they being externally validated?

Part II contains prediction models developed for use in the clinical practice of a spine surgeon. All models were meant to be used before surgery allowing patients and surgeons to be aided in decision-making or counseling. In Part III we focus on (external) validation and implementation of the prediction models in clinical practice. While the developed prediction models may show great results validated in the same patient group, determining whether the prediction model performs well outside of the group it was developed on, is essential before widespread implementation.

Part I – Quality of Prediction Models

The relative novelty of the ML field itself combined with the unfamiliarity of the subject among medical researchers, doctors, and reviewers has caused a wide variety in methodology.[19,20] In all prediction-modelling (including non-ML models) there is a tendency to overfocus on area under the curve (AUC), while essential elements for making prediction models relevant for clinical practice (i.e. calibration, external validation, to be explained later) are often ignored or simply overlooked.[31] The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement aimed to set recommendations for reporting of prediction models based on 22 items.[32] At the time of publication of the TRIPOD statement more than half of the items considered essential were not addressed in published prediction models and more than 80% of the models lacked model specifications and/or performance metrics.[33] In **Chapter 1,** the current methodologies are determined as well as uses of ML prediction models in orthopedic surgery and in **Chapter 2** we assess the quality of these models.

Furthermore, another critical element for actually implementing all these prediction models is external validation. In external validation the model is assessed on its performance on patients outside the patient set used for development.[34] Datasets are often from a single institution or from national databases which may not be representative of patients throughout the world. Furthermore, not just patient characteristics but healthcare systems differ dramatically across countries adding an additional lack of generalizability. While undoubtedly essential, external validation is not routinely performed for most prediction models throughout medicine.[35–37]

Part II – Development of Prediction Models

In **Chapter 3**, a nomogram is presented for predicting nonoperative failure in patients with spinal epidural abscess. It is important to note that this nomogram was developed before our group transitioned to using machine learning as the foundation for constructing prediction models.

Chapters 4, 5 and 6 showcase a number of prediction models based on ML. Several steps are involved in the creation of ML prediction models, as illustrated in Figure 1. The construction of these models relies on a form of ML known as supervised learning. In this approach the model is trained on data in which the desired outcome variable is already known, i.e survival or estimated blood loss. This data can be derived from an institution's own database or from national databases such as the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) or the National Inpatient Sample (NIS). Missing data should first be handled appropriately, for instance by doing complete-case analysis or by imputing values.[38] The data is then split into a training set and a testing set. The size of each set is decided by the developers. The training set is used to develop the prediction model; the testing set is reserved to validate the model later in the process. Subsequently, the most relevant predictor variables (e.g. gender, age) for the model need to be selected, which can be done by clinical judgment or with statistical methods pointing out the most important variables in the dataset. A ML algorithm is then picked for model development depending on size and type of data and the preference of the developers. Figure 2 shows commonly used algorithms with their respective pros and cons. In the training process the algorithm uses the outcome variable to build a combination of the predictor variables to predict the outcome in new, unseen data. This is markedly different from traditional programming in which the instructions for the model are written by the programmer.[39]

A pitfall for all prediction models, whether based on ML or not, is overfitting. Overfitting can arise when a model becomes overly proficient at learning the details of the training data, capturing noise and random variations instead of discerning the fundamental patterns in the data. This leads to a lack in generalizability; i.e. a suboptimal performance when applying the algorithm to new, unseen data.

Critical therefore in evaluating prediction models is the validation process. The most common method involves employing the testing set, the portion of the dataset reserved from the start and excluded during the algorithm's training, to validate the model. A different method, which can be used concurrently, is cross-validation in which the data is divided into a number of groups and each group is used as testing set in different cycles with the others combined as training set. The model is then constructed using the averages of each of these cycles. If the model excels on the training set but performs poorly on the testing set, there's a risk of overfitting to the training data. In such cases, adjustments can be implemented to enhance the model's ability to generalize effectively.

Discrimination and calibration are used to describe the model performance. [40] Discrimination refers to the ability of the model to distinguish patients with the outcome variable from those without the outcome variable. Discrimination can be assessed graphically with a Receiver Operating Curve (ROC) and numerically with the Area Under the Curve (AUC). An AUC value of 0.5 means no discriminatory ability and 1.0 means perfect discrimination.[31,41] Calibration measures how well the model assigns predicted probabilities compared to the actual probabilities. The overall calibration intercept and slope are used for numerical assessment. The calibration intercept measures whether the model generally over- or underestimates the probabilities; the value for a perfect model would be 0. The calibration slope measures whether predictor effects in the model are too extreme or too moderate. A value above 1 means probabilities are too high for patients at high risk and too low for patients a low risk. A value below 1 means the probabilities are too moderate. [42] Calibration plots can be used to graphically show how well the model is calibrated. Figure 3 is a calibration plot which shows the model underestimating the risk of an outcome in patient over the age of 45. Overall performance can be expressed with the Brier score, which represents the error the model makes per individual prediction.[43] It is calculated by taking the mean-squared error between the predicted probabilities and the actual observed values.

Major criticism of ML models is the purported black box nature . Their lack of odds ratios per individual variable, like in regression modelling, supposedly makes them harder to interpret. Therefore, variable importance plots have been developed which can show the overall relative importance of each variable in the study population.[44] For the individual patient-specific prediction, local explanation shows how much variables contributed to the estimated prediction (Figure 4).[45] To assess usefulness in clinical practice, decision-curve analyses have been added to models.[46] Changing management for all? patients based on the prediction models can either harm or benefit patients depending on whether the prediction model was right. Assessing this tradeoff can be done by providing the net benefit, which is the weighted sum of those benefiting and those being harmed. Since it is not feasible to include every single possible change in management per patient, decision-curve analysis calculates this net benefit over the full-range of predictions in the study population.

Part III – Validation and Implementation

Prior to employing the prediction model in clinical practice, external validation should be performed to assess the generalizability outside the dataset used for development. The developed prediction model is evaluated using new, unseen data distinct from the original dataset. This data isn't limited to different geographical areas but can also span various time periods and settings.[47] In **Chapter 7** a previously American-made prediction model was externally validated in the country of Taiwan. The final **Chapter 8** includes a study on the impact of predictions made on decision-making. Humans are known to be poor with estimating and understanding risks and are prone to a number of biases when it comes to decision-making. The aim is to find out how the addition of a prediction percentage from a prediction model changes clinical decision-making.

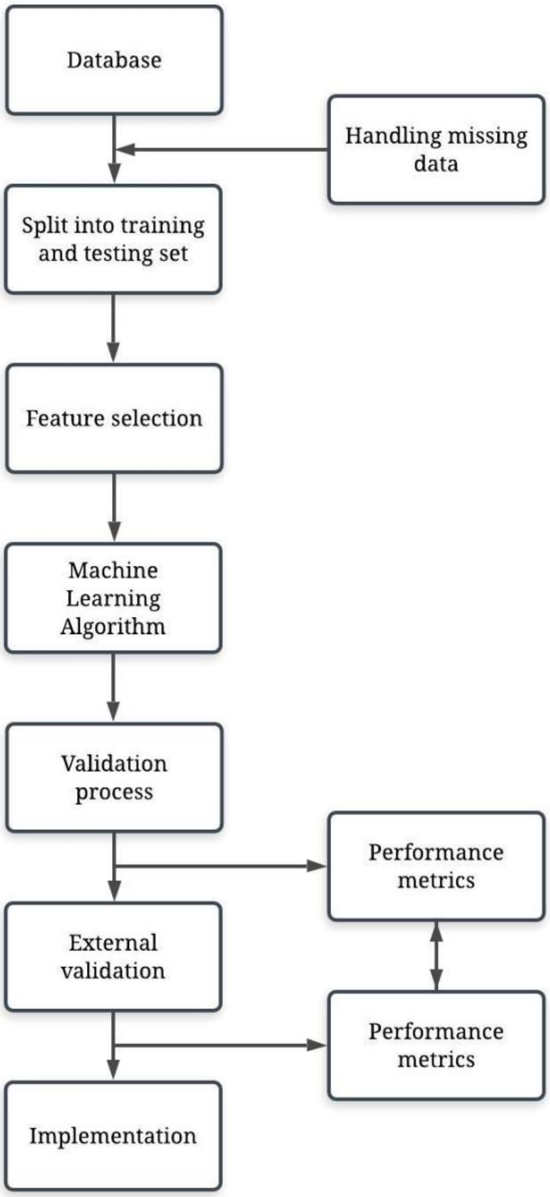


Figure 1. General flowchart of development of a ML prediction model

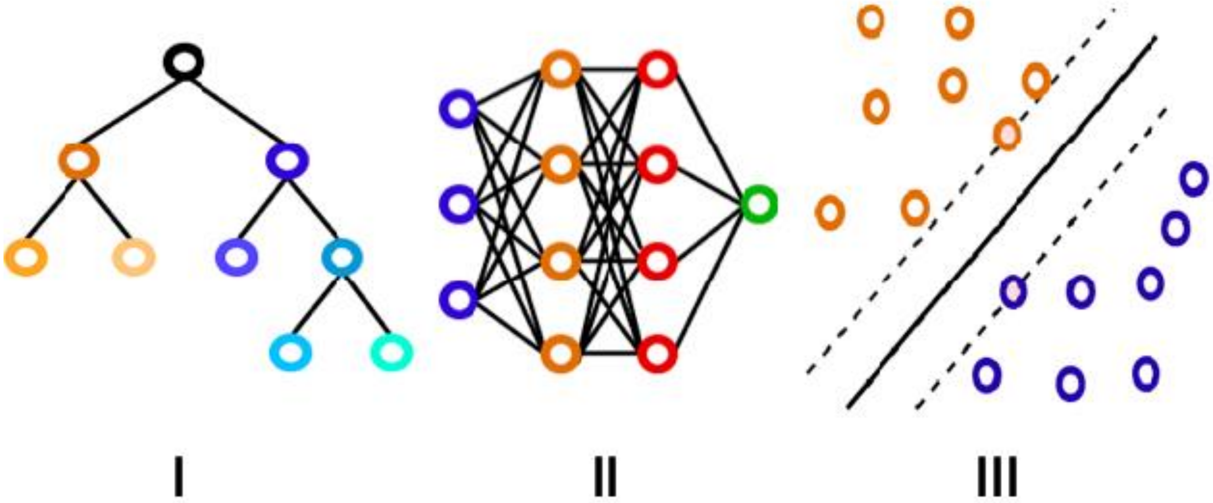


Figure 2. (I) Decision trees are hierarchical structures in which each node performs a test on the input value with the subsequent branches representing the outcomes. Their graphical representation as seen above make them easy to understand and interpret.

(II) Neural networks are based on interconnected nodes. The input features are represented by the first (blue) layer. The designated outcome is represented by the final (green) layer. The middle hidden layers (orange and red) base their output on the input they get from prior layers. Neural networks have been around for a long time and are often used, but interpretation of the relationships between the different layers remains difficult to this day.

(III) Support Vector Machines (SVMs) perform classification by determining the optimal separating hyperplane between datapoints which maximizes the distance between the 2 closest points of either group. This can be used very effectively in non-linear classification.

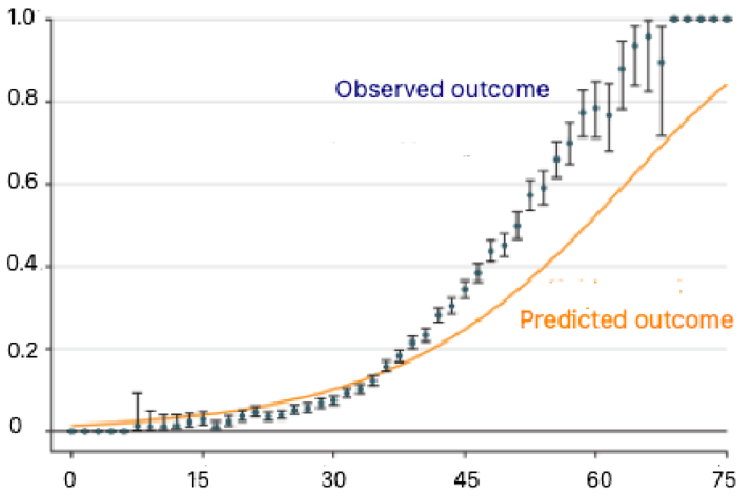


Figure 3. Calibration plot illustrates the connection between the observed outcome (mean indicated by data markers, with error bars representing the 95% CI) and predicted outcome. This model underestimated the outcome in patients aged 45 years and older.

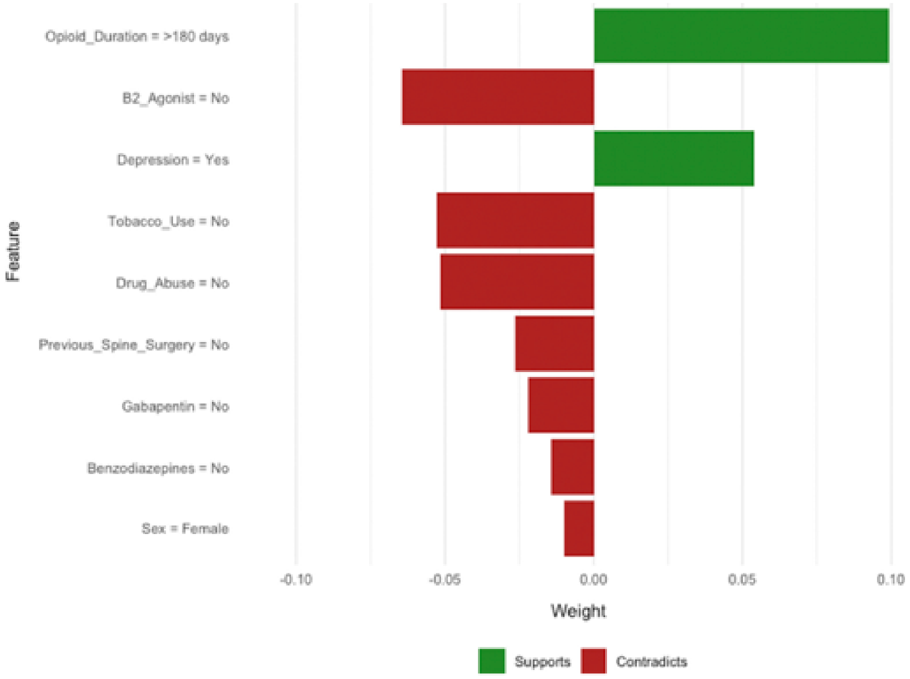


Figure 4. A local explanation graph depicting an example of a patient-specific explanation for a generated prediction.

Part I

Quality of Prediction Models

Chapter 1

Wide Range of Applications for Machine Learning Prediction Models in Orthopedic Surgical Outcome: A Systematic Review

Ogink PT, Groot OQ, Karhade AV, Bongers MER, Oner FC, Verlaan JJ, Schwab JH

Acta Orthop. 2021 Oct;92(5):526-531

Abstract

Background and purpose Advancements in software and hardware have enabled the rise of clinical prediction models based on machine learning (ML) in orthopedic surgery. Given their growing popularity and their likely implementation in clinical practice we evaluated which outcomes these new models have focused and what methodologies are being employed.

Methods We performed a systematic search in PubMed, Embase, and Cochrane Library for studies published up to June 18th 2020. Studies reporting on non-ML prediction models or non-orthopedic outcomes were excluded. After screening 7138 studies, 59 studies reporting on 77 prediction models were included. We extracted data regarding outcome, study design, and reported performance metrics.

Results Of the 77 identified ML prediction models the most commonly reported outcome domain was medical management (17/77). Spinal surgery was the most commonly involved orthopedic subspecialty (28/77). The most frequently employed algorithm was neural networks (42/77). Median size of datasets was 5507 [IQR 635 – 26]. The median area under the curve (AUC) was 0.80 (IQR 0.73 – 0.86). Calibration was reported for 26 of the models and 14 provided decision-curve analysis.

Interpretation ML prediction models have been developed for a wide variety of topics in orthopedics. Topics regarding medical management were the most commonly studied. Heterogeneity between studies is based on study size, algorithm, and time-point of outcome. Calibration and decision-curve analysis were generally poorly reported.

Introduction

Surgical decision-making in orthopaedic surgery involves weighing the benefits of an intervention against its inherent risks. Prognostic scoring tools have been devised to individualize risk prediction and thus improve surgical decision-making.[1–3] Although clinical prediction models are not new, recent advancements in artificial intelligence have created a host of prediction models based on machine learning (ML).[4]

ML is a branch of artificial intelligence which enables computer algorithms to learn from experience from large datasets without explicit programming. Existing reviews of machine learning in studies have provided a broad overview of applications ranging from vision, natural language processing, and predictive analytics.[4] To our knowledge, there is no study that has critically assessed the body of studies focused on ML prediction models for surgical outcome in orthopaedics. These types of prediction models are most likely the first branch of artificial intelligence to be employed in clinical practice.[5] Therefore, familiarizing practicing orthopaedic surgeons with ML's concepts and the topics these new methods have focused on can optimize their implementation in clinic.

As such, the purpose of this systematic review is to 1) evaluate which surgical outcomes orthopaedic clinical prediction models have focused on, and 2) determine which techniques current prediction models use for development and validation.

Methods

Systematic literature search

Adhering to the 2009 PRISMA guidelines a systematic search was performed in Pubmed, Embase and the Cochrane Library for articles published up to June 18st 2020.[6] Two different domains of medical subject headings (MeSH) terms and keywords were combined with "AND" and within the two domains the terms were combined with "OR". The first domain included words related to ML and the second domain related to possible orthopaedic specialties (Appendix 1). Terms were restricted to MeSH, title, abstract, and keywords. Two reviewers (PTO, OQG) independently screened all titles and abstracts for eligible articles based on predefined criteria. Eligible full-text articles were evaluated and cross-referenced for potentially relevant articles not identified by the initial search (Figure 2). Discrepancies between the two reviewers were adjudicated by the senior author (JHS).

Eligibility criteria

Included were studies reporting on ML based prediction models addressing orthopaedic surgical outcomes. All intraoperative and postoperative outcomes were included. The surgical orthopaedic population was defined as disorders of the bones, joints, ligaments, tendons, or muscles treated by any type of operation. Excluded were studies (1) that did not include at least 1 ML based prediction models for surgical outcome (e.g. logistic regression based models), (2) non-English studies, (3) lack of full text, and (4) non-relevant study types such as animal studies, letters to the editors, and case-reports.

Assessment of methodological quality

Quality assessment was performed based on a modified nine-item Methodological Index for Non-Randomized Studies (MINORS) checklist.[7] We made it applicable for our systematic review by including disclosure, study aim, input feature, output feature, validation method, dataset distribution, performance metric, and explanation of the used AI model.[8] These nine items were scored on a binary scale; 0 (not reported or unclear) and 1 (reported and adequate).

Data extraction

Table 1 lists the data we extracted from each study. For this review, six main orthopaedic surgical outcome domains were identified, consisting of (1) intraoperative complications (e.g. blood transfusion, prolonged operative time), (2) postoperative complications (e.g. venous thromboembolism), (3) survival, (4) Patient Reported Outcome Measures (PROM), (5) medical management (e.g. hospitalization), and (6) other. For studies reporting the performance of multiple ML models, the best performing ML model was used. Thirteen studies provided multiple models for multiple surgical outcomes; these were extracted separately resulting in more ML models than studies. Only the two performance measures AUC and accuracy were extracted as they were most the commonly reported results.

Study characteristics

After screening of titles and abstracts, 758 full-text articles were assessed for eligibility and ultimately 59 articles were included reporting on 77 ML prediction models (Table 1). Median sample size was 5818 (Interquartile range [IQR] 635 – 26,869) Using the MINORS criteria, all 59 articles were found to be of similar quality. All included a minimum of eight out of nine appraisal items (Appendix 2).

Statistical Analysis

AUC scores and accuracies in tables are expressed as they were originally reported. For studies that reported multiple results within a single outcome domain (e.g. multiple different postoperative PROMs with each an independent AUC) averages were taken. The sizes of the training, validation, and test sets are reported as percentages of the total dataset. No meta-analysis was performed because of obvious heterogeneity between studies and in orthopaedic applications. However, to summarize the findings in some quantitative form, the median AUC and accuracy of the prediction performance were calculated for all studies.

We used Microsoft Excel (Version 16.31; Microsoft Inc, Redmond, WA, USA) for standardized forms for data extraction and quality assessment, and Mendeley as reference management software.

Results

Study design

More than half of all models was developed with data from national databases or registries (55% [42/77]) (Table 3). Median number of predictor variables used in the ML model was 10 (IQR 8 – 15). Models using national data did not include more variables; 10 (IQR 8 – 13). Ninety-two percent [68/77] of the models had a binary distribution of the outcome variable. Most frequently employed algorithms were neural networks (55% [42/77]) and random forests (39% [30/77]). Thirty-six of the neural networks were single-layer, 5 deep learning, and 1 convolutional. The median size of the number of patients used was 5507 [interquartile range (IQR) 635 – 26.364]. Median AUC was 0.80 (IQR 0.73 – 0.86) and median accuracy was 79% (IQR 75% - 88%). Calibration was reported for 34% [26/77] of the models and 30% [23/77] provided Brier scores. Decision-curve analysis was employed in 18% [14/77. Twenty-three percent [18/77] provided a digital application for their prediction model.

Outcome

The most commonly reported outcome domains were medical management (22% [17/77]) and survival (22% [16/77]). Medical management mostly focused on discharge destination (41%[7/17]) and hospitalization (24%[4/17]). The studies on survival all addressed patient survival. Six (38%) survival studies were in orthopedic oncology and 5 (31%) in orthopaedic trauma. Both medical management and survival had a higher median AUC (0.82 and 0,84 than overall median AUC). Spinal surgery was the most commonly involved subspecialty (36% [28/77]).

Discussion

Recent years have seen an increasing interest in artificial intelligence and ML in orthopaedics.[9,10] With this systematic review we aimed to provide an introduction into the main concepts of developing ML models for orthopaedic surgeons and analyze the current application and design of these models in orthopaedic surgery. We found a wide range of potential applications ranging from predicting survival in spinal metastases, clinical outcome after shoulder arthroplasty, and hospitalization after hip fracture surgery.

This systematic review has a number of limitations. First, due to the relative novelty of this field of research in orthopaedic surgery the variety in study designs renders comparisons and comprehensive quantitative analysis difficult. We therefore opted to perform a qualitative analysis of the current publications. Hopefully, the increasing familiarity with these types of studies will lead to better reporting and open up the possibility to perform quantitative analyses. Second, this review is likely influenced by publication bias. ML prediction models with good performance are more likely to be published than models with mediocre or poor performance. This positive publication bias has been shown both in medicine and computational sciences.[11] The performance measures presented here were therefore likely to be more favorable than those of all developed models. Third, despite our efforts to perform a search across multiple online libraries, we have missed a number of studies reporting ML prediction models. Whilst unfortunate, we do no not think these omissions will significantly alter our findings on research topics or most utilized methodology as this review included nearly 60 studies.

This systematic review shows that ML models have been developed for a wide variety of topics across all subspecialties within orthopaedics. Perhaps surprisingly, medical management was the most studied domain with the majority of models focusing on readmissions and discharge placement. Both readmissions and discharge delays impose a heavy burden on healthcare costs.[12] Healthcare expenditure has risen steadily throughout the developed world in recent decades.[13] While there is enormous variation in healthcare systems, government institutions in virtually all countries have looked at improving medical management to help curb costs.[14] Papanicolas et al. found activities relating to planning, regulating and managing health services was a major factor in the difference in healthcare expenditure between the US and 10 other high-income countries.[15] Shrank et al. concluded failure of care coordination, leading to unnecessary readmissions among other things, amounts to $78 billion of waste in the US.[16] To address this problem the Centers for Medicare and Medicaid Services started the Hospital Readmissions Reduction Program in 2012, incentivizing hospitals to lower readmission rates. Knowing in advance which patients are at risk of being readmitted within 30 days after discharge is crucial, which is a possible explanation why so many prediction models focus on this topic. Similarly, knowing in advance where patients are likely to be discharged to, makes preventing delayed discharge a lot easier than the other interventions tried over the years.[17,18] Furthermore, the databases available in the studies on medical management appear to be larger, enabling researchers to include more variables and create better performing prediction models. These models are more likely to be published as evidenced by the higher AUC for medical management compared to overall AUC.

Survival was the other commonly studied outcome domain. Accurately estimating remaining life-expectancy is an important feature in medical decision-making in orthopaedic oncology.[1] In a patient group with only limited life-span remaining, the aim of treatment is to preserve quality of life. Accurate survival estimations can guide decision-making; whether or not to perform surgery and if so, which operative treatment should be opted for.[19] With an ageing population and cancer patients surviving longer, the incidence of bone metastases will continue to rise and prediction models will likely play an increasing role in this field.[19]

The AAOS Census 2018 showed only 8.3% of orthopaedic surgeons' primary specialty area was in spine, while 36% of the prediction models was linked to spinal surgery.[20] Cost reduction may also be the driving factor in the overrepresentation of spinal surgery prediction models; the economic cost of spinal surgery is large and growing with spinal fusions alone costing $30 billion annually in the US.[21] Prediction models could play a role in curbing costs by improving patient selection and surgical decision-making, although this could be said for all other subspecialties. Another possible explanation for the disproportionate number is the overlap with neurosurgery. The neurosurgical field was relatively quicker to use ML to develop prediction models and had developed several models in spinal surgery earlier on.[22] Finally, the field of prediction models is expanding but still small. A significant portion of the prediction models are developed by a few research groups who happen to focus on spine surgery. With the field expanding as fast as it is with new prediction models being published every month, we expect the overrepresentation of spine surgery to be temporary in a field in its infancy.

While there is wide variation in study design, certain study design elements are fairly similar across most studies. The most common designs are comprised of binary outcomes; either a 70:30 or 80:20 split between training and test set; and 10-FCV as method of internal validation. Wide variety exists in study size, time-point of outcome, and choice of ML algorithms. Study size is mostly defined by whether a national database or registry was used for model development. These quality improvement databases offer a large number of datapoints with a variety of variables of a diverse group of hospitals enabling the creation of prediction models. However, these databases are sometimes flawed by errors and their generalizability is also yet to be assessed.[23] External validation remains crucial considering generalizability outside the geographical origin of the database is not ensured.[24] Institutional databases offer the advantage of more veracious data, for instance including PROM data, which can extend over longer periods of time, but often lack adequate size.

Which ML algorithm is chosen seems highly random. While studies do list the pros and cons of certain algorithms, no study elaborates on why specifically those algorithms were chosen. A potential reason neural networks and random forests are selected so often is due to the familiarity of these algorithms. Neural networks have been around for decades, but were limited by lagging computational power.[25] The increase in computational power has led to a significant expansion of what neural networks can process and scientist have been able to build on the work of previous decades.[26] Future research should report on multiple ML algorithms and provide the performance measures of all models, thus enabling comparison between different approaches.

Despite the importance of performance metrics, a mere 34% of prediction models included information on calibration; similar to prior studies assessing prediction models in multiple medical domains.[27,28] Calibration is important to evaluate if the model is under- or overestimating the risk regardless of the discriminative abilities. Systematically underestimating risk can lead to undertreatment, while overestimating risk can cause overtreatment.[29,30] To improve the quality of reporting of clinical prediction models Collins et al.[31] published the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement. While not tailored for ML prediction models this guideline can provide a framework for researcher to use during development. Hopefully, a more widespread adaptation of the TRIPOD statement can lead to less variation in study designs and better reporting of performance metrics.

Only 23% of prediction models has a digital application available. The purpose of prediction models is to aid clinicians and patient in decision-making, which can only be achieved if the models are available for use. Otherwise, predictive analytics based on ML will remain a mere theoretical exercise. Furthermore, researchers should be encouraged to not only provide a digital application of their prediction model, but share their code as well. With a field in its infancy, providing code of more experienced researchers can guide beginning research groups in their endeavors. Additionally, this can greatly increase the small number of external validation studies being performed.

ML prediction models have been developed for a wide variety of topics in orthopaedic surgery. Topics regarding medical management and survival were the most commonly studied and spine surgery was the most involved subspecialty. Heterogeneity between studies is mostly based on study size, choice of ML algorithm, and time-point of outcome. Most published prediction models showed fair to good discriminative abilities, while calibration was poorly reported. Future studies should preferably include more multi-institutional, prospective databases and develop multiple models enabling comparison between different ML approaches. Also, important performance measures such as calibration should be reported to accurately evaluate the prediction model.

References

1. Pereira NRP, Janssen SJ, Van Dijk E, et al. Development of a prognostic survival algorithm for patients with metastatic spine disease. *J. Bone Jt. Surg. - Am. Vol.* 2016;98(21):1767–1776.

2. Shah AA, Ogink PT, Nelson SB, et al. Nonoperative management of spinal epidural abscess development of a predictive algorithm for failure. *J. Bone Jt. Surg. - Am. Vol.* 2018;100(7):546–555.

3. Janssen SJ, van der Heijden AS, van Dijke M, et al. 2015 Marshall Urist Young Investigator Award: Prognostication in Patients With Long Bone Metastases: Does a Boosting Algorithm Improve Survival Estimates? *Clin. Orthop. Relat. Res.* 2015;473(10):3112–3121.

4. Cabitza F, Locoro A, Banfi G. Machine Learning in Orthopedics: A Literature Review. *Front. Bioeng. Biotechnol.* 2018;6(June).

5. Staartjes VE, Stumpo V, Kernbach JM, et al. Machine learning in neurosurgery: a global survey. *Acta Neurochir. (Wien).* 2020.

6. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med.* 2009;6(7).

7. Slim K, Nini E, Forestier D, et al. Methodological index for non-randomized studies (Minors): Development and validation of a new instrument. *ANZ J. Surg.* 2003;73(9):712–716.

8. Langerhuizen DWG, Janssen SJ, Mallee WH, et al. What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review. *Clin. Orthop. Relat. Res.* 2019;477(11):2482–2491.

9. Jayakumar P, Moore MLG, Bozic KJ. Value-based Healthcare: Can Artificial Intelligence Provide Value in Orthopaedic Surgery? *Clin. Orthop. Relat. Res.* 2019;477(8):1777–1780.

10. Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *J. Arthroplasty.* 2018;33(8):2358–2361.

11. Boulesteix AL, Stierle V, Hapfelmeier A. Publication bias in methodological computational research. *Cancer Inform.* 2015;14(Suppl 5):11–19.

12. Wan H, Zhang L, Witz S, et al. A literature review of preventable hospital readmissions: Preceding the Readmissions Reduction Act. *IIE Trans. Healthc. Syst. Eng.* 2016;6(4):193–211.

13. Health resources - Health spending - OECD Data.

14. Schwierz C. *Cost-Containment in the European Union.*; 2016.

15. Papanicolas I, Woskie LR, Jha AK. Health care spending in the United States and other high-

income countries. *JAMA - J. Am. Med. Assoc.* 2018;319(10):1024–1039.

16. Shrank WH, Rogstad TL, Parekh N. Waste in the US Health Care System: Estimated Costs and Potential for Savings. *JAMA - J. Am. Med. Assoc.* 2019;322(15):1501–1509.

17. Bryan K. Policies for reducing delayed discharge from hospital. *Br. Med. Bull.* 2010;95(1):33–46.

18. Ou L, Chen J, Young L, et al. Effective discharge planning - timely assignment of an estimated date of discharge. *Aust. Heal. Rev.* 2011;35(3):357.

19. Quinn RH, Randall RL, Benevenia J, et al. Contemporary management of metastatic bone disease: tips and tools of the trade for general practitioners. *Instr Course Lect.* 2014;63:431–441.

20. AAOS Department of Clinical Quality and Value. Orthopaedic Pactice in the U.S 2018. 2019;(January):1–68.

21. Johnson WC, Seifi A. Trends of the neurosurgical economy in the United States. *J. Clin. Neurosci.* 2018;53(2018):20–26.

22. Senders JT, Staples PC, Karhade A V., et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg.* 2018;109(Ml):476-486.e1.

23. Rolston JD, Han SJ, Chang EF. Systemic inaccuracies in the National Surgical Quality Improvement Program database: Implications for accuracy and validity for neurosurgery outcomes research. *J. Clin. Neurosci.* 2017;37(2017):44–47.

24. Janssen DMC, van Kuijk SMJ, D'Aumerie BB, et al. External validation of a prediction model for surgical site infection after thoracolumbar spine surgery in a Western European cohort. *J. Orthop. Surg. Res.* 2018;13(1):114.

25. Hopfield JJ. Artificial Neural Networks. *IEEE Circuits Devices Mag.* 1988;4(5):3–10.

26. Schmidhuber J. Deep Learning in neural networks: An overview. *Neural Networks.* 2015;61:85–117.

27. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Med.* 2018;16(120).

28. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. *PLoS Med.* 2012;9(5).

29. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):1–7.

30. Van Calster B, Vickers AJ. Calibration of risk prediction models: Impact on decision-analytic performance. *Med. Decis. Mak.* 2015;35(2):162–169.

31. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur. Urol.* 2015;67(6):1142–1151.

Table 1. Data extracted from each study

1	Year of publication
2	First author
3	Disease condition
4	Type of surgery
5	Input feature
6	Number of features in final model
7	Type of outcome
8	Time points of outcome
9	Number of output classes
10	ML algorithm used
11	Number of patient
12	Distribution between training, validation, and test set
13	Validation method
14	AUC and accuracy of model
15	Reporting of calibration and Brier score
16	Decision-curve analysis
17	Digital application of the model

Table 2. Studies evaluating ML models for orthopedic surgical outcome prediction

First author, year of publication	Disease condition	Operation	Input features	Number of features	Output	Output: time points	Number of classes	Machine learning model*	Number of patients	Size training set	Validation method/size	Size test set	AUC	ACC
					Intraoperative complications									
Durand, 2018	Spinal deformity	Nos	Clin, Sur, HRF	4	Intraoperative	3d	2	**RF**, DT	1029	80%	10-FCV	20%	0,85	
Huang, 2018	NA	THA, TKA	Clin, Surg	7	Intraoperative	NA	2	**RF**, LR	15187	100%	5-FCV	NA	0,84	
Siccoli, 2019	Spinal stenosis	Decompression	Clin	15	Intraoperative	45min	2	**RF**, XGB, BDT, KNN, ANN, GLM, BGLM	635	70%	NA	30%	0,54	78%
					Postoperative complications									
Arvind, 2018	NA	ACDF	Clin		Complications	NA	2	**ANN**, SVM, RF	20879	70%	5-FCV	30%	0,65	
Fatima, 2020	Degenerative spondylolisthesis	Nos	Clin, Sur	10	Complications	1m	2	**LR**, LASSO	80610	70%	10-FCV	30%	0,70	
Gowd, 2019	Shoulder arthritis	Total shoulder arthroplasty	Clin, Sur		Complications	1m	2	**LR,** GBM, RF, KNN, DT, NB	17119	80%	CV (nos)	20%	0,71	95%
Han, 2019	Spinal pathology	Spinal surgery	Clin, Sur	274	Complications	1m	2	**LR,** LASSO	1104233	70%	10-FCV	30%	0,70	
Harris, 2018	Osteoarthritis	THA, TKA	Clin, Sur	13	Complications	1m	2	**BR**, LASSO	70569	100%	10-FCV	NA	0,70	
Harris, 2019	Nonemergent primary	THA, TKA	Clin		Complications	1m	2	**LASSO**	107792	100%	10-FCV	NA	0,64	
Hopkins, 2020	Spinal pathology	Posterior spinal fusion	Clin, Sur, HRF		Complications	NA		**NN**	4046	75%	CV (nos)	25%	0,79	
Karhade6, 2020	Spinal pathology	ALIF	Clin, Sur	6	Complications	intra-operative		**EPLR,** SGB, RF, SVM, NN	1035	75%	CV (nos)	25%	0,73	
Kim1, 2018	Spinal deformity	Nos	Clin	12	Complications	NA	2	**ANN**, LR	5818	70%	5-FCV	30%	0,64	
Kim2, 2018	Degenerative spine pathology	PLIF	Clin	12	Complications	NA	2	**ANN**, LR	22629	70%	NA	30%	0,63	

Kukar, 1996	Femur fracture	Nos	Clin	17	Complications	24m	2	**Backpropagation**, ANN, NB, KNN, LFC, DT	151	70%	10-FCV	30%		71%
				17			5	**Semi NB**, ANN, NB, KNN, LFC, DT	151	70%	10-FCV	30%		67%
Pua, 2019	Knee osteoarthritis	TKA	Clin		Complications	6m	2	**LR**, RF, GBM	4026	70%	Inner cross-validation loop	30%	0,75	
Scheer, 2017	Adult spinal deformity	Nos	Clin, Sur, Rad	20	Complications	1.5m	2	**RT**	557	70%	NA	30%	0,89	88%
Wu, 2016	Lower extremities (nos)	Nos (including PCEA)	Clin, Sur	9	Complications	NA	2	**SVM**, LR	195	75%	CV (nos)	25%	0,93	88%
					Medical management									
Gabriel, 2018	Osteoarthritis	THA	Clin	9	Hospitalization	≦3 days	2	**RR**, LASSO, RF, MLR	960	67%	NA	33%	0,76	
Goyal, 2019	Spinal pathology	Spinal fusion	Clin		Non-home discharge	1m	2	**GLM**, NB, ANN, RF, GBM, LDA	59145	100%	10-FCV	NA	0,87	79%
Gowd, 2019	Shoulder arthritis	Total shoulder arthroplasty	Clin, Sur		Extended LOS	1m	2	**GBM**, RF, KNN, DT, NB, LR	17119	80%	CV (nos)	20%	0,68	82%
Karhade1, 2018	LDDD	Nos	Clin	10	Non-home discharge	NA	2	**NN**, BPM, BDT, SVM	26364	80%	10-FCV	20%	0,82	
Karnuta, 2019	Hip fracture	Nos	Clin	7	Hospitalization	NA	4	**NB**	98562	90%	10-FCV	10%	0,88	77%
		Nos	Clin	7	Cost	NA	3	**NB**	98562	90%	10-FCV	10%	0,89	79%
Karnuta, 2020	Spinal pathology	Spinal fusion	Clin	8	Cost	NA	3	**NB**	38070	100%	10-FCV	NA	0,88	80%
		Spinal fusion	Clin	8	LOS	NA	3	**NB**	38070	100%	10-FCV	NA	0,94	87%
		Spinal fusion	Clin	8	Non-home discharge	NA	3	**NB**	38070	100%	10-FCV	NA	0,91	88%
Merrill, 2018	Ankle fracture	ORIF	Clin	9	Hospitalization	3d	2	**Bo**, LR	16501	70%	CV (nos)	30%	0,76	72%
Ogink1, 2019	Spinal stenosis	Surgery	Clin	10	Non-home discharge	NA	2	**ANN**, SVM, BPM, BDT	28600	80%	10-FCV	20%	0,74	
Ogink2, 2019	Degenerative spondylolisthesis	Surgery	Clin	10	Non-home discharge	NA	2	**BPM,** ANN, SVM, BDT	9338	80%	10-FCV	20%	0,75	
Ottenbacher, 2004	Hip fracture	Nos	Clin, Rad	6	Non-home discharge	80d	2	**ANN**, LR	3708	67%	3-FCV	33%	0,73	

Ramkumar, 2019	Osteoarthritis	THA	Clin, HRF	15	LOS	NA	2	**ANN**	78335	100%	10-FCV		0,82	75%
		THA	Clin, HRF	15	Charges	NA	2	**ANN**	78335	100%	10-FCV		0,83	76%
		THA	Clin, HRF	15	Non-home discharge	NA	2	**ANN**	78335	100%	10-FCV		0,79	72%
Siccoli, 2019	Spinal stenosis	Decompression	Clin	15	Hospitalization	28h	2	**XGB**, RF BDT, KNN, ANN, GLM, BGLM	635	70%	NA	30%	0,58	77%
					PROMs									
Azimi, 2014	LSS	Nos	Clin	7	PROM	24m	2	**ANN**, LR	168	50%	25%	25%	0,80	97%
Fontana, 2018	Osteoarthritis	THA, TKA	Clin, Sur, HRF		PROM	24m	2	**LASSO**, RF, SVM	13719	80%	5-FCV	20%	0,80	
Huber, 2018	Osteoarthritis	THA, TKA	Clin		PROM	NA	2	**XGB**, ANN, KNN, NB , RF, MSAENET, LM, LB	66356	97%	5-FCV	3%	0,81	75%
Khan,2019	DCM	Nos	Clin	28	PROM	12m	2	**MARS,** CT, SVM, PLS, GBoM, GAM, RF, LR	193	75%	10-FCV	25%	0,78	71%
Kumar, 2020	Shoulder pathology	aTSA	Clin, Sur	291	PROM	1y, 2-3y, 3-5y,>5y	2	**NN,** LM, DT	4782	67%	NA	33%	0,86	91%
		rTSA	Clin, Sur	291	PROM	1y, 2-3y, 3-5y,>5y	2	**NN,** LM, DT	4782	67%	NA	33%	0,88	94%
Kunze, 2020	Osteoarthritis	THA	Clin	8	PROM	24m		**RF,** SGB, SVM, NN, EPLR	616	80%	CV (nos)	20%	0,97	
Lungu, 2015	Osteoarthritis	THA	Clin	6	PROM	12m, 24m	2	**RF**	265	100%	bootstrap resampling	NA		89%
Merali, 2019	DCM	Decompression	Clin, Surg	5	PROM	6m, 12m, 24m	2	**RF**	605	70%	10-FCV	30%	0,72	71%
Nwachukwu, 2020	Femoroacetabular Impingement	Hip arthroscopy	Clin	5	PROM	24m	2	**LR**	1103	100%	10-FCV	NA	0,86	
Siccoli, 2019	Spinal stenosis	Decompression	Clin	15	PROM	1,5m, 3m	2	**BDT,** RF, XGB, KNN, ANN, GLM, BGLM	635	70%	NA	30%	0,86	76%
Schwartz, 1997	Osteoarthritis	THA	Clin	14	PROM	12m	2	**ANN**, LR	221	95%	LOOCV	5%	0,79	

					Survival									
Arvind, 2018	NA	ACDF	Clin		Survival	NA	2	**ANN**, SVM, RF	20879	70%	5-FCV	30%	0,98	
Chen, 2020	Hip fracture	Nos	Clin, HRF	11	Survival	NA	2	**ANN**	10534	70%	15%	15%	0,93	93%
Forsberg, 2011	Bone metastases	Nos	Clin		Survival	3m, 12m	2	**BNN**	189	90%	10-FCV	10%	0,84	
Harris, 2018	Osteoarthritis	THA, TKA	Clin, Sur	13	Survival	1m	2	**BR**, LASSO	70569	100%	10-FCV	NA	0,73	
Harris, 2019	Nonemergent primary	THA, TKA	Clin		Survival	1m	2	**LASSO**	107792	100%	10-FCV	NA	0,73	
Karhade2, 2018	Spine metastasis	Nos	Clin	7	Survival	1m	2	**BPM**, NN, DT, SVM	1790	80%	10-FCV	20%	0,78	
Karhade3, 2018	Spinal chordoma	Nos	Clin, Surg	5	Survival	60m	2	**BPM**, BDT, SVM, ANN	265	100%	10-FCV	0%	0,80	
Karhade5, 2019	Spine metastasis	Nos	Clin	17	Survival	3m, 12m	2	**SGB**, PLR, RF, NN, SVM	732	80%	10-FCV	20%	0,86	
Kim1, 2018	Spinal deformity	Nos	Clin	12	Survival	NA	2	**ANN**, LR	5818	70%	5-FCV	30%	0,84	69%
Kim2, 2018	Various degenerative diseases	PLIF	Clin	12	Survival	NA	2	**ANN**, LR	22629	70%	NA	30%	0,70	60%
Lin, 2010	Femur fracture	Various	Clin, Rad	11	Survival	12m	2	**ANN**	286	70%	NA	30%	0,95	96%
Merrill, 2018	Ankle fracture	ORIF	Clin	9	Survival	NA	2	**Bo**, LR	16501	70%	CV (nos)	30%	0,74	85%
Pereira, 2016	Spine metastasis	Various	Clin	9	Survival	1m, 3m, 12m	2	**Nomogram**, Bo	649	80%	5-FCV	20%	0,74	75%
Shi, 2013	Femur fracture	DHS	Clin	9	Survival	12m	2	**ANN**, LR	2150	67%	NA	33%	0,87	86%
Thio, 2020	Extremity Metastatic Disease	Nos	Clin	15	Survival	3m, 12m	2	**SGB**, RF, SVN, NN, PLR	1090	80%	10-FCV	20%	0,86	
Zhang, 2019	Pertrochanteric fracture	PFNA	Clin, HRF	14	Survival	12m	2	**BNN**	448	100%	10-FCV	NA	0,85	
					Other									
Anderson, 2020	ACL rupture	ACL reconstruction	Clin		Sustained opioid use	3m	2	**GBM,** LR, BNN, RF	10919	80%	CV (nos)	20%	0,77	
Azimi, 2015	LDH	Microdiscecto-my	Clin	14	Recurrence	NA	2	**ANN**, LR	402	50%	NA	25%	0,83	94%
Bevevino, 2014	Calcaneus fracture	Limb salvage	Clin, Rad	8	Amputation	NA	2	**ANN**, LR	155	100%	10-FCV	NA	0,80	79%
Hopkins2, 2020	Spinal pathology	posterior spinal fusion	Clin, Sur, HRF	177	Readmission	1m	2	**ANN**	23264	75%	Cv (nos)	25%	0,81	79%

Kalagara, 2018	NA	Lumbar laminectomy	Clin, Surg, HRF	13	Readmission	1m	2	**GBM**	26869	85%	10-FCV	15%	0,81	95%
Karhade4, 2019	Cervical pathology	ACDF	Clin, Surg	10	Sustained opioid use	3m	2	**SGB**, RF, NN, SVM, EPLR	2737	80%	10-FCV	20%	0,81	
Karhade7, 2019	Hip arthritis	THA	Clin	7	Sustained opioid use	3m	2	**EPLR**, SGB, RF, SVM, ANN	5507	80%	10-FCV	20%	0,77	
Karhade8, 2019	LDH	Nos	Clin	9	Sustained opioid use	6m	2	**EPLR**, RF, SGB, ANN, SVM	5413	80%	10-FCV	20%	0,81	
Karhade9,2020	LDH, spinal stenosis, spondylolisthesis	Decompression and/or fusion	Clin, Sur	6	Sustained opioid use	3m	2	**EPLR**, SGB, RF, SVM, ANN	8435	80%	10-FCV	20%	0,70	
Katakam, 2020	Knee osteoarthritis	TKA	Clin	9	Sustained opioid use	6m	2	**SGB**, RF, SVM, ANN, EPLR	12542	80%	CV (nos)	20%	0,76	
Martini, 2020	Degenerative spine pathology	Nos	Clin, Sur	30	Readmission	1m	2	**RF**	11150	75%	5-FCV	25%	0,75	
Merrill, 2018	Ankle fracture	ORIF	Clin	9	Readmission	1m	2	**Bo**, LR	33504	70%	CV (nos)	30%	0,70	85%
Siccoli, 2019	Spinal stenosis	Decompression	Clin	15	Reoperations	NA	2	**XGB**, RF, BDT, KNN, ANN, GLM, BGLM	635	70%	NA	30%	0,66	69%
Zhang, 2020	Low back and lower extremity pain	Thoracic or lumbar surgery	Clin, Sur	9	Sustained opioid use	12m	2	**LR**, RF, SGB, SVM, NN	903	80%	NA	20%	0,85	

❋ Best performing ML model is highlighted in bold. ACDF = anterior cervical discectomy and fusion, ANN = artificial neural network, AUC = area under the curve, BDT = boosted decision tree, BGLM = Batesian generalized linear models, BNN = bayesian belief network, Bo = boosting, BPM = Bayes point machine, BR = boosting regression, Clin = clinical, Sur = surgical, CHAID = chi-square automatic interaction dector, CV (nos) = cross-validation not otherwise specified, CT = classification tree, DCM = degenerative cervical myelopathy, DHS = dynamic hip screws, DT = decision tree, EPLR = elastic-net penalized logistic regression, FCM = fuzzy C-means, FCV = fold cross validation, FIS = fuzzy inference system, GAM = generalized additive models, GBM = gradient boosting machine, GboM = generalized boosted models, GLM = generalized linear models, HRF = hospital-related factors (surgeon volume, hospital volume), KNN = K-nearest neighbors, LASSO = least absolute shrinkage and selection operator, LB = logistic boost, LDA = linear discirminant analysis, LFC = lookahead feature construction, LM = linear model, LOOCV = leave-one-out cross validation, LOS = length of stay, LR = logistic regression, MARS = multivariable adaptive regression splines, MLR = multivariable logistic regression, MSAENET = multi-step elastic-net, NA = not available, NB = naïve Bayes, ORIF = open reduction and internal fixation, PCA = principal component analysis, PCEA = patient-controlled epidural analgesia, PLIF = posterior lumbar spine fusion, PLR = penalized logistic regression, PLS = partial least squares, RF = random forests, RT = random trees, RR = ridge regression, SGB = stochastic gradient boosting, SVM = support vector model, SVR = support vector regression, XGB = extreme gradient boosting, LSS = lumbar spinal stenosis, LDH = lumbar disc herniation, LDDD = lumbar degenerative disc disease, nos, not otherwise specified, THA = total hip arthroplasty, TKA = total knee arthroplasty.

Table 3. Characteristics of studies (n=77)	
Variables	
	median (IQR)
Sample size	5818 (635-26.364)
Predictors included in final model[a]	10 (8-15)
	n (%)
Outcome domain	
Medical management	17 (22)
Survival	16 (21)
Complication	15 (19)
PROMs	12 (16)
Intraoperative complication	3 (3.9)
Other	14 (18)
Orthopedic subspecialty	
Spine	28 (36)
Arthroplasty	21 (27)
Trauma	13 (17)
Oncology	6 (7.8)
Other	9 (12)
National/Registry database[b]	42 (55)
Split sample	
70-30	22 (29)
80-20	19 (25)
Other	36 (46)
ML algorithm[C]	
Neural network	42 (55)
Single layer	36 (47)

Deep learning	6 (8)
Convolutional	1 (1)
Random forest	30 (39)
Support vector machine	19 (25)
Naive bayes	11 (14)
Stochastic gradient boosting	10 (13)
Performance metric[c]	
AUC	74 (96)
Accuracy	39 (51)
Brier score	23 (30)
Calibration	26 (34)
Model explanation	
Global	34 (44)
Local	17 (22)
Decision curve analysis	14 (18)
Digital application available	18 (23)
AUC = Area Under the Curve, IQR = Interquartile range, ML = machine learning, PROM = Patient Reported Outcome Measure	
a The amount of predictors that were included in the final, best performing machine learning algorithm. In 16% (13/81) this could not be extracted from the study or was unclear.	
b This includes databases such as Surveillance, Epidemiology, and End Results (SEER) or American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP).	
c Not mutually exclusie	

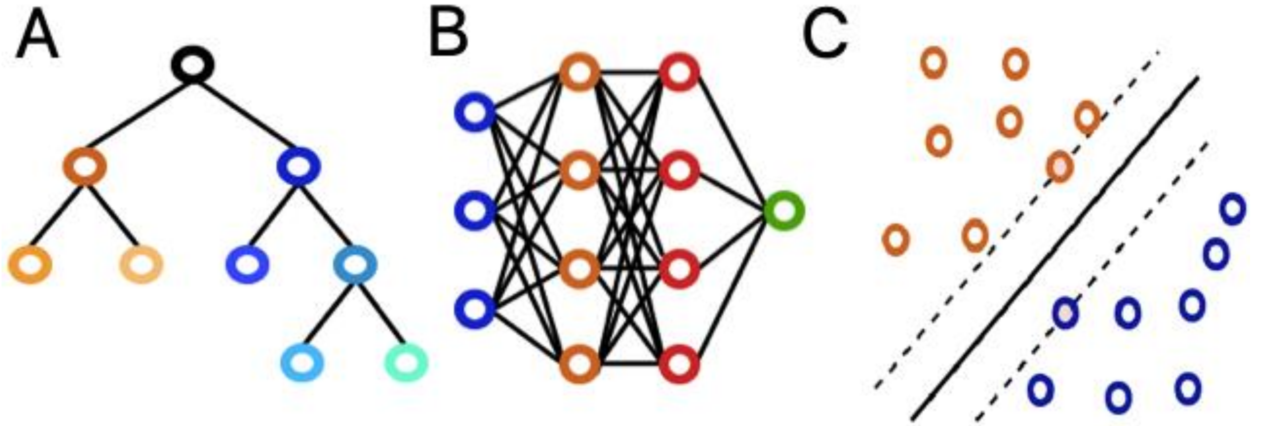


Figure 1. (I) Decision trees are hierarchical structures in which each node performs a test on the input value with the subsequent branches representing the outcomes. Their graphical representation as seen above make them easy to understand and interpret. However, they are prone to overfitting.

(II) Neural networks are based on interconnected nodes. The input features are represented by the first (blue) layer. The designated outcome is represented by the final (green) layer. The middle hidden layers (blue and orange) base their output on the input they get from prior layers. Neural networks have been around for a long time and offer good discriminative abilities, but interpretation of the relationships between the different layers remains difficult.

(III) Support Vector Machines (SVMs) perform classification by determining the optimal separating hyperplane between datapoints which maximizes the distance between the 2 closest points of either group. They can be used for both linear and nonlinear relationship. While they remain effective in data with a great number of features, they do not work well in larger datasets.

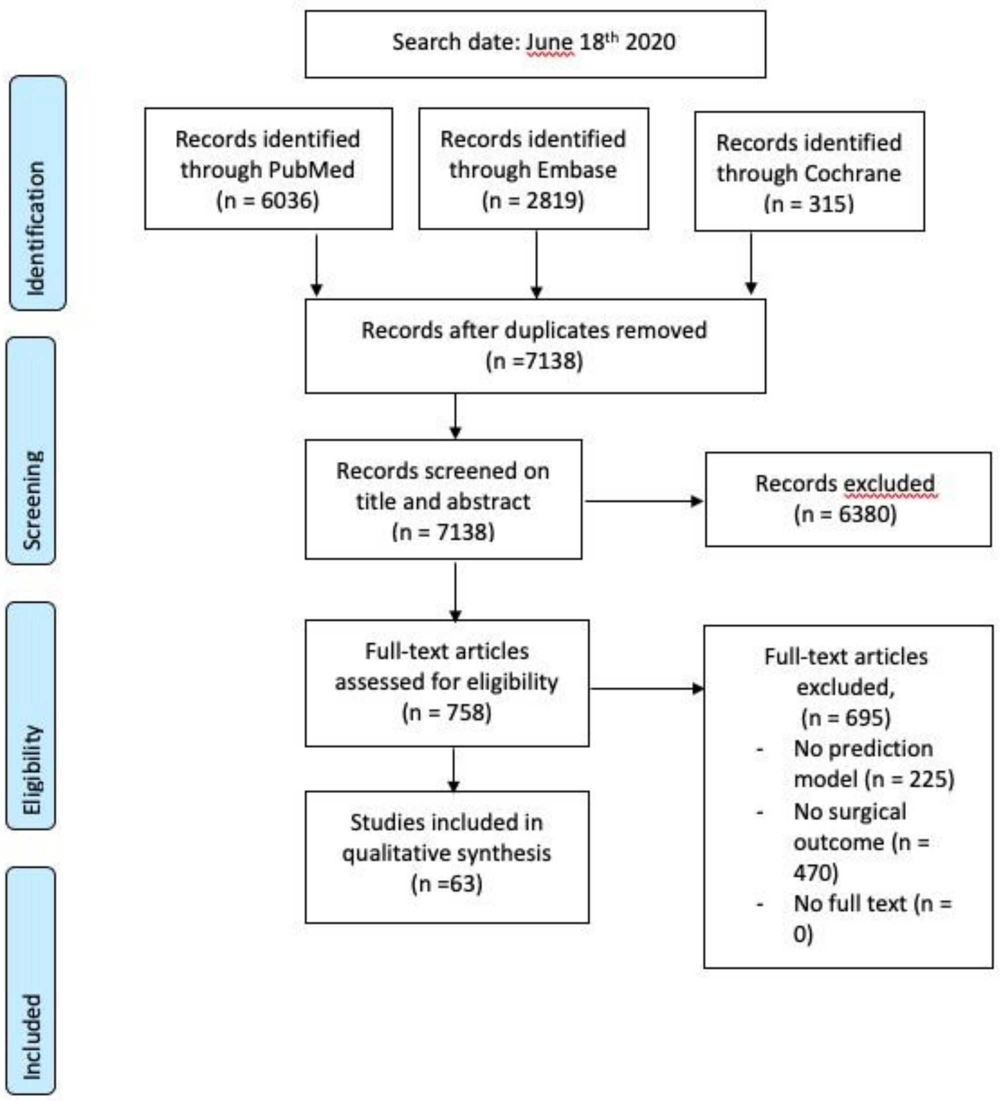

Search date: June 18th 2020
Identification
Records identified through PubMed (n = 6036)
Records identified through Embase (n = 2819)
Records identified through Cochrane (n = 315)
Records after duplicates removed (n =7138)
Screening
Records screened on title and abstract (n = 7138)
Records excluded (n = 6380)
Eligibility
Full-text articles assessed for eligibility (n = 758)
Full-text articles excluded, (n = 695)
- No prediction model (n = 225)
- No surgical outcome (n = 470)
- No full text (n = 0)
Included
Studies included in qualitative synthesis (n =63)

Figure 2. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flowchart of study inclusions and exclusions.

Chapter 2

Machine learning prediction models in orthopedic surgery: A systematic review in transparent reporting

Ogink PT[1], Groot OQ[1], Lans A, Twining PK, Kapoor ND, DiGiovanni W, Bindels BJJ, Bongers MER, Oosterhoff JHF, Karhade AV, Oner FC, Verlaan JJ, Schwab JH.
[1] *Shared first authorship*

Journal of Orthop Res. 2022 Feb;40(2):475-483

ABSTRACT

Background Machine learning (ML) studies are becoming increasingly popular in orthopaedics but lack a critically appraisal of their adherence to peer-reviewed guidelines.

Objectives (1) Evaluate quality and transparent reporting of machine ML prediction models in orthopaedic surgery based on the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement;

(2) Assess the risk of bias with the Prediction model Risk Of Bias ASsessment Tool (PROBAST) guidelines.

Design Systematic review

Methods A systematic review was performed to identify all ML prediction studies published in orthopaedic surgery through June 18th, 2020. Studies were included if they evaluated ML models for any prediction in an orthopaedic surgery outcome such as survival, patient reported outcomes measures (PROMs), or complications. Exclusion criteria were (1) non-ML techniques (such as multivariable regression analysis), (2) conference abstracts, (3) non-English studies, (4) lack of full-text, and (5) non-relevant study types such as animal studies, letters to the editors, and case-reports. Two reviewers independently extracted data and discrepancies were resolved by discussion with at least two additional reviewers present.

Results After screening 7138 studies, 59 studies met the study criteria and were included. Across all studies, the overall median completeness for the TRIPOD checklist was 53% (interquartile range 47%-60%). TRIPOD items that were reported in less than 10% of studies were abstract (3%), model-building procedures (3%), and model specifications (8%). TRIPOD items that were reported in more than 90% of studies were data source (100%), overall interpretation (98%), limitations (97%), and specifying the objective (95%). As assessed by PROBAST, the overall risk of bias was low in 44% (n=26), high in 41% (n=24), and unclear in 15% (n=9). High overall risk of bias was driven by incomplete reporting of performance measures, inadequate handling of missing data, and use of small datasets with not enough number of outcomes.

Conclusion Although the number of ML studies in orthopaedic surgery is increasing rapidly, over 40% of the existing models are at high risk of bias. Furthermore, over half incompletely reported their methods and/or performance measures. Until these issues are adequately addressed to give patients and providers trust in ML models, a considerable gap remains between the development of ML prediction models and their implementation in orthopaedic practice.

Introduction

Prediction models for orthopaedic surgical outcomes based on machine learning (ML) are rapidly emerging. Such models, if adequately reported, can guide treatment decision making, predict adverse outcomes, and streamline perioperative healthcare management. However, transparent and complete reporting is required to allow the reader to critically assess the presence of bias, facilitate study replication, and correctly interpret study results. Unfortunately, previous studies have suggested that prediction models demonstrate incomplete, untransparent reporting of items such as study design, patient selection, variable definitions and performance measures.[1,2] To our knowledge, there is no systematic review that has assessed the completeness of reporting for the currently available prognostic ML models in orthopaedic surgery.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement was published in 2015 to improve the quality of reporting of prediction models.[3,4] It provides a guideline for essential elements of prediction model studies. The statement is endorsed by over ten leading medical journals and has been cited thousands of times. The Prediction model Risk Of Bias ASsessment Tool (PROBAST) was developed to assess risk of bias in prediction models by the Cochrane Prognosis group in 2019, and has been successfully piloted.[5] Both the PROBAST and TRIPOD had yet to be published at the time several ML prediction models for orthopaedic surgical outcome were developed; nonetheless, we believe they can be used as benchmarks for measuring quality of reporting and bias even if the prediction models were published before their introduction.

In this systematic review, we (1) evaluate the quality and completeness of reporting of prediction model studies based on ML for prognosis of surgical outcomes in orthopaedics according to their adherence to the TRIPOD statement, and (2) assess the risk of bias with the PROBAST.

Methods

Systematic Literature Search

Registration in the PROSPERO international prospective register of systematic reviews was performed prior to study initiation and can be found online (registration number CRD42020206522). The study is reported according to the 2009 PRISMA guidelines.[6] A systematic search, in collaboration with a medical professional librarian, of the available literature was performed in PubMed, Embase, and the Cochrane Library for studies published up to June 18th, 2020. Different domains of medical subject headings (MeSH) terms and keywords were combined with 'AND'. Two domains with all related words were included in our search: ML and all possible orthopaedic specialties (Appendix 1). Two reviewers (PTO, OQG) independently screened and assessed all eligible studies based on predefined criteria (Figure 1).

Eligibility Criteria

Studies were included if they evaluated ML models for any prediction in an orthopaedic surgery outcome such as survival, patient reported outcomes measures (PROMs), or complications. Exclusion criteria were (1) non-ML techniques (such as logistic or linear regression analysis), (2) conference abstracts, (3) non-English studies, (4) lack of full-text, and (5) non-relevant study types such as animal studies, letters to the editors, and case-reports. Orthopaedic specialties were defined as any operation for patients with musculoskeletal disorders.

Data Extraction

Six reviewers (PTO, OQG, AL, PT, NDK, BBJ) independently assessed the first 10% of studies. All extracted data were then discussed during a group session with the principal investigator (PI) (JHS) to ensure quality and consistency. Any questions about discrepancies in the extracted data were resolved by the PI. After this quality training, the same six reviewers split up in pairs of two and each pair independently assessed the remaining 90% of studies which were evenly distributed among the three formed pairs. Each pair consisted of a research fellow with a medical doctor degree and a medical student. Disagreements within a pair were resolved during a consensus meeting with at least two other reviewers present. All six reviewers and the PI previously worked on and/or published ML prediction models in orthopaedic surgical outcomes.

For each included study, we extracted the following information: journal, prospective study design (yes/no), use of national or registry database (yes/no), size of total dataset, number of predictors used in final ML model, predicted outcome, mention of adherence to TRIPOD guideline in study (yes/no), access to ML algorithm (yes/no), TRIPOD items and PROBAST domains. The TRIPOD items and PROBAST domains are explained in more detail below.

The TRIPOD statement consists of 22 main items, of which two main items (12 and 17) refer to model updating or external validation studies, leaving 20 main items to be extracted for prognostic prediction modeling studies[4]. These main items were transformed into an adherence assessment form by the statement developers. Of the 20 main items, 11 had no subitems (1, 2, 8, 9, 11, 16, 18, 19, 20, 21, and 22), seven were divided into two subitems (e.g. 3a and 3b; 3, 4, 6, 7, 13, 14, and 15), and two into three subitems (e.g. 5a, 5b, 5c; 5 and 10). Four subitems (10c, 10e, 13c, and 19a) were, together with the two main items (12 and 17), not extracted because they did not refer to developmental studies (e.g. 10c "For validation, describe how the predictions were calculated"; Appendix 2). Hereafter, subitems and main items are defined under one nomenclature "items" (e.g. main item 3 consists of two items; 3a and 3b). In total, 29, 30, or 31 potential items could be assessed per study. This total number of items varied between 29 and 31 because some items could be scored with "not applicable" (e.g. 14b "if nothing on univariable analysis (in methods or results) is reported, score not applicable") and this was excluded when calculating the completeness of reporting. Also, some items could be scored with "referenced" (e.g. item 6a) Referenced was considered "completed" and included when calculating the completeness of reporting.

Each item may consist of multiple elements. Both elements must be scored "yes" for the item to be scored "completed." To calculate the completeness of reporting of TRIPOD items, the number of completely reported TRIPOD items was divided by the total number of TRIPOD items for that study. If a study reported on multiple prediction models (e.g. prediction model for 90-day and 1-year survival), we extracted data only on the best performing model.

PROBAST assesses the risk of bias in prognostic prediction model studies[5]. This tool consists of 20 signaling questions across four domains: participants selection (1), predictors (2), outcome (3), and analysis (4). Each domain is rated "low", "high", or "unclear" risk of bias. 'Unclear" indicates that the reported information is insufficient – no reliable judgement on low or high risk of bias can be made with the information provided. Participants selection (1) covers potential sources of bias in the origin of data and criteria for participant selection – are all patients included and excluded appropriately? Predictors (2) should include a list of all considered predictors, a clear definition and timing of measurement. An outcome (3) should include clear definitions and timing of measurements, and a description of the time interval between predictor assessment and outcome determination. Lasty, analysis (4) covers potential sources of bias related to inappropriate analysis methods or omission of key performance measures such as discrimination and calibration.

The ratings of the four domains resulted in an overall judgement about risk of bias. Low overall risk of bias was assigned if each domain scored low. High overall risk of bias was assigned if at least one domain was judged to be high risk of bias. Unclear overall risk of bias was noted if at least one domain was judged unclear and all other domains low. The four domains and the overall judgement were reported – not every signaling question.

Statistical Analysis

Completeness of reporting of TRIPOD items and PROBAST domains were visualized by bar graphs. We used Microsoft Excel Version 19.11 (Microsoft Inc, Redmond, WA, USA) to extract and record data using standardized forms, Stata® 14.0 (StataCorp LP, College Station, TX, USA) for the statistical analyses, and Mendeley Desktop Version 1.19.4 (Mendeley Ltd, London, UK) as reference management software.

Results

The conducted search yielded 7,138 unique studies. Seven hundred and fifty-eight potential studies were selected by title and abstract screening, of which 59 remained after full-text screening (Appendix 3). Table 1 lists the study characteristics of the included study. The majority (83%; 49/59) was published after the launch of the TRIPOD statement (see Appendix 4). The 59 studies were published in 33 different medical journals of which three journals published 31% of all included studies (18/59). None of the studies were published in a journal that requested adherence to the TRIPOD guidelines in their instructions to authors.

TRIPOD

Among all studies, the overall median completeness for the TRIPOD items was 53% (IQR 47%-60%; see Figure 2 and Appendix 5). Eight items were reported in over 75% of studies and seven items in less than 25% (Table 2). The abstract (2) and the model-building procedure (10b) were the most poorly reported items with only 3% (2/59). Source of data (4a) was reported in all studies (100%; 59/59).

PROBAST

The overall risk of bias was low in 44% (26/59), high in 41% (24/59), and unclear in 15% (9/59) of the studies (Figure 3.). The studies that rated highly for overall risk of bias were mainly rated this way due to bias in the analysis domain, (as opposed to the other three domains) incomplete reporting of performance measures, inadequate handling of missing data, or use of small datasets with low number of outcomes. Most notable was the lack of adequate reporting of performance measures such as calibration results, Brier scores, or decision-curves. Unclear risk of bias in the analysis domain was scored in 20% (12/59), mainly due to the lack of mention as to how continuous and categorical predictors were handled or how the handling of complexities in the data was reported (e.g. competing risk analysis).

Discussion

In this systematic review we aimed to assess the quality and transparency of reporting of currently published ML prediction models in surgical outcome in orthopaedics using the TRIPOD and PROBAST guidelines. The reporting of the study abstract had the worst adherence in existing models. According to the PROBAST, 41% of the studies displayed a high risk of bias, primarily due to risk of bias in the analysis domain. ML prediction models may support clinical decision making, but future studies should adhere to recognized methodological standards in order to develop ML prediction models of clinically significant value to healthcare professionals.

This review has several limitations. First, despite using a comprehensive search term in multiple online medical libraries, we may have missed some publications. However, we do not believe that these missed publications would have had a profound impact on the completeness of our reporting or on the final conclusions. Considering the large number of included studies, adding potentially missed studies would most likely not change our main conclusions that the overall adherence is poor. Second, TRIPOD guidelines were employed as a reporting benchmark. However, the relative importance of each item and what composes an acceptable score is up for debate. Third, a strict adherence to scoring was implemented on all elements of a TRIPOD item. For example, item 2 "Abstract" consists of 12 elements which all have to be fulfilled in order for item 2 to be marked as "completely reported". Also, authors as well as journal reviewers might have good reasons to exclude certain TRIPOD information. For example, one may not report regression coefficients in item 15 "model specifications" or provide "the potential clinical use of the model" in item 20 if they believe that their prediction model is not fit for clinical use. Nonetheless, we scored these items in this current study as "incomplete". This rigorous method of scoring is in line with the nature of the TRIPOD guideline and is deemed essential for consistent and transparent reporting of prediction models. In addition, most journals require a maximum word count or prescribe specific requirement. These restrictions could potentially prevent authors from including all 12 elements. Despite these limitations, this review provides the first comprehensive overview of completeness of transparent reporting for ML prediction models in orthopaedics. Illustrating poor reporting of TRIPOD items identifies current hurdles and may improve future transparent reporting.

The TRIPOD statement was published in 2015 to provide a framework for transparent reporting and quality of prediction models. Despite being published in 11 medical journals and being well-referenced 24% [12/49] of included studies published after the TRIPOD statement referenced TRIPOD. A possible explanation is the usual slow implementation of guidelines after publication.[7–12] Although the 11 medical journals are leading, high impact journals, none are orthopaedic specific journals so they may have been missed by the orthopaedic community. Another reason could be that authors of ML models have been dissuaded to adhere to TRIPOD doubting its applicability to their study. The explanatory documents of the TRIPOD statement focus on models based on regression techniques and several items do not fully pertain to ML, e.g. item 15a on regression coefficients. The authors of the TRIPOD statement recently acknowledged this drawback and have announced the development of a version specific to ML, TRIPOD-ML, similar to the CONSORT-AI extension.[13,14]

Alternative reasons for incomplete items are reviewers demanding different information than the items in TRIPOD, journal format and maximum word count limiting the number of items to mention, or researchers only using reporting guidelines near the end when writing up the manuscript. A study by Agha et al. [15] found considerable improvement in reporting was achieved after a surgical journal started mandating reporting guideline checklists to be included in the submission to the editor and reviewers. This could trigger researchers to include reporting guidelines like TRIPOD or ARRIVE (Animal Research: Reporting In Vivo Experiments)[16] in the early stages of study design instead of during manuscript writing, which according to Dewey et al. led to increased perceived value of the reporting guidelines.[17] However, adherence to TRIPOD is not a panacea. Logullo et al.[18] argue adherence to guidelines does not equal quality despite often being interpreted that way. For the TRIPOD statement it is important to stress the relative importance of each item as well as what constitutes a "good" score is debatable. For example, the omission of any calibration measure is arguably worse than incomplete reporting of the title. Nonetheless, in this relatively new research field it is a useful framework for standardization of reporting and researchers should strive to adhere to the TRIPOD statement.

According to the PROBAST assessment numerous studies were at high risk of bias. Predominantly, three area in the analysis domain were poorly scored. First, most models were built on databases with missing values, mostly due to use of national or registry databases such as NSQIP. Most often, predictors with incomplete data were excluded in the model building process. Both may lead to confounding or selection bias.[19,20] In other words, variables with a strong predictive accuracy may be missed or misinterpreted. This highlights the importance of preferably using prospective, complete datasets, and when missing data are present, processing them appropriately through techniques such as multiple imputation.[21]

A second issue is the incomplete reporting of performance measures. The vast majority of studies describe discrimination measures, predominantly area under the curve, while only a minority report calibration measure. Calibration is an essential element of describing the performance of ML models and its importance has extensively been discussed in earlier reviews.[22–24] The frequent omission of calibration renders assessment of performance virtually impossible and is in line with previous literature on prediction models.[2,25,26]

Finally, the small sample sizes with often small outcome numbers introduce risk of overfitting. Overfitting refers to including too many prognostic factors relative to the number of cases. This may improve the prediction performance in the dataset but reduces the generalizability outside the training dataset. While the use of national databases may circumvent the issue of small sample sizes, they have the disadvantage of oftentimes less granular data (e.g., lacking PROM scores), missing data, as highlighted earlier, and may lack important predictors such as laboratory values.[27]

Our findings lead to some careful recommendations for researchers developing ML prediction models. First, authors should mind all the necessary steps in model development and reporting, starting at the early stages of study design; the TRIPOD checklist can be a guiding tool to this end. Second, next to discrimination and calibration, model performance should always include a measure of clinical utility for decision-making. Decision-making analysis has been around for a significant amount of time, but has only recently started gaining popularity as a valuable tool in prediction models.[22,28] In short, decision-making analysis measures the net benefit of using the ML model prediction across the entire spectrum of predictions by weighing both the benefits for certain patients (true-positives) and the harm for other patients (false-positives). This is preferably assessed and visualized using decision curve analysis.[29]

Third, mere development of clinical prediction models is not the end goal, as they are eventually intended to be used in clinical practice. Prior to utilization by the medical community, extensive external validation is required to ensure robustness of the model outside the database used for development. However, less than half of the published studies offered means to calculate predictions through web calculators or in-study formulas, making external validation and individual predictions difficult.[30] Ideally, the algorithms are published online to facilitate sharing and collaboration.

Conclusion

Prognostic surgical outcome models are rapidly entering the orthopaedic field to guide treatment decision making. This review indicates that numerous studies display poor reporting and are at high risk of bias. Future studies aimed at developing prognostic models should explicitly address the concerns raised, such as incomplete reporting of performance measures, inadequate handling of missing data, and not providing means to make individual predictions. Collaboration for sharing data and expertise is needed not just for developing more reliable prediction models, but also for validating current models. Methodological guidance such as the TRIPOD statement should be followed, for unreliable prediction models can cause more harm than benefit when guiding medical decision making.

References

1. Groot OQ, Bongers MER, Ogink PT, et al. Does artificial intelligence outperform natural intelligence in interpreting musculoskeletal radiological studies? A systematic review. *Clin Orthop Relat Res*. 2020;478:2751-2764

2. Wang W, Kiik M, Peek N, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One*. 2020;15(6):e0234722.

3. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol*. 2015;67(6):1142–1151.

4. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Med*. 2018;16(120).

5. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–58.

6. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015;4(1):1.

7. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA*. 2001;285(15):1992–1995.

8. Korevaar DA, van Enst WA, Spijker R, et al. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evid Based Med*. 2014;19(2):47–54.

9. Sekula P, Mallett S, Altman DG, et al. Did the reporting of prognostic studies of tumour markers improve since the introduction of REMARK guideline? A comparison of reporting in published articles. *PLoS One*. 2017;12(6):e0178531.

10. Smidt N, Rutjes AWS, van der Windt DAWM, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology*. 2006;67(5):792–797.

11. Turner L, Shamseer L, Altman DG, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev*. 2012;11(11):MR000030.

12. Chan A-W, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet (London, England)*. 2005;365(9465):1159–1162.

13. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*. 2020;370:m3164.

14. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet (London, England)*. 2019;393(10181):1577–1579.

15. Agha RA, Fowler AJ, Limb C, et al. Impact of the mandatory implementation of reporting guidelines on reporting quality in a surgical journal: A before and after study. *Int J Surg*. 2016;30:169–172.

16. Percie du Sert N, Hurst V, Ahluwalia A, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *J Physiol*. 2020;598(18):3793–3801.

17. Dewey M, Levine D, Bossuyt PM, et al. Impact and perceived value of journal reporting guidelines among Radiology authors and reviewers. *Eur Radiol*. 2019;29(8):3986–3995.

18. Logullo P, MacCarthy A, Kirtley S, et al. Reporting guideline checklists are not quality evaluation forms: they are guidance for writing. *Heal Sci reports*. 2020;3(2):e165.

19. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA Symp*. 2013;2013:1109–1115.

20. Skelly AC, Dettori JR, Brodt ED. Assessing bias: the importance of considering confounding. *Evid Based Spine Care J*. 2012;3(1):9–12.

21. Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data. *JAMA*. 2015;314(18):1966–1967.

22. Karhade AV, Schwab JH. CORR synthesis: when should we be skeptical of clinical prediction models? *Clin Orthop Relat Res*. 2020;478:2722-2728

23. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.

24. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928–935.

25. Hodgson A, Helmy N, Masri BA, et al. Comparative repeatability of guide-pin axis positioning in computer-assisted and manual femoral head resurfacing arthroplasty. *Proc Inst Mech Eng H*. 2007;221(7):713–724.

26. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ*. 2020;369:m1328.

27. Janssen DMC, van Kuijk SMJ, d'Aumerie BB, et al. External validation of a prediction model for surgical site infection after thoracolumbar spine surgery in a Western European cohort. *J Orthop Surg Res*. 2018;13(1):114.

28. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565–574.

29. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925–1931.

30. Groot OQ, Bindels BJJ, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop*. 2021:1–9.

Table 1. Characteristics of included studies (n=59)

Variables	**Median (IQR)**
Sample size	4782 (616-23.264)
Predictors included in final model[a]	10 (7-14)
	% (n)
Year of publication	
<2015 (prior to TRIPOD guideline)	17 (10)
>2016	83 (49)
Number of publications per journal	
<5 publications per journal	69 (41)
>5 publications per journal	31 (18)
Prospective database	3 (5)
National/Registry database[b]	51 (30)
Mention of using TRIPOD	20 (12)
Predicted outcome	
Complications	24 (14)
PROM	20 (12)
Mortality	19 (11)
Health management	19 (11)
Other	19 (11)

TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; ML=machine learning; PROM=Patient Reported Outcome Measure;

a The amount of predictors that were included in the final, best performing machine learning algorithm. In 14% (8/59) this could not be extracted from the study or was unclear.

b This includes databases such as Surveillance, Epidemiology, and End Results (SEER) or American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP).

Table 2. Individual TRIPOD items sorted by completeness of reporting over 75% and under 25%.

Complete reporting > 75%			**Complete reporting < 25%**		
TRIPOD item	*TRIPOD description*	*% (n)*	*TRIPOD item*	*TRIPOD description*	*% (n)*
4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data).	100% (59)	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	3% (2)
19b	Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence.	98% (58)	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	3% (2)
18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	97% (57)	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	8% (5)
3b	Specify the objectives, including whether the study describes the development of the model.	95% (56)	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	19% (11)
3a	Explain the medical context and rationale for developing the multivariable prediction model, including references to existing models.	85% (50)	14a	Specify the number of participants and outcome events in each analysis.	20% (12)
5b	Describe eligibility criteria for participants.	83% (49)	1	Identify the study as developing a multivariable prediction model, the target population, and the outcome to be predicted.	20% (12)
5c*	Give details of treatments received, if relevant.	81% (48)	14b*	If done, report the unadjusted association between each candidate predictor and outcome.	24% (11)
8	Explain how the study size was arrived at.	76% (45)			

TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.
**All items consisted of 59 datapoints, except for 5c (58) and 14b (45) due to "Not applicable" option.*

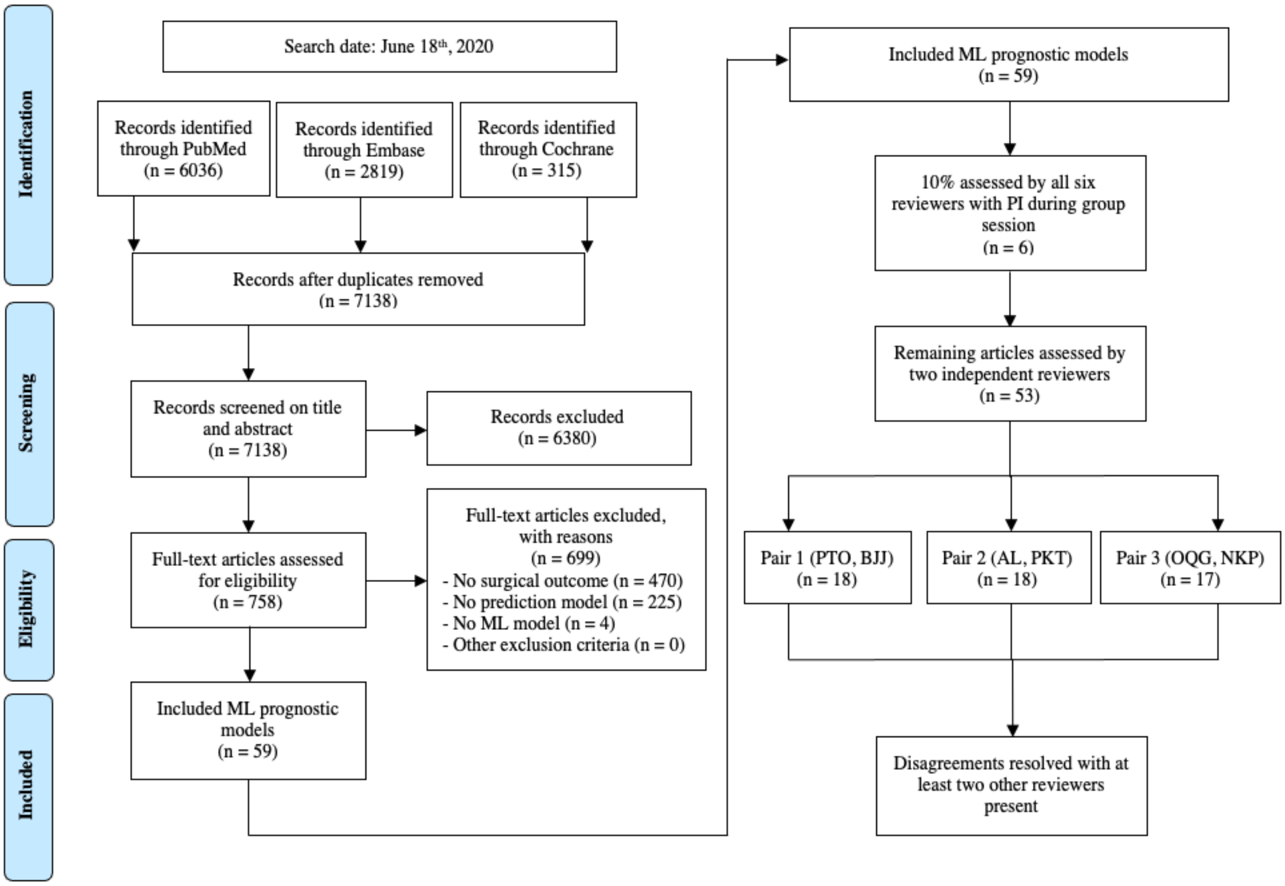

Identification
Screening
Eligibility
Included
Search date: June 18th, 2020
Records identified through PubMed (n = 6036)
Records identified through Embase (n = 2819)
Records identified through Cochrane (n = 315)
Records after duplicates removed (n = 7138)
Records screened on title and abstract (n = 7138)
Records excluded (n = 6380)
Full-text articles assessed for eligibility (n = 758)
Full-text articles excluded, with reasons (n = 699)
- No surgical outcome (n = 470)
- No prediction model (n = 225)
- No ML model (n = 4)
- Other exclusion criteria (n = 0)
Included ML prognostic models (n = 59)
Included ML prognostic models (n = 59)
10% assessed by all six reviewers with PI during group session (n = 6)
Remaining articles assessed by two independent reviewers (n = 53)
Pair 1 (PTO, BJJ) (n = 18)
Pair 2 (AL, PKT) (n = 18)
Pair 3 (OQG, NKP) (n = 17)
Disagreements resolved with at least two other reviewers present

Figure 1. PRISMA flowchart of study inclusions and exclusions. ML=machine learning; PI=principal investigator.

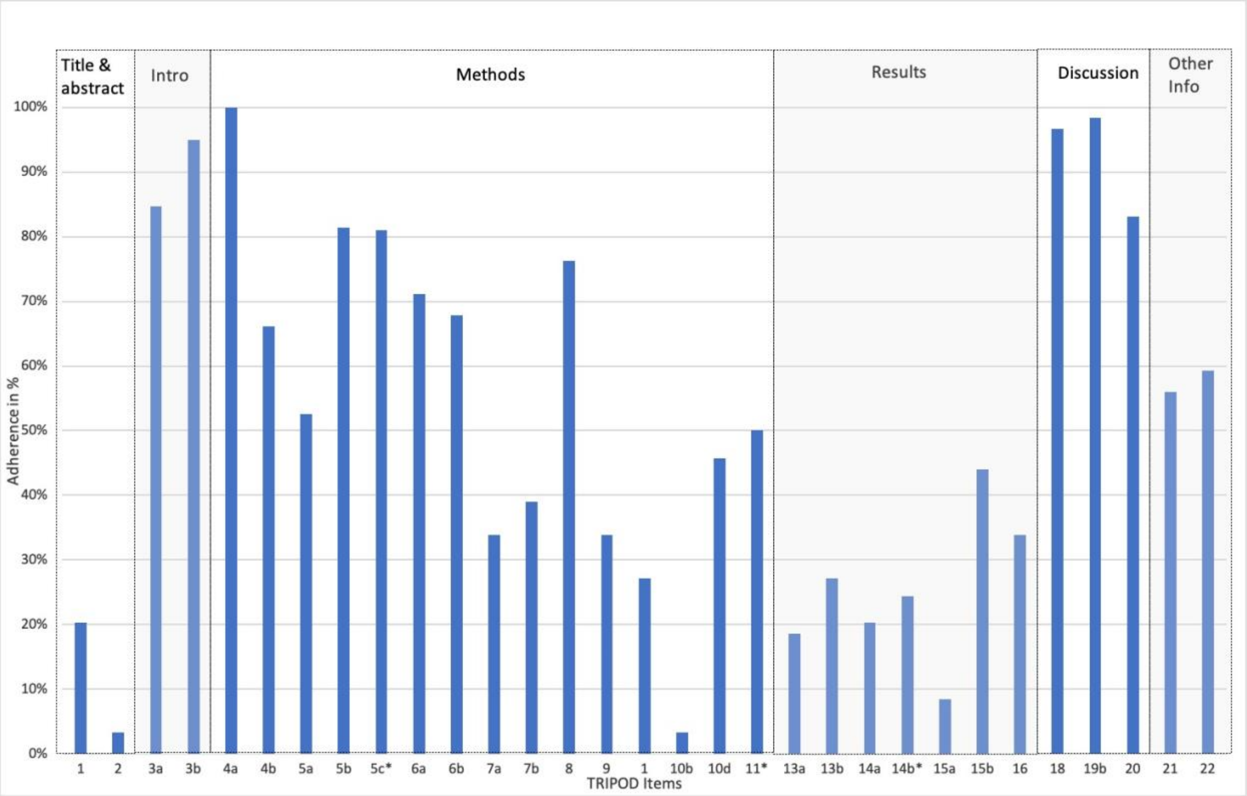


Figure 2. Overall adherence per TRIPOD item. *All items consisted of 59 datapoints, except for item 5c (58), item 11 (4) and item 14b (45) due to the "Not applicable" option.

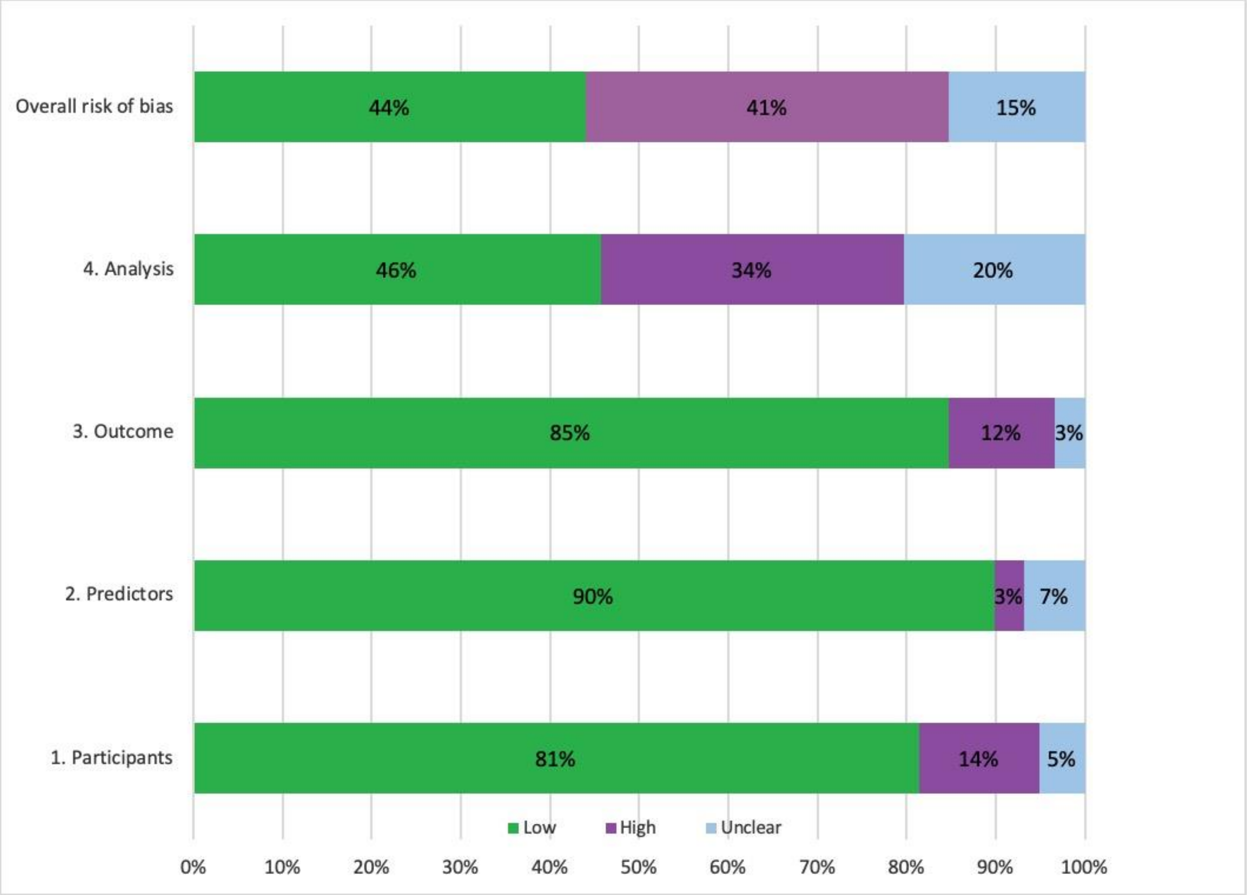


Figure 3. PROBAST results for all included studies (n=59).

Part II

Development of Prediction Models

Chapter 3

Nonoperative Management of Spinal Epidural Abscess: Development of a Predictive Algorithm for Failure

Shah AA, Ogink PT, Nelson SB, Harris MB, Schwab JH

J Bone Joint Surg Am. 2018 Apr 4;100(7):546-555

Abstract

Background Prompt diagnosis and treatment are critical in spinal epidural abscess, as delay can lead to paralysis or death. The initial management decision for spinal epidural abscess is not always clear, with the literature showing conflicting results. When considering nonoperative management, it is crucial to avoid failure of treatment, given the neurologic compromise incurred through failure. Unfortunately, data regarding risk factors associated with failure are scarce.

Methods All patients admitted to our hospital system with a diagnosis of spinal epidural abscess from 1993 to 2016 were identified. Patients who were ≥18 years of age and were initially managed nonoperatively were included. Explanatory variables and outcomes were collected retrospectively. Bivariate and multivariable analyses were performed on these variables to identify independent predictors of failure of nonoperative treatment. A nomogram was constructed to generate a risk of failure based on these predictors.

Results We identified 367 patients who initially underwent nonoperative management. Of these, 99 patients underwent medical management that failed. Multivariable logistic regression yielded 6 independent predictors of failure: a presenting motor deficit, pathologic or compression fracture in affected levels, active malignancy, diabetes mellitus, sensory changes, and dorsal location of abscess. We constructed a nomogram that generates a probability of failure based on the presence of these factors.

Conclusions By quantifying the risk of failure on the basis of the presence of 6 independent predictors of treatment failure, our nomogram may provide a useful tool for the treatment team when weighing the risks and benefits of initial nonoperative treatment compared with operative management.

Introduction

The diagnosis and management of spinal epidural abscess are challenging because of its low incidence, insidious presentation, and nonspecific presenting symptoms. A delay in diagnosis and treatment can be dangerous, potentially leading to neurologic impairment or paralysis[1,2].

For much of the twentieth century, urgent surgical decompression with intravenous antibiotics was the gold standard for spinal epidural abscess management[3-9]. Some authors have suggested that with earlier diagnosis afforded by advances in imaging, nonoperative management may be a valid treatment option for spinal epidural abscess[6-8,10,11]. Indeed, there have been a number of reports of successful medical management[4-10,12,19]. Nonetheless, these studies recommended closely following patients who are managed nonoperatively. Disease progression, including neurologic compromise, spinal instability, severe spinal angulation, and sepsis, can be precipitous and unpredictable[2,8].

The initial treatment modality (i.e., operative compared with nonoperative) is of the utmost importance in spinal epidural abscess. It is crucial to avoid failure of nonoperative management, given the risk of neurologic compromise incurred through failure.[20] Data with regard to risk factors associated with failure are scarce. Studies by Patel et al.[20] and Kim et al.[21] have identified potential predictors of failure of nonoperative management and have demonstrated the importance of stratifying patients to determine who is most likely to undergo failed nonoperative management.

We primarily aimed to identify independent risk factors for failure of nonoperative management, providing guidance for when it is acceptable to opt for nonoperative management. Secondarily, we aimed to develop a nomogram that generates a probability of failure of nonoperative management based on the presence of independent risk factors.

Materials and methods

Study Design and Subjects

Our institutional review board approved a waiver of consent for this retrospective study. We included patients who were ≥18 years of age diagnosed with spinal epidural abscess by magnetic resonance imaging (MRI) or computed tomography (CT) in our hospital system of 2 tertiary academic medical centers and 3 regional community hospitals. We excluded patients who were initially treated operatively or who began treatment at an outside institution.

We identified our cohort by performing a computer query search of all patients admitted to our institution between 1993 and 2016 for International Classification of Diseases, Ninth Revision (ICD-9) codes for spinal epidural abscess and synonyms (ICD-9 324.1 and ICD-10 G06.1). We also performed a computer query search for Current Procedural Terminology (CPT) codes for "laminectomy for excision or evacuation of intraspinal lesion other than neoplasm, extradural" (CPT 63275 to 63278). This initial search yielded 2,756 unique patients. Screening these medical records yielded 1,053 potentially eligible patients. For patients who presented with spinal epidural abscess on >1 occasion, the first encounter of nonoperative management was included.

Of these 1,053 patients, 472 were initially treated nonoperatively at the discretion of the primary attending physician. We defined nonoperative treatment as systemic antibiotic therapy with or without CT-assisted percutaneous drainage. Treatment groups were defined by the intention of the treating team: we considered the patient to have been treated nonoperatively if the team initially elected for nonoperative management. We excluded patients who were treated nonoperatively for palliation or because they were too ill to undergo a surgical procedure. Only patients for whom the primary team decided nonoperative management was the best treatment modality for eradicating the infection were included.

To avoid patients being prematurely labeled as having not undergone failed treatment compared with being in the process of treatment failure when they were lost to follow-up, we only included patients without documented treatment failure if they had ≥60 days of follow-up from initiation of treatment. If patients had follow-up of <60 days but had a documented treatment failure, they were included. This yielded 367 patients (Fig. 1).

Outcome and Other Variables

Our primary outcome measure was failure of nonoperative management. Failure was defined as neurologic deterioration, worsened back and/or radicular pain, persistent symptoms, or progression on serial imaging despite initiation of antibiotic therapy. Nonoperative management was only considered to have failed if it was initiated with the goal of successfully eradicating the infection. Radiographic progression of disease or worsening symptoms in a patient who was treated nonoperatively for palliation or because of an inability to undergo a surgical procedure was not considered a failure.

For patients with multiple presentations for spinal epidural abscess, we carefully analyzed subsequent presentations to ensure that these did not represent treatment failures. Seven patients managed nonoperatively had subsequent presentations for spinal epidural abscess not due to treatment failure, with a median time between presentations of 48 weeks. Four subsequent abscesses were at different locations from the original presentation, 1 was due to a different microorganism, and 1 was an abscess that developed after a surgical procedure for a spinal epidural abscess that underwent failed nonoperative management. The remaining subsequent presentation was due to reseeding of the epidural space by a fistula in the setting of metastatic rectal cancer.

We extracted the following explanatory factors from review of clinical notes: age, sex, body mass index (BMI), social habits, medical comorbidities, previous spinal procedures or instrumentation, concurrent infections, back or radicular pain, presenting motor function, bowel or bladder dysfunction, and sensory dysfunction. In terms of laboratory values, we collected data on white blood-cell (WBC) count, erythrocyte sedimentation rate (ESR), and C-reactive protein (CRP). Motor status was determined using the American Spinal Injury Association (ASIA) Scale.[22] Abscess anatomy and presence of concurrent spinal infections were determined from radiology reports. Blood and tissue culture data were obtained through microbiology reports.

Motor or nonmotor neurologic deficits were scored as positive only if these were new symptoms. We define sensory changes to include frank sensory deficit and subjective paresthesias. Patients were considered to be immunocompromised if they had an immunosuppressive condition or were on immunosuppressive medications (e.g., chemotherapy and corticosteroids). Previous spinal procedures within 1 year prior to admission included any spinal surgical procedure, implantation of epidural devices, and/or epidural corticosteroid injection. An abscess was considered to be above the level of the conus medullaris if the most caudal level of the abscess was above L1. If different organisms grew out of blood cultures and wound cultures, we deferred to wound culture data. An infectious diseases specialist reviewed all cultures containing organisms that were potentially contaminants to ensure that they were the likely or confirmed spinal epidural abscess pathogen.

Statistical Analysis

Categorical variables are provided with frequencies and percentages, and continuous variables are provided with medians and interquartile ranges (IQRs). For nomogram construction, we randomly selected 80% (294 patients) of the cohort of 367 patients to serve as our learning cohort and reserved the remaining 20% (73 patients) as a validation cohort for internal validation. Bivariate logistic regression was used to determine variables associated with failure of nonoperative management. Stepwise backward logistic regression on bootstrap samples of the learning cohort (100 replications, full sample, with replacement) was used to determine variables eligible for inclusion in the multivariable model. Minimum Akaike information criterion (AIC) fit values were used to select the optimum multivariable model. We constructed the nomogram to predict a binary outcome of nonoperative management failure using the mean β coefficients of each predictor, determined by bootstrap analysis of the learning cohort (1,000 replications, full sample, with replacement). Nonsignificant variables in the multivariable model were included to avoid overestimation of the significant variables and to preserve predictive accuracy[23-25]. Model discrimination and calibration were determined using the area under the receiver operating characteristic curve (AUC) and the Hosmer-Lemeshow test.

We performed internal validation using bootstrap analysis of the validation cohort with equally sized random samples of the learning cohort (1,000 replications, full sample, with replacement). Internal validation was achieved if the AUC and the regression coefficients for the validation sample fell within the 95% confidence interval (CI) of the primary sample.

Significance was set at $p < 0.05$. We used Stata (version 12 SE; StataCorp) for statistical analyses and nomogram construction.

Results

Demographic Characteristics

Our cohort of 367 patients had a median age of 59 years (IQR, 49 to 71 years) with 237 male patients (65%). The most common observed medical comorbidity was diabetes mellitus, in 82 patients (22%). Twenty-five patients (6.8%) had an active malignancy at the time of presentation. Seventeen patients (4.6%) had a pathologic or compression fracture in the affected area (Table I). The median WBC count was 10.4 × 10^3 cells/µL (IQR, 7.6 to 14.1 × 10^3 cells/µL). The median levels of inflammatory markers were elevated at 87 mm/hr (IQR, 53 to 106 mm/hr) for ESR and 100 mg/L (IQR, 32 to 164 mg/L) for CRP.

Presentation and Abscess Characteristics

On presentation, 353 patients (96%) had back pain, 83 patients (23%) were febrile, 303 patients (83%) had normal motor function, and 54 patients (15%) had a motor deficit. With respect to nonmotor neurologic symptoms, 43 patients (12%) had sensory changes and 23 (6.3%) had bowel and/or bladder dysfunction (Table I).

Abscesses spanned a median of 2 vertebral levels (IQR, 1 to 4 vertebral levels) and were most commonly located in the lumbar spine, with 135 lumbar abscesses (37%). There were 108 abscesses (29%) located above the conus medullaris. With respect to location within the spinal canal, 243 abscesses (66%) were located ventral to the thecal sac, 59 abscesses (16%) were exclusively dorsal abscesses, and 26 abscesses (7.1%) circumferentially surrounded the thecal sac. In this study, 213 patients (58%) had positive blood cultures, and 114 patients (31%) had cultures from samples retrieved using radiographic guidance. Ninety-two percent of cultures were obtained prior to the initiation of antibiotic therapy. The most common causative organism was methicillin-sensitive Staphylococcus aureus (MSSA), with 124 cases (34%), 9 patients (2.5%) had cultures that grew multiple organisms, and 84 patients (23%) had sterile cultures (Table I).

Failure of Nonoperative Management

There were 99 patients (27%) who underwent failed nonoperative management. Of those patients, 65 (66%) subsequently underwent a surgical procedure. The most common reason for a surgical procedure following failure was radiographic disease progression (46%), followed by neurologic deterioration (25%). Indications for a surgical procedure were persistent or worsening symptoms in 20% and progressive deformity or instability in 9.2% of those requiring a surgical procedure after treatment failure. The median time to failure was 25 days (IQR, 11 to 37 days) (Table II).

Bivariate and Multivariable Analysis

We performed bivariate logistic regression to assess the association between explanatory variables and failure of nonoperative management (Table III). Minimum AIC fit values were used to select the best model, with an AIC value of 275. Multivariable analysis using the model selected by minimum AIC fit values yielded 6 independent predictors of nonoperative management failure (Table IV). Motor deficit at presentation ($p < 0.001$), pathologic or compression fracture ($p = 0.003$), active malignancy ($p = 0.028$), diabetes mellitus ($p = 0.001$), and sensory changes ($p = 0.005$) were positive predictors of failure. Dorsal location of the abscess relative to the thecal sac was a negative predictor of failure ($p = 0.014$) (Table IV).

Nomogram

We generated a nomogram using these 6 independent predictors from the multivariable analysis. Each binary independent predictor is assigned a point value (Table VI). Although used to construct the nomogram, nonsignificant factors were not assigned point values. The points are summed and the total is converted to a probability of nonoperative management failure, calculated by the following algorithm:

The constant is −1.95 and the points coefficient is 0.21.

Internal Validation

The AUC for the primary sample was 0.82 (95% CI, 0.73 to 0.90), and the AUC for the validation sample was 0.82 (95% CI, 0.75 to 0.89). The Hosmer-Lemeshow test for goodness of fit was 0.3969. All regression coefficients of the bootstrap sample were within the 95% CI of the primary sample.

Discussion

To our knowledge, our cohort of 367 patients represents the largest series of medically managed patients with spinal epidural abscess in a single cohort. We collected data from 24 years of admissions at our hospital system, composed of 2 tertiary academic medical centers and 3 regional community hospitals.

This study had limitations. The first was its retrospective design. The second was that extenuating circumstances sometimes dictated medical management. For instance, if a patient declined a surgical procedure, the treatment team was forced to opt for nonoperative management. This may have introduced selection bias in which patients underwent nonoperative management. Finally, because most radiographic images were not available for review in our electronic medical record prior to 2007, abscess region and location relative to the thecal sac in these cases were based solely on radiology reports.

The mainstay of spinal epidural abscess treatment has long involved prompt surgical decompression with drainage of pus and/or debridement of infected granulation tissue followed by systemic antibiotics[2,5,7,9,26]. With advances in antibiotic therapy and the feasibility of following disease progression with serial MRI, nonoperative management of spinal epidural abscess has become a viable treatment option[4,10,12-19]. Nonetheless, data comparing nonoperative and operative management are not conclusive. A number of studies have compared them, with some advocating for surgical decompression, others advocating for nonoperative management, and still others claiming no difference between operative and nonoperative management[4,5,7,9,14,16,19,27–29.] Complicating the initial management decision is the dire prognosis of patients who undergo failed nonoperative management[5,20]. Failure rates ranged from 6% to 49% in a systematic review by Arko et al.[3]; in the current study, our rate of 27% is in line with this. It is essential that clinicians are cognizant of the risk factors for failure. In a rigorously performed analysis of 142 nonoperatively managed patients, Kim et al. identified 4 independent predictors of failure: age of >65 years, diabetes mellitus, methicillin-resistant Staphylococcus aureus (MRSA) infection, and neurologic deficit involving the spinal cord[21]. Patel et al.[20] also identified 4 independent predictors of failure: diabetes, WBC count > 12.5 × 10^3 cells/µL, positive blood cultures, and CRP > 115 mg/L.

Using a multivariable logistic regression model, we identified 6 predictors of failure of nonoperative management. Like Kim et al., we identified a pretreatment motor deficit as a risk factor for nonoperative failure. Neurologic status is a key prognostic factor in spinal epidural abscess, with poorer outcomes observed in patients who present with a motor deficit[1,6–9,21,30,31]. We also identified the presence of sensory changes as a risk factor for failure. Similar to motor weakness, the presence of either paresthesias or frank sensory deficit represents an advanced stage of disease with spinal cord injury[7].

Consistent with Patel et al.[20] and Kim et al.21, we found that diabetes is predictive of failure. Poor glycemic control has also been demonstrated to correlate with poor motor recovery after surgical treatment of spinal epidural abscess[32,33]. Diabetes may adversely affect outcomes by impairing immune response and diminishing spinal microvasculature integrity[29,31,34]. We also found that an active malignancy at presentation is a predictor of nonoperative management failure. Similar to diabetes, malignancy has an established immunosuppressive effect. Several tumor-derived factors inhibit dendritic cell maturation and T-cell activation[35], potentially complicating efforts to fully eradicate the infection.

Mechanical and anatomical factors have not been previously linked to failure of nonoperative management. A mechanical factor that we found to be predictive of failure is the presence of a local pathologic or compression fracture. Pathologic or compression fractures can cause local kyphotic deformity[36]. Focal kyphosis may reduce the size of the epidural space in that area, allowing for purulent expansion to more readily cause neurologic dysfunction. Furthermore, nonpenetrating trauma may cause local inflammation or hematoma that can serve as a nidus for infection[13,31,37,38].

The location of the abscess within the spinal canal is a significant predictor of treatment failure. Some authors have found that dorsally located abscesses are independently associated with poor prognosis[19,28], although others have found no impact of abscess location[21,39]. Contrary to these studies, we have found that an exclusively dorsal abscess is a negative predictor of failure. Disruption of the anterior spinal artery, the dominant spinal cord supply vessel, by a ventral abscess may cause cord ischemia and worsened disease[40]. An exclusively dorsal abscess may thus be protective against poor outcomes.

Once the diagnosis of spinal epidural abscess is confirmed, a pressing question that the clinician must answer is which treatment modality to pursue. Given the scarcity of data with regard to failure of medical management, it is difficult for clinicians to make a data-driven treatment decision. Using 6 independent risk factors of failure of nonoperative management, we have constructed a nomogram that generates an individualized probability of treatment failure for a given patient with spinal epidural abscess.

To illustrate the utility of the nomogram, we provide a patient example of how a treatment team could use it. The patient is a 68-year-old woman with metastatic breast cancer who presents with 1 week of mid-back pain. She has full strength in the upper and lower extremities bilaterally. MRI reveals a T8 to T10 circumferential abscess with a T9 compression fracture. The patient receives 8.8 points for a local compression fracture and 5.6 points for active malignancy. This gives a total of 14.4 points. Inserting this into the algorithm yields a 75% risk of nonoperative management failure. Even in the absence of a motor deficit, there may be a risk of failure in the presence of other independent risk factors. This is a notable finding, because the lack of a presenting motor deficit is often considered an indication for nonoperative management[2,8,12,41].

It should be noted that our analysis does not make any conclusions with regard to the efficacy of surgical management. Our study does not and cannot demonstrate that a surgical procedure would be more successful than nonoperative management in those patients found to have a high probability of treatment failure.

In conclusion, with a cohort of 367 patients with spinal epidural abscess, we identified 6 independent predictors of failure of nonoperative management. These factors include measures of the patient's general health and neurologic status at the time of presentation as well as radiographic data and local abscess anatomy. We included these factors in the construction of an algorithm that generates a patient-specific probability of treatment failure. By quantifying the risk of failure of nonoperative management based on the presence or absence of independent risk factors, we are confident that our nomogram will provide a useful tool for the treatment team when weighing nonoperative management for spinal epidural abscess.

References

1. Davis DP, Wold RM, Patel RJ, Tran AJ, Tokhi RN, Chan TC, Vilke GM. The clinical presentation and impact of diagnostic delays on emergency department patients with spinal epidural abscess. J Emerg Med. 2004 Apr;26(3):285-91.

2. Darouiche RO. Spinal epidural abscess. N Engl J Med. 2006 Nov 9;355(19):2012-20.

3. Arko L 4th, Quach E, Nguyen V, Chang D, Sukul V, Kim BS. Medical and surgical management of spinal epidural abscess: a systematic review. Neurosurg Focus. 2014 Aug;37(2):E4.

4. Connor DE Jr, Chittiboina P, Caldito G, Nanda A. Comparison of operative and nonoperative management of spinal epidural abscess: a retrospective review of clinical and laboratory predictors of neurological outcome. J Neurosurg Spine. 2013 Jul;19(1):119-27. Epub 2013 May 10.

5. Curry WT Jr, Hoh BL, Amin-Hanjani S, Eskandar EN. Spinal epidural abscess: clinical presentation, management, and outcome. Surg Neurol. 2005 Apr;63(4):364-71; discussion 371.

6. Danner RL, Hartman BJ. Update on spinal epidural abscess: 35 cases and review of the literature. Rev Infect Dis. 1987 Mar-Apr;9(2):265-74.

7. Darouiche RO, Hamill RJ, Greenberg SB, Weathers SW, Musher DM. Bacterial spinal epidural abscess. Review of 43 cases and literature survey. Medicine (Baltimore). 1992 Nov;71(6):369-85.

8. Hlavin ML, Kaminski HJ, Ross JS, Ganz E. Spinal epidural abscess: a ten-year perspective. Neurosurgery. 1990 Aug;27(2):177-84.

9. Rigamonti D, Liem L, Sampath P, Knoller N, Namaguchi Y, Schreibman DL, Sloan MA, Wolf A, Zeidman S. Spinal epidural abscess: contemporary trends in etiology, evaluation, and management. Surg Neurol. 1999 Aug;52(2):189-96; discussion 197.

10. Wheeler D, Keiser P, Rigamonti D, Keay S. Medical management of spinal epidural abscesses: case report and review. Clin Infect Dis. 1992 Jul;15(1):22-7.

11. Hanigan WC, Asner NG, Elwood PW. Magnetic resonance imaging and the nonoperative treatment of spinal epidural abscess. Surg Neurol. 1990 Dec;34(6):408-13.

12. Leys D, Lesoin F, Viaud C, Pasquier F, Rousseaux M, Jomin M, Petit H. Decreased morbidity from acute bacterial spinal epidural abscesses using computed tomography and nonsurgical treatment in selected patients. Ann Neurol. 1985 Apr;17(4):350-5.

13. Mampalam TJ, Rosegay H, Andrews BT, Rosenblum ML, Pitts LH. Nonoperative treatment of spinal epidural infections. J Neurosurg. 1989 Aug;71(2):208-10.

14. Savage K, Holtom PD, Zalavras CG. Spinal epidural abscess: early clinical outcome in patients treated medically. Clin Orthop Relat Res. 2005 Oct;439:56-60.

15. Tang HJ, Lin HJ, Liu YC, Li CM. Spinal epidural abscess—experience with 46 patients and evaluation of prognostic factors. J Infect. 2002 Aug;45(2):76-81.

16. Siddiq F, Chowfin A, Tight R, Sahmoun AE, Smego RA Jr. Medical vs surgical management of spinal epidural abscess. Arch Intern Med. 2004 Dec 13-27;164(22):2409-12.

17. Bamberger DM. Outcome of medical treatment of bacterial abscesses without therapeutic drainage: review of cases reported in the literature. Clin Infect Dis. 1996 Sep;23(3):592-603.

18. Rust TM, Kohan S, Steel T, Lonergan R. CT guided aspiration of a cervical spinal epidural abscess. J Clin Neurosci. 2005 May;12(4):453-6.

19. Adogwa O, Karikari IO, Carr KR, Krucoff M, Ajay D, Fatemi P, Perez EL, Cheng JS, Bagley CA, Isaacs RE. Spontaneous spinal epidural abscess in patients 50 years of age and older: a 15-year institutional perspective and review of the literature: clinical article. J Neurosurg Spine. 2014 Mar;20(3):344-9. Epub 2013 Dec 20.

20. Patel AR, Alton TB, Bransford RJ, Lee MJ, Bellabarba CB, Chapman JR. Spinal epidural abscesses: risk factors, medical versus surgical management, a retrospective review of 128 cases. Spine J. 2014 Feb 1;14(2):326-30. Epub 2013 Nov 12.

21. Kim SD, Melikian R, Ju KL, Zurakowski D, Wood KB, Bono CM, Harris MB. Independent predictors of failure of nonoperative management of spinal epidural abscesses. Spine J. 2014 Aug 1;14(8):1673-9. Epub 2013 Oct 30.

22. Kirshblum SC, Burns SP, Biering-Sorensen F, Donovan W, Graves DE, Jha A, Johansen M, Jones L, Krassioukov A, Mulcahey MJ, Schmidt-Read M, Waring W. International standards for neurological classification of spinal cord injury (revised 2011). J Spinal Cord Med. 2011 Nov;34(6):535-46.

23. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996 Feb 28;15(4):361-87.

24. Kattan MW, Reuter V, Motzer RJ, Katz J, Russo P. A postoperative prognostic nomogram for renal cell carcinoma. J Urol. 2001 Jul;166(1):63-7.

25. Sorbellini M, Kattan MW, Snyder ME, Reuter V, Motzer R, Goetzl M, McKiernan J, Russo P. A postoperative prognostic nomogram predicting recurrence for patients with conventional clear cell renal cell carcinoma. J Urol. 2005 Jan;173(1):48-51.

26. Khanna RK, Malik GM, Rock JP, Rosenblum ML. Spinal epidural abscess: evaluation of factors influencing outcome. Neurosurgery. 1996 Nov;39(5):958-64.

27. Alton TB, Patel AR, Bransford RJ, Bellabarba C, Lee MJ, Chapman JR. Is there a difference in neurologic outcome in medical versus early operative management of cervical epidural abscesses? Spine J. 2015 Jan 1;15(1):10-7. Epub 2014 Jun 14.

28. Karikari IO, Powers CJ, Reynolds RM, Mehta AI, Isaacs RE. Management of a spontaneous spinal epidural abscess: a single-center 10-year experience. Neurosurgery. 2009 Nov;65(5):919-23; discussion 923–4.

29. Pradilla G, Ardila GP, Hsu W, Rigamonti D. Epidural abscesses of the CNS. Lancet Neurol. 2009 Mar;8(3):292-300.

30. Lu CH, Chang WN, Lui CC, Lee PY, Chang HW. Adult spinal epidural abscess: clinical features and prognostic factors. Clin Neurol Neurosurg. 2002 Sep;104(4):306-10.

31. Reihsaus E, Waldbaur H, Seeling W. Spinal epidural abscess: a meta-analysis of 915 patients. Neurosurg Rev. 2000 Dec;23(4):175-204; discussion 205.

32. Wang TC, Lu MS, Yang JT, Weng HH, Cheng YK, Lin MH, Su CH, Lee MH. Motor function improvement in patients undergoing surgery for spinal epidural abscess. Neurosurgery. 2010 May;66(5):910-6.

33. Huang PY, Chen SF, Chang WN, Lu CH, Chuang YC, Tsai NW, Chang CC, Wang HC, Chien CC, Chen SH, Huang CR. Spinal epidural abscess in adults caused by Staphylococcus aureus: clinical characteristics and prognostic factors. Clin Neurol Neurosurg. 2012 Jul;114(6):572-6. Epub 2011 Dec 27.

34. Broner FA, Garland DE, Zigler JE. Spinal infections in the immunocompromised host. Orthop Clin North Am. 1996 Jan;27(1):37-46.

35. Kim R, Emi M, Tanabe K. Cancer immunosuppression and autoimmune disease: beyond immunosuppressive networks for tumour immunity. Immunology. 2006 Oct;119(2):254-64.

36. Pradhan BB, Bae HW, Kropf MA, Patel VV, Delamarter RB. Kyphoplasty reduction of osteoporotic vertebral compression fractures: correction of local kyphosis versus overall sagittal alignment. Spine (Phila Pa 1976). 2006 Feb 15;31(4):435-41.

37. Korovessis P, Sidiropoulos P, Piperos G, Karagiannis A. Spinal epidural abscess complicated closed vertebral fracture. A case report and review of literature. Spine (Phila Pa 1976). 1993 Apr;18(5):671-4.

38. Nussbaum ES, Rigamonti D, Standiford H, Numaguchi Y, Wolf AL, Robinson WL. Spinal epidural abscess: a report of 40 cases and review. Surg Neurol. 1992 Sep;38(3):225-31.

39. Soehle M, Wallenfang T. Spinal epidural abscesses: clinical manifestations, prognostic factors, and outcomes. Neurosurgery. 2002 Jul;51(1):79-85; discussion 86–7.

40. Colman MW, Hornicek FJ, Schwab JH. Spinal cord blood supply and its surgical implications. J Am Acad Orthop Surg. 2015 Oct;23(10):581-91. Epub 2015 Sep 16.

41. Harrington P, Millner PA, Veale D. Inappropriate medical management of spinal epidural abscess. Ann Rheum Dis. 2001 Mar;60(3):218-22.

Table 1. Observational data

Variable	**All Patients (n = 472)**
Demographics	
	Median (IQR)
Age (years)	59 (49 - 71)
	Number (%)
Male	302 (64)
Body mass index (in kg/m^2)†	
< 18.5	6 (1.3)
18.5 - 30	127 (27)
> 30	55 (12)
Habits	
Smoking	227 (48)
Intravenous drug use	89 (19)
Alcohol use	68 (14)
Medical comorbidities	
Diabetes mellitus	109 (23)
Immunocompromised	76 (16)
Active malignancy	42 (8.9)
Hemodialysis	29 (6.1)
HIV positive	15 (3.2)
Spinal instrumentation in place	27 (5.7)
Spinal procedure within 5 years prior to presentation	74 (16)
Spinal trauma within 5 years prior to presentation	
Mechanical injury with no fracture	26 (5.5)
Pathologic/compression fracture	22 (4.7)
Mechanical fracture	11 (2.3)
	Median (IQR)
Laboratory values†	
White blood cell count (10/μL)	10.4 (7.6 - 14.3)
Erythrocyte sedimentation rate (mm/h)	87.5 (54.5 - 106)
C-reactive protein (mg/L)	104 (34.4 - 165)
Number of affected levels	2 (1 - 4)
Hospitalization duration (days)	11 (7 - 19)
Presentation	
Back pain	450 (95)
Fever	108 (23)
Motor function†	
Normal (ASIA E)	386 (82)
Incomplete injury (ASIA B, C, D)	60 (13)

Complete injury (ASIA A)	7 (1.5)
Sedated/existing deficit	18 (3.8)
Non-motor neurologic symptoms	170 (36)
Radicular pain	122 (26)
Sensory changes	62 (13)
Urinary incontinence/retention	16 (3.4)
Fecal incontinence/retention	10 (2.1)
Symptom duration prior to presentation†	
≤24 hours	33 (7.0)
24 - 72 hours	54 (11)
72 hours - 2 weeks	209 (44)
>2 weeks	176 (37)
Bacteremia	284 (60)
Fungemia	2 (0.42)
Abscess characteristics	
Region of spine	
Cervical	46 (9.8)
Cervicothoracic	17 (3.6)
Thoracic	83 (18)
Thoracolumbar	35 (7.4)
Lumbar	168 (36)
Lumbosacral	98 (21)
Sacral	2 (0.42)
Multifocal/non-contiguous	18 (3.8)
>2 contiguous regions	5 (1.1)
Above conus medullaris	142 (30)
Location of abscess relative to spinal cord†	
Anterior	303 (64)
Posterior	83 (18)
Circumferential	37 (7.8)
Multiple locations	45 (9.5)
Ventral component to abscess	385 (82)
Organism	
No growth	112 (24)
Methicillin-sensitive staphylococcus aureus	157 (33)
Methicillin-resistant staphylococcous aureus	59 (13)
Streptococci	48 (10)
Coagulase-negative staphylococci	23 (4.9)
Multiple organisms	15 (3.2)
Escherichia coli	14 (3.0)

Mycobacteria	13 (2.8)
Enterococcus	9 (1.9)
Anaerobe	4 (0.85)
Candida	3 (0.64)
Pseudomonas aeruginosa	2 (0.42)
Other	13 (2.8)
Local spinal infections	
Spondylodiscitis	264 (56)
Psoas/paraspinal abscess	234 (50)
Vertebral osteomyelitis	61 (13)
Prevertebral abscess/retropharyngeal abscess	41 (8.7)
Discitis	24 (5.1)
Wound infection	20 (4.2)
Local non-spinal infections	
Infectious endocarditis	31 (6.6)
Non-spinal abscess/cellulitis	29 (6.1)
Septic arthritis	25 (5.3)
Pneumonia/empyema	18 (3.8)
Meningitis	9 (1.9)
Non-vertebral osteomyelitis	7 (1.5)
Other	17 (3.6)
	Number (%)
Outcomes	
Failure of non-operative management	99 (21)
Treatment after non-operative failure	
Operative	64 (65)
Non-operative	35 (35)
	Median (IQR)
Follow-up (weeks)	22 (7 - 81)
Time to failure (days)	21 (10 - 36)

IQR = Interquartile range; mg/L = milligrams per liter; µL = microliter; mm/h = millimeters per hour; kg/m2 = kilogram per square meter; L = liter

† Body mass index was available in 188 cases (40%), ASIA scores were available in 471 cases (99.8%), location of abscess relative to the spinal cord was available in 468 cases (99%), erythryocyte sedimentation rate was available in 396 cases (84%), C-reactive protein was available in 310 cases (66%), ASIA scores could be compared in 449 cases (95%).

Table 2. Bivariate logistic regression assessing risk factors for failure of non-operative management

Explanatory variables (n = 424)	Odds Ratio (95% CI)	Standard Error	*p* value
Demographics			
Age (years)	1.01 (1.00 - 1.03)	0.01	0.155
Male	0.77 (0.48 - 1.21)	0.18	0.256
Body mass index (in kg/m^2)†			
< 18.5	2.68 (0.46 - 15.6)	2.41	0.274
18.5 - 30	Reference value		
> 30	0.78 (0.31 - 1.97)	0.37	0.599
Habits			
Smoking	0.98 (0.62 - 1.54)	0.23	0.927
Intravenous drug use	0.82 (0.45 - 1.49)	0.25	0.513
Alcohol use	2.16 (1.21 - 3.86)	0.64	**0.010**
Medical comorbidities			
Diabetes mellitus	2.58 (1.58 - 4.23)	0.65	**<0.001**
Immunocompromised	1.52 (0.86 - 2.68)	0.44	0.152
Active malignancy	1.64 (0.77 - 3.50)	0.63	0.199
Hemodialysis	1.30 (0.53 - 3.21)	0.60	0.572
HIV positive	1.33 (0.41 - 4.33)	0.80	0.640
Spinal instrumentation in place	1.47 (0.59 - 3.68)	0.69	0.411
Spinal procedure within 5 years prior to presentation	1.68 (0.4 - 2.98)	0.49	**0.079**
Spinal trauma within 5 years prior to presentation			
Mechanical injury without fracture	0.91 (0.33 - 2.51)	0.47	0.851
Pathologic/compression fracture	3.95 (1.56 - 10.0)	1.87	**0.004**
Mechanical fracture	1.10 (0.22 - 5.52)	0.90	0.911

Presentation			
Back pain	1.66 (0.47 - 5.81)	1.06	0.430
Fever	0.67 (0.38 - 1.18)	0.19	0.167
Motor deficit at presentation	8.64 (4.74 - 15.8)	2.65	**<0.001**
Non-motor neurologic symptoms			
Radicular pain	1.25 (0.76 - 2.06)	0.32	0.386
Sensory changes	3.34 (1.81 - 6.15)	1.04	**<0.001**
Urinary incontinence/retention	9.02 (2.76 - 29.4)	5.44	**<0.001**
Fecal incontinence/retention	24.7 (2.99 - 203)	26.5	**0.003**
Symptom duration prior to presentation†			
<24 hours	Reference value		
24 - 72 hours	1.06 (0.34 - 3.27)	0.61	0.917
72 hours - 2 weeks	0.79 (0.30 - 2.09)	0.39	0.629
>2 weeks	1.62 (0.62 - 4.24)	0.80	0.327
Bacteremia	0.86 (0.54 - 1.36)	0.20	0.515
Fungemia	1		
Abscess characteristics			
Region of spine			
Cervical	0.33 (0.11 - 0.94)	0.18	**0.039**
Cervicothoracic	2.27 (0.79 - 6.53)	1.22	0.130
Thoracic	1.61 (0.93 - 2.77)	0.45	**0.088**
Thoracolumbar	1.21 (0.52 -2 .81)	0.52	0.656
Lumbar	0.77 (0.48 - 1.24)	0.19	0.283
Lumbosacral	1.17 (0.67 - 2.01)	0.33	0.584
Sacral	1		

Multifocal/non-contiguous	0.82 (0.23 - 2.95)	0.53	0.755
>2 contiguous regions	0.82 (0.09 - 7.41)	0.92	0.859
Above conus medullaris	1.10 (0.68 - 1.79)	0.27	0.695
Location of abscess relative to spinal cord†			
Anterior	Reference value		
Posterior	0.44 (0.21 - 0.93)	0.17	**0.031**
Circumferential	1.51 (0.70 - 3.26)	0.59	0.296
Multiple locations	1.37 (0.14 - 17.7)	1.94	0.712
Ventral component to abscess	2.31 (1.14 - 4.69)	0.83	**0.020**
Organism			
No growth	0.82 (0.47 - 1.42)	0.23	0.470
Methicillin-sensitive staphylococcus aureus	0.78 (0.48 - 1.27)	0.19	0.313
Methicillin-resistant staphylococcus aureus	0.96 (0.48 - 1.90)	0.33	0.896
Streptococci	0.93 (0.44 - 1.95)	0.35	0.850
Coagulase negative staphylococci	2.85 (1.15 - 7.10)	1.33	**0.024**
Multiple organisms	1.87 (0.61 - 5.71)	1.06	0.273
Escherichia coli	0.65 (0.14 - 3.01)	0.51	0.582
Mycobacteria	1.10 (0.29 - 4.13)	0.74	0.891
Enterococcus	1.10 (0.22 - 5.52)	0.9	0.911
Anaerobe	1.10 (0.11 - 10.6)	1.27	0.938
Candida	1.65 (0.15 - 18.4)	2.03	0.685
Pseudomonas aeruginosa	1		
Other	1.10 (0.29 - 4.13)	0.74	0.891
Local spinal infections			
Spondylodiscitis	1.31 (0.83 - 2.07)	0.31	0.251
Psoas/paraspinal abscess	1.24 (0.79 - 1.94)	0.29	0.355

Vertebral osteomyelitis	1.05 (0.54 - 2.04)	0.36	0.893
Prevertebral abscess/retropharyngeal abscess	0.67 (0.29 - 1.57)	0.29	0.361
Discitis	1.69 (0.66 - 4.31)	0.81	0.272
Wound infection	1.68 (0.61 - 4.61)	0.86	0.311
Local non-spinal infections			
Infectious endocarditis	0.36 (0.11 - 1.21)	0.22	**0.099**
Non-spinal abscess/cellulitis	0.93 (0.37 - 2.38)	0.45	0.886
Septic arthritis	1.43 (0.54 - 3.83)	0.72	0.473
Pneumonia/empyema	0.65 (0.18 - 2.28)	0.42	0.497
Meningitis	0.54 (0.06 - 4.56)	0.59	0.573
Non-vertebral osteomyelitis	1		
Other	0.75 (0.21 - 2.69)	0.49	0.659
White blood cell count (10/µL)	0.99 (0.95 - 1.02)	0.02	0.432
Erythrocyte sedimentation rate (mm/h)	1.01 (1.00 - 1.01)	0.003	0.106
C-reactive protein (mg/L)	1.00 (1.00 - 1.00)	0.002	0.259
Number of affected levels	1.01 (0.93 - 1.09)	0.04	0.904

Bold indicates significance (P value < 0.05). CI = Confidence Interval; mg/L = milligrams per liter; µL = microliter; mm = millimeter; mm/h = millimeters per hour; kg/m2 = kilogram per square meter

† Body mass index was available in 175 cases (41%), ASIA scores were available in 423 cases (99.8%), location of abscess relative to the spinal cord was available in 421 cases (99%), erythrocyte sedimentation rate was available in 359 cases (85%), C-reactive protein was available in 285 cases (67%).

Chapter 4

Machine Learning Algorithm Predicting Discharge Placement after Elective Surgery for Lumbar Spinal Stenosis

Ogink PT, Karhade AV, Thio QCBS, Gormley WB, Oner FC, Verlaan JJ, Schwab JH

Eur Spine J. 2019 Jun;28(6):1433-1440

Abstract

Purpose An excessive amount of total hospitalization is caused by delays due to patients waiting to be placed in a rehabilitation facility or skilled nursing facility (RF/SNF). An accurate preoperative prediction of who would need a RF/SNF place after surgery could reduce costs and allow more efficient organizational planning. We aimed to develop a machine learning algorithm that predicts non-home discharge after elective surgery for lumbar spinal stenosis.

Methods We used the American College of Surgeons National Surgical Quality Improvement Program (ACS - NSQIP) to select patient that underwent elective surgery for lumbar spinal stenosis between 2009 and 2016. The primary outcome measure for the algorithm was non-home discharge. Four machine-learning algorithms were developed to predict non-home discharge. Performance of the algorithms was measured with discrimination, calibration, and an overall performance score.

Results We included 28,600 patients with a median age of 67 (Interquartile range [IQR] 58 – 74). The non-home discharge rate was 18.2%. Our final model consisted of the following variables: age, sex, body mass index, diabetes, functional status, ASA class, level, fusion, preoperative hematocrit, and preoperative serum creatinine. The Neural Network was the best model based on discrimination (c-statistic = 0.751), calibration (slope = 0.933; intercept = 0.037) , and overall performance (Brier score = 0.131).

Conclusions A machine learning algorithm is able to predict discharge placement after surgery for lumbar spinal stenosis with both good discrimination and calibration. Implementing this type of algorithm in clinical practice could avert risks associated with delayed discharge and lower costs.

Introduction

In recent years there has been a trend towards quicker discharges after orthopedic surgery, which does not seem to affect patients' outcomes inordinately.[1,2] However, an excessive amount of total hospitalization is caused by delays due to patients waiting to be placed in a rehabilitation facility or skilled nursing facility (RF/SNF).[3–8] Not only does this incur unnecessary costs and hamper efficient delivery of care, but more importantly delayed discharges are detrimental to the patient.[9] Increased length of stay has been associated with hospital acquired infections and adverse drug events.[7,9–11] Although increasing the number of facilities seems the obvious solution, a study by Gaughan et al.[12] found that this would only have a small effect on delayed discharges and would actually cost more.

Previous studies have determined risk factors for non-home discharge placement. Some have developed scoring systems based on these risk factors aiming to predict who will likely not be discharged home after spine surgery.[13–16] However, no studies have looked at employing machine learning (ML) algorithms. The increased amounts of available data combined with more computational hardware is currently causing a rapid expansion of ML in medicine. ML is a form of artificial intelligence which allows algorithms to learn and self-improve from experience without explicit programming by a data scientist. The capacity of these algorithms to handle large datasets and incorporate nonlinear interactions allows for more accurate and personalized prediction than regular statistical methods.

An accurate personal preoperative prediction of who would need a RF/SNF place would allow reservation of a place in advance and earlier insurance precertification. This could reduce costs and avoid the risks of (unnecessary) prolonged hospitalization.

Lumbar spinal stenosis is a relatively common degenerative spine condition for which the SPORT trial has indicated surgical treatment to be superior to nonsurgical treatment. Currently, it is one of the most common indications for spine surgery.[17,18]

In this study, we aim to develop a prediction tool using ML algorithms to predict discharge to a RF/SNF after elective surgery for lumbar spinal stenosis for patients living at home preoperatively. Second, we aim to select the best performing algorithm and develop an application to enable healthcare providers to arrange a place in a RF/SNF well in advance.

Methods

Data Source

We used the American College of Surgeons National Surgical Quality Improvement Program (ACS - NSQIP) as our main data source. The ACS-NSQIP is a large clinical database with data of more than 680 US hospitals combined and has often been used in the spine literature.

We included patients based on the following criteria: 1) International Classification of Disease Ninth Revision (ICD-9) code 724.02 or 724.03 for lumbar spinal stenosis; 2) year of surgery between 2009 and 2016; 3) Current Procedural Terminology (CPT) codes for decompression, fusion, or fixation. We included 28,600 patients in our dataset to train and test the algorithms.

Data Analysis

Our primary outcome measure was non-home discharge defined as all discharges not to home. This variable was created by grouping together discharges to rehabilitation facilities, skilled nursing facilities, and unskilled nursing facilities. Variable selection for our algorithm was performed by entering all available variables in a Random Forest regression, which then ranks variables according to their predictive power for the outcome variable.[19]

We performed a stratified 80:20 split of the dataset into a training set and a testing set. We used the training set for algorithm training and assessment of performance by tenfold cross validation. Cross validation means dividing the data into a selected number of groups, named *folds*. Each *fold* is withheld once and treated as the test set while the other *folds* together are treated as the training set. Results are subsequently averaged across all repetitions of this sequence.[20]

Four different algorithms (Neural Network, Support Vector Machine, Bayes Point Machine, Boosted Decision Tree) were trained using these variables to predict non-home discharge. We choose these four because they each have different merits for prediction (Appendix 1). Senders et al.[20] provides an accessible overview of the most commonly used algorithms. The model with the best performance was subsequently used in the testing set to predict discharge placement. These predictions were then compared with the actual outcomes of the testing set to assess the performance of the algorithm outside of the training set.

Model Assessment

Performance of the algorithms was measured with discrimination, calibration, and an overall performance score.[21,22]

Discrimination is the algorithm's ability to distinguish patients who were discharged to home from patients who were not discharged to home. We assessed discrimination with receiver-operating curves (ROC) and c-statistics. A c-statistic of 1.0 indicates perfect discrimination while a c-statistic of 0.5 indicates discrimination similar to chance.[23]

Calibration determines if the predicted probabilities of the algorithm are similar to the actual observed events. The calibration intercept determines whether the algorithm is over- or underestimating the probabilities; the calibration slope determines if the predictor effects are similar in the training and the testing set. A perfect model has an intercept score of 0.0 and a slope score of 1.0.

Overall model performance was assessed with the Brier score, calculated by obtaining the mean squared error between the observed events in the testing set and the predictions given by the algorithm. A perfect algorithm would have a score of 0. The Brier score combines discrimination and calibration characteristics, but must always be interpreted in the context of the prevalence of the predicted outcome – in our study non-home discharge.[22] If the prevalence of the outcome variable is lower, the maximum score of a poor algorithm is lower as well. Therefore, the Brier score must be compared with the *null* Brier score, which is calculated by assigning each patient a probability equivalent to the prevalence of non-home discharge. Steyerberg et al.[22] offers a detailed framework of all performance metrics.

Web-based Application

The algorithm with the best performance based on discrimination, calibration, and overall performance was subsequently incorporated in a web-based application. This application is built to input the variable values collected by a healthcare provider into the algorithm, calculate the probability, and output the result to the healthcare provider in real-time.

Microsoft Azure, STATA 13 (StataCorp LP, College Station, TX, USA), RStudio version 1.0.153, and Python version 3.6 (Python Software Foundation) (Anaconda distribution) were used for data analysis, model creation, and application development.

Results

Of the 28,600 patients 18.2% were not discharged to home. Baseline characteristics are shown in Table 1. The following variables were included after variable selection: age (years), sex (male/female), body mass index (BMI), American Society of Anesthesiologists (ASA) class (I/II/III/IV), functional status (independent/dependent), number of levels included in surgery (1 or 2 levels/3 or more levels), fusion (yes/no),), diabetes (no/oral medication/insulin-dependent), preoperative hematocrit (vol%), and preoperative serum creatinine (mg/dL).

Table 2 lists the AUC, calibration slope and intercept, and Brier score for the four algorithms. The null Brier score was 0.150. Based on numerical and graphical assessment of these metrics the Neural Network algorithm was chosen as the final model with a c-statistic of 0.751, a calibration slope of 0.933, a calibration intercept of 0.037, and a Brier score of 0.130 (Figure 1).

When evaluating the Neural Network algorithm on the testing set a c-statistic of 0.744, a calibration slope of 0.915, a calibration intercept of -0.131, and a Brier score of 0.131 were achieved (Figure 2 and 3).

The web application based on the Neural Network can be accessed at https://sorg-apps.shinyapps.io/stenosisdisposition/. As an example, a 75-year-old male is scheduled for two level surgery with fusion. He has a BMI of 34 and is classified as ASA II; he lives independently at home and does not have diabetes. His preoperative creatinine level is 2.9 mg/dL and preoperative hematocrit level is 34%. After filling out these values in the algorithm, this patient has a 24.4% chance of non-home discharge.

Discussion

We aimed to develop an ML algorithm that can predict discharge to a RF/SNF after elective surgery for lumbar stenosis. Our algorithm included age, sex, BMI, functional status, ASA class, level, fusion, diabetes, preoperative hematocrit, and preoperative serum creatinine. The Neural Network was picked as the best algorithm based on discrimination (c-statistic = 0.752), calibration (intercept = -1.27 x 10^{-5}; slope = 0.996), and overall performance (Brier score = 0.1257) in the training set and subsequent performance on internal validation.

Our study has several limitations. First, studies using a large clinical database are always affected by miscoding and other inaccuracies. Although widely used, few studies have assessed the actual accuracy of the NSQIP database. Rolston et al.[24] found many internal inconsistancies between procedure CPT codes and postoperative ICD-9 codes in neurosurgery. However, the codes for lumbar stenosis and lumbar surgery are more straightforward so we estimate that potential miscoding will not severely affect our algorithm. Second, certain variables of interest are not always available in the ACS-NSQIP. Considering preoperative patient reported outcomes are known to be predictors of discharge placement after spine surgery, we consider this a major limitation of our work.[25] While the current AUC of 0.751 is fair, the algorithm could potentially be improved by adding these and other relevant variables. Third, although the ACS-NSQIP database consists of data of 680 US hospitals, these results may not be applicable to all the patients it is intended for due to differences in demographic or clinical characteristics. Fourth, the differences between the algorithms are small, which makes the choice for a neural network somewhat arbitrary. However, settling on an algorithm based on numerical and graphical assessment is the most reproducible method. Finally, it must be emphasized that this study focuses on accurate prediction of a, rather simple, prespecified outcome (here 'non-home discharge') in contrast to the explanation of this outcome, which is the focus of the vast majority of medical research. The variables in our model cannot simply be interpreted as independent explanatory variables.

Age, sex, diabetes, functional status, fusion, and preoperative hematocrit have previously been identified in other (explanatory) studies on discharge placement after spine surgery.[26–28] The inclusion of most variables in our model can likely be attributed to being independent risk factors for major complications after surgery for lumbar stenosis. Age, diabetes, BMI, functional status, ASA class, preoperative hematocrit, and preoperative creatinine have all been shown to be associated with major complications.[29–32] Number of levels and fusion are likely surrogates for longer procedural time which is also implicated in postoperative morbidity.[30,31]

The importance of eliminating delayed discharges for patients lies in averting the risks associated with longer hospitalization and the advantages of starting rehabilitation earlier. Umarji et al. found that 58% of patients with a hip fracture acquire nosocomial infections when discharge was delayed beyond 8 days.[11] Hauck et al. found that each additional night in hospital increases the risk by 0.5% for adverse drug events and 1.6% for infections.[33] With regard to rehabilitation, other studies have found worse post-rehabilitation scores for patients with delays in discharge.[34,35] While those studies did not necessarily focus on elective spine patients, other spine centers have acknowledged the problem and aimed to contruct risk scores for predicting discharge placement. McGirt et al.[15] created the Carolina-Semmes Grading score for all degenerative lumbar spine surgery based on logistic regression. They included the variables age, ASA class, fusion, Oswestry Disbility Index score, ambulation, and nonprivate insurance and achieved an area under the curve (AUC) of 0.731. Kanaan et al.[14] used age, prior level of function, and gait distance to create a model for discharge placement after lumbar laminectomy and achieved an AUC of 0.80. Slover et al.[13] stratified spine patients in low, medium, and high risk based on points for age, sex, walking distance, gait aid, community support, and availability of caregiver at home. They did not report an AUC. None of the abovementioned studies assessed calibration. Although often-overlooked, assessment of calibration is an essential feature of studies creating prediction models. In our study the Neural Network and the Bayes Point Machine had highly similar performance metrics. However, on graphical assessment the calibration of the Bayes Point Machine was slightly inaccurate between the predicted probabilities of 0.15 and 0.50, which represent a significant part of the study population (Fig. 3). This deviation means the algorithm slightly underestimates the chance of discharge to a RF/SNF, which for some patients would mean no placement has been arranged before surgery- the situation as it is right now. Assessing calibration over the full range of predictions is crucial in ensuring the model is useful.[23] Future studies aiming to create models should always feature a numerical and graphical assessment of calibration. As depicted in the calibration subplot in Figure 3, the vast majority of patients has a 10% to 40% chance of discharge to an RF/SNF, as can be expected for an elective spine procedure. The algorithm is meant to trace and designate higher risk patients so their potential discharge delay might be avoided.

Where hospitals set their threshold to arrange an RF/SNF placement in advance would differ per health system. There are major differences in the availability of RF/SNF beds, insurance regulations, and discharge practices between countries.[36–38] Length of stay for deforming dorsopathies ranges from 4.6 to 27 within Europe. American patients are 3 times more likely to be discharged to RF/SNF than Canadian patients with a hip fracture.[39] While these complex differences do exist, delayed discharges are a problem for patients and hospitals around the world.[9] Mirroring the differences between countries a wide variety of policies have been implemented internationally to try to lower amount and duration of these delayed discharges.[40,41] In Great Britain imposing fines has reduced the number of delayed discharges, but simultaneously rising readmission rates brought up questions about the quality of discharges.[42] Sweden tried making local municipalities financially responsible for the care of elderly.[43] Others focused on developing allocation decision tools or the effect of increasing nursing home supply.[12,44]

At the very core of all these suggested policies, regardless of health system, is the inability to make an accurate assessment of who will need a RF/SNF placement with enough time to set things in motion. An ML algorithm can give an individualized prediction. Thorough external validation needs to be performed along with an assessment of where to place the threshold before these algorithms can be implemented, especially if the algorithm were to be used outside the US.

Nevertheless, considering the risks for patients and the unnecessary costs involved with longer hospitalization due to delayed discharges, the use of predictive algorithms could be worth the initial effort.

Conclusion

A prediction tool based on an ML algorithm is able to predict discharge placement after surgery for lumbar spinal stenosis with both good discrimination and calibration. This methodology can be implemented for a variety of other diseases and elective treatments, which could avoid risks associated with delayed discharge and lower costs.

References

1. Regenbogen SE, Cain-Nielsen AH, Norton EC, et al. Costs and consequences of early hospital discharge after major inpatient surgery in older adults. *JAMA Surg.* 2017;152(5):e170123.

2. Basques BA, Tetreault MW, Della Valle CJ. Same-Day Discharge Compared with Inpatient Hospitalization Following Hip and Knee Arthroplasty. *J. Bone Joint Surg. Am.* 2017;99(23):1969–1977.

3. Hwabejire JO, Kaafarani HMA, Imam AM, et al. Excessively long hospital stays after trauma are not related to the severity of illness: Let's aim to the right target! *JAMA Surg.* 2013;148(10):956–961.

4. Watkins JR, Soto JR, Bankhead-Kendall B, et al. What's the hold up? Factors contributing to delays in discharge of trauma patients after medical clearance. *Am. J. Surg.* 2014;208(6):969–973.

5. Costa AP, Poss JW, Peirce T, et al. Acute care inpatients with long-term delayed discharge: evidence from a {Canadian} health region. *BMC Health Serv. Res.* 2012;12(In press):6–11.

6. Smith AL, Kulhari A, Wolfram JA, et al. Impact of Insurance Precertification on Discharge of Stroke Patients to Acute Rehabilitation or Skilled Nursing Facility. *J. Stroke Cerebrovasc. Dis.* 2017;26(4):711–716.

7. Rosman M, Rachminov O, Segal O, et al. Prolonged patients' In-Hospital Waiting Period after discharge eligibility is associated with increased risk of infection, morbidity and mortality: A retrospective cohort analysis. *BMC Health Serv. Res.* 2015;15(1):1–5.

8. New PW, Andrianopoulos N, Cameron PA, et al. Reducing the length of stay for acute hospital patients needing admission into inpatient rehabilitation: a multicentre study of process barriers. *Intern. Med. J.* 2013;43(9):1005–11.

9. Rojas-García A, Turner S, Pizzo E, et al. Impact and experiences of delayed discharge: A mixed-studies systematic review. *Heal. Expect.* 2018;21(1):41–56.

10. Härkänen M, Kervinen M, Ahonen J, et al. Patient-specific risk factors of adverse drug events in adult inpatients - evidence detected using the Global Trigger Tool method. *J. Clin. Nurs.*

2015;24(3–4):582–591.

11. Umarji SIM, Lankester BJA, Prothero D, et al. Recovery after hip fracture. *Injury.* 2006;37(8):712–717.

12. Gaughan, James; Gravelle Hugh; Siciliani L. Testing the bed-blocking hypothesis: does nursing and care home supply reduce delayed hospital discharges? *Health Econ.* 2015;24:32–44.

13. Slover J, Mullaly K, Karia R, et al. The use of the Risk Assessment and Prediction Tool in surgical patients in a bundled payment program. *Int. J. Surg.* 2017;38:119–122.

14. Kanaan SF, Yeh H-W, Waitman RL, et al. Predicting discharge placement and health care needs after lumbar spine laminectomy. *J. Allied Health.* 2014;43(2):88–97.

15. McGirt MJ, Parker SL, Chotai S, et al. Predictors of extended length of stay, discharge to inpatient rehab, and hospital readmission following elective lumbar spine surgery: introduction of the Carolina-Semmes Grading Scale. *J. Neurosurg. Spine.* 2017;27(4):382–390.

16. Niedermeier S, Przybylowicz R, Virk SS, et al. Predictors of discharge to an inpatient rehabilitation facility after a single-level posterior spinal fusion procedure. *Eur. Spine J.* 2017;26(3):771–776.

17. Weinstein JJN, Tosteson TTD, Lurie JD, et al. Surgical versus Nonsurgical Therapy for Lumbar Spinal Stenosis. *N. Engl. J. Med.* 2008;358(8):794–810.

18. Weinstein JN, Tosteson TD, Lurie JD, et al. Surgical versus nonoperative treatment for lumbar spinal stenosis four-year results of the Spine Patient Outcomes Research Trial. *Spine (Phila. Pa. 1976).* 2010;35(14):1329–1338.

19. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief. Bioinform.* 2017.

20. Senders JT, Staples PC, Karhade A V., et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg.* 2018;109(Ml):476-486.e1.

21. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models. *Jama.* 2017;318(14):1377.

22. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models : A framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.

23. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928–935.

24. Rolston JD, Han SJ, Chang EF. Systemic inaccuracies in the National Surgical Quality Improvement Program database: Implications for accuracy and validity for neurosurgery outcomes research. *J. Clin. Neurosci.* 2017;37(2017):44–47.

25. Mancuso CA, Duculan R, Craig CM, et al. Psychosocial Variables Contribute to Length of Stay and Discharge Destination after Lumbar Surgery Independent of Demographic and Clinical Variables. *Spine (Phila. Pa. 1976)*. 2018;43(4):281–286.

26. Best MJ, Buller LT, Falakassa J, et al. Risk Factors for Nonroutine Discharge in Patients Undergoing Spinal Fusion for Intervertebral Disc Disorders. *Iowa Orthop. J.* 2015;35(305):147–155.

27. Abt NB, McCutcheon BA, Kerezoudis P, et al. Discharge to a rehabilitation facility is associated with decreased 30-day readmission in elective spinal surgery. *J. Clin. Neurosci.* 2017;36(2017):37–42.

28. Murphy ME, Gilder H, Maloney PR, et al. Lumbar decompression in the elderly: increased age as a risk factor for complications and nonhome discharge. *J. Neurosurg. Spine*. 2017;26(3):353–362.

29. Deyo RA, Hickam D, Duckart JP, et al. Complications after surgery for lumbar stenosis in a veteran population. *Spine (Phila. Pa. 1976)*. 2013;38(19):1695–1702.

30. Schoenfeld AJ, Carey PA, Cleveland AW, et al. Patient factors, comorbidities, and surgical characteristics that increase mortality and complication risk after spinal arthrodesis: A prognostic study based on 5,887 patients. *Spine J.* 2013;13(10):1171–1179.

31. Veeravagu A, Patil CG, Lad SP, et al. Risk factors for postoperative spinal wound infections after spinal decompression and fusion surgeries. *Spine (Phila. Pa. 1976)*. 2009;34(17):1869–1872.

32. Lakomkin N, Goz V, Cheng JS, et al. The utility of preoperative laboratories in predicting postoperative complications following posterolateral lumbar fusion. *Spine J.* 2018;18(6):993–997.

33. Hauck K, Zhao X. How Dangerous is a Day in Hospital? *Med. Care.* 2011;49(12):1068–1075.

34. Young J, Green J. Effects of delays in transfer on independence outcomes for older people requiring postacute care in community hospitals in England. *J. Clin. Gerontol. Geriatr.* 2010;1(2):48–52.

35. Sirois MJ, Lavoie A, Dionne CE. Impact of Transfer Delays to Rehabilitation in Patients with Severe Trauma. *Arch. Phys. Med. Rehabil.* 2004;85(2):184–191.

36. Kondo A, Zierler BK, Isokawa Y, et al. Comparison of lengths of hospital stay after surgery and mortality in elderly hip fracture patients between Japan and the United States - The relationship between the lengths of hospital stay after surgery and mortality. *Disabil. Rehabil.* 2010;32(10):826–835.

37. Nikkel LE, Kates SL, Schreck M, et al. Length of hospital stay after hip fracture and risk of early mortality after discharge in New York state: Retrospective cohort study. *BMJ.* 2015;351(December):1–10.

38. M.W. R, G. L, K. S, et al. Nursing homes in 10 nations: a comparison between countries and settings. *Age Ageing.* 1997;26(SUPPL. 2):3–12.

39. Beaupre LA, Wai EK, Hoover DR, et al. A comparison of outcomes between Canada and the United States in patients recovering from hip fracture repair: Secondary analysis of the FOCUS trial. *Int. J. Qual. Heal. Care.* 2018;30(2):97–103.

40. Bryan K. Policies for reducing delayed discharge from hospital. *Br. Med. Bull.* 2010;95(1):33–46.

41. Ou L, Chen J, Young L, et al. Effective discharge planning - timely assignment of an estimated date of discharge. *Aust. Heal. Rev.* 2011;35(3):357.

42. McCoy D, Godden S, Pollock AM, et al. Carrot and sticks? The Community Care Act (2003) and the effect of financial incentives on delays in discharge from hospitals in England. *J. Public*

Health (Bangkok). 2007;29(3):281–287.

43. Styrborn K, Thorslund M. “Bed-blockers”: Delayed discharge of hospital patients in a nationwide perspective in Sweden. *Health Policy (New. York).* 1993;26(2):155–170.

44. Zychlinski N, Mandelbaum A, Momˇ P. Bed Blocking in Hospitals due to Scarce Capacity in Geriatric Institutions – Cost Minimization via Fluid Models. *Time-Varying Fluid Networks with Blocking Model. Support. Patient Flow Anal. Hosp.* 2017:1–41.

Table 1. Baseline characteristics of patients, n = 28,600

Variable	**Definition**	
Age	Median (IQR)	67 (58-74)
	Missing, n (%)	113 (0.37)
Sex	Female	13518 (47.3)
	Male	15082 (52.7)
BMI	Median (IQR)	30.09 (26.58-34.43)
	Missing, n (%)	125 (0.34)
Functional Status	Independent	27917 (97.6)
	Dependent	508 (1.8)
	Missing, n (%)	175 (0.6)
Fusion	No	13053 (45.6)
	Yes	15547 (54.4)
Approach	Posterior	26633 (93.1)
	Anterior	682 (2.4)
	Combined	1285 (4.5)
Level	One or Two Levels	14618 (41.5)
	Three or More Levels	20638 (58.5)
Instrumentation	No	15973 (55.8)
	Yes	12627 (44.2)
ASA Class	I	475 (1.7)
	II	12281 (42.9)
	III	15079 (52.7)
	IV	765 (2.7)
Hematocrit	Median (IQR)	41.1 (38.4-43.8)
	Missing, n (%)	1926 (6.7)
White Blood Cell	Median (IQR)	7.0 (5.8-8.4)
	Missing, n (%)	2260 (7.9)
Platelet	Median (IQR)	232 (194-276)
	Missing, n (%)	2285 (7.9)
Sodium	Median (IQR)	140 (138-141)
	Missing, n (%)	3401 (11.9)
Creatinine	Median (IQR)	0.90 (0.77-1.09)
	Missing, n (%)	3270 (11.4)
Diabetes	No	22488 (78.6)
	Oral Medication	4111 (14.4)
	Insulin Dependent	2001 (7.0)
Chronic Obstructive Pulmonary Disease		1493 (5.2)
Chronic Steroid Use		1189 (4.2)
Bleeding disorder		571 (2.0)

BMI = Body Mass Index; ASA = American Society of Anesthesiologists Classification; IQR = Interquartile Range

Table 2. Model performance for discharge disposition prediction on training set

Performance Metric	Neural Network	Support Vector Machine	Bayes Point Machine	Boosted Decision Tree
C-statistic	0.751	0.743	0.752	0.747
Calibration slope	0.933	0.996	1.038	0.694
Calibration intercept	0.037	5.2 x10^{-4}	-3.57 x10^{-4}	4.58 x10^{-3}
Brier Score	0.130	0.131	0.131	0.133
Null Model Brier Score	0.150			

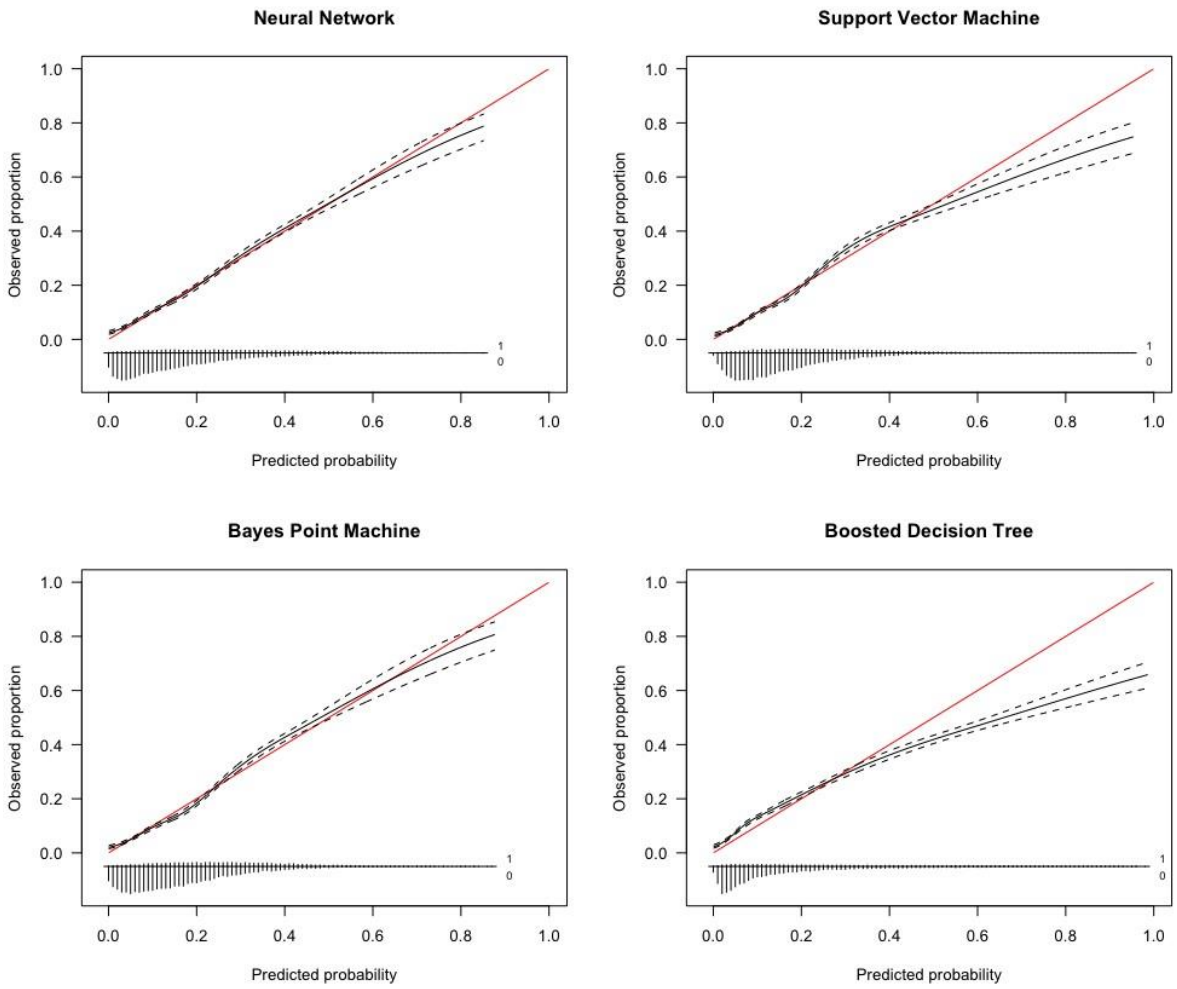


Figure 1. Calibration curve per model for prediction of non-home discharge in the training set

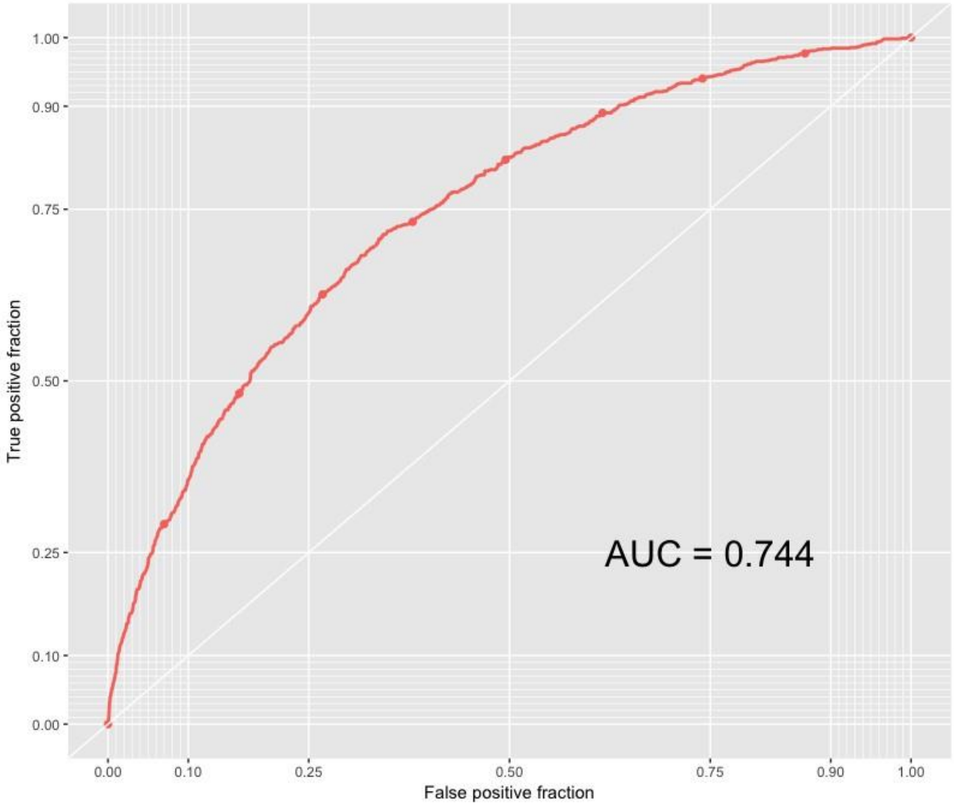


Figure 2. Receiver operating curve of the neural network in the testing set

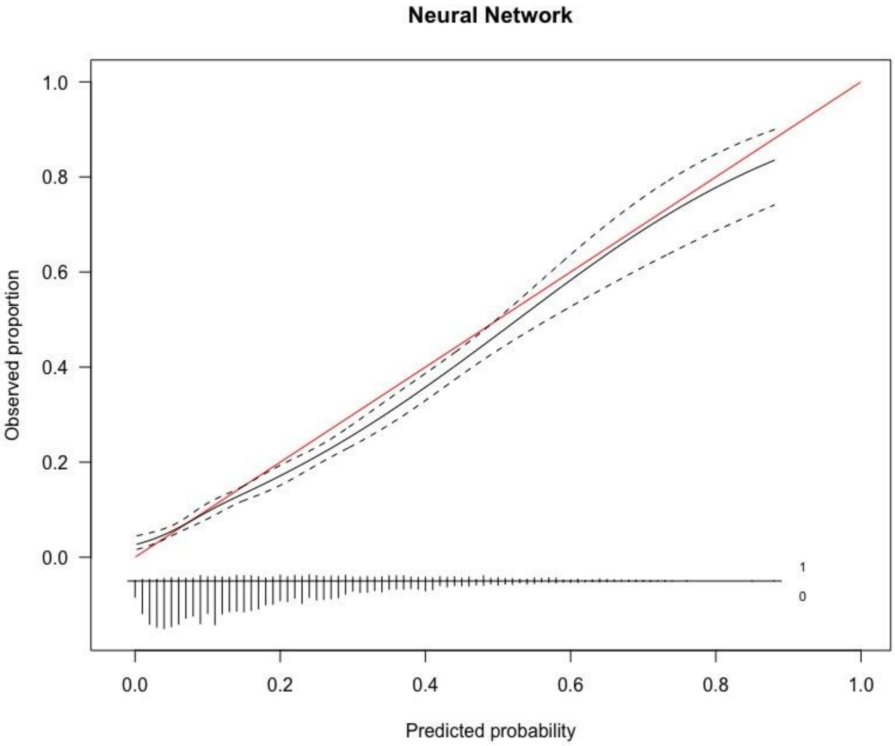


Figure 3. Calibration curve of the neural network in the testing set

Chapter 5

Development of a Machine Learning Algorithm Predicting Discharge Placement after Surgery for Spondylolisthesis

Ogink PT, Karhade AV, Thio QCBS, Hershman SH, Cha TD, Bono CM, Schwab JH

Eur Spine J. 2019 Aug;28(8):1775-1782

Abstract

Purpose We aimed to develop a machine learning algorithm that can accurately predict discharge placement in patients undergoing elective surgery for degenerative spondylolisthesis.

Methods The National Surgical Quality Improvement Program (NSQIP) database was used to select patients that underwent surgical treatment for degenerative spondylolisthesis between 2009 and 2016. Our primary outcome measure was non-home discharge which was defined as any discharge not to home for which we grouped together all non-home discharge destinations including rehabilitation facility, skilled nursing facility, and unskilled nursing facility. We used Akaike Information Criterion to select the most appropriate model based on the outcomes of the stepwise backward logistic regression. Four machine-learning algorithms were developed to predict discharge placement and were assessed by discrimination, calibration, and overall performance.

Results Nine thousand three hundred and thirty-eight patients were included. Median age was 63 (Interquartile range [IQR] 54 - 71) and 63% (n=5,887) were female. The non-home discharge rate was 18.6%. Our models included age, sex, diabetes, elective surgery, BMI, procedure, number of levels, ASA class, preoperative white blood cell count, and preoperative creatinine. The Bayes Point Machine was considered the best model based on discrimination (AUC = 0.753), calibration (slope = 1.111;

intercept = -0.002), and overall model performance (Brier score = 0.132).

Conclusion This study has shown that it is possible to create a predictive machine learning algorithm with both good accuracy and calibration to predict discharge placement. Using our methodology this type of model can be developed for many other conditions and (elective) treatments.

Introduction

Decreasing the length of hospitalization can significantly reduce healthcare costs; each extra day in hospital costs roughly between $424 in Spain and $5,220 in the US depending on the type of hospital and the hospital's location.[1] Despite a recent trend towards early discharges after major orthopedic surgery, patients' outcomes have largely remained similar.[2,3] Recent studies have demonstrated that a disproportionate part of patients' total hospitalization is related to delays while patients wait to be discharged to a rehabilitation facility (RF) or skilled nursing facility (SNF).[4,5] Hwabejire et al.[6] consulted their institution's case management department and found that 47% of prolonged hospitalizations were due to difficulties in RF placement. More importantly, prolonged hospitalization is a known risk factor for adverse events, such as venous thromboembolism and hospital-acquired infections.[7,8]

Some studies have looked at which variables affect discharge to RF/SNF following spine surgery[9–12] and others have developed grading scales which can help predict patient disposition [13–15], however, to our knowledge there have been no studies utilizing machine learning algorithms to help predict discharge placement. Pre-operatively determining which patients are likely to require RF/SNF placement can reduce the risk of prolonged hospitalization and can potentially allow healthcare personnel to reserve a place in a RF/SNF well in advance of a patient's discharge. This could therefore minimize many of the risks associated with extended hospital stay, allow for more efficient departmental planning, and can potentially increase surgical volume.

In this proof-of-concept application of machine learning for predicting disposition we aim to develop a machine learning algorithm using the ACS-NSQIP database that can accurately predict discharge placement in patients with degenerative spondylolisthesis. Machine learning lies at the intersection of statistics and computer science, and is increasingly being used in medicine to develop prediction models and decision-making tools from large datasets.[16,17] We selected degenerative spondylolisthesis because 1) this group represents a sizeable portion of the spine surgery population 2) patients are relatively older and thereby at risk for discharge to RF/SNF and 3) most surgeries are elective which means there is time to arrange an RF/SNF placement if we were to develop a useful predictive algorithm.

Methods

Patient selection

We selected all patients from the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) database ACS-NSQIP database. The ACS-NSQIP database is a multi-institutional database that has previously been used in numerous spine studies.[18–21] The database consists of prospectively collected patient demographics, comorbidities, laboratory values, and peri-operative and post-operative outcomes in the 30 days following surgery. Unlike administrative databases the data is collected by trained reviewers leading to better registration of outcomes.[22,23] The American College of Surgeon aims to ensure data reliability by training participating hospitals, ongoing education and systematic audits.

We included patients who met the following criteria: 1) International Classification of Disease Ninth Revision (ICD-9) code 738.4 for acquired spondylolisthesis or ICD-10 code M43.10 2) Current Procedural Terminology (CPT) codes for decompression, fusion, or fixation 3) year of surgical treatment between 2009 and 2016. Ultimately, 9,338 patients were included for development of the algorithm.

Variable Selection and Data Analysis

The primary outcome measure was non-home discharge which was defined as any discharge not to home for which we grouped together all non-home discharge destinations including rehabilitation facility, skilled nursing facility, and unskilled nursing facility. We performed a bootstrap stepwise backward logistic regression with the following candidate variables: Age [years], sex [male, female], body mass index (BMI) [kg/m^2], race [Caucasian, African-American, other], diabetes [no, oral medication, insulin-dependent], anti-hypertensive medication [yes, no], specialty [neurosurgery, orthopedics], ASA class [I,II,III,IV], elective surgery [yes, no], type of procedure [decompression, fusion, decompression and fusion], approach [anterior, posterior], number of levels [1 or 2, 3 or more], preoperative white blood cell count [10^3/µL], preoperative creatine[mg/dL], preoperative platelets [10^3/$mm^{3]}$, preoperative albumin [g/dL], preoperative blood urea nitrogen [mg/dL], and preoperative sodium [mEq/L].

We subsequently used Akaike Information Criterion to select the most appropriate model with independently significant variables based on the outcomes of the stepwise backward logistic regression.[24] This model included age, sex, diabetes, elective surgery, BMI, procedure, number of levels, ASA class, preoperative white blood cell count, and preoperative creatinine.

Boosted Decision Tree, Support Vector Machine, Bayes Point Machine and Neural Network algorithms were trained with these variables to predict which patients were not discharged home. We did a stratified 80:20 split of the dataset into a training set and a test set. We used the training set for algorithm training and assessment of performance by cross validation (10x). The algorithms were subsequently used in the test set to make predictions on discharge placement. The predictions were then compared with the actual outcomes of the test set to assess the performances outside the training set.

Model Performance

Model performance was measured with the following three metrics: discrimination, calibration, and overall model performance.

Discrimination is the ability to distinguish patients discharged to home from patients who were not discharged to home. We assessed discrimination with receiver-operating curves (ROC) and with the c-statistic. Models with discrimination similar to chance have a c-statistic of 0.5 and models with a perfect discrimination have a c-statistic of 1.0.

Calibration shows how well the model's predicted probabilities are in line with the actual observed occurrences in the test set. The calibration intercept measures whether the model is over- or underestimating the probabilities and the calibration slope measures whether the predictor effects in the training and test set are the same. A perfect model has an intercept value of 0 and a slope value of 1. The Brier score was used to assess overall model performance. It combines discrimination and calibration and is calculated by obtaining the mean squared error between the probabilities given by the model and the actual observed values. Smaller Brier scores (closer to zero) indicate better overall performance. However, the Brier score must always take into account the prevalence of the outcome in the patient sample. Therefore, the null Brier score was determined by assigning probabilities to every patient similar to the prevalence of the outcome.

Application

The model was subsequently developed into a web-based application making it accessible on smartphones, computers, and tablets. The application is designed to let the user input the necessary variables, calculate the scores using the selected algorithm, and output the results. Microsoft Azure, STATA 13 (StataCorp LP, College Station, TX, USA) RStudio version 1.0.153, and Python version 3.6 (Python Software Foundation) (Anaconda distribution) were used for data analysis, model creation, and application development.

Results

The non-home discharge rate was 18.6% for the 9,338 included patients. Median age was 63 (Interquartile range [IQR] 54 – 71) and 63% (n=5,887) were female. Baseline characteristics are shown in Table 1. The c-statistics of the 4 models ranged from 0.733 for the Boosted Decision Tree to 0.755 for the Neural Network (Table 2; Figure 1). Calibration slope values ranged from 0.459 for the Boosted Decision Tree to 1.111 for the Bayes Point Machine while calibration intercept values ranged from -0.015 for the Boosted Decision Tree to 0.123 for the Neural Network (Table 2; Figure 2).

Overall model performance, based on the Brier score, ranged from 0.132 for the Bayes Point Machine and Neural Network to 0.146 for the Boosted Decision Tree. The null Brier model performance was 0.152. Considering the performance in calibration and overall assessment the Bayes Point Machine was chosen as the final model. The web application based on the Bayes Point Machine model can be accessed at **https://sorg-apps.shinyapps.io/spondydisposition/.**

Discussion

Unexpected non-home discharge is a potential cause for extended length of stay and subsequent adverse events for patients. We aimed to develop a machine learning algorithm to predict which patients are likely to have a non-home discharge.

Our model included age, sex, diabetes, elective surgery, BMI, procedure, number of levels, ASA class, preoperative white blood cell count, and preoperative creatinine. Of the four tested machine learning algorithms the Bayes Point Machine was considered the best model based on discrimination (AUC = 0.753), calibration (slope = 1.111; intercept = -0.002), and overall model performance (Brier score = 0.132).

This study has limitations. First, despite ACS-NSQIP database being frequently-used and the rigorous oversight, it comes with the inherent limitations of potential miscoding and missing values. A study by Rolston et al.[25] found varying degrees of miscoding in neurosurgical outcomes. These inaccuracies may bias our outcomes and thus our algorithm. Furthermore, the database does not contain all the variables of interest that other studies identified as risk factors for non-home discharge. For instance, insurance status, employment status, and preoperative patient reported outcomes scores, which have all been established as important predictors for discharge placement, were not available. Despite the lack of these variables and potential miscoding the

ACS-NSQIP database provides a large patient set, which is required for predictive algorithms, with a sizeable number of important variables. External validation of the algorithm is crucial to check its applicability. Second, although the database is constructed with data from 690 hospitals, the patient population may not be reflective of all the patients for which it's intended, especially if this model were to be used outside the US.

The variables included in our model have been identified through other studies examining risk factors for non-home discharges in degenerative spine surgery. Murphy et al.[26] found age, BMI, number of levels, ASA class, diabetes, and female gender to be predictors of not being discharged home after decompression without fusion. Abt et al.[27] similarly found age, BMI, ASA class, and diabetes to be predictors but they found male gender rather than female gender was more likely to suggest a non-home discharge. Best et al.[11] concluded that age > 65 years was by far the greatest predictor for non-routine discharge after fusion for intervertebral disc disorders. The only variables in our model not featured in these studies are preoperative white blood cell count and creatinine although preoperative creatinine has been identified as predictor of discharge status in other specialties.[28,29] While adding intraoperative and immediate postoperative outcomes, e.g. operation time or complications, would likely have increased the model's performance we opted to include only preoperatively known variables. Otherwise, the prediction would lose much of its value considering the window of opportunity for all of the potential benefits – preoperative arrangements and education- would be gone. We envision the model being used after the visit to the surgeon and anesthesiologist, when all variables are known, in a preoperative visit with a nurse practitioner or case management. This would allow for education and potentially initiate arrangements to be made.

Two previous studies tested predictive grading scales in all degenerative spine patients. McGirt et al.[13] constructed a grading system for extended length of stay, discharge to rehab, and hospital readmission after elective spine surgery. Their grading system included the variables age > 70 years, ASA class > III, Oswestry Disability Index, diabetes, non-independent ambulation, and non-private insurance. While age, ASA class, and diabetes are included in our model, ambulation and insurance status were not available in the NSQIP database. Furthermore, they reported a c-statistic of 0.731 compared to our 0.755. Slover et al. [14] tested the Risk Assessment and Prediction Tool – which stratifies patients into high and low-risk groups for non-home discharge after total joint replacement- on spinal fusion patients. This tool is based on points for age and sex – also present in our study- combined with walking distance, use of gait aid, community support, and a caregiver at home. They did not report any c-statistic. Importantly, both these studies fail to report calibration metrics, which are essential to determine whether the predictive models/scoring systems are useful. While most studies being published on predictive models report c-statistics, many of them lack assessment of calibration. A predictive model may be able to discriminate well between those who will be discharged home and those who will not, but give inaccurate risk estimates for individual patients due to poor calibration. Clinically useful prediction tools need to discriminate well and be well calibrated in order to make an accurate risk assessment.[30]

In our study, the model with the best discrimination, the Neural Network, was inferior to the Bayes Point Machine with respect to calibration over the full range of prediction (Figure 2). Future studies on predictive models should assess calibration graphically and numerically to determine model performance.

Waiting on RF/SNF placement has been determined to be a major factor in delayed discharges[4,6,31], unfortunately, simply increasing capacity is not the answer. Gaughan et al.[32] studied whether increasing the supply of nursing home beds would reduce the number of delayed discharges. They determined that this would only reduce delayed discharges by 6-9% and that this small effect would make this intervention more costly instead of reducing cost. Implementation of our predictive model could potentially prevent some delayed discharges without incurring additional costs. However, implementation of predictive models in clinical practice is difficult and has not been done on a large scale yet, despite the multitude of models currently available.[33,34]

First and foremost, rigorous testing of a model's predictive ability and external validation should be performed to prevent unintended consequences. Nonetheless, with the pressure of reducing costs, the obvious role that waiting time plays in delaying discharges, and the increasing use of predictive analytics by caregivers, implementing a model which can predict discharge placement may be worth pursuing.

Conclusion

This study has shown that it is possible to create a predictive machine learning algorithm with both good discrimination and calibration to predict discharge placement. Using our methodology this type of model can be developed for many other conditions and (elective) treatments. Integrating these models into practice could potentially make hospitals more efficient, save unnecessary healthcare costs, and minimize adverse events for patients due to delayed discharges.

References

1. International Federation of Health Plans. Variation in Medical and Hospital Prices by Country. *Comp. Price Rep.* 2015.

2. Regenbogen SE, Cain-Nielsen AH, Norton EC, et al. Costs and consequences of early hospital discharge after major inpatient surgery in older adults. *JAMA Surg.* 2017;152(5):e170123.

3. Basques BA, Tetreault MW, Della Valle CJ. Same-Day Discharge Compared with Inpatient Hospitalization Following Hip and Knee Arthroplasty. *J. Bone Joint Surg. Am.* 2017;99(23):1969–1977.

4. Costa AP, Poss JW, Peirce T, et al. Acute care inpatients with long-term delayed discharge: evidence from a {Canadian} health region. *BMC Health Serv. Res.* 2012;12(In press):6–11.

5. Watkins JR, Soto JR, Bankhead-Kendall B, et al. What's the hold up? Factors contributing to delays in discharge of trauma patients after medical clearance. *Am. J. Surg.* 2014;208(6):969–973.

6. Hwabejire JO, Kaafarani HMA, Imam AM, et al. Excessively long hospital stays after trauma are not related to the severity of illness: Let's aim to the right target! *JAMA Surg.* 2013;148(10):956–961.

7. Andrews LB, Stocking C, Krizek T, et al. An alternative strategy for studying adverse events in medical care. *Lancet.* 1997;349(9048):309–313.

8. Hauck K, Zhao X. How Dangerous is a Day in Hospital? *Med. Care.* 2011;49(12):1068–1075.

9. Gruskay JA, Fu M, Bohl DD, et al. Factors affecting length of stay after elective posterior lumbar spine surgery: A multivariate analysis. *Spine J.* 2015;15(6):1188–1195.

10. Sharma M, Sonig A, Ambekar S, et al. Discharge dispositions, complications, and costs of hospitalization in spinal cord tumor surgery: analysis of data from the United States Nationwide Inpatient Sample, 2003–2010. *J. Neurosurg. Spine.* 2014;20(2):125–141.

11. Best MJ, Buller LT, Falakassa J, et al. Risk Factors for Nonroutine Discharge in Patients Undergoing Spinal Fusion for Intervertebral Disc Disorders. *Iowa Orthop. J.* 2015;35(305):147–155.

12. Niedermeier S, Przybylowicz R, Virk SS, et al. Predictors of discharge to an inpatient rehabilitation facility after a single-level posterior spinal fusion procedure. *Eur. Spine J.* 2017;26(3):771–776.

13. McGirt MJ, Parker SL, Chotai S, et al. Predictors of extended length of stay, discharge to inpatient rehab, and hospital readmission following elective lumbar spine surgery: introduction of the Carolina-Semmes Grading Scale. *J. Neurosurg. Spine*. 2017;27(4):382–390.

14. Slover J, Mullaly K, Karia R, et al. The use of the Risk Assessment and Prediction Tool in surgical patients in a bundled payment program. *Int. J. Surg.* 2017;38:119–122.

15. Kanaan SF, Yeh H-W, Waitman RL, et al. Predicting discharge placement and health care needs after lumbar spine laminectomy. *J. Allied Health.* 2014;43(2):88–97.

16. Jordan M, Mitchell T. Machine learning: Trends, perspectives, and prospects. *Science (80-.).* 2015;349(6245):255–260.

17. Senders JT, Staples PC, Karhade A V., et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg.* 2018;109(Ml):476-486.e1.

18. Schoenfeld AJ, Le H V., Marjoua Y, et al. Assessing the utility of a clinical prediction score regarding 30-day morbidity and mortality following metastatic spinal surgery: The New England Spinal Metastasis Score (NESMS). *Spine J.* 2015;16(4):482–490.

19. Bekelis K, Desai A, Bakhoum SF, et al. A predictive model of complications after spine surgery: The National Surgical Quality Improvement Program (NSQIP) 2005-2010. *Spine J.* 2014;14(7):1247–1255.

20. Veeravagu A, Li A, Swinney C, et al. Predicting complication risk in spine surgery: a prospective analysis of a novel risk assessment tool. *J. Neurosurg. Spine*. 2017;27(1):81–91.

21. Sebastian A, Huddleston P, Kakar S, et al. Risk factors for surgical site infection after posterior cervical spine surgery: an analysis of 5,441 patients from the ACS NSQIP 2005–2012. *Spine J.* 2016;16(4):504–509.

22. Steinberg SM, Popa MR, Michalek JA, et al. Comparison of risk adjustment methodologies in

surgical quality improvement. *Surgery.* 2008;144(4):662–7; discussion 662-7.

23. Davenport DL, Holsapple CW, Conigliaro J. Assessing Surgical Quality Using Administrative and Clinical Data Sets: A Direct Comparison of the University HealthSystem Consortium Clinical Database and the National Surgical Quality Improvement Program Data Set. *Am. J. Med. Qual.* 2009;24(5):395–402.

24. Posada D, Buckley TR. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Thorne J, ed. *Syst. Biol.* 2004;53(5):793–808.

25. Rolston JD, Han SJ, Chang EF. Systemic inaccuracies in the National Surgical Quality Improvement Program database: Implications for accuracy and validity for neurosurgery outcomes research. *J. Clin. Neurosci.* 2017;37(2017):44–47.

26. Murphy ME, Maloney PR, McCutcheon BA, et al. Predictors of Discharge to a Nonhome Facility in Patients Undergoing Lumbar Decompression Without Fusion for Degenerative Spine Disease. *Neurosurgery.* 2017;81(4):638–649.

27. Abt NB, McCutcheon BA, Kerezoudis P, et al. Discharge to a rehabilitation facility is associated with decreased 30-day readmission in elective spinal surgery. *J. Clin. Neurosci.* 2017;36(2017):37–42.

28. Kelly DM, Bennett R, Brown N, et al. Predicting the discharge status after liver transplantation at a single center: A new approach for a new era. *Liver Transplant.* 2012;18(7):796–802.

29. Tong MZ, Pattakos G, He J, et al. Sequentially updated discharge model for optimizing hospital resource use and surgical patients' satisfaction. *Ann. Thorac. Surg.* 2015;100(6):2174–2181.

30. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* 2007;115(7):928–935.

31. Benson RT, Drew JC, Galland RB. A waiting list to go home: An analysis of delayed discharges from surgical beds. *Ann. R. Coll. Surg. Engl.* 2006;88(7):650–652.

32. Gaughan, James; Gravelle Hugh; Siciliani L. Testing the bed-blocking hypothesis: does nursing and care home supply reduce delayed hospital discharges? *Health Econ*. 2015;24:32–44.

33. Harris AHS. Path from predictive analytics to improved patient outcomes. *Ann. Surg*. 2017;265(3):461–463.

34. Parikh RB, Kakad M, Bates DW. Integrating Predictive Analytics Into High-Value Care. *Jama*. 2016;315(7):651.

Table 1. Baseline characteristics of patients, n = 9,338

Variable	**All patients** **(n = 9,338)**
	Median (IQR)
Age (years)	63 (54 – 71)
BMI (kg^2/m^2)	30 (26 – 34)
Creatinine levels (mg/dL)	0.86 (0.72 – 1.00)
White blood cell count ($10^3/\mu L$)	6.9 (5.7 – 8.3)
	Number (%)
Female	5,887 (63)
Race	
Caucasian	8,369 (90)
African-American	695 (7.4)
Other	274 (2.9)
Procedure	
Decompression and fusion	5,897 (63)
Fusion	2,857 (31)
Decompression	584 (6.3)
ASA Class	
I	238 (2.6)
II	4,632 (50)
III	4,288 (46)
IV	180 (1.9)
Diabetes	
Oral diabetics	1,123 (12)
Insulin dependent	494 (5.3)
Elective surgery	9,114 (98)
Anti-hypertensive medication	5,580 (60)

BMI = Body Mass Index; ASA = American Society of Anesthesiologists

Table 2. Machine learning model performance for discharge disposition prediction in patients undergoing surgery for degenerative spondylolisthesis

Method	**Metric**	**Machine Learning Algorithm**			
		Boosted Decision Tree	**Support Vector Machine**	**Bayes Point Machine**	**Neural Network**
Discrimination	C-statistic	0.733	0.742	0.753	0.755
	AUC-PR	0.388	0.399	0.417	0.420
Calibration	Calibration Slope	0.459	0.994	1.111	0.970
	Calibration Intercept	-0.015	-0.003	-0.002	0.123
Overall	Brier Score	0.146	0.134	0.132	0.132
	Null Model Brier Score	0.152			

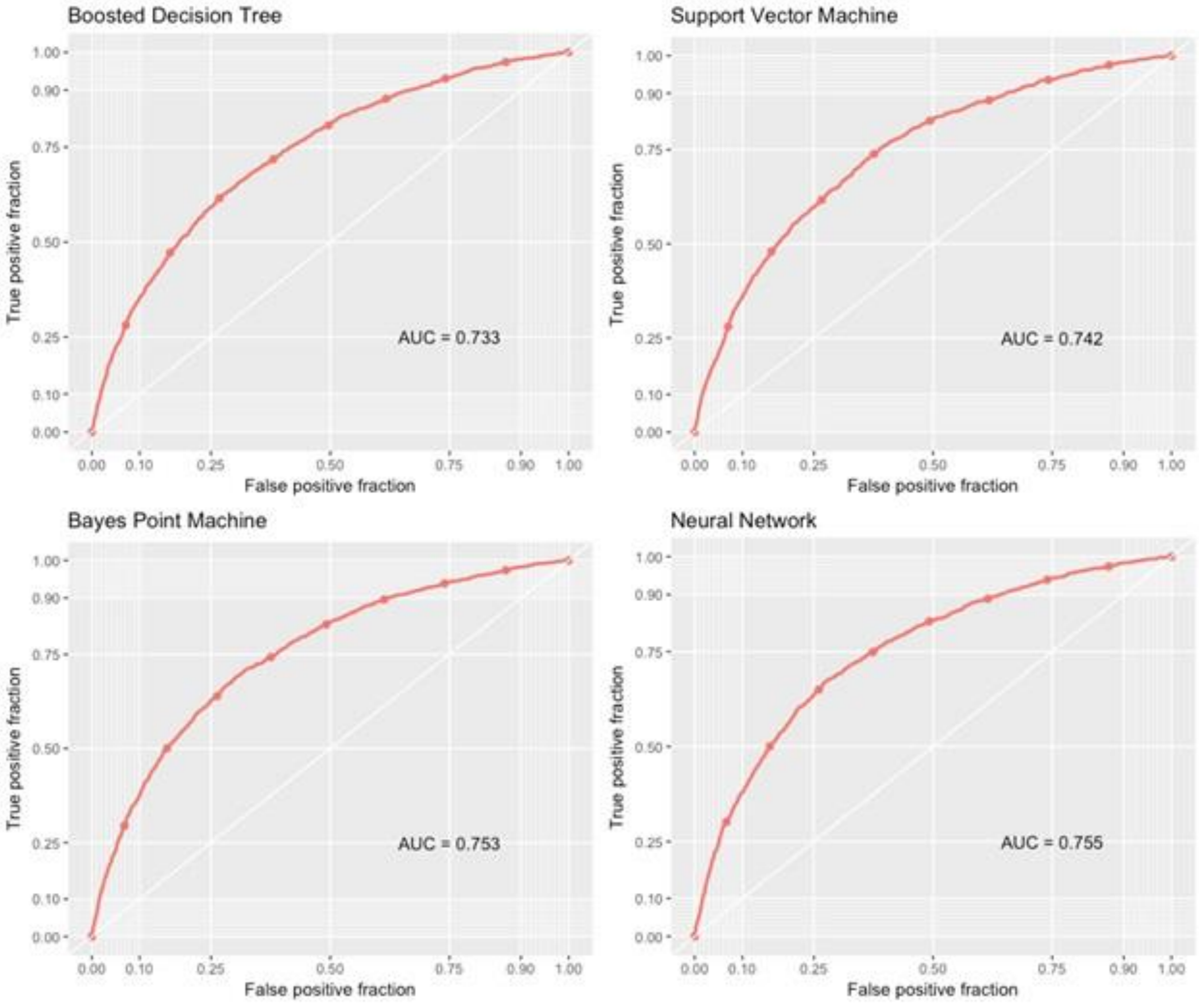


Figure 1. Receiver operating curve by model for prediction of discharge disposition

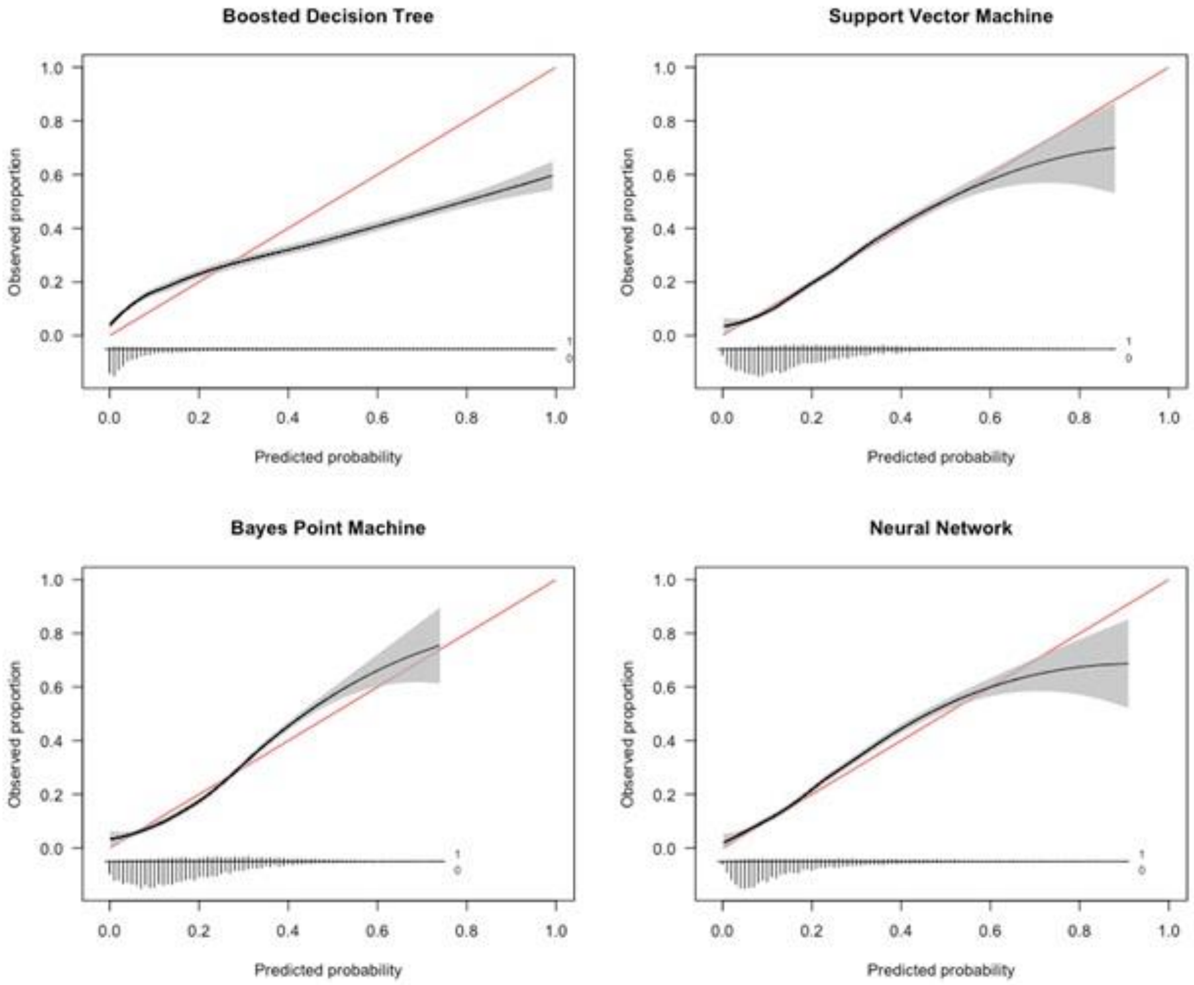


Figure 2. Calibration curve by model for prediction of thirty-day mortality in the test set

Chapter 6

Predicting Sustained Opioid Dependence after Surgery for Degenerative Spondylolisthesis using Machine Learning Algorithms

Ogink PT, Karhade AV, van Stein NJ, Groot OQ, Oner FC, Verlaan JJ, Schwab JH

Under Review The Spine Journal

Abstract

Introduction Spinal surgery is known to have high rates of postoperative opioid prescriptions, which can lead to dependence and abuse. An individual preoperative prediction of who is at increased risk of prolonged opioid use could allow for earlier targeted counseling on pain medication. We aimed to develop and internally validate a machine learning prediction model for prolonged opioid use following surgery for degenerative spondylolisthesis.

Methods We used International Classification of Disease 9th Version (ICD-9) code 738.4 and ICD-10 code M43.16 to include all patients 18 years or older who underwent surgery for degenerative spondylolisthesis between 2009 and 2016 in 2 tertiary care referral spine centers. Our primary outcome measure was prolonged opioid use defined as opioid prescriptions 90-180 days after the index surgery. Four machine-learning algorithms were developed to predict prolonged opioid use and were assessed by discrimination, calibration, and overall performance.

Results Eight hundred seventy-eight patients were included in our study for development of the algorithm. The rate of prolonged opioid use was 10% (88/878). The following variables were selected and subsequently used for algorithm development: age, BMI, duration of symptoms, preoperative laboratory values for hemoglobin and white blood cell count, and preoperative use of opioids, statins and antihypertensive drugs. The Random Forest algorithm was the best model based on discrimination (AUC = 0.80, calibration (slope = 0.86 ; intercept = 0.01 , and overall performance (Brier score = 0.09).

Discussion and conclusion In this study we have developed a predictive machine learning algorithm with good discrimination as well as calibration to predict sustained opioid use after surgery for degenerative spondylolisthesis. Utilization of this model could aid clinicians and healthcare organizations in tailoring their strategies and policies for reducing sustained opioid use to high-risk patients.

Introduction

The current opioid epidemic is having enormous impact on society.[1–3] Postoperative opioid prescriptions play a significant role in this epidemic.[4] Orthopedic surgery is considered a specialty at increased risk of misuse of opioid medication.[5–8] As a subspecialty within orthopedics, spinal surgery is known to have high rates of postoperative opioid prescriptions. A study by Adogwa et al. found 83% of all patients who underwent spine surgery utilized opioids within the 2-year postoperative window and 66% had continued opioid use 1 year after surgery.[9] This prolonged postoperative use can lead to dependence and abuse.

Although several studies have focused on risk factors for prolonged opioid use after spine surgery,[9–11] to our knowledge, none have tried to develop a prediction model for prolonged opioid use after lumbar spine surgery. An individual preoperative assessment of who is at increased risk of prolonged opioid use could allow for earlier targeted counseling on pain medication as well as extra caution during the postoperative visits.[12,13]

In recent years there has been a renewed interest in medicine in artificial intelligence, and more specifically supervised machine learning.[14] Machine learning stands at the intersection of statistics and computer science, and is increasingly being used in orthopedic surgery to develop prediction models and decision-making tools.[15] In this study we aimed to develop and internally validate a machine learning prediction model for prolonged opioid use after surgery for degenerative spondylolisthesis.

Methods

Patient selection

This study was approved by our Internal Review Board. We used International Classification of Disease 9th Version (ICD-9) code 738.4 and ICD-10 code M43.16 to include all patients 18 years or older who underwent surgery for degenerative spondylolisthesis between 2009 and 2016 in 2 tertiary care referral spine centers. We excluded the patients who had prior lumbar surgery on the same level, fusion in the lumbar spine, a malignancy, or a vertebral fracture within a 1-year period before or after diagnosis. Ultimately, 878 patients were included in our study for development of the algorithm. Our primary outcome measure was prolonged opioid use as defined by Brummett et al. as opioid prescriptions 90-180 days after the index surgery.[4] Appendix 1 lists all medications included as opioids.

We manually extracted the following potential explanatory variables from our electronic health record: age[years], gender[male/female], body mass index (BMI) [kg/m^2], race, median household income [$] based on the US Censuses Bureau American Community Survey zip code data, history of substance abuse, number of levels involved, degree of largest slip [Meyerding classification I/II/III/IV], duration of symptoms [days] , preoperative medication use (opioids, ACE-inhibitors, antidepressants, antipsychotics, beta-2-agonists, beta-blockers, benzodiazepines, gabapentin, immunosuppressants, nonsteroidal anti-inflammatory drugs). Furthermore, we collected preoperative laboratory values within 1 week of surgery: hemoglobin [g/dL], white blood cell count [10^3/µL], creatinine [mg/dL], platelets [10^3/mm$^{3]}$, albumin [g/dL], blood urea nitrogen [mg/dL], and sodium [mEq/L].

Variable selection and Data Analysis

Variable selection for the ML algorithms was performed by entering all available variables in a random forest regression, which ranks variables according to their respective prediction power for the outcome. The following variables were selected and subsequently used for algorithm development: age, BMI, duration of symptoms, preoperative laboratory values for hemoglobin and white blood cell count, and preoperative use of opioids, statins and antihypertensive drugs. A stratified 80:20 split of the dataset into a training set and testing set was performed. The training set was used for algorithm training and assessment of performance using tenfold cross validation. Four algorithms were trained using the selected variables to predict prolonged opioid use: Neural Network, Support Vector Machine, Random Forest, Gradient Boosting Machine.[16]

The four developed models were then used on the patients of the testing set to predict prolonged opioid use. These predictions were subsequently compared to the real outcomes of the testing set to determine the performance of the algorithm outside the training set.

Model Assessment and explanation

Performance of ML algorithms is often measured with the following metrics: discrimination, calibration and overall model performance.[17] Discrimination is the algorithm's ability to distinguish between patients with prolonged opioid use and those without. We used receiver-operating curves and the c-statistic to assess discrimination. Algorithms with a c-statistic of 0.5 are no better than chance and a score of 1.0 signifies a perfect algorithm. Calibration depicts how well the predictions of the model correspond with the actual observed occurrences in the testing set. The calibration slope measures whether the predictor effects in the training and testing set are the same whereas the calibration intercept measures if the algorithm is over- or underestimating the probabilities of prolonged opioid use. Ideally, the slope has a value of 1 and the intercept a value of 0. Overall performance was assessed using the Brier score.

The model with the best performance according to discrimination, calibration, and overall performance was selected and given explanations at both the global level, i.e. variable-importance across the entire database, and local level, i.e. patient-specific variable-importance. Global explanation is shown by averaging the importance of individual variables across all patients.[18] At the local level, explanation is done by showing how each variable influences the patient-specific final outcome prediction.[19] Furthermore, this selected model was used to build an open-access web application. STATA 13 (StataCorp LP, College Station, TX, USA) RStudio version 1.0.153, and Python version 3.6 (Python Software Foundation) (Anaconda distribution) were used for data analysis and model creation.

Results

The rate of prolonged opioid use was 10% (88/878). Median age was 67 (interquartile range [IQR] 60 – 74) and 32% (278/878) were men. Other baseline characteristics are shown in Table 1. The c-statistics of the 4 algorithms in the testing set ranged from 0.589 (Support Vector Machine) to 0.818 (Gradient Boosting Machine; Table 2). The calibration slope values ranged from 0.640 (Neural Network) to 11.0 (Support Vector Machine), while the calibration intercept ranged from -0.55 (Neural Network) to 21.8 (Support Vector Machine).

The null Brier score was 0.088, with the Brier score of the 4 models ranging from 0.076 (Gradient Boosting Machine) to 0.087 (Support Vector Machine). Based on discrimination, calibration, and overall performance the Random Forest algorithm provided the best prediction results. The web application based on the Random Forest algorithm is available on **https://sorg-apps.shinyapps.io/spondyopioid/.**

Discussion

We aimed to develop a machine learning algorithm to predict which patients are at risk for prolonged opioid use after surgery for spondylolisthesis to tailor treatment and educational effort towards high-risk patients. The factors included in our model included age, BMI, duration of symptoms, preoperative laboratory values (hemoglobin, platelet, and white blood cell count), and preoperative medication use (opioids, anti-hypertensive medication, and statins). The Random Forest algorithm was the best model based on discrimination (AUC = 0.80, calibration (slope = 0.86 ; intercept = 0.01, and overall performance (Brier score = 0.079).

Preoperative opioid use [20–23], age [5,22–25], and BMI [26] have all been identified in previous studies as risk factors for (sustained) opioid use. Other variables are likely vectors for other previously identified risk factors; anti-hypertensive medication, statins, and laboratory values for the presence of comorbidities [4,27], duration of symptoms for the duration of preoperative opioid use [28]. Despite already having been identified earlier, our algorithm combines these risk factors to a patient-specific prediction including a breakdown of how much each variable contributes to that prediction.

Our study has several limitations. First, the data we used to develop the algorithm was collected from two tertiary care referral centers in the same city. This patient population may not be reflective of the patient population in the US, let alone the global patient population. In order to effectively use these algorithms outside these hospitals external validation is crucial. While there are many predictions models being created, far too few are being externally validated.[29] Especially for opioid (mis)use, where differences in socioeconomic factors, medical culture, and societal norms play a profound role, any predictive model should be externally validated on a regional or local level before implementation. Second, the definition of what should be considered sustained opioid use remains unclear. While we have aimed to predict sustained opioid *use*, we do not have the clinical information whether these patient are chemically dependent. However, the literature published clearly demonstrates a link between sustained opioid use and dependence. Third, we used hospital level prescription data and cannot be sure whether patients were actually adhering to the prescribed medication. Furthermore, this study is also not able to identify those patients seeking opioids from outside sources, which is a significant factor in the current opioid epidemic.[30]

A perceived obstacle in implementation of prediction models based on ML algorithms are its purported 'black box' aspects.[31] The use of global and local explanation can help alleviate these concerns and enhance the usability of ML algorithms in practice. Healthcare institutions and regulatory agencies can focus on the variables deemed most important by global explanation while physicians can use the local explanation during patient visits. Not only can high-risk patients be made aware of their risks, but the graphical representation of how each variables influences the outcome can show patients which variables they can positively influence to help improve their outcome.

Those extra features make our prediction model a potentially effective tool in lowering the number of patients who use opioids for a prolonged period after surgery. Patient education can play a significant role in efforts to lower opioid consumption. Syed et al. used preoperative patient education in patients undergoing rotator cuff repair and found preoperatively counseled patient were more than two times more likely to discontinue opioid use after 3 months.[12] Tailoring educational efforts to high-risk patients can aid surgical departments considering the limited time and resources healthcare organizations have for preoperative education. Additionally, clinicians could opt to have more frequent postoperative visits with high-risk patients, considering anxiety and catastrophic thinking have been shown to be associated with sustained opioid use in orthopedic surgery.[25,32]

Conclusion

In our study we have developed a predictive machine learning algorithm with good discrimination as well as calibration to predict sustained opioid use after surgery for degenerative spondylolisthesis. Utilizing this model could aid clinicians and healthcare organizations in tailoring their strategies and policies for reducing sustained opiod use to high-risk patients.

References

1. Florence CS, Zhou C, Luo F, et al. The Economic Burden of Prescription Opioid Overdose, Abuse, and Dependence in the United States, 2013. *Med. Care.* 2016;54(10):901–906.

2. Gomes T, Tadrous M, Mamdani MM, et al. The Burden of Opioid-Related Mortality in the United States. *JAMA Netw. Open.* 2018;1(2):e180217.

3. Gomes T, Khuu W, Martins D, et al. Contributions of prescribed and non-prescribed opioids to opioid related deaths: population based cohort study in Ontario, Canada. *BMJ.* 2018;362:k3207.

4. Brummett CM, Waljee JF, Goesling J, et al. New Persistent Opioid Use After Minor and Major Surgical Procedures in US Adults. *JAMA Surg.* 2017;152(6):e170504.

5. Jiang X, Orton M, Feng R, et al. Chronic Opioid Usage in Surgical Patients in a Large Academic Center. *Ann Surg.* 2017;265(4):722–727.

6. Cauley CE, Anderson G, Haynes AB, et al. Predictors of in-hospital postoperative opioid overdose after major elective operations. *Ann. Surg.* 2017;265(4):702–708.

7. Jena AB, Goldman D, Karaca-mandic P, et al. Hospital Prescribing of Opioids to Medicare Beneficiaries. *JAMA Intern Med.* 2011;176(7):990–997.

8. Menendez ME, Ring D, Bateman BT. Preoperative Opioid Misuse is Associated With Increased Morbidity and Mortality After Elective Orthopaedic Surgery. *Clin. Orthop. Relat. Res.* 2015;473(7):2402–2412.

9. Adogwa O, Davison MA, Vuong VD, et al. Regional Variation in Opioid Use After Lumbar Spine Surgery. *World Neurosurg.* 2018.

10. Schoenfeld AJ, Nwosu K, Jiang W, et al. Risk Factors for Prolonged Opioid Use Following Spine Surgery, and the Association with Surgical Intensity, among Opioid-Naive Patients. *J. Bone Jt. Surg. - Am. Vol.* 2017;99(15):1247–1252.

11. Yang S, Werner BC. Risk Factors for Prolonged Postoperative Opioid Use After Spinal Fusion for Adolescent Idiopathic Scoliosis. *J. Pediatr. Orthop.* 2018;00(00):1–7.

12. Syed UAM, Aleem AW, Wowkanech C, et al. Neer Award 2018: the effect of preoperative education on opioid consumption in patients undergoing arthroscopic rotator cuff repair: a prospective, randomized clinical trial. *J. Shoulder Elb. Surg.* 2018;27(6):962–967.

13. Alter TH, Ilyas AM. A Prospective Randomized Study Analyzing Preoperative Opioid Counseling in Pain Management After Carpal Tunnel Release Surgery. *J. Hand Surg. Am.* 2017;42(10):810–815.

14. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* 2016;375(13):1216–1219.

15. Cabitza F, Locoro A, Banfi G. Machine Learning in Orthopedics: A Literature Review. *Front. Bioeng. Biotechnol.* 2018;6(June).

16. Senders JT, Staples PC, Karhade A V., et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg.* 2018;109(Ml):476-486.e1.

17. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models : A framework for some traditional and novel measures. *Epidemiology.* 2010;21(1):128–138.

18. Greenwell BM, Boehmke BC, McCarthy AJ. A Simple and Effective Model-Based Variable Importance Measure. 2018:1–27.

19. Ribeiro MT, Singh S, Guestrin C. Model-Agnostic Interpretability of Machine Learning. 2016;(Whi).

20. Jain N, Brock JL, Phillips FM, et al. Chronic preoperative opioid use is a risk factor for increased complications, resource use, and costs after cervical fusion. *Spine J.* 2018.

21. Kalakoti P, Hendrickson NR, Bedard NA, et al. Opioid Utilization Following Lumbar Arthrodesis. *Spine (Phila. Pa. 1976).* 2018;43(17):1208–1216.

22. Sun EC, Darnall BD, Baker LC, et al. Incidence of and Risk Factors for Chronic Opioid Use Among Opioid-Naive Patients in the Postoperative Period. *JAMA Intern. Med.* 2016;176(9):1286.

23. Pugely AJ, Bedard NA, Kalakoti P, et al. Opioid use following cervical spine surgery: trends and factors associated with long-term use. *Spine J.* 2018;7828.

24. Singh JA, Lewallen DG. Predictors of use of pain medications for persistent knee pain after primary Total Knee Arthroplasty: a cohort study using an institutional joint registry. *Arthritis Res. Ther.* 2012;14(6):R248.

25. Armaghani SJ, Lee DS, Bible JE, et al. Preoperative opioid use and its association with perioperative opioid demand and postoperative opioid independence in patients undergoing spine surgery. *Spine (Phila. Pa. 1976).* 2014;39(25):E1524–E1530.

26. Jiang X, Orton M, Feng R, et al. Chronic opioid usage in surgical patients in a large academic center. *Ann. Surg.* 2017;265(4):722–727.

27. Han B, Compton WM, Blanco C, et al. Prescription opioid use, misuse, and use disorders in U.S. Adults: 2015 national survey on drug use and health. *Ann. Intern. Med.* 2017;167(5):293–301.

28. Schoenfeld AJ, Belmont PJ, Blucher JA, et al. Sustained Preoperative Opioid Use Is a Predictor of Continued Use Following Spine Surgery. *J. Bone Jt. Surg.* 2018;100(11):914–921.

29. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Med.* 2018;16(120).

30. Gangavalli A, Malige A, Terres G, et al. Misuse of Opioids in Orthopaedic Postoperative Patients. *J. Orthop. Trauma.* 2017;31(4):e103–e109.

31. Gui C, Chan V. Machine learning in medicine. *Univ. West. Ont. Med. J.* 2017;86(2):76–78.

32. Helmerhorst GTT, Vranceanu AM, Vrahas M, et al. Risk factors for continued opioid use one to two months after surgery for musculoskeletal trauma. *J. Bone Jt. Surg. - Ser. A.* 2014;96(6):495–499.

Table 1. Baseline characteristics	
Variable	**All Patients** **(n = 878)**
Preoperative variables	
	Median (IQR)
Age (years)	67 (60 - 74)
Hospitalization (days)	4 (3 - 5)
Estimated blood loss (mL)†	250 (100 - 460)
Body mass index (in kg/m^2)†	29 (25 - 34)
Hemoglobin levels (g/dL)†	13 (10 - 13)
White blood cell count (10^3/µL)†	6.9 (5.7 - 8.4)
Creatinine levels (mg/dL)†	0.88 (0.75 - 1.1)
Platelet count (10^3/mm^3)†	251 (209 - 296)
Albumin levels (g/dL)†	3.8 (3.4 - 4.2)
Median household income ($)	79,914 (63,740 – 102,008)
	N (%)
Men	278 (32)
Race	
Caucasian	781 (91)
African-American	39 (4.6)
Hispanic	17 (2.0)
Asian	13 (1.5)
Other	5 (0.6)
Smoking	
Yes	44 (6.9)
No	292 (45)
Former	302 (47)
Marital status	
Married	519 (62)
Single	110 (13)
Widowed	79 (9.5)
ASA Class	
I	11 (1.9)
II	378 (64)
III	201 (34)
IV	2 (0.3)
Prior injection	212 (24)
Prior physical therapy	248 (28)

Highest Meyerding grade	
I	798 (91)
II	79 (9.0)
Medication use	
Angiotensin-converting enzyme	101 (12)
Antidepressant	19 (2.2)
Beta-blocker	122 (14)
Benzodiazepines	49 (5.6)
Preoperative opioid	303 (35)
Statin	195 (22)
Steroid	19 (2.2)
Comorbidities	
Diabetes	133 (15)
Renal failure	77 (8.8)
Depression	166 (19)
Myocardial infarction	76 (8.7)
Congestive heart failure	14 (1.6)
HIV	14 (1.6)
Drug abuse	24 (2.7)
Alcohol abuse	166 (19)

ASA = American Society of Anesthesiologists; g/dL = gram per deciliter; HIV = Human Immunodeficiency Virus; IQR = Interquartile range; kg/m2 = kilogram per square meter; mg/dL = milligram per deciliter; mL = milliliter; mm3 = cubic millimeter; µL = microliter;

** Hospitalization was missing in 2 cases (0.2%), estimated blood loss in 55 cases (6.3%), body mass index in 2 cases (0.2%), hemoglobin levels in 148 cases (17%), white blood cell count in 175 cases (20%), creatinine levels in 174 cases (20%), platelet count in 177 cases (20%), albumin levels in 627 (71%), median household income in 20 cases (2.3%), race in 23 cases (2.6%), and ASA class in 285 cases (32%)*

Table 2. Model performance for prolonged opioid dependence on testing set

Performance Metric	Neural Network	Gradient Boosting Machine	Random Forest	Support Vector Machine
C-statistic	0.765	0.82	0.80	0.59
Calibration slope	0.64	1.3	0.86	11.0
Calibration intercept	-0.55	0.66	0.10	21.8
Brier Score	0.083	0.076	0.079	0.087
Null Model Brier Score	0.088			

Performance Metric	Neural Network	Gradient Boosting Machine	Random Forest	Support Vector Machine
C-statistic	0.75	0.82	0.81	0.59
Calibration slope	0.64	1.4	1.00	11.0
Calibration intercept	0.18	-0.04	0.22	-0.04
Brier Score	0.083	0.076	0.079	0.087
Null Model Brier Score	0.088			

Part III

Validation and Implementation

Chapter 7

A Machine Learning Algorithm for Predicting Prolonged Postoperative Opioid Prescription after Lumbar Disc Herniation Surgery. An External Validation Study using 1,316 Patients from a Taiwanese Cohort

Yen HK, Ogink PT, Huang CC, Groot OQ, Su CC, Chen SF, Chen CW, Karhade AV, Peng KP, Lin WH, Chiang H, Yang JJ, Dai SH, Yen MH, Verlaan JJ, Schwab JH, Wong TH, Yang SH, Hu MH

Spine J. 2022 Jul;22(7):1119-1130

Abstract

Objective Preoperative prediction of prolonged postoperative opioid prescription helps identify patients for increased surveillance after surgery. The SORG machine learning model has been developed and successfully tested using 5,413 patients from the United States (US) to predict the risk of prolonged opioid prescription after lumbar discectomy. However, external validation is an often-overlooked element in the process of incorporating prediction models in current clinical practice. This cannot be stressed enough in prediction models where medicolegal and cultural differences may play a major role. Therefore, the authors aimed to investigate the generalizability of the American prediction model SORG to a Taiwanese patient cohort.

Methods Retrospective review was conducted at a large academic medical center in Taiwan to identify patients 18 years or older undergoing initial operative management for lumbar disc herniation between 2010 and 2018. The primary outcome of interest was prolonged opioid prescription defined as continuing opioid prescription to at least 90 to 180 days after the index surgery**.** Discrimination (area under the curve [AUROC] and precision-recall curve [AUPRC]), calibration, overall performance (Brier score), and decision curve analysis were used to assess the performance of the SORG ML algorithm in the validation cohort.

Results Overall, 1.1316 patients were identified with sustained postoperative opioid prescription in 41 (3.1%) patients. The validation cohort differed from the development cohort on several variables including 93% of Taiwanese patients receiving NSAIDS preoperatively compared with 22% of American patients, while 30% of Taiwanese patients received opioids versus 25% in the US. Despite these differences, the SORG prediction model retained good discrimination (AUROC of 0.76 and AUPRC of 0.33) and good overall performance (Brier score of 0.028 compared with null model Brier score of 0.030) while somewhat overestimating the chance of prolonged opioid use (calibration slope of 1.07 and calibration intercept of -0.87). Decision-curve analysis showed the SORG model was suitable for clinical use.

Conclusions Despite differences at baseline and a very strict opioid policy the SORG algorithm for prolonged opioid use after surgery for lumbar discectomy has good discriminative abilities and good overall performance in a Han Chinese patient group in Taiwan. This freely available digital application can be used to identify high-risk patients and tailor prevention policies for these patients that may mitigate the long-term adverse consequence of opioid dependence: https://sorg-apps.shinyapps.io/lumbardiscopioid/.

Introduction

Lumbar disc herniations (LDH) are the third-most-common etiology for low back pain, following lumbar strain and degenerative processes of discs and facets.[3,10] The general treatment of LDH starts with rest, physical therapy, and appropriate pain management. Surgical intervention is indicated for patients who do not respond to conservative treatment after a minimum of 6 weeks, or patients accompanied with progressive neurological deficit.[43]

Patients who have had surgical intervention for LDH are often prescribed opioids postoperatively.[4,28,32,58,63] Despite the obvious benefit of pain control, opioids may have several side effects, including increased risk of ileus, increased infection risk, prolonged hospital stay, and higher readmission rate.[12,16,48,55] These adverse events could dramatically decrease the patients' life quality and lead to unnecessary medical cost.[11,30,46,52] Importantly, prolonged opioid use is associated with future opioid dependence, which has an enormous societal impact.[15,27] Therefore, it could be beneficial to identify patients at risk for prolonged opioid use[20,31,35,38,40,47,59]. Furthermore, a preoperative prediction tool for evaluating the probability of prolonged opioid use could aid in making more individualized treatment plans[2,1462].

Karhade et al. used machine learning to develop the Skeletal Oncology Research Group (SORG) prediction model which predicts the risk of prolonged opioid prescription after surgery for LDH using 5,413 patients from the US.[25] The prediction model displayed excellent performance on internal validation. However, the use of opioids is affected by a multitude of factors that can substantially differ across countries.[39] Therefore, external validation with local patient cohorts is critical before implementation. Our primary aim of this study was to externally validate the SORG prediction model with a Taiwanese patient cohort. We hypothesize the prediction model does not perform well on external validation considering there are very strict opioid regulations in Taiwan and a substantial lower opioid consumption compared to the US and Europe.[6,44] Additionally, only 7.7% of the patients in the development study had prolonged postoperative opioid prescription, which suggests an imbalanced outcome. Karhade et al. only reported the area under the receiver operating characteristic curve (AUROC) to evaluate the discriminatory ability. However, AUROC might give an overoptimistic view of the performance of the diagnostic test when the outcome is uncommon; the area under the precision-recall curve (AUPRC) should be also reported according to earlier studies.[37,42] Therefore, our second aim was to determine the discriminatory ability of the SORG prediction model by AUPRC. The tertiary aim was to provide an easy step-by-step guide for validating open accessible prediction models that other institutions can use.

Materials and Methods

Guideline

This retrospective external validation study was performed under the guidance of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement.[9,29] This study was approved by our institutional review board (202105061RINA) and the design did not violate the Declaration of Helsinki.[13]

Participants

The index surgery was defined as the first surgery for lumbar disc herniation at our institution. Inclusion criteria of patients in this study include: (1) patients 20 years or older, and (2) lumbar spine surgery with diagnosis of disc herniation between January 1st, 2010, and December 31st, 2018. Exclusion criteria include: (1) index surgery performed with an additional diagnosis of trauma, infection, tumor, inflammatory status, pseudoarthrosis, scoliosis, and spondylolisthesis; and (2) patients without complete surgery.

Outcome

The prolonged opioid prescription was defined as continuing opioid prescription to at least 90 to 180 days after the index surgery.[5,27] The full list of opioids medications is provided in the development study.[25]

Predictors

The following predictors were retrieved retrospectively from the medical records: age (years), gender, marital status, veteran, ethnicity, procedure method (fusion, anterior approach, multilevel surgery, instrumentation), previous spinal surgery, white blood cell count ($10^3/\mu L$), hemoglobin (g/dL), platelet count (10^3 /mL), creatinine (mg/dL), neighborhood characteristics from Taiwan National Department of Household Registration online database based on patient's living area zip code (median household income, educational level, median age, neighborhood unemployment rate, population density), preoperative opioids, other preoperative medications (angiotensin-converting enzyme inhibitor, angiotensin receptor blocker, antidepressants, anti-psychotics, beta-2 agonists, beta-blockers, benzodiazepines, gabapentin and pregabalin, immunosuppressants, nonsteroidal anti-inflammatory drugs), and preoperative comorbidities (tobacco use, drug abuse, diabetes, arrhythmias, valvular heart disease, peripheral vascular disease, renal failure, liver disease, solid tumors, depression, psychoses, myocardial infarction, congestive heart failure, cerebrovascular accidents, chronic obstructive pulmonary disease). Preoperative opioid use was divided into three categories: continuous prescription more than 180 days before the index surgery, continuous prescription less than 180 days before the index surgery, and no opioid prescription within a year prior to index surgery. To avoid bias, records of opioid or other medications of interest use in 2009 and 2019 were also retrieved. The same definitions for outcome and predictor variables were used as the original development study.[25] The authors of the development study were not part of the data extraction or analysis.

Data Collection and Missing Data

The proportions of missing data were all less than 30%[49], and the MissForest algorithm was applied to impute the missing data for the following variables: hemoglobin in 28 patients (2.1%); white blood cell in 28 patients (2.1%); platelet in 28 patients (2.1%); creatinine in 32 patients (2.4%); median age of neighborhood in 7 patients (0.5%); and unemployment rate in 2 patients (0.2%).

Statistical Analysis and Methods

All predictions were retrieved from the online SORG-MLA model at **https://sorg-apps.shinyapps.io/lumbardiscopioid/.** The website allows to fill out all variables in order to get a prediction. In this example we used a male, without previous spine surgery, who has used opioids >180 days in addition to gabapentin and benzodiazepines. He smokes, has no history of drug abuse and his hemoglobin level is 13g/dL and white blood cell count is 6 10^3/µL. Figure 1 shows he has a 29% chance of prolonged opioid use postoperatively with certain predictors supporting prolonged opioid use and others opposing.

A comparison of baseline characteristics between the validation cohort and the developmental cohort was performed. Continuous baseline characteristics were compared by one-way median tests and categorical baseline characteristics were analyzed by chi-square tests with Yate's correction (if applicable). The actual and average predicted prolonged opioid prescription rates were compared with one sample t-tests. We also applied discrimination (measured by AUROC and AUPRC), calibration, overall performance, and decision curve analysis (DCA) to externally evaluate SORG-MLA's performance. A two-tailed p value = 0.05 was considered significant. R for Mac (version 4.0.4, R Core Team), along with its packages of missForest, risk model decision analysis (rmda), Precision-Recall and ROC Curves for Weighted and Unweighted Data (PRROC), Tools for Descriptive Statistics (DescTools), and CalibrationCurves (downloaded through Github), was used for all statistical analyses.

The measurement of how well a model separates risk is referred to as discrimination. AUROC is usually a useful first step in discrimination analysis and the interpretation is relatively simple. An AUROC of 0.7 is usually considered clinically acceptable, while an AUROC of 0.5 indicates a prediction no better than a random guess.[19] However, an AUROC might give an overly optimistic evaluation when the event of interest is rare since the huge number of true-negative cases might dilute and mask the influence of false-positive cases. Unlike AUROC focused on sensitivity and specificity, AUPRC concerns the tradeoff between precision (also known as sensitivity) and recall (also known as positive predictive value), and the false-positive cases were evaluated along with the true-positive cases to avoid masking and diluting the influence of false-positive cases. Therefore, it is believed that the interpretation of AUPRC might avoid such misevaluation and might also be a suitable metric for discrimination analysis.[37,64] Whereas, the interpretation of an AUPRC is relatively complex. An AUPRC equal to the prevalence of the outcome indicates the baseline curve, and an AUPRC of 1 suggests perfect a discrimination. The interpretation should be derived from the improvement from the baseline AUPRC, namely the observed prevalence of the outcome (e.g. the prevalence of prolonged opioid prescription rate of 3.1% in this cohort).

Calibration concerns average risk in a population and a well-calibrated model predicts close to *x* individuals have the event of interest, for every 100 individuals given a probability of *x*%. A calibration analysis could be performed by plotting a calibration plot, and a perfect calibration has an intercept of 0 and a slope of 1. A positive intercept indicates the underestimation by the model while a negative intercept indicates overestimating the outcome of interest.[28] Overall performance analysis is measured by Brier scores, which capture both the discrimination and calibration. Brier scores of 0 suggest the best prediction while 1 suggests the worst. A null-model Brier score, which gives a default prediction equal to the prevalence of the outcome, should be considered as the benchmark.

A decision curve was plotted concerning the clinical net benefit over different threshold probabilities. It provides a more comprehensive evaluation of a model's clinical utility by taking the clinical cost and benefit into account. It could also be used in the shared-decision making process by evaluating each patient's will and value and gives an appropriate treatment plan based on their personalized threshold probability. The user of the model can decide which threshold probability (such as, the ratio of potential risk to the potential benefit) of a treatment is important or applicable and determine whether the model is valuable at that threshold and see what the predicted net benefit would be. In general, if the potential risks associated with a treatment are high, such as, performing extensive surgery in a fragile patient, a higher threshold possibility should be chosen for decision-making.[26,53,54] In contrast, if the harm of a treatment modality is relatively limited, for example, antibiotics for infection, the clinician may choose a lower threshold probability. However, since there is no consensus method to compare two decision curves, whether the interpretation of DCA should prioritize calibration or discrimination analysis is still debatable.[51,56]

Results

Overall, 1,316 patients underwent surgery for lumbar disc herniation and 41 (3.3%) patients had sustained postoperative opioid prescription at 90 to 180 days after surgery. Five hundred (38%) patients were female, and the median age was 53 (interquartile range [IQR]=39-63). NSAIDs were the most preoperative prescribed medicine in 1,227 (93%) patients. Preoperative opioid use was none in 927 (70%) patients, 180 days or less in 382 (29%) patients, and greater than 180 days in 7 (0.5%) patients (Table 1).

In comparison to the developmental cohort for the SORG-ML algorithm, the population in this validation study were older, more inpatient dispositions, lower income, higher preoperative NSAID use, and less depression as comorbidity. All patients in the validation cohort were on national health insurance compared with the American development cohort that consisted of various types of insurance.

The SORG-ML algorithm achieved an AUROC of 0.76 (Figure 2) and AUPRC of 0.33 (Figure 3). The algorithm overestimated the observed proportion of patients with sustained opioid prescription (Figure 4). This was reflected in the negative calibration intercept of -0.87 with calibration slope of 1.07. The actual prolonged opioid prescribed rate was also lower than the predicted prolonged opioid prescribed rate (3.1% versus 6.8%; one-sample t-test $p < 0.01$). The raw Brier score was 0.028 relative to the null model Brier score of 0.030 (Table 2). On decision curve analysis in the validation cohort, the algorithm provided greater net benefit than the default strategies of changing management for no patients or all patients across all threshold probabilities (Figure 5).

An easy step-by-step guide for validation of open accessible prediction models is provided in the supplementary material using a dummy dataset. Other institutions can implement this easy-to-use roadmap to facilitate external validation of prediction models (Supplementary material).

Discussion

We sought to assess how the SORG algorithm predicting prolonged opioid use after surgery for LDH performed in a dataset consisting of patients from outside the US, specifically Taiwan. We found that the SORG prediction model retained good discriminative ability, good overall performance while somewhat overestimating the chance of prolonged opioid use. Furthermore, decision-curve analysis shows this model is suitable for clinical use. External validation is an often-overlooked element in the process of incorporating prediction models in current clinical practice.[34,41,50] While there has been an abundance of ML prediction models in orthopedics as of late,[23,24,36,53,61] a study by Groot et al.[17] concluded a mere 10 of 59 available prediction models for orthopedic surgical outcomes were externally validated; none of which were performed on the available opioid use prediction models. Similar to those 10 available external validations our external validation showed an AUROC > 0.70 and less than 0.10 decreased performance compared to the development study. Furthermore, the discrimination analysis, measured by AUPRC, also revealed well-performed results compared to the benchmark (the prevalence of prolonged opioid use in this cohort).

While it is preferable to externally validate all prediction models, this cannot be stressed enough in prediction models where medicolegal and cultural differences play a major role. There are major differences in opioid prescription and use between countries and even within countries.[1,57] There is a very strict drug policy having already put fierce restrictions on opioid use in Taiwan since 1996.[7,22,27] For instance, an oral high-potency opioid use, such as oxycodone, for non-cancer patients is generally not recommended. Furthermore, hospitals must submit a patient evaluation every three months with regard to chronic opioid treatment to the Taiwanese Food and Drug Administration. The effects of these policies can be seen in the baseline table comparing the patient groups. Ninety-three percent of patients were receiving NSAIDs preoperatively in Taiwan compared with 22% in the US, while 30% of Taiwanese patients received opioids versus 35% in the US[21,33,45]. Furthermore, there was a substantial difference in the duration of opioid prescription preoperatively between the two countries. Besides, due to the strict policy, it was almost impossible to get any opioids drug, other than codeine, from pharmacies in Taiwan. All such drugs for pain relief must be evaluated by physicians. Also, due to the flat medical transfer system and the more affordable medical cost to visit a tertiary center in Taiwan (around 500 NTD, which is only three-times the price of a Big Mac combo in Taiwan), most patients were postoperatively followed up at tertiary centers instead of local hospitals. These might make the outcome evaluation of prolonged opioid use more precise in Taiwan compared to the evaluation in the USA.[25]

Besides duration of preoperative opioid prescription, two other variables were deemed particularly important by the development study: the use of instrumentation and depression as a comorbidity. While there was no statistically significant difference in the use of instrumentation, US patients were 3.5 times more likely to have depression at baseline. Remarkably, despite these distinct difference in 2 of the 3 important variables at baseline, the SORG algorithm prediction model was still able to retain a good discrimination and overall performance. The slight overestimation of prolonged opioid use by the SORG algorithm can perhaps be explained by the aforementioned quarterly reports that have to be submitted in Taiwan. Possibly, these types of evaluations stimulate physicians to rethink their (postoperative) pain regimen. Another strength of this study is the use of the area under curve precision-recall plot (AUPRC). Saito et al. showed that the AUROC can give overoptimistic estimations if the outcome is unbalanced.[42] Considering the outcome of prolonged opioid use is unbalanced in this cohort (7% of patients had a positive outcome) using the AUPRC is more apt to evaluate this algorithm.

Our study also has several limitations. First, the initial study of Karhade et al. covered the time period of 2000 until 2018. Our validation cohort was based on patients who had surgery between 2010 and 2018. This means a larger section in our validation cohort was having surgery during a timespan in which the opioid crisis was already being brought to the medical world's attention. During the last decade prescription policies have already changed in the US as well, as evidenced by the peak in 2010 of morphine milligram equivalents per capita.[18] This could be another potential cause of the algorithm overestimating the chance of long-term opioid use not just in this Taiwanese population but in more contemporary American patient groups as well. Despite this difference in time span, the SORG algorithm has retained a good performance. It would be of interest to perform a temporal validation in a US cohort as well. Second, this validation is performed in a single institution in Taiwan with a predominantly Han Chinese patient cohort. Considering the strict limitations on opioid prescription in Taiwan, the generalizability of this validation in other Han Chinese patient groups or other Asian countries in general is questionable. An outcome that is so dependent on societal and medicolegal difference should always be externally validated before being implemented. Therefore, external validation remains to be performed in non-American and non-Taiwanese patient groups. Third, while we found a number of differences in demographic features between the validation and development cohort, a substantial difference we could not assess may have been the timing of surgical treatment between the US and Taiwan. Historically, there has been a higher rate of spine surgery in the US compared to other countries.[8,60] Taiwanese surgeons potentially continued conservative treatment for a longer period with a group of patients' complaints subsiding without surgery. However, despite the inability to accurately assess disease characteristics the model performed well in our validation group.

Conclusion

Despite major differences at baseline and a very strict opioid policy the SORG algorithm for prolonged opioid use after surgery for LDH has good discriminative abilities and good overall performance in a Han Chinese patient group in Taiwan. Clinicians in Taiwan can use this algorithm to identify high-risk patients and tailor prevention policies to these patients.

References

1. Adogwa O, Davison MA, Vuong VD, Desai SA, Lilly DT, Moreno J, et al: Regional Variation in Opioid Use After Lumbar Spine Surgery. World Neurosurg 121:e691-e699, 2019
2. Anderson AB, Grazal CF, Balazs GC, Potter BK, Dickens JF, Forsberg JA: Can Predictive Modeling Tools Identify Patients at High Risk of Prolonged Opioid Use After ACL Reconstruction? Clin Orthop Relat Res 478:0-1618, 2020
3. Andersson GB: Epidemiological features of chronic low-back pain. Lancet 354:581-585, 1999
4. Bot AG, Bekkers S, Arnstein PM, Smith RM, Ring D: Opioid use after fracture surgery correlates with pain intensity and satisfaction with pain relief. Clin Orthop Relat Res 472:2542-2549, 2014
5. Brummett CM, Waljee JF, Goesling J, Moser S, Lin P, Englesbe MJ, et al: New Persistent Opioid Use After Minor and Major Surgical Procedures in US Adults. JAMA Surg 152:e170504, 2017
6. Chen TC, Wang TC, Lin CP, Bonar K, Ashcroft DM, Chan KA, et al: Increasing tramadol utilisation under strict regulatory control of opioid prescribing - A cross-sectional study in Taiwan from 2002 through 2016. J Formos Med Assoc 120:810-818, 2021
7. Cheng IC, Chang CS, Tsay WI: Long-term usage of narcotic analgesics by chronic intractable noncancer pain patients in Taiwan from 2003 to 2012. J Formos Med Assoc 115:773-778, 2016
8. Cherkin DC, Deyo RA, Loeser JD, Bush T, Waddell G: An international comparison of back surgery rates. Spine (Phila Pa 1976) 19:1201-1206, 1994

9. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 13:1, 2015

10. Deyo RA, Weinstein JN: Low back pain. N Engl J Med 344:363-370, 2001

11. DiGiorgio AM, Stein R, Morrow KD, Robichaux JM, Crutcher CL, Tender GC: The increasing frequency of intravenous drug abuse-associated spinal epidural abscesses: a case series. Neurosurg Focus 46:E4, 2019

12. Dunn LK, Durieux ME, Fernandez LG, Tsang S, Smith-Straesser EE, Jhaveri HF, et al: Influence of catastrophizing, anxiety, and depression on in-hospital opioid consumption, pain, and quality of recovery after adult spine surgery. J Neurosurg Spine 28:119-126, 2018

13. Gandevia B, Tovell A: Declaration of Helsinki. Med J Aust 2:320-321, 1964

14. Gifford C, Minnema AJ, Baum J, Humeidan ML, Vazquez DE, Farhadi HF: Development of a postoperative ileus risk assessment scale: identification of intraoperative opioid exposure as a significant predictor after spinal surgery. J Neurosurg Spine:1-8, 2019

15. Gomes T, Tadrous M, Mamdani MM, Paterson JM, Juurlink DN: The Burden of Opioid-Related Mortality in the United States. JAMA Netw Open 1:e180217, 2018

16. Goyal A, Payne S, Sangaralingham LR, Jeffery MM, Naessens JM, Gazelka HM, et al: Incidence and risk factors for prolonged postoperative opioid use following lumbar spine surgery: a cohort study. J Neurosurg Spine:1-9, 2021

17. Groot OQ, Bindels BJJ, Ogink PT, Kapoor ND, Twining PK, Collins AK, et al: Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. Acta Orthop 92:385-393, 2021

18. Guy GP, Jr., Zhang K, Bohm MK, Losby J, Lewis B, Young R, et al: Vital Signs: Changes in Opioid Prescribing in the United States, 2006-2015. MMWR Morb Mortal Wkly Rep 66:697-704, 2017

19. Holmberg L, Vickers A: Evaluation of prediction models for decision-making: beyond calibration and discrimination. PLoS Med 10:e1001491, 2013

20. Hozack BA, Rivlin M, Lutsky KF, Graham J, Lucenti L, Foltz C, et al: Preoperative Exposure to Benzodiazepines or Sedative/hypnotics Increases the Risk of Greater Filled Opioid Prescriptions After Surgery. Clin Orthop Relat Res 477:1482-1488, 2019

21. Hung WT, Chen HM, Wu CH, Hsu WM, Lin JW, Chen JS: Recurrence rate and risk factors for recurrence after thoracoscopic surgery for primary spontaneous pneumothorax: A nationwide population-based study. J Formos Med Assoc 120:1890-1896, 2021

22. Kang KH, Kuo LF, Cheng IC, Chang CS, Tsay WI: Trends in major opioid analgesic consumption in Taiwan, 2002-2014. J Formos Med Assoc 116:529-535, 2017

23. Karhade AV, Cha TD, Fogel HA, Hershman SH, Tobert DG, Schoenfeld AJ, et al: Predicting prolonged opioid prescriptions in opioid-naive lumbar spine surgery patients. Spine J 20:888-895, 2020

24. Karhade AV, Chaudhary MA, Bono CM, Kang JD, Schwab JH, Schoenfeld AJ: Validating the Stopping Opioids after Surgery (SOS) score for sustained postoperative prescription opioid use in spine surgical patients. Spine J 19:1666-1671, 2019

25. Karhade AV, Ogink PT, Thio Q, Cha TD, Gormley WB, Hershman SH, et al: Development of machine learning algorithms for prediction of prolonged opioid prescription after surgery for lumbar disc herniation. Spine J 19:1764-1771, 2019

26. Karhade AV, Thio Q, Ogink PT, Bono CM, Ferrone ML, Oh KS, et al: Predicting 90-Day and 1-Year Mortality in Spinal Metastatic Disease: Development and Internal Validation. Neurosurgery 85:E671-E681, 2019

27. Lin TC, Ger LP, Pergolizzi JV, Jr., Raffa RB, Wang JO, Ho ST: Long-term use of opioids in 210 officially registered patients with chronic noncancer pain in Taiwan: A cross-sectional study. J Formos Med Assoc 116:257-265, 2017

28. Lovecchio F, Premkumar A, Stepan JG, Mejia D, Stein D, Patel DV, et al: Opioid Consumption Patterns After Lumbar Microdiscectomy or Decompression. Spine (Phila Pa 1976) 44:1599-1605, 2019

29. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al: Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res 18:e323, 2016

30. Many GM, Drazin D: Editorial. Is the rise in spinal infections an unexpected consequence of the opioid epidemic? Neurosurg Focus 46:E5, 2019

31. Massie L, Gunaseelan V, Waljee J, Brummett C, Schwalb JM: Relationship between initial opioid prescription size and likelihood of refill after spine surgery. Spine J 21:772-778, 2021

32. Mastronardi L, Pappagallo M, Puzzilli F, Tatta C: Efficacy of the morphine-Adcon-L compound in the management of postoperative pain after lumbar microdiscectomy. Neurosurgery 50:518-524; discussion 524-515, 2002

33. McDonald EL, Daniel JN, Rogero RG, Shakked RJ, Nicholson K, Pedowitz DI, et al: How Does Perioperative Ketorolac Affect Opioid Consumption and Pain Management After Ankle Fracture Surgery? Clin Orthop Relat Res 478:144-151, 2020

34. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al: Risk prediction models: II. External validation, model updating, and impact assessment. Heart 98:691-698, 2012

35. O'Connell C, Azad TD, Mittal V, Vail D, Johnson E, Desai A, et al: Preoperative depression, lumbar fusion, and opioid use: an assessment of postoperative prescription, quality, and economic outcomes. Neurosurg Focus 44:E5, 2018

36. Ogink PT, Groot OQ, Karhade AV, Bongers MER, Oner FC, Verlaan JJ, et al: Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review. Acta Orthop 92:526-531, 2021

37. Ozenne B, Subtil F, Maucort-Boulch D: The precision--recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 68:855-859, 2015

38. Patel AA, Walker CT, Prendergast V, Radosevich JJ, Grimm D, Godzik J, et al: Patient-Controlled Analgesia Following Lumbar Spinal Fusion Surgery Is Associated With Increased Opioid Consumption and Opioid-Related Adverse Events. Neurosurgery 87:592-601, 2020

39. Rahavard BB, Candido KD, Knezevic NN: Different pain responses to chronic and acute pain in various ethnic/racial groups. Pain Manag 7:427-453, 2017

40. Rajamaki TJ, Moilanen T, Puolakka PA, Hietaharju A, Jamsen E: Is the Preoperative Use of Antidepressants and Benzodiazepines Associated with Opioid and Other Analgesic Use After Hip and Knee Arthroplasty? Clin Orthop Relat Res 479:2268-2280, 2021

41. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M: External validation of prognostic models: what, why, how, when and where? Clin Kidney J 14:49-58, 2021

42. Saito T, Rehmsmeier M: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10:e0118432, 2015

43. Schoenfeld AJ, Weiner BK: Treatment of lumbar disc herniation: Evidence-based practice. Int J Gen Med 3:209-214, 2010

44. Scholten W: Improving access to adequate pain management in Taiwan. Acta Anaesthesiol Taiwan 53:62-65, 2015

45. Shao CH, Tai CH, Lin FJ, Wu CC, Wang JT, Wang CC: Comparison of risk of acute kidney injury between patients receiving the combination of teicoplanin and

piperacillin/tazobactam versus vancomycin and piperacillin/tazobactam. J Formos Med Assoc, 2021

46. Sharma M, Ugiliweneza B, Aljuboori Z, Boakye M: Health care utilization and overall costs based on opioid dependence in patients undergoing surgery for degenerative spondylolisthesis. Neurosurg Focus 44:E14, 2018

47. Sharma M, Ugiliweneza B, Aljuboori Z, Nuno MA, Drazin D, Boakye M: Factors predicting opioid dependence in patients undergoing surgery for degenerative spondylolisthesis: analysis from the MarketScan databases. J Neurosurg Spine 29:271-278, 2018

48. Sodhi N, Anis HK, Acuna AJ, Vakharia RM, Gold PA, Garbarino LJ, et al: Opioid Use Disorder Is Associated with an Increased Risk of Infection after Total Joint Arthroplasty: A Large Database Study. Clin Orthop Relat Res 478:1752-1759, 2020

49. Stekhoven DJ, Buhlmann P: MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 28:112-118, 2012

50. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al: Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21:128-138, 2010

51. Talluri R, Shete S: Using the weighted area under the net benefit curve for decision curve analysis. BMC Med Inform Decis Mak 16:94, 2016

52. Tank A, Hobbs J, Ramos E, Rubin DS: Opioid Dependence and Prolonged Length of Stay in Lumbar Fusion: A Retrospective Study Utilizing the National Inpatient Sample 2003-2014. Spine (Phila Pa 1976) 43:1739-1745, 2018

53. Thio Q, Karhade AV, Bindels BJJ, Ogink PT, Bramer JAM, Ferrone ML, et al: Development and Internal Validation of Machine Learning Algorithms for Preoperative Survival Prediction of Extremity Metastatic Disease. Clin Orthop Relat Res 478:322-333, 2020

54. Tseng TE, Lee CC, Yen HK, Groot OQ, Hou CH, Lin SY, et al: International Validation of the SORG Machine-learning Algorithm for Predicting the Survival of Patients with Extremity Metastases Undergoing Surgical Treatment. Clin Orthop Relat Res, 2021

55. Turcotte J, Sanford Z, Broda A, Patton C: Centers for Medicare & Medicaid Services Hierarchical Condition Category score as a predictor of readmission and reoperation following elective inpatient spine surgery. J Neurosurg Spine:1-7, 2019

56. Vickers AJ, Elkin EB: Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26:565-574, 2006

57. Wagemaakers FN, Hollingworth SA, Kreijkamp-Kaspers S, Tee EHL, Leendertse AJ, van Driel ML: Opioid analgesic use in Australia and The Netherlands: a cross-country comparison. Int J Clin Pharm 39:874-880, 2017

58. Walker CT, Gullotti DM, Prendergast V, Radosevich J, Grimm D, Cole TS, et al: Implementation of a Standardized Multimodal Postoperative Analgesia Protocol Improves Pain Control, Reduces Opioid Consumption, and Shortens Length of Hospital Stay After Posterior Lumbar Spinal Fusion. Neurosurgery 87:130-136, 2020

59. Wang MC, Lozen AM, Laud PW, Nattinger AB, Krebs EE: Factors associated with chronic opioid use after cervical spine surgery for degenerative conditions. J Neurosurg Spine:1-8, 2019

60. Weeks WB, Paraponaris A, Ventelou B: Geographic variation in rates of common surgical procedures in France in 2008-2010, and comparison to the US and Britain. Health Policy 118:215-221, 2014

61. Yang JJ, Chen CW, Fourman MS, Bongers MER, Karhade AV, Groot OQ, et al: International external validation of the SORG machine learning algorithms for predicting 90-day and one-year survival of patients with spine metastases using a Taiwanese cohort. Spine J 21:1670-1678, 2021

62. Yang MMH, Riva-Cambrin J, Cunningham J, Jette N, Sajobi TT, Soroceanu A, et al: Development and validation of a clinical prediction score for poor postoperative pain control following elective spine surgery. J Neurosurg Spine:1-10, 2020

63. Yerneni K, Nichols N, Abecassis ZA, Karras CL, Tan LA: Preoperative Opioid Use and Clinical Outcomes in Spine Surgery: A Systematic Review. Neurosurgery 86:E490-E507, 2020

64. Zhou QM, Zhe L, Brooke RJ, Hudson MM, Yuan Y: A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve. Diagn Progn Res 5:13, 2021

Table 1. Comparison between the external validation cohort and the developmental cohort

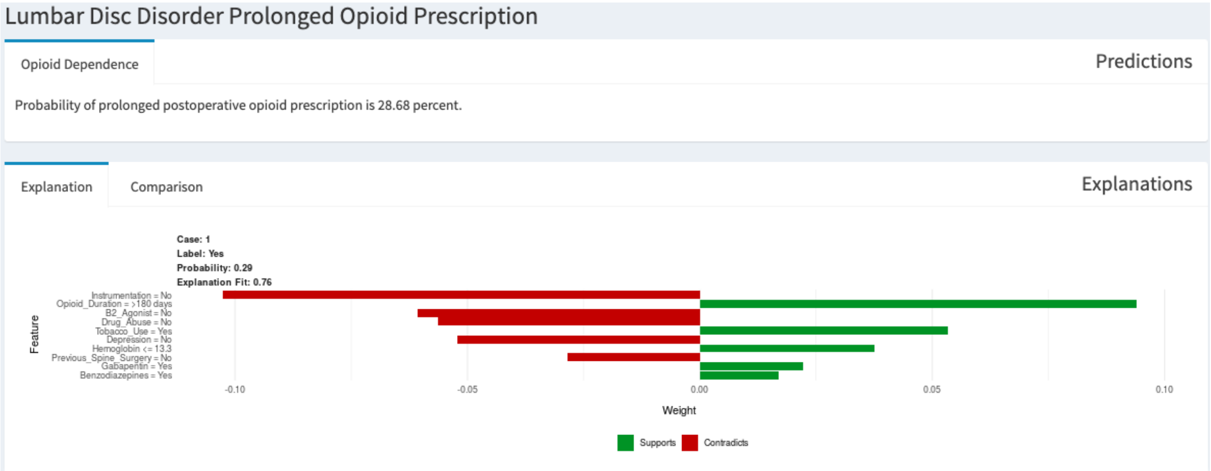
Variable	**n (%) | median (IQR)**		**P value**
	Validation cohort (n=5413)	Developmental cohort (n=1316)	
Age	46.0 (37.0-58.0)	53.0 (39.0-63.0)	<0.01
Female sex	2,424 (44.8)	500 (38.0)	<0.01
Race			
Non-White	581 (11.1) †	1316 (100.0)	
White	4,657 (88.9) †	-	
Ethnicity			
Hispanic	199 (3.8)	-	
Non-Hispanic	5,039 (96.2)	1316 (100.0)	
Married	3,188 (61.5) †	913 (69.4)	<0.01
Veteran	408 (8.0)	9 (0.7)	<0.01
Disposition			
Inpatient	4,177 (77.2)	1316 (100.0)	
Outpatient	1,236 (22.8)	-	
Surgical factors			
Fusion	488 (9.0)	127 (9.7)	0.47
Anterior approach	144 (2.7)	71 (5.4)	<0.01
Instrumentation	446 (8.2)	127 (9.7)	0.10
Multi-level	747 (13.8)	236 (17.9)	<0.01
Previous spine surgery	161 (3.0)	50 (3.8)	0.12
Preoperative lab values			
Hemoglobin (g/dL)	14.1 (13.2-15.1)†	14.3 (13.1-15.4)※	<0.01
White blood cell (103/uL)	7.37 (6.01-8.90)†	6.89 (5.76-8.39)※	<0.01
Platelet (103/uL)	264.0 (222.0-313.0)†	241.0 (206.0-282.0)※	<0.01
Creatinine (mg/dL)	0.90 (0.78-1.01)†	0.90 (0.70-1.00)※	<0.01
Insurance			
Medicaid	375 (6.9)	-	
Medicare	761 (14.1)	-	
Workers compensation	68 (1.3)	-	
NIH	-	1316 (100.0)	
Uninsured	223 (4.1)	-	
Neighborhood characteristics			
Median household income (USD)	80,139 (61,527-99,924)	21,267 (20,000-23,267)	<0.01
Median age (y)	41.1 (36.3-44.5)†	42.0 (40.0-44.0)※	<0.01
High school graduation rate (%)	24 (16-30)†	79 (75-83)	<0.01
Unemployment rate (%)	5.7 (4.6-7.2)†	3.8 (3.8-3.8)※	<0.01
Population density (per square mile)	2,336 (862-7,069)†	43069 (10478-61577)	<0.01
Preoperative medications			
ACEi	251 (4.6)	26 (2.0)	<0.01
ARB	97 (1.8)	161 (12.2)	<0.01
Anti-depressants	523 (9.7)	384 (29.2)	<0.01

Beta-2-agonists	214 (4.0)	6 (0.5)	<0.01
Beta-blockers	260 (4.8)	131 (10.0)	<0.01
Benzodiazepines	787 (14.5)	286 (21.7)	<0.01
Gabapentin	823 (15.2)	121 (9.2)	<0.01
Immunosuppressant	824 (15.2)	202 (15.3)	0.91
NSAID	1,198 (22.1)	1227 (93.2)	<0.01
Opioid	1,874 (34.6)	389 (29.6)	<0.01
Anti-psychotic	129 (2.4)	147 (11.2)	<0.01
Preoperative opioid duration			
Greater than 180 days	1,122 (20.7)	7 (0.5)	<0.01
180 days or less	752 (11.7)	382 (29.0)	<0.01
None	3,656 (67.5)	927 (70.4)	0.04
Comorbidities			
Tobacco use	595 (11.0)	309 (23.5)	<0.01
Drug abuse	114 (2.1)	3 (0.2)	<0.01
Diabetes	428 (7.9)	152 (11.6)	<0.01
Renal failure	92 (1.7)	29 (2.2)	0.22
Depression	713 (13.2)	49 (3.7)	<0.01
Psychoses	38 (0.7)	10 (0.8)	0.82
Myocardial infarction	114 (2.1)	111 (8.4)	<0.01
Congestive heart failure	93 (1.7)	21 (1.6)	0.76
PVD	107 (2.0)	37 (2.8)	0.06
Cerebrovascular accident	100 (1.8)	59 (4.5)	<0.01
COPD	626 (11.6)	69 (5.2)	<0.01
Arrythmias	415 (7.7)	43 (3.3)	<0.01
Valvular heart disease	142 (2.6)	21 (1.6)	0.03
Liver disease	139 (2.6)	60 (4.6)	<0.01
Solid tumors	116 (2.1)	174 (13.2)	<0.01
Prolonged opioid use	359 (4.3)	41 (3.1)	<0.01

NIH, National Health Insurance; ACEI, angiotensin converting enzyme inhibitors; ARB, angiotensin receptor blockers; NSAID, non-steroidal anti-inflammatory drugs; PVD, peripheral vascular; COPD, chronic obstructive pulmonary disease
disease.

[†] The patient number and rate of missing data in the developmental cohort are as follow: race = 175 (3.2%); marital status = 230 (4.2%); white blood cell = 1,303 (23.5%); hemoglobin = 1,236 (22.3%); platelet = 1,306 (23.6%); creatinine = 1,593 (28.8%); median age of neighborhood= 85 (1.5%); high school graduation rate = 78 (1.4%); unemployment rate = 81 (1.5%); population density 104 (1.9%).

[※] The patient number and rate of missing data in the validation cohort are as follow: hemoglobin = 28 (2.1%); white blood cell = 28 (2.1%); platelet =28 (2.1%); creatinine = 32 (2.4%); median age of neighborhood = 7 (0.5%); unemployment rate = 2 (0.2%).



Lumbar Disc Disorder Prolonged Opioid Prescription

Probability of prolonged postoperative opioid prescription is 28.68 percent.

Case: 1
Label: Yes
Probability: 0.29
Explanation Fit: 0.76

Figure 1. Individual explanation of the probability of prolonged postoperative opioid prescription.

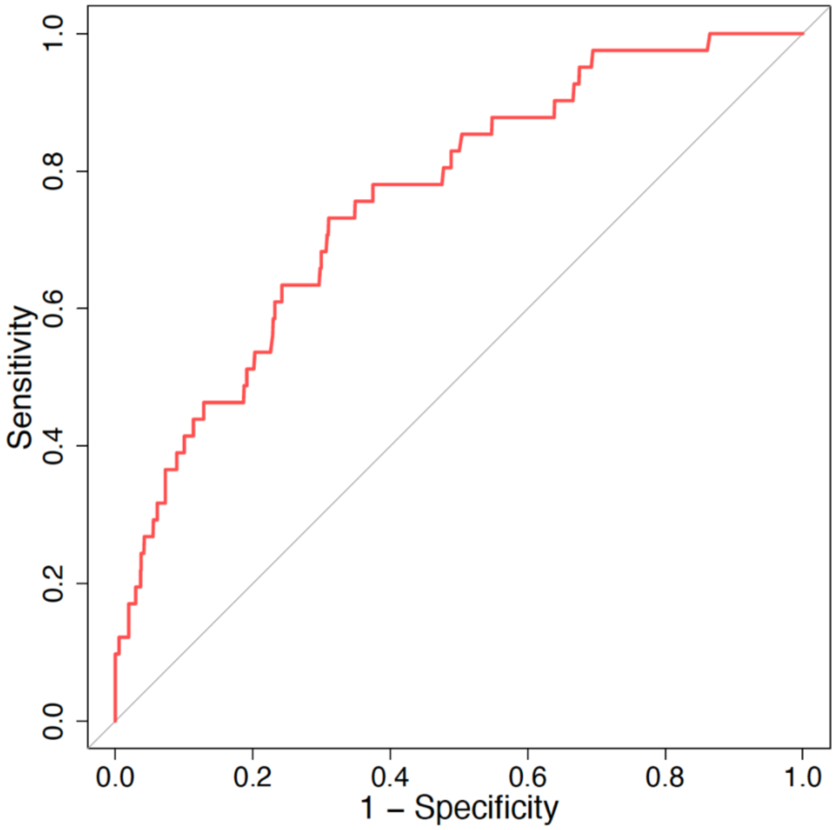


Figure 2. Area under the receiver operating characteristic curve (AUROC) for Skeletal Oncology Research Group machine learning algorithm.

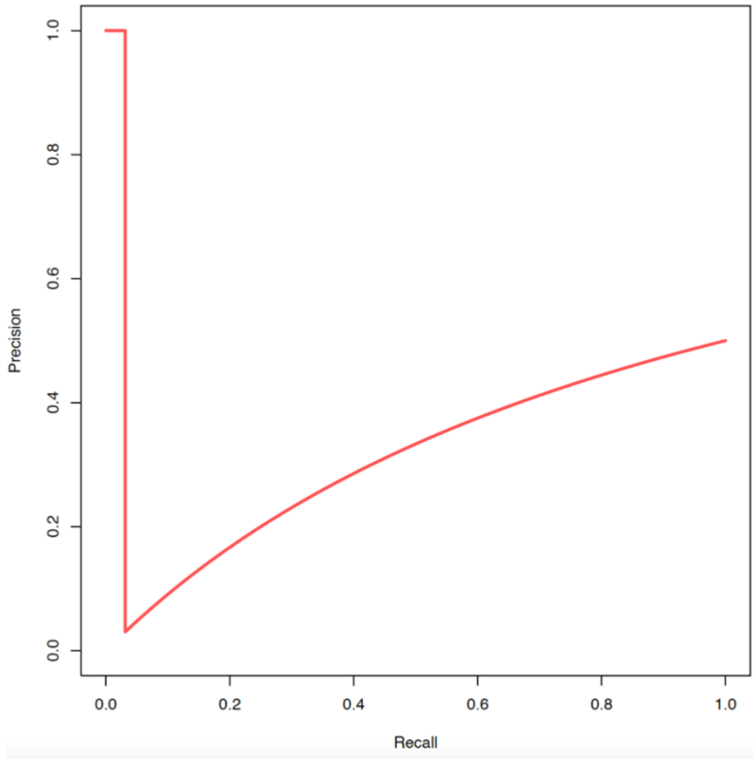


Figure 3. Area under the precision-recall curve (AUORC) for Skeletal Oncology Research Group machine learning algorithm.

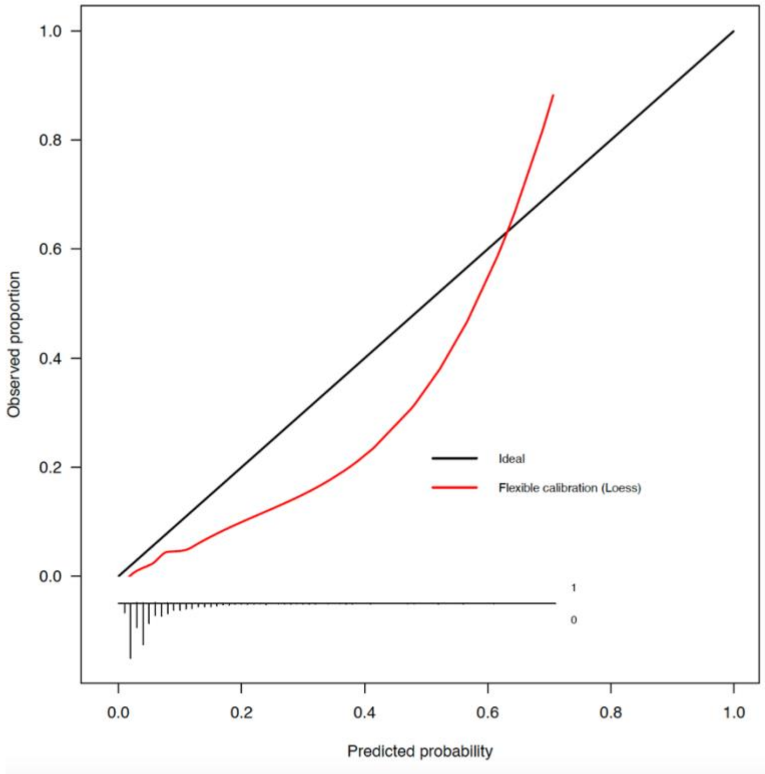


Figure 4. Calibration plot for Skeletal Oncology Research Group machine learning algorithm. Calibration intercept = -0.87, calibration intercept = 1.07.

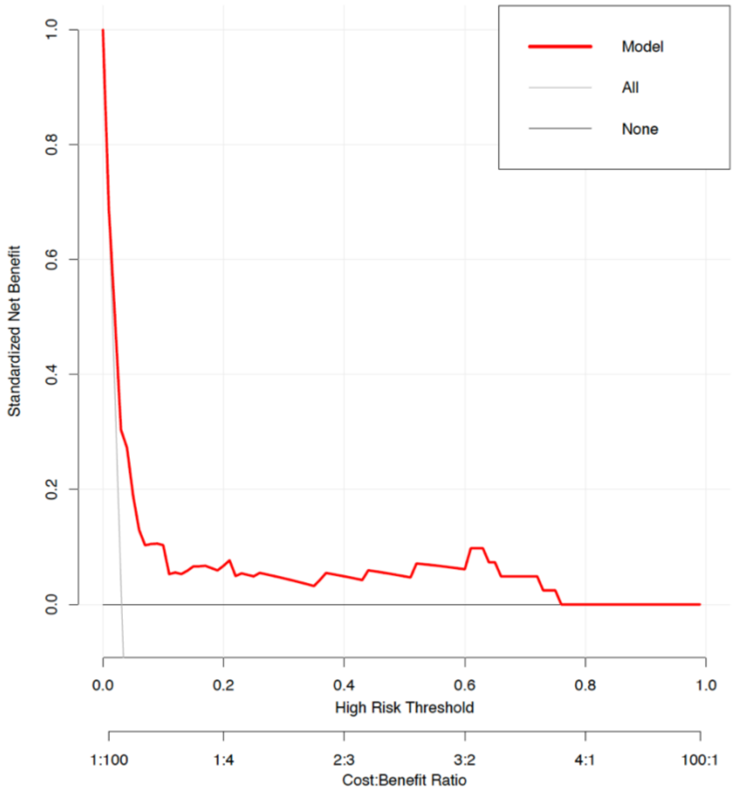


Figure 5. Decision curve analysis with standardized net benefit by threshold probability.

Chapter 8

Impact of Machine Learning Prediction Models on Surgical Decision-Making in Spinal Epidural Abscess

PT Ogink, OQ Groot, A Karhade, A Shah, M Bongers, FC Oner, JJ Verlaan, JH Schwab

In production

Background

Spinal epidural abscess (SEA) is relatively rare spinal condition with an often insidious presentation and nonspecific presenting symptoms, which makes diagnosis and treatment challenging.[1] Due to its low incidence most spine surgeons and/or internists, the 2 specialists that usually deal with SEA in the ER, do not have extensive clinical experience in treating these patients. This constitutes a major problem in these patients considering it is imperative to choose the right treatment modality immediately.[2] A recent study by Karhade et al. (in press) used a machine learning algorithms to develop a prediction model for failure of nonoperative (antibiotic) treatment. This model would allow physicians to consult the application and derive a risk percentage of failure of antibiotic treatment. A subsequent change of the initial treatment could avoid the detrimental effects of switching later on or excessive surgery.

The recent renewed interest in artificial intelligence – and more specifically its subset machine learning (ML)- has led to a significant increase in developed prediction models.[3,4]

The consecutive stages of producing a prognostic model are development, validation, and impact studies. Impact studies aim to quantify if using a predictive model in daily practice actually changes physicians' behavior and improves decision-making.[5] Ideally, impact studies are designed as a randomized trial with a group utilizing the prediction model and the control group providing usual care.[6,7] However, some conditions are pretty rare, which makes a randomized trial design troublesome. One such condition is SEA. Nonetheless, the impact of these prediction models on daily practice needs to be assessed. We propose conducting interobserver survey studies to assess how these models of relatively rare conditions impact decision-making.

The primary aim of this study is to determine the difference in treatment recommendations for SEA between physicians who are aided by the machine learning prediction model and physicians who are not aided by that model. The secondary aim is to compare the change in treatment recommendation by the prediction model between the 2 physicians groups mostly involved with SEA – internists and spine surgeons.

Methods

The spine surgeons and internists will separately be randomized in two groups.

Using an online survey tool, REDCap, both groups will be asked to complete a survey containing 20 cases of patients presenting with SEA. We will use case examples with MRI imaging, which illustrate cases from normal daily clinical practice. We aim to provide cases with varying risks of failure of antibiotic treatment as outputted by the ML algorithm. Group 1 will receive patient cases with the added information of the risk percentage of failure of antibiotic treatment given by the ML algorithm. In the second group, the participants will complete the same case examples but without the risk percentage of the ML algorithm. We will identify potential subjects through personal network and database of Nederlandsche Orthopedische Vereniging (NOV), North American Spine Society (NASS), and other orthopedic societies. The approach will be by email and in total two reminders will be sent out (one and two months after the initial email). After collecting all email addresses, we will randomly assign group 1 or 2 to every email address using randomizer.org. The corresponding links of the different surveys will be sent out.

Study aims

To determine any difference in treatment recommendation for spinal epidural abscess (SEA) between doctors who are aided by the risk assessment of a machine learning (ML) prediction model and physicians who are not aided by that prediction model

To compare the change in treatment recommendation between spine surgeons and internists treating SEA using a ML prediction model.

Survey document

General first page: three questions about the clinician

1. What is your specialty?
    1. Spine surgeon
    2. Spine surgeon – fellow
    3. Internist
    4. Internist – fellow

2. Location of practice?
    1. Europe
    2. Australia
    3. North America
    4. South America
    5. Asia
    6. Africa

3. How many years of practice?
… years

Cases
Case description with deidentified information including patient age/gender/BMI/comorbidities, disease specifics (size/location of abscess), motor and/or neurologic deficit at presentation. Also, deidentified MRI images will be provided.

Questions

- Group 1 (aided by the ML algorithm):

Based on the ML algorithm, the non-operative failure is … %.

1. What treatment do you recommend ?

A. Operative B. Non-operative

- Group 2:

1. What treatment do you recommend (not aided by the ML algorithm)?

A. Operative B. Non-operative

2. Would you change therapy if an ML algorithm predicted otherwise?

A. Yes B. No

Statistical Analysis

We will use frequencies and percentages to describe categorical variables and mean with standard deviation for continuous variables. A two-sample test of proportion

Fisher's exact test will be used for categorical variables and Student's t-test for continuous variables. P-values < 0.05 will be considered statistically significant.

References

1. Darouiche R. Spinal Epidural Abscess. *N Engl J Med* 2006;355:2012–20.
2. Davis DP, Wold RM, Patel RJ, et al. The clinical presentation and impact of diagnostic delays on emergency department patients with spinal epidural abscess. *J Emerg Med* 2004;26:285–91.
3. Cabitza F, Locoro A, Banfi G. Machine Learning in Orthopedics: A Literature Review. *Front Bioeng Biotechnol*;6. Epub ahead of print 2018. DOI: 10.3389/fbioe.2018.00075.
4. Senders JT, Staples PC, Karhade A V., et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg* 2018;109:476–486.e1.
5. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
6. Foy R, Penney GC, Grimshaw JM, et al. A randomised controlled trial of a tailored multifaceted strategy to promote implementation of a clinical guideline on induced abortion care. *BJOG An Int J Obstet Gynaecol* 2004;111:726–33.
7. Meyer G, Köpke S, Bender R, et al. Predicting the risk of falling - Efficacy of a risk assessment tool compared to nurses' judgement: A cluster-randomised controlled trial [ISRCTN37794278]. *BMC Geriatr* 2005;5:1–4.

Part IV

Summary and General Discussion

Summary

Part I

Chapter 1.

Given the increasing popularity of ML prediction models and their potential implementation in clinical practice, an assessment was conducted to examine the outcomes these new models have targeted and the methodologies being employed. Despite the relative novelty of the field, ML prediction models have been developed for a wide variety of topics in orthopedic surgery. Medical management and survival were the most commonly studied subjects and spine surgery was the most studied subspecialty. Variations among studies is mostly based on study size, choice of ML algorithm, and timepoint of outcome. Most published prediction models showed fair to good discriminative abilities, while calibration was poorly reported. Future studies should preferably include more multi-institutional, prospective data and develop multiple models enabling comparison between different ML approaches.

Chapter 2

Previous studies have suggested that prediction models demonstrate incomplete, untransparent reporting of items, such as study design, patient selection, variable definitions and performance measures. This systematic review evaluates the quality and completeness of reporting in ML prediction models for surgical outcomes in orthopedic surgery, assessing their adherence to the TRIPOD statement and evaluating the risk of bias using PROBAST. Many studies displayed poor reporting and are at high risk of bias. Future studies aimed at developing prognostic models should explicitly address the concerns raised, such as incomplete reporting of performance measures, inadequate handling of missing data, and not providing means to make individual predictions. Adherence to methodological guidance, such as the TRIPOD statement, is crucial.

Unreliable prediction models have the potential to cause more harm than good when influencing medical decision-making.

Part II

Chapter 3

The choice between operative and nonoperative treatment modalities is critical in the management of spinal epidural abscess. Preventing the failure of nonoperative approaches is paramount, as failure poses a significant risk of neurological compromise. In Chapter 3 a nomogram was developed which could aid in clinical decision-making in this relatively uncommon pathology. Six independent predictors of failure of nonoperative management were identified which included measures of the patient's general health and neurologic status at the time of presentation as well as radiographic data and local abscess anatomy.

Chapter 4

Lumbar spinal stenosis is one of the most common indications for spine surgery. An accurate personal preoperative prediction of who would need a RF/SNF place could reduce costs and avoid the risks of (unnecessary) pro- longed hospitalization. In Chapter 4 a prediction model was developed able to predict discharge placement with both good discrimination and calibration based on a neural network algorithm. The inclusion of most variables in our model can likely be attributed to being independent risk factors for major complications after surgery for lumbar stenosis.

Chapter 5

Degenerative spondylolisthesis is a spinal pathology which represents a sizeable portion of the spine surgery population with relatively older patients . They are at an increased risk of discharge to a RF/SNF place. In Chapter 5 a prediction model based on a Bayes Point Machine algorithm was developed using data from the The National Surgical Quality Improvement Program (NSQIP) database. The model demonstrated not only good discrimination but also reliable calibration. Prior studies on this subject showed worse discriminative abilities without reporting any calibration measures.

Chapter 6

Spinal surgery is known to have high rates of postoperative opioid prescriptions, which can lead to dependence and abuse. An individual preoperative prediction of who is at increased risk of prolonged opioid use could allow for earlier targeted counseling on pain medication. In Chapter 6 the aim was to develop and internally validate a ML prediction model for prolonged opioid use following surgery for degenerative spondylolisthesis. The Random Forest algorithm was selected with good discrimination, calibration, and overall performance. Important variables such as preoperative opioid use, age , and BMI have all been identified in previous studies as risk factors for (sustained) opioid use. Other variables are likely vectors for other previously identified risk factors; anti-hypertensive medication, statins, and laboratory values for the presence of comorbidities and overall health condition, and duration of symptoms for the duration of preoperative opioid use.

Part III

Chapter 7

As mentioned in the previous Chapter preoperative prediction of prolonged postoperative opioid prescription can help identify patients for increased surveillance after surgery. External validation is an often-overlooked element in the process of incorporating prediction models in current clinical practice. This cannot be stressed enough in prediction models where medicolegal and cultural differences may play a major role. Therefore, in Chapter 7 the SORG algorithm for prolonged opioid use after surgery for lumbar discectomy was externally validated in Taiwan.

Taiwan has enforced strict opioid use regulations since 1996, with strict limitations on opioids like oxycodone for non-cancer patients. Additionally, hospitals in Taiwan must regularly submit patient evaluations on chronic opioid treatment to the Taiwanese Food and Drug Administration, leading to notable differences in preoperative medication patterns compared to the US, with a higher prevalence of NSAID use and a lower percentage of opioid use in Taiwanese patients.

In comparison to the developmental cohort for the SORG algorithm, the patients in this cohort were older, had more inpatient dispositions, lower income, as expected higher preoperative NSAID use, and less depression as comorbidity. All patients in the validation cohort were on national health insurance compared with the American development cohort that consisted of various types of insurance. Despite these obvious differences at baseline and a very strict national opioid policy in Taiwan the SORG algorithm has good discriminative abilities and good overall performance in a Han Chinese patient group.

Chapter 8

Spinal epidural abscess poses diagnostic and treatment challenges due to its rarity and nonspecific symptoms, leading to limited clinical experience among treating physicians, particularly spine surgeons and internists. Karhade et al. developed a prediction model for nonoperative treatment failure, building on the nomogram from Chapter 3. Despite the growing number of prediction models, assessing the impact of such models on decision-making is crucial. In Chapter 8 an interobserver survey study is currently being performed to evaluate how machine learning models influence treatment recommendations, focusing on the determining differences in recommendations between physicians aided by the model and those without its assistance.

General Discussion

Part I: Quality of prediction models

"Quality is not an act, it's a habit" – Aristotle

Part I of this thesis discussed the current uses of ML models in orthopedics, the quality of reporting, and the availability of external validations. Quality of reporting is not just an exercise in ticking a statistician's box, but a quintessential element for prediction models. Reporting key information is necessary not just to assess a models predictive capabilities but also to judge the methodology and adequately convey the model's target population. The goal should be for all models to be easily interpreted, (externally) validated, and ultimately implemented. Failing to do so can casts doubts on their usefulness and potentially even cause harm.

The most glaring omission in reporting that was found is the lack of calibration. This omission has the potential to actually hurt patients if the model is used in clinical decision-making. Poor calibration may lead to under- or overestimation of the assessed risk causing potential under- or overtreatment. As an example, a patient with spine metastasis in which survival is severely overestimated while the risk of having a major complication is underrated may lead to unwarranted surgery and this patient spending the last weeks of his/her life being operated on and ultimately dying of a major complication. Why this specific metric is absent in so many studies remains a mystery. This omission plays a major role in the astonishing 41% of all models rated as having a high risk of bias. Inadequate handling of missing data and small datasets were the other major driving factors in this outcome.

However, spine surgery, or orthopedic surgery in general for that matter, are not the only medical specialties suffering from this omission. Throughout medicine there is an abundance of very poorly made ML prediction models.[48–51] Similar to our findings in Chapter 2, the most important flaws can be found in poor reporting of outcomes, small sample sizes, and poor handling of missing data. Additionally, Navarro et al.[52] and Dhiman et al.[49] suggest many researchers are overselling their models by making unjustified claims of quality and advising clinical use without any external validation.

External validations are not exempt from the mistakes made in developmental studies. Algorithms based on data from one group can have very poor predictive abilities in other groups and lead to poor decision-making as in the earlier mentioned example. External validations seem to be unpopular to perform and submit. First of all, there is a reluctance in this somewhat novel research field to simply share code with other groups, necessary if an external validation is to be performed. Second of all, doing an external validation using someone else's prediction model is not very attractive. Most researchers interested in this field will find it far more interesting to make a model themselves instead of performing an external validation. Simply put: despite its importance, external validations just are not very appealing.

Our findings in Part I suggest emphasis should be put on adequate reporting of ML prediction models and their external validations and not merely for scientific reasons.

The rise of AI in medicine, coupled with its limited backing by evidence, has attracted the attention of regulatory agencies. Medical devices in the EU are governed by Regulation 2017/745 on Medical Devices (MDR). It's definition of a medical device clearly includes ML prediction models: 'Any instrument, apparatus, appliance, *software*, implant, reagent, material or other article, intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes [. . .] diagnosis, prevention, monitoring, *prediction*, *prognosis* of disease'. Concerning which MDR risk class prediction models fall in the EU stipulates: 'Software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, except if such decisions have an impact that may cause: death or an irreversible deterioration of a person's state of health, in which case it is in class III". Considering the aforementioned serious harm poorly calibrated models may cause, they can very well be considered to be in class III, which is the highest risk class. The proposed EU AI act aims to introduce a regulatory framework and legal framework for AI. For prediction models based on AI it would mandate among other things a registration in an EU database and adherence to quality criteria related to the training and validation of datasets.[53] The FDA in the US has similarly proposed a new regulatory framework. [54] Therefore, it seems crucial for researchers to seriously improve the quality (of reporting) of ML prediction models if they are ever to be implemented.

Luckily, researchers do not have to start from scratch in this matter. In 2015, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement was published providing a 22-item checklist for prediction models in general.[32] In 2021, an update to the statement was announced to be more specific for AI-based prediction models.[55] This checklist provides an easy to follow and concise list of items that should be included in every published model. Ideally, reviewers of medical journals should no longer accept models not following these items. Furthermore, all research groups with the slightest interest in ML should help and contribute to the execution of external validations. Making connections with research groups, both domestically as well as internationally, can help foster an international ML community that can become something better than the sum of its parts.

Part II Development of prediction models

"Data is the new gold" – Clive Humby

Part II showcases a number of examples ML can be used for in spine surgery. The wide range of topics in Part II signifies the potential (ML) prediction models have for our clinical practice; from helping us choose to operate or not in infectious diseases to managing logistical problems like delayed discharges after elective surgeries.

While these are interesting and useful examples, the topics of our prediction models are very limited with respect to the questions surgeons ask themselves of the future. How do we provide a framework for evermore models assisting us in clinical practice? First and foremost, referencing back to the quote above, we must collect data. While current EHR's provide excellent opportunities for data collection incomparable to the times before EHR's there is still a lot to be improved upon. First and foremost, a prominent place should be reserved for Patient Reported Outcomes (PROMS) in our future efforts. Ultimately, if spine surgeons had to pick one prediction of all the potential predictions out there, they would undoubtedly choose a model telling them whether the surgery will be beneficial to the patient's quality of life. While there has been an uptick in papers focusing on PROMS, there are still only a handful of papers available in the orthopedic literature, scattered over the subspecialties.[56–58] Spine surgery should be at the forefront of this effort considering the inherent difficulties of patient selection for surgery compared to, for instance, total hip arthroplasty. Diligently setting up a system of letting patients fill out PROMS questionnaire before and after surgery is an enormous task initially, but can pay itself back tenfold when used in the future to significantly improve patient selection.

Improving EHR's, collecting more PROMS data and developing more and more ML prediction models is just the beginning, however. Wearable devices like smart watches, smartphones and home sensors can add an enormous amount of data to use. Future prediction will not only use this data to construct new prediction models but can actively warn that something might be wrong. For example, the smartwatch of a patient who recently had lumbar fusion is actively feeding data into an algorithm. It detects a slightly raised body temperature, a lower blood pressure and a marked drop in distance walked the last 24 hours. A prediction is made that a postoperative infection may be imminent and the patient is advised to call the doctor's office. Far-fetched perhaps now, but real-time prediction-making has already been developed predicting exacerbations for patients with COPD and for predicting seizures in epilepsia.[59,60] Besides adding data to feed into the algorithms, AI can also help updating the prediction model continuously circumventing the problem called concept drift.[61] Concept drift refers to the changing relationship between the input variables and the outcome as time pass. In other words, over time the model becomes obsolete and no longer functions due to changes in the real world. For instance, a model predicting survival in spinal metastasis will likely perform significantly worse if a new and effective immunotherapy is introduced by the oncologists. Instead of having to completely update the model every single time, AI can be instructed to constantly update the algorithms based on new data coming in all the time both from the wearables devices and the EHR.

Part III Validation and Implementation

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write!" – HG Wells

In Part III an external validation of a prediction model was performed in an entirely different continent, and questioned the impact of machine learning models on surgeon's decision-making. What if we have accurate, well-constructed and externally validated prediction models, developed on solid data from across the socio-economic spectrum, but practicing doctors do not have a clue how to use them or simply ignore them? The ability to understand, interpret and use numbers, percentages and orders of magnitudes is a critical skill in modern healthcare for both patients and doctors, but both groups are found lacking.

Humans in general are particularly poor with numbers and risk assessment. A simple enough concept as a weather forecast is too difficult to comprehend for a large percentage of society.[62] Medical students and doctors do not fare much better. [63,64] On top of that certain other human tendencies interfere with decision-making; recall bias, conformation bias, and commission bias are but a few biases doctors and patients suffer from in decision-making. No clear cut solution exists for this human deficit. While there have been suggestions made to, for instance, improve the presentation of statistical information for patients, and improve numeracy in general in patients, there is a scarcity in the literature of any numeracy intervention and its effectiveness. This area is where much of the future efforts in prediction modelling should be done. While researchers across the globe are churning out prediction model after prediction model, albeit of dubious quality, we are not even sure whether they influence decision-making at all and, if so, how. In other words, lots of serious effort may all be for naught. In Chapter 8 a framework is given of what these studies may look like. At the time of writing this thesis, the results of this study are not yet known. But whatever the outcome may be, improving numeracy in both patients and doctors should be stimulated and additional research on the effect of predictive modelling on decision-making should expand.

Challenges and Future Perspectives

"Now I am become Death, the destroyer of worlds" – Robert Oppenheimer

Stories about developments and inventions that turn out better than expected are abundant. Other inventions turn out worse than intended. Mark Zuckerberg probably didn't expect Facebook to become a source of misinformation when he founded the site in his dorm room many years ago. Others are quicker to see the potential harms of their creation. Robert Oppenheimer thought of the Hindu verse "Now I am become Death, the destroyer of worlds" upon seeing the first atomic bomb explode on which he had worked for three years. While ML prediction models will not destroy worlds, it is important to acknowledge the limitations they have and the potential harm they can do.

First, by bringing improvement to only certain select groups, ML predictions have the ability to widen an already existing health(care) inequality. The data used in our studies are from tertiary care hospitals in the northeastern United States. Not only is the United States an economically wealthy nation, but the patients in this hospital system are disproportionately well off socioeconomically compared with the average US citizen. Because of the necessity of vast amounts of data to develop these tools, the algorithms will solely be based on patients in advanced nations treated in medical centers with the resources to collect these data and use it for model development. This lack of equity may look decidedly trivial at this point considering algorithms are not widely used yet. However, once they start being incorporated and improve patient outcomes, the already existing gaps in healthcare outcomes will only be exacerbated. While well-off patients receive individualized predictions enhancing their care, patients with lower socioeconomic status who already have worse health outcomes, both in general and specifically after orthopedic surgery, lag even further behind. Case in point is the opioid crisis which featured in Chapter 4.[65,66] The opioid crisis is the result of a potpourri of failures combining false marketing, lack of oversight, poor health literacy, stigmatization and negligence of (people in) certain geographical areas, and plain fraud to produce the biggest disaster of modern medicine.[67] The model, developed using data from patients with high socioeconomic status in the greater Boston area, which is generally less affected than many other areas by opioid overdose hospitalizations, may not adequately represent the patients most at risk. Potentially, this model can help prevent opioid addictions, but its effectiveness may be limited to patients who can afford hospital expenses, and its applicability may vary significantly in more severely affected areas outside of the Boston region. Therefore, it is crucial for implementation of artificial intelligence in healthcare to incorporate fairness in our efforts. We need to put effort into including all socioeconomic groups from every area in our prediction models. An obvious hurdle is of course the necessity of data. There simply is no EHR data if a patient does not visit a hospital. And even if they do, if they only visit a pain clinic once or twice, get prescribed a large dose of opioids, and subsequently turn to the streets for additional opioids, there is no way of including them. The best we can do know is at least acknowledge their omission in datasets and try and find ways we can incorporate them as well.

While disproportionate improvement of certain groups is one potential negative aspect, negative effects for everybody are, of course, even worse. As previously discussed, the developed models currently lack a staggering number of features to assess their quality. If, at some point, these predictions are widely trusted and used, the models must not do more harm than good. To avoid these perils, newly developed models must be held to the highest standards when it comes to their reporting and external validation should be required before implementation in clinical practice. The proposed frameworks by the EU and FDA as discussed previously may begin to enforce these requirements.

While doctors and patients will, and should, drive the development and stand for the quality of more and better ML models, there are a other contenders such as regulatory agencies and insurance companies. Healthcare costs have risen steadily in the last decades and constitute a major share of government expenditure in all developed countries. With ageing populations in these developed countries and ever-expanding medical solutions to erstwhile incurable diseases, this growth does not look like stopping anytime soon and will put increasing pressure on government budgets. Considering the current geopolitical situation and the perils of climate change, healthcare expenditure will be in much tougher competition with expenditures on defense and CO2 reduction than in the previous 30 years. This situation is further complicated by the fact that aging populations will result in a diminishing workforce, providing less tax revenue to cover these expenses and a lack of manpower to care for them. In other words, governments will take a look at one of their biggest expenses and start cutting, or at the very least, not let the healthcare budget grow out of control. Regulatory agencies and insurance companies, as instruments of national governments striving to control costs, will recognize the potential of AI, whether in predictive modelling or imaging analysis.

However, the biggest competitor to doctors will likely be for-profit healthcare companies. Their well-funded organizations can beat doctors to the punch in terms of developing elaborate prediction models and implementing advanced healthcare technologies. Doctors, who are spending most of their time on their patients, may struggle in competing against the resources and manpower of healthcare companies. However, it is key for doctors to stay in the driver's seat. Allowing medical companies to control development of AI in healthcare could lead to the prioritization of profit over patient welfare. A recent publication in JAMA showing an increase in adverse advents after private equity took over hospitals, proves this is not just a hypothetical concern.[68] The goal of ML-based prediction models is to provide unbiased personalized predictions to enhance decision-making, not to nudge doctors and patients in the direction of what is most profitable for company X/Y/Z. Given the medical industry's history of downplaying risks and exaggerating the success of their products, tampering with prediction models would merely become the latest method in this behavior. [69–71] The general lack of knowledge with regards to AI among doctors will make discovering any tampering or bias even harder than it already is in "normal" studies. Aside from these pessimistic considerations, there are numerous positive reasons doctors should play a central role in the development and implementation of AI-based predictive models in healthcare. Doctors are the ones who engage in medical decision-making with patients. Therefore, they understand the nuances of medical decision-making which are invaluable for developing models that reflect real-world scenarios; and can help determine which scenario would actually benefit from an accurate prediction. Second, it is the responsibility of doctors to explain ML-based predictions to patients, address their concerns, and ensure that patients understand what they are choosing. Third, doctors can provide distinct real-life feedback for the improvement of AI models. This feedback from the field is essential for continuous refinement to ensure prediction models keep evolving to meet the changing needs.

In order to achieve this, orthopedic surgeons need to join forces to share data, externally validate each other's prediction models and drive these developments to where we want them to go. The Machine Learning Consortium is a Netherlands-based, international organization aiming to do just this with trauma surgeons, residents, and PhD candidates joining forces in the field of orthopedic trauma. Considering the size of spinal surgery as a subspecialty and the large amount of financial resources it represents, a similar organization should be construed focusing on predictive modelling for spine surgery.

And finally, the inevitable question asked in any piece written about any field of AI: will *it* replace *us?* What if *they* get too good? What if computers do not just beat us at chess, but start beating us at being doctors. While prediction and prognosis are only a portion of the entire treatment of a patient, it does not take that much imagination to see how it goes. Doctors start using AI for diagnostic and predictive purposes, enhancing daily workflow. A potential risk arises as reliance on AI becomes excessive, leading to a potential degradation of clinical intuition. As people get more comfortable with AI running their daily lives, -driving their car, paying their bills and choosing what to wear- the human doctors with all their skills begin to fade out as well. The argument of diminishing clinical skill due to advancing technology is not new. Many doctors and residents who work in developing countries as part of an aid programme claim their clinical skills over there are better, since they do not have the luxury of technologies like a CAT scan; clinical skills supposedly follow the 'use it or lose it' adage. While this claim has been repeated often and can be heard in every hospital in the Western world, evidence is lacking. [72,73] The sentiment expressed in these remarks represents a distinct reservation about technology only exacerbated by the potential of AI. Ted Kaczynski, more widely known as the Unabomber, expressed an extreme version of this sentiment in his manifesto. While denounced because of his atrocities, certain aspects of his manifesto have become somewhat commonplace; technological advancement as an opponent of human nature, forcing us into a proverbial cage instead of representing an enhancing feature of human development. Perhaps at this point it is indeed hard to see what AI will mean for the medical practice of let’s say 50 years from now. However, we should not be denying patients and doctors accurate information and subsequently better care out of fear for technology. The stethoscope was once a high-tech tool enhancing doctor’s abilities to listen to heart sounds replacing direct auscultation with their ears. Similarly, for now, ML prediction tools are not meant, or able, to replace the doctor but are simply a high tech tool to enhance their ability to make better decisions. Let us use it wisely.

References

1. Hippocrates, Prognosticon, PART 1.

2. Van Nuffelen P. Galen, divination and the status of medicine. *Class. Q.* 2014;64(1):337–352.

3. Horden P. What's Wrong with Early Medieval Medicine? *Soc. Hist. Med.* 2011;24(1):5.

4. Meehl PE. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. *Clin. versus Stat. Predict. A Theor. Anal. a Rev. evidence.* 2006.

5. Ægisdóttir S, White MJ, Spengler PM, et al. The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *Couns. Psychol.* 2006;34(3):341–382.

6. Grove WM. Clinical versus statistical prediction: The contribution of Paul E. Meehl. *J. Clin. Psychol.* 2005;61(10):1233–1243.

7. Joyner MJ, Paneth N. Seven Questions for Personalized Medicine. *Jama.* 2015;55905:2015–2016.

8. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Med. Genomics.* 2015;8(1):1–12.

9. Jameson L, Longo D. Precision Medicine — Personalized, Problematic, and Promising. *N Engl J Med.* 2015;372(23):2229–2234.

10. Shapiro SD. The future of medicine is very personal. *Am. J. Respir. Crit. Care Med.* 2012;185(9):903–904.

11. Esplin ED, Oei L, Snyder MP. Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease. *Pharmacogenomics.* 2014;15(14):1771–1790.

12. Antman JMMSE. Population and Personalized Medicine in the Modern Era. *JAMA.* 2014;02115(19):1969–1970.

13. James MT, Pannu N, Hemmelgarn BR, et al. Derivation and External Validation of Prediction Models for Advanced Chronic Kidney Disease Following Acute Kidney Injury. *JAMA.* 2017;318(18):1787.

14. Sultan AA, West J, Grainge MJ, et al. Development and validation of risk prediction model for venous thromboembolism in postpartum women: multinational cohort study. *BMJ.* 2016;355:i6253.

15. Lamberink HJ, Otte WM, Geerts AT, et al. Individualised prediction model of seizure recurrence and long-term outcomes after withdrawal of antiepileptic drugs in seizure-free patients: a systematic review

and individual participant data meta-analysis. *Lancet. Neurol.* 2017;16(7):523–531.

16. Yeh RW, Secemsky EA, Kereiakes DJ, et al. Development and Validation of a Prediction Rule for Benefit and Harm of Dual Antiplatelet Therapy Beyond 1 Year After Percutaneous Coronary Intervention. *JAMA.* 2016;315(16):1735.

17. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models. *Jama.* 2017;318(14):1377.

18. Obermeyer Z. TL. Lost in Thought — The Limits of the Human Mind and the Future of Medicine. *N. Engl. J. Med.* 2010;363(1):1–3.

19. Cabitza F, Locoro A, Banfi G. Machine Learning in Orthopedics: A Literature Review. *Front. Bioeng. Biotechnol.* 2018;6(June).

20. Senders JT, Staples PC, Karhade A V., et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg.* 2018;109(MI):476-486.e1.

21. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial Intelligence in Cardiology. *J. Am. Coll. Cardiol.* 2018;71(23):2668–2679.

22. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 2015;13:8–17.

23. Computing after Moore's Law - Scientific American.

24. Jordan M, Mitchell T. Machine learning: Trends, perspectives, and prospects. *Science (80-.).* 2015;349(6245):255–260.

25. Tu J V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* 1996;49(11):1225–1231.

26. Direct Medical Costs | BMUS: The Burden of Musculoskeletal Diseases in the United States.

27. Martin BI, Deyo RA, Mirza SK, et al. Expenditures and health status among adults with back and neck problems. *JAMA - J. Am. Med. Assoc.* 2008;299(6):656–664.

28. Martin BI, Tosteson ANA, Lurie JD, et al. Variation in the care of surgical conditions: Spinal stenosis. A Dartmouth Atlas of health care series. 2014;(603):1–48.

29. Lønne G, Schoenfeld AJ, Cha TD, et al. Variation in selection criteria and approaches to surgery for Lumbar Spinal Stenosis among patients treated in Boston and Norway. *Clin. Neurol. Neurosurg.*

2017;156(2017):77–82.

30. Ogink PT, van Wulfften Palthe O, Teunis T, et al. Practice Variation Among Surgeons Treating Lumbar Spinal Stenosis in a Single Institution. *Spine (Phila. Pa. 1976).* 2018.

31. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928–935.

32. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur. Urol.* 2015;67(6):1142–1151.

33. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Med.* 2018;16(120).

34. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605.

35. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416.

36. Alblas M, Velt KB, Pashayan N, et al. Prediction models for endometrial cancer for the general population or symptomatic women: A systematic review. *Crit. Rev. Oncol. Hematol.* 2018;126:92–99.

37. Hodgson LE, Sarnowski A, Roderick PJ, et al. Systematic review of prognostic prediction models for acute kidney injury (AKI) in general hospital populations. *BMJ Open*. 2017;7(9):e016591.

38. Janssen KJM, Vergouwe Y, Donders ART, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin. Chem.* 2009;55(5):994–1001.

39. Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*. 2015;521(7553):452–459.

40. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models : A framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.

41. Carter J V., Pan J, Rai SN, et al. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surg. (United States)*. 2016;159(6):1638–1645.

42. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):1–7.

43. BRIER GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Mon. Weather Rev.* 1950;78(1):1–3.

44. Greenwell BM, Boehmke BC, McCarthy AJ. A Simple and Effective Model-Based Variable Importance Measure. 2018:1–27.

45. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2020;2(1):56–67.

46. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* 2008;26(6):565–574.

47. Siontis GCM, Tzoulaki I, Castaldi PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J. Clin. Epidemiol.* 2015;68(1):25–34.

48. Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J. Clin. Epidemiol.* 2021;138:60–72.

49. Dhiman P, Ma J, Andaur Navarro CL, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *J. Clin. Epidemiol.* 2023;157.

50. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;375.

51. Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med. Res. Methodol.* 2022;22(1).

52. Andaur Navarro CL, Damen JAA, Takada T, et al. Systematic review finds “spin” practices and poor reporting standards in studies on machine learning-based prediction models. *J. Clin. Epidemiol.* 2023;158:99–110.

53. Fraser AG, Biasin E, Bijnens B, et al. Expert Review of Medical Devices ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ierd20 Artificial intelligence in medical device software and high-risk medical devices-a review of definitions, expert recommendations and regulatory initiatives Artificial intelligence in medical device software and high-risk medical devices-a review of definitions, expert recommendations and regulatory initiatives. 2023.

54. Software as a Medical Device (SaMD) Action Plan. 2021.

55. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol: Protocol for development of a reporting

guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):48008.

56. Kunze KN, Krivicich LM, Clapp IM, et al. Machine Learning Algorithms Predict Achievement of Clinically Significant Outcomes After Orthopaedic Surgery: A Systematic Review. *Arthroscopy*. 2022;38(6):2090–2105.

57. Kunze KN, Karhade A V., Sadauskas AJ, et al. Development of Machine Learning Algorithms to Predict Clinically Meaningful Improvement for the Patient-Reported Health State After Total Hip Arthroplasty. *J. Arthroplasty*. 2020;35(8):2119–2123.

58. Karhade A V., Fogel HA, Cha TD, et al. Development of prediction models for clinically meaningful improvement in PROMIS scores after lumbar decompression. *Spine J.* 2021;21(3):397–404.

59. Wu CT, Li GH, Huang CT, et al. Acute Exacerbation of a Chronic Obstructive Pulmonary Disease Prediction System Using Wearable Device Data, Machine Learning, and Deep Learning: Development and Cohort Study. *JMIR mHealth uHealth*. 2021;9(5).

60. Beniczky S, Karoly P, Nurse E, et al. Machine learning and wearable devices of the future. *Epilepsia*. 2021;62 Suppl 2(S2):S116–S124.

61. Pechenizkiy M. Žliobaitė et al. 2016 - An Overview of Concept Drift Applications.pdf. 2016:1–24.

62. Gigerenzer G, Hertwig R, Van Den Broek E, et al. "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Anal.* 2005;25(3):623–629.

63. Friederichs H, Birkenstein R, Becker JC, et al. Risk literacy assessment of general practitioners and medical students using the Berlin Numeracy Test. *BMC Fam. Pract.* 2020;21(1).

64. Sheridan SL, Pignone M. Numeracy and the medical student's ability to interpret data. *Eff. Clin. Pract.* 2002;5(1):35–40.

65. Goodman SM, Mandl LA, Parks ML, et al. Disparities in TKA Outcomes: Census Tract Data Show Interactions Between Race and Poverty. *Clin. Orthop. Relat. Res.* 2016;474(9):1986–1995.

66. Mackenbach JP, Stirbu I, Roskam A-JR, et al. Socioeconomic inequalities in health in 22 European countries. *N. Engl. J. Med.* 2008;358(23):2468–2481.

67. Keefe PR. *Empire of pain : the secret history of the Sackler dynasty*. First edition. New York: Doubleday; 2021.

68. Kannan S, Bruch JD, Song Z. Changes in Hospital Adverse Events and Patient Outcomes Associated

With Private Equity Acquisition. *JAMA*. 2023;330(24):2365–2375.

69. Lundh A, Lexchin J, Mintzes B, et al. Industry sponsorship and research outcome. *Cochrane database Syst. Rev.* 2017;2(2).

70. Fabbri A, Lai A, Grundy Q, et al. The Influence of Industry Sponsorship on the Research Agenda: A Scoping Review.

71. Van Zee A. The Promotion and Marketing of OxyContin: Commercial Triumph, Public Health Tragedy. *Am. J. Public Health*. 2009;99(2):221.

72. van den Hombergh P, de Wit NJ, van Balen FA. Experience as a doctor in the developing world: does it benefit the clinical and organisational performance in general practice? 2009.

73. Gupta AR, Wells CK, Horwitz RI, et al. The International Health Program: the fifteen-year experience with Yale University's Internal Medicine Residency Program. *Am. J. Trop. Med. Hyg.* 1999;61(6):1019–1023.

Main Study Questions and Conclusions

Part I – Quality of Prediction Models

Chapter 1

What outcomes and methodologies are being employed in machine learning (ML) prediction models for orthopedic surgery?

ML models in orthopedics predominantly focus on medical management outcomes, with neural networks as the most used algorithm, though reporting on calibration and decision-curve analysis remains inconsistent.

Chapter 2

How well do ML prediction models in orthopedic surgery adhere to transparent reporting and bias assessment guidelines?

Only 44% of ML studies had low risk of bias, with 53% median adherence to TRIPOD reporting, highlighting significant gaps in methodology and transparency that hinder clinical implementation.

Part II – Development of Prediction Models

Chapter 3

What are the risk factors for failure of nonoperative treatment in patients with spinal epidural abscess (SEA)?

Six independent predictors of treatment failure were identified, and a nomogram was created to quantify failure risk, aiding in treatment decision-making between nonoperative and operative management.

Chapter 4

Can a machine learning algorithm accurately predict non-home discharge after elective surgery for lumbar spinal stenosis?

A machine learning model, based on a Neural Network, demonstrated good discrimination and calibration in predicting non-home discharge, enabling improved discharge planning and cost reduction in clinical practice.

Chapter 5

Can a machine learning algorithm be developed to accurately predict discharge placement (home vs. non-home) for patients undergoing elective surgery for degenerative spondylolisthesis?

The study successfully developed a predictive machine learning algorithm (Bayes Point Machine) that demonstrated good accuracy, calibration, and overall performance for predicting discharge placement. The methodology used can be adapted for other conditions and treatments.

Chapter 6

Can a machine learning model be developed to predict the risk of prolonged opioid use (90–180 days post-surgery) in patients undergoing surgery for degenerative spondylolisthesis?

The study successfully developed a Random Forest machine learning model with good discrimination, calibration, and overall performance to predict prolonged opioid use after surgery. This model can help healthcare providers target high-risk patients with tailored strategies and policies to reduce sustained opioid use.

Part III – Validation and Implementation

Chapter 7

Can the SORG machine learning model for predicting prolonged postoperative opioid prescription after lumbar discectomy, developed in the U.S., be externally validated and generalized to a Taiwanese patient cohort?

The SORG model demonstrated good discrimination and overall performance in predicting prolonged opioid use in a Taiwanese patient group, despite differences in baseline characteristics and opioid policies. The freely available digital tool can help identify high-risk patients and support the development of targeted prevention strategies.

Chapter 8

Does the use of a machine learning (ML) prediction model for assessing the risk of failure of non-operative treatment in spinal epidural abscess (SEA) influence treatment recommendations compared to traditional decision-making without the model?

In Production

Appendices

Nederlandstalige Samenvatting

Deel I – Kwaliteit van Voorspellingsmodellen

Hoofdstuk 1

Gezien de toenemende populariteit van ML-voorspellingsmodellen en hun potentiële implementatie in de klinische praktijk, onderzochten we waar deze nieuwe modellen zich op richten en de toegepaste methodologieën. Ondanks dat het vakgebied in de kinderschoenen staat, zijn ML-voorspellingsmodellen ontwikkeld voor een breed scala aan onderwerpen in de orthopedische chirurgie. Medisch management en overleving waren de meest bestudeerde onderwerpen, en wervelkolomchirurgie was de meest bestudeerde subspecialisatie. Variaties tussen studies zijn voornamelijk gebaseerd op de omvang van de studie, de keuze van het ML-algoritme en het gekozen eindpunt van de uitkomst. De meeste gepubliceerde voorspellingsmodellen toonden redelijk goede discriminatieve vermogens, terwijl kalibratie slecht werd gerapporteerd. Toekomstige studies zouden bij voorkeur meer multi-institutionele, prospectieve gegevens moeten opnemen en meerdere modellen moeten ontwikkelen om vergelijking tussen verschillende ML-benaderingen mogelijk te maken.

Hoofdstuk 2

Eerdere studies suggereren dat voorspellingsmodellen onvolledige en niet-transparante zijn wat betreft het rapporten van studiedesign, patiëntenselectie, definities van variabelen en uitkomsten. Deze systematische review evalueert de kwaliteit en volledigheid van rapporten in ML-voorspellingsmodellen voor chirurgische uitkomsten in de orthopedische chirurgie, beoordeelt hun naleving van de TRIPOD-richtlijn en evalueert het risico op bias met behulp van PROBAST. Veel studies vertoonden een matige methodologie en hadden een hoog risico op bias. Toekomstige studies gericht op het ontwikkelen van prognostische modellen moeten expliciet deze tekortkomingen aanpakken. Het naleven van methodologische richtlijnen, zoals de

TRIPOD-richtlijn, is cruciaal. Onbetrouwbare voorspellingsmodellen kunnen meer kwaad dan goed doen bij het beïnvloeden van medische besluitvorming.

Deel II – Ontwikkeling van Voorspellingsmodellen

Hoofstuk 3

De keuze tussen operatieve en niet-operatieve behandelingsopties is cruciaal bij de behandeling van een spinaal epiduraal abces. Het voorkomen van falen van niet-operatieve benaderingen is van het grootste belang, aangezien falen een aanzienlijk risico op neurologische complicaties met zich meebrengt. In hoofdstuk 3 werd een nomogram ontwikkeld dat kan helpen bij klinische besluitvorming bij deze relatief zeldzame pathologie. Zes onafhankelijke voorspellers van falen van niet-operatief beheer werden geïdentificeerd, waaronder metingen van de algemene gezondheid en neurologische status van de patiënt op het moment van presentatie, evenals radiologische data en de lokale anatomie van het abces.

Hoofdstuk 4

Lumbale spinale stenose is een van de meest voorkomende indicaties voor wervelkolomchirurgie. Een nauwkeurige persoonlijke preoperatieve voorspelling van wie een revalidatiecentrum of verpleeghuis nodig heeft postoperatief, kan kosten verlagen en de risico's van (onnodige) langdurige ziekenhuisopname vermijden. In hoofdstuk 4 werd een voorspellingsmodel ontwikkeld dat ontslagbestemming kon voorspellen met zowel goede discriminatie als kalibratie, gebaseerd op een neuraal netwerk. Voor de meeste variabelen in ons model geldt dat zij onafhankelijke risicofactoren voor grote complicaties na een operatie voor lumbale stenose zijn.

Hoofdstuk 5

Degeneratieve spondylolisthesis is een aandoening wervelkolom die een aanzienlijk deel van de wervelkolomchirurgie vertegenwoordigt, met relatief oudere patiënten. Deze patiënten hebben een verhoogd risico om te worden ontslagen naar een revalidatiecentrum of verpleeghuis. In hoofdstuk 5 werd een voorspellingsmodel ontwikkeld op basis van een Bayes Point Machine-algoritme, met gebruik van gegevens uit de National Surgical Quality Improvement Program (NSQIP)-database. Het model toonde niet alleen goede discriminatie, maar ook betrouwbare kalibratie. Eerdere studies over dit onderwerp lieten slechtere discriminatieve vermogens zien en rapporteerden geen enkele kalibratiemaatregel.

Hoofdstuk 6

Wervelkolomchirurgie staat bekend om de hoge aantallen postoperatieve opioïden, wat kan leiden tot afhankelijkheid en misbruik. Een individuele preoperatieve voorspelling van wie een verhoogd risico heeft op langdurig opioïdgebruik kan vroegtijdige en gerichte counseling over pijnmedicatie mogelijk maken. In hoofdstuk 6 was het doel om een ML-voorspellingsmodel te ontwikkelen en intern te valideren voor langdurig opioïdgebruik na een operatie voor degeneratieve spondylolisthesis. Het Random Forest-algoritme werd geselecteerd vanwege de goede discriminatie, kalibratie en algehele prestaties. Belangrijke variabelen zoals preoperatief opioïdgebruik, leeftijd en BMI zijn in eerdere studies geïdentificeerd als risicofactoren voor (langdurig) opioïdgebruik. Andere variabelen zijn waarschijnlijk indicatoren van andere eerder geïdentificeerde risicofactoren, zoals antihypertensieve medicatie, statines en laboratoriumwaarden voor de algehele gezondheidstoestand, en de duur van symptomen voor de duur van het preoperatieve opioïdgebruik.

Deel III – Validatie en Implementatie

Hoofdstuk 7

Zoals vermeld in het vorige hoofdstuk kan een preoperatieve voorspelling van langdurige postoperatieve opioïdgebruik helpen om patiënten te identificeren die na de operatie extra aandacht nodig hebben. Externe validatie is een vaak over het hoofd gezien element in het proces van voorspellingsmodellen. Dit is van bijzonder belang bij voorspellingsmodellen waarin medicolegale en culturele verschillen een grote rol kunnen spelen. Daarom werd in hoofdstuk 7 het SORG-algoritme voor langdurig opioïdgebruik na een operatie voor lumbale discectomie extern gevalideerd in Taiwan. Taiwan handhaaft sinds 1996 strikte regelgeving voor opioïdgebruik, met strenge beperkingen op opioïden zoals oxycodon voor niet-kankerpatiënten. Bovendien moeten ziekenhuizen in Taiwan regelmatig patiëntbeoordelingen van chronische opioïdbehandelingen indienen bij de Taiwanese voedsel- en warenautoriteit, wat leidt tot aanzienlijke verschillen in preoperatieve medicatiepatronen in vergelijking met de VS, met een hogere prevalentie van NSAID-gebruik en een lager percentage opioïdgebruik bij Taiwanese patiënten.

In vergelijking met de ontwikkelingscohort voor het SORG-algoritme waren de patiënten in deze cohort ouder, hadden ze meer opnames, een lager inkomen, zoals verwacht een hoger preoperatief NSAID-gebruik en minder depressie als comorbiditeit. Alle patiënten in de validatiecohort hadden een nationale ziektekostenverzekering, terwijl de Amerikaanse ontwikkelingscohort uit verschillende soorten verzekeringen bestond. Ondanks deze duidelijke verschillen in basiskenmerken en een zeer streng nationaal opioïdbeleid in Taiwan heeft het SORG-algoritme goede discriminatieve vermogens en een goede algehele prestatie in een Han-Chinese patiëntengroep.

Hoofdstuk 8

Een epiduraal spinaal abces brengt diagnostische en behandelingsuitdagingen met zich mee vanwege de zeldzaamheid en niet-specifieke symptomen, wat leidt tot beperkte klinische ervaring bij behandelende artsen, met name wervelkolomchirurgen en internisten. Karhade et al. ontwikkelden een voorspellingsmodel voor het falen van niet-operatieve behandeling, voortbouwend op het nomogram uit hoofdstuk 3. Ondanks het groeiende aantal voorspellingsmodellen is het essentieel om de impact van dergelijke modellen op de besluitvorming te beoordelen. In hoofdstuk 8 wordt momenteel een interobserver-onderzoek uitgevoerd om te evalueren hoe machine learning-modellen behandelaanbevelingen beïnvloeden, met de nadruk op het vaststellen van verschillen in aanbevelingen tussen artsen die gebruikmaken van het model en artsen zonder ondersteuning van het model.

List of Publications

External validation of machine learning algorithm predicting prolonged opioid prescriptions in opioid-naïve lumbar spine surgery patients using a Taiwanese cohort.
Chen SF, Su CC, Huang CC, **Ogink PT**, Yen HK, Groot OQ, Hu MH.
J Formos Med Assoc. 2023 Dec;122(12):1321-1330.

The Skeletal Oncology Research Group Machine Learning Algorithm (SORG-MLA) for predicting prolonged postoperative opioid prescription after total knee arthroplasty: an international validation study using 3,495 patients from a Taiwanese cohort.
Tsai CC, Huang CC, Lin CW, **Ogink PT**, Su CC, Chen SF, Yen MH, Verlaan JJ, Schwab JH, Wang CT, Groot OQ, Hu MH, Chiang H.
BMC Musculoskelet Disord. 2023 Jul 5;24(1):553.

Social determinants of health in prognostic machine learning models for orthopaedic outcomes: A systematic review.
Lans A, Kanbier LN, Bernstein DN, Groot OQ, **Ogink PT**, Tobert DG, Verlaan JJ, Schwab JH.
J Eval Clin Pract. 2023 Mar;29(2):292-299.

Preoperative embolization in surgical treatment of spinal metastases originating from non-hypervascular primary tumors: a propensity score matched study using 495 patients.
Groot OQ, van Steijn NJ, **Ogink PT**, Pierik RJ, Bongers MER, Zijlstra H, de Groot TM, An TJ, Rabinov JD, Verlaan JJ, Schwab JH.
Spine J. 2022 Aug;22(8):1334-1344.

Practice Variation Within a Single Institution in Management of Degenerative Spondylolisthesis.
Ogink PT, Groot OQ, van Steijn N, Im GH, Cha TD, Hershman SH, Bono CM, Schwab JH.
Clin Spine Surg. 2022 Jul 1;35(6):E546-E550.

A machine learning algorithm for predicting prolonged postoperative opioid prescription after lumbar disc herniation surgery. An external validation study using 1,316 patients from a Taiwanese cohort.
Yen HK, **Ogink PT**, Huang CC, Groot OQ, Su CC, Chen SF, Chen CW, Karhade AV, Peng KP, Lin WH, Chiang H, Yang JJ, Dai SH, Yen MH, Verlaan JJ, Schwab JH, Wong TH, Yang SH, Hu MH.
Spine J. 2022 Jul;22(7):1119-1130.

Albumin and Survival in Extremity Metastatic Bone Disease: An Analysis of Two Independent Datasets.
Thio QCBS, Karhade AV, Pham A, **Ogink PT**, Ferrone ML, Schwab JH.
Nutr Cancer. 2022;74(6):1986-1993.

Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review.
Ogink PT, Groot OQ, Karhade AV, Bongers MER, Oner FC, Verlaan JJ, Schwab JH.
Acta Orthop. 2021 Oct;92(5):526-531.

Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review.
Groot OQ, Bindels BJJ, **Ogink PT,** Kapoor ND, Twining PK, Collins AK, Bongers MER, Lans A, Oosterhoff JHF, Karhade AV, Verlaan JJ, Schwab JH.

Acta Orthop. 2021 Aug;92(4):385-393.

Machine learning prediction models in orthopedic surgery: A systematic review in transparent reporting.
Groot OQ, **Ogink PT**, Lans A, Twining PK, Kapoor ND, DiGiovanni W, Bindels BJJ, Bongers MER, Oosterhoff JHF, Karhade AV, Oner FC, Verlaan JJ, Schwab JH.
J Orthop Res. 2022 Feb;40(2):475-483.

Do Cohabitants Reliably Complete Questionnaires for Patients in a Terminal Cancer Stage when Assessing Quality of Life, Pain, Depression, and Anxiety?
Groot OQ, Paulino Pereira NR, Bongers MER, **Ogink PT**, Newman ET, Verlaan JJ, Raskin KA, Lozano-Calderon SA, Schwab JH.
Clin Orthop Relat Res. 2021 Apr 1;479(4):792-801.

The use of autologous free vascularized fibula grafts in reconstruction of the mobile spine following tumor resection: surgical technique and outcomes.
Bongers MER, **Ogink PT**, Chu KF, Patel A, Rosenthal B, Shin JH, Lee SG, Hornicek FJ, Schwab JH.
J Neurosurg Spine. 2020 Nov 6;34(2):283-292.

Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? A Systematic Review.
Groot OQ, Bongers MER, **Ogink PT**, Senders JT, Karhade AV, Bramer JAM, Verlaan JJ, Schwab JH.
Clin Orthop Relat Res. 2020 Dec;478(12):2751-2764.

The Prevalence of Calcifications at the Origin of the Extensor Carpi Radialis Brevis Increases with Age.
Tarabochia M, Janssen SJ, **Ogink PT**, Ring D, Chen NC.
Arch Bone Jt Surg. 2020 Jan;8(1):21-26.

Development and Internal Validation of Machine Learning Algorithms for Preoperative Survival Prediction of Extremity Metastatic Disease.
Thio QCBS, Karhade AV, Bindels BJJ, **Ogink PT,** Bramer JAM, Ferrone ML, Calderón SL, Raskin KA, Schwab JH.
Clin Orthop Relat Res. 2020 Feb;478(2):322-333.

Discharge Disposition After Anterior Cervical Discectomy and Fusion.
Karhade AV, **Ogink PT**, Thio QCBS, Cha TD, Hershman SH, Schoenfeld AJ, Bono CM, Schwab JH.
World Neurosurg. 2019 Dec;132:e14-e20.

External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease.
Karhade AV, Ahmed AK, Pennington Z, Chara A, Schilling A, Thio QCBS, **Ogink PT**, Sciubba DM, Schwab JH.
Spine J. 2020 Jan;20(1):14-21.

Development of machine learning algorithms for prediction of prolonged opioid prescription after surgery for lumbar disc herniation.

Karhade AV, **Ogink PT**, Thio QCBS, Cha TD, Gormley WB, Hershman SH, Smith TR, Mao J, Schoenfeld AJ, Bono CM, Schwab JH.
Spine J. 2019 Nov;19(11):1764-1771.

High Risk of Symptomatic Venous Thromboembolism After Surgery for Spine Metastatic Bone Lesions: A Retrospective Study.
Groot OQ, **Ogink PT**, Paulino Pereira NR, Ferrone ML, Harris MB, Lozano-Calderon SA, Schoenfeld AJ, Schwab JH.
Clin Orthop Relat Res. 2019 Jul;477(7):1674-1686.

Sagittal spinal parameters after en bloc resection of mobile spine tumors.
Massier JRA, **Ogink PT**, Schlösser TPC, Ferrone ML, Hershman SH, Cha TD, Shin JH, Schwab JH.
Spine J. 2019 Oct;19(10):1606-1612.

Predicting discharge placement after elective surgery for lumbar spinal stenosis using machine learning methods.
Ogink PT, Karhade AV, Thio QCBS, Gormley WB, Oner FC, Verlaan JJ, Schwab JH.
Eur Spine J. 2019 Jun;28(6):1433-1440.

Development of a machine learning algorithm predicting discharge placement after surgery for spondylolisthesis.
Ogink PT, Karhade AV, Thio QCBS, Hershman SH, Cha TD, Bono CM, Schwab JH.
Eur Spine J. 2019 Aug;28(8):1775-1782.

Predicting 90-Day and 1-Year Mortality in Spinal Metastatic Disease: Development and Internal Validation.
Karhade AV, Thio QCBS, **Ogink PT**, Bono CM, Ferrone ML, Oh KS, Saylor PJ, Schoenfeld AJ, Shin JH, Harris MB, Schwab JH.
Neurosurgery. 2019 Oct 1;85(4):E671-E681.

Prognostic value of serum alkaline phosphatase in spinal metastatic disease.
Karhade AV, Thio QCBS, Kuverji M, **Ogink PT**, Ferrone ML, Schwab JH.
Br J Cancer. 2019 Mar;120(6):640-646.

Neutrophil to lymphocyte ratio and mortality in spinal epidural abscess.
Karhade AV, Shah KC, Shah AA, **Ogink PT**, Nelson SB, Schwab JH.
Spine J. 2019 Jul;19(7):1180-1185.

Machine learning for prediction of sustained opioid prescription after anterior cervical discectomy and fusion.
Karhade AV, **Ogink PT**, Thio QCBS, Broekman MLD, Cha TD, Hershman SH, Mao J, Peul WC, Schoenfeld AJ, Bono CM, Schwab JH.
Spine J. 2019 Jun;19(6):976-983.

Allograft reconstruction of the humerus: Complications and revision surgery.
Ogink PT, Teunissen FR, Massier JR, Raskin KA, Schwab JH, Lozano-Calderon SA.
J Surg Oncol. 2019 Mar;119(3):329-335.

Albumin and Spinal Epidural Abscess: Derivation and Validation in Two Independent Data Sets.
Karhade AV, Shah AA, Lin KY, **Ogink PT**, Shah KC, Nelson SB, Schwab JH.

World Neurosurg. 2019 Mar;123:e416-e426.

Development of Machine Learning Algorithms for Prediction of 30-Day Mortality After Surgery for Spinal Metastasis.
Karhade AV, Thio QCBS, **Ogink PT**, Shah AA, Bono CM, Oh KS, Saylor PJ, Schoenfeld AJ, Shin JH, Harris MB, Schwab JH.
Neurosurgery. 2019 Jul 1;85(1):E83-E91.

Development of machine learning algorithms for prediction of discharge disposition after elective inpatient surgery for lumbar degenerative disc disorders.
Karhade AV, **Ogink P**, Thio Q, Broekman M, Cha T, Gormley WB, Hershman S, Peul WC, Bono CM, Schwab JH.
Neurosurg Focus. 2018 Nov 1;45(5):E6.

Practice Variation Among Surgeons Treating Lumbar Spinal Stenosis in a Single Institution.
Ogink PT, van Wulfften Palthe O, Teunis T, Bono CM, Harris MB, Schwab JH, Cha TD.
Spine (Phila Pa 1976). 2019 Apr 1;44(7):510-516.

Can Machine-learning Techniques Be Used for 5-year Survival Prediction of Patients With Chondrosarcoma?
Thio QCBS, Karhade AV, **Ogink PT**, Raskin KA, De Amorim Bernstein K, Lozano Calderon SA, Schwab JH.
Clin Orthop Relat Res. 2018 Oct;476(10):2040-2048.

High Risk of Venous Thromboembolism After Surgery for Long Bone Metastases: A Retrospective Study of 682 Patients.
Groot OQ, **Ogink PT**, Janssen SJ, Paulino Pereira NR, Lozano-Calderon S, Raskin K, Hornicek F, Schwab JH.
Clin Orthop Relat Res. 2018 Oct;476(10):2052-2061.

Serum alkaline phosphatase and 30-day mortality after surgery for spinal metastatic disease.
Karhade AV, Thio QCBS, **Ogink PT**, Schwab JH.
J Neurooncol. 2018 Oct;140(1):165-171.

Development of Machine Learning Algorithms for Prediction of 5-Year Spinal Chordoma Survival.
Karhade AV, Thio Q, **Ogink P**, Kim J, Lozano-Calderon S, Raskin K, Schwab JH.
World Neurosurg. 2018 Nov;119:e842-e847.

Development of Predictive Algorithms for Pre-Treatment Motor Deficit and 90-Day Mortality in Spinal Epidural Abscess.
Shah AA, **Ogink PT**, Harris MB, Schwab JH.
J Bone Joint Surg Am. 2018 Jun 20;100(12):1030-1038.

Complications and reoperations after surgery for 647 patients with spine metastatic disease.
Paulino Pereira NR, **Ogink PT**, Groot OQ, Ferrone ML, Hornicek FJ, van Dijk CN, Bramer JAM, Schwab JH.
Spine J. 2019 Jan;19(1):144-156.

The Prevalence of Incidental and Symptomatic Lumbar Synovial Facet Cysts.
Janssen SJ, **Ogink PT**, Schwab JH.

Clin Spine Surg. 2018 Jun;31(5):E296-E301

Independent predictors of spinal epidural abscess recurrence.
Shah AA, Yang H, **Ogink PT**, Schwab JH.
Spine J. 2018 Oct;18(10):1837-1844.

Nonoperative Management of Spinal Epidural Abscess: Development of a Predictive Algorithm for Failure.
Shah AA, **Ogink PT**, Nelson SB, Harris MB, Schwab JH.
J Bone Joint Surg Am. 2018 Apr 4;100(7):546-555.

Variation in costs among surgeons for lumbar spinal stenosis.
Ogink PT, Teunis T, van Wulfften Palthe O, Sepucha K, Bono CM, Schwab JH, Cha TD.
Spine J. 2018 Sep;18(9):1584-1591.

Clinical Outcome After Arthroscopic Debridement and Microfracture for Osteochondritis Dissecans of the Capitellum.
Bexkens R, van den Ende KIM, **Ogink PT**, van Bergen CJA, van den Bekerom MPJ, Eygendaal D.
Am J Sports Med. 2017 Aug;45(10):2312-2318.

Cognitive intrusion of pain and catastrophic thinking independently explain interference of pain in the activities of daily living.
Talaei-Khoei M, **Ogink PT**, Jha R, Ring D, Chen N, Vranceanu AM.
J Psychiatr Res. 2017 Aug;91:156-163.

Donor-site morbidity after osteochondral autologous transplantation for osteochondritis dissecans of the capitellum: a systematic review and meta-analysis.
Bexkens R, **Ogink PT,** Doornberg JN, Kerkhoffs GMMJ, Eygendaal D, Oh LS, van den Bekerom MPJ.
Knee Surg Sports Traumatol Arthrosc. 2017 Jul;25(7):2237-2246.

Reoperation After Combined Injury of the Index Finger: Repair Versus Immediate Amputation.
Wilkens SC, Claessen FM, **Ogink PT**, Moradi A, Ring D.
J Hand Surg Am. 2016 Mar;41(3):436-40.e4.

Acknowledgments

This chapter marks the end of my long PhD journey. After years of jokingly threatening to write a book about research and residency life in the style of The House of God, I've decided to let you all off the hook. What happens in research life must stay in research life.

As I reflect on this journey, I do so with profound pleasure and gratitude for everything it has brought me—most importantly, the incredible people I met along the way. Some of you I haven't seen in quite a while, but that doesn't mean I've forgotten about you. I hope our paths cross again somewhere, someday soon.

Geachte prof. Verlaan, beste JJ

Toen ik voor het eerst langskwam in Utrecht om het idee van mijn promotie te pitchen moest ik weer even wennen. Na alleen maar 'that's awesome' en 'incredible stuff' was jouw 'mwah niet slecht' op mijn promotieplan een kleine cultuurschok. Je hebt er vrij lang op moeten wachten en ongetwijfeld wel eens gedacht dat het er niet meer van zou komen. Maar het is toch echt gelukt. Ik wil je ten eerste bedanken voor de heldere begeleiding en feedback die je de afgelopen jaren hebt gegeven. Zowel manuscripten als presentaties kwamen altijd met ijzersterk commentaar terug. Je bent met je bedrijf en de manier waarop je de metastasenzorg hebt vormgegeven een voorbeeld voor velen, waaronder mij, dat er grotere dingen zijn dan alleen protheses knallen op OK. Ik zal echter ook niet ontkennen dat we als AIOS op weg terug naar de assistentenkamer na een ellenlange overdracht soms verzuchtten "was JJ er nog maar". Ik wens je het allerbeste in het persoonlijke en het professionele. Dank voor alles!

Dear Dr. Schwab, dear Joe

It has been a while since I arrived in Boston for the first time. When Nuno approached me to come join your team for a full 2 years I was a little intimated at first. This was the big leagues. As I entered your office it was filled to the brim with manuscripts, articles and a significant number of history books of which 1776 by David Mccullough always stood out. I thought about the decision that followed a lot recently. My life took a different course and I consider it to be a far better course than it would have been otherwise. You have played a pivotal role in all this. Your work attitude, intellectual curiosity, positivity and kindness is an example to us all. My sincere gratitude for everything you've done.

Beste Quirina

We kwamen tegelijk aan in Boston en gingen ook ongeveer tegelijk weg. Jij bent m'n ultieme MGH-buddy en zonder jou was alles anders geweest. We hebben onze onderzoekskinders goed opgevoed en menigeen is nu zelf al gepromoveerd. Nu eindelijk dan ook ik. We hebben ons in die jaren moeten navigeren door een eindeloos stroom meetings, intelligente en minder intelligente short-termers, fitties met Tina, opstandige research coördinators en de continue stress op de achtergrond van het "in beeld blijven voor de opleiding" een oceaan verderop. Ik ben blij voor je dat je je plek bij de sportgeneeskunde hebt gevonden en natuurlijk ook met je jonge gezin. Dankje voor alles!

Beste Olivier

Begonnen als mijn onderzoeksstudent en nu zelf al jaren gepromoveerd. Het kan verkeren. We hebben sinds de Boston jaren het een en ander meegemaakt; teveel om op te noemen en vooral ook onverstandig om hier op te noemen. Ik kan wel eerlijk zeggen dat dit boekje er niet was gekomen zonder jou. Bij terugkomst in Nederland was het onderzoeksvuur bij mij nogal gedoofd, maar die werd door jou persoonlijk met een vlammenwerper geregeld weer geactiveerd. Ik ben erg trots op het feit dat jouw onderzoeksvuur nog steeds niet gedoofd is, maar vooral natuurlijk dat je in opleiding bent gekomen. We gaan elkaar de komende jaren nog eindeloos vaak spreken, maar bij dezen dank voor alles. Go Pats y Venga!

Dear Aditya

Thank you very much for everything you did for me and our research team. After your "transfer" from neurosurgery we were a little unsure what you wanted and what you could contribute, but that took about 5 minutes. I'm pretty sure it's the biggest steal since Brady got picked in the 6th round. You have moved from orthopedic surgery to bigger and better things but I hope you haven't forgotten about us.

Dear Jason and Sarah

Thank you for all your support in helping me navigate the wonderful world of MGH. Like many others on this journey, I haven't seen either of you in quite some time, but I hope our paths cross again soon. Wishing you both the very best of luck!

Beste Bianca Verbeek, Nuno Rui Paulino Pereira, Stein Janssen, Olivier van Wulfften Palthe, Bart Lubberts, Kamil Oflazoglu, Rens Bexkens, Alex Kernkamp, Nick Hilgersom, Joeky Senders, Vincent Groot, Bastiaan van Hoorn, Tom Kootstra, Reinout Heijboer, David Langerhuizen, Reinier Beks en Rens Varkevisser

Zonder jullie was mijn tijd in Boston niet hetzelfde geweest. Ik wil jullie ten eerste bedanken voor het samenwerken en/of begeleiden van mijn eerste stapjes in de wondere wereld van de research. Maar bovenal wil ik jullie bedanken voor alles daaromheen. Huisfeestjes aan Harvard Street, kampioen in Austin, Brahmin, puntengrafiek, 3e pagina op PubMed, Marco Island, Victoria's Secret op de beamer, Chicago, HBS rugby, de Coogans, Thai Fri, Red Sox, bakkies in Baltimore, Shoulders the Feeling hitje, Super Bowls: het was geweldig en dat was het dankzij jullie.

Beste Julie Massier, Nicole van Steijn, Merel Stor, Bas Bindels, Florine Binnendijk, Kayoumars Azizpour en Alexander Goudriaan

Lieve onderzoekskinders, Quirina, Bianca en ik zijn nog steeds trots op jullie. Dank voor al jullie inzet en gezelligheid tijdens jullie tijd in MGH.

Beste Michiel Bongers

Je bent een waardig opvolger gebleken. Niet alleen qua onderzoek maar ook dat je het stokje bij HBS Rugby hebt overgenomen. Dat je de Harvard Street hebt laten lopen heb ik je ondertussen vergeven. Ik zal verder ook nooit meer schoenen op een tafel doen.

Dear dr. Cha, dr. Lozano-Calderon and dr. Hershman

My sincere gratitude for your help and guidance during my time at MGH. I hope you are all doing well and hope to see you at future orthopedic conferences.

Chers Hadrien Laubie, Pierre Baduel, Julie Vu et Valentin Geber

Quand on part pour la première fois en Amérique et qu'on décide même d'y vivre, on ne s'attend pas forcément à se retrouver plongé dans le monde merveilleux des Français. Pourtant, c'est heureusement ce qui m'est arrivé. Je vous remercie chaleureusement pour l'accueil dans la maison franco-américaine au 24 Mag, ainsi que pour l'introduction à votre enclave française à Boston. J'espère vous revoir très bientôt!

Beste Christel Braaksma, Mathilde Tol, Bibian Schaffer, Jup Kuipers, Joost van Erp, Thom Snijders, Jos Oudeman, Reiner Spek, Anneke Voorhuis, Roel Janssens, Steven de Reuver, WP Gielis, Ran Hendrix, Daniel Verstift, Justin Lemans, Roderick Piekaar, Anouk van de Kuit, Anouk van der Vossen, Yordi de Wit, Soufyan Kalaai Mara van der Valk, Fien de Nies, Sonny Hopman, Sophie Uittenbogaard, Ted van Iersel, Pascal van Diepen, Johan Heemskerk, Jan Jaap de Graeff, Marianne Koolen, Jonneke Kuperus, Dino Colo, Huub de Visser, Mark Flipsen, Dirk ter Meulen, Kiran Mahabier, Marrit Hoekstra, Said Sadiqi en alle andere collega ANIOS/AIOS van de afgelopen jaren die ik vergeten ben

Ondanks dat ik het niet altijd laat merken ga ik eigenlijk altijd met een grijns op m'n gezicht naar werk. Dat is voor een groot deel aan jullie te danken. De assistentenkamer in UMC/OLVG/Anton voelt dankzij jullie als een huiskamer. Zak chips open, poten op tafel en eindeloos ouwehoeren. Al gaat het richting het einde, ik blijf jullie hopelijk de komende decennia natuurlijk zien op congressen.

Beste Martijn van Dijk, Diyar Delawi, Matthijs Krijnen, Derek van Deurzen, Nienke van Egmond en Bart van de Wal

Ook jullie wil ik graag bedanken. Ik kan me simpelweg niet meer op mijn gemak voelen dan in deze ROGO. Jullie zijn altijd betrokken, persoonlijk geïnteresseerd en geven ons AIOS het broodnodige vertrouwen wat deze onzekere generatie soms wel eens extra nodig heeft. Ik spreek namens alle AIOS dat ik jullie wil bedanken voor jullie begeleiding als opleiders.

Beste Mees Emmelot en Bas Bindels, mentorkindjes gedraag je tijdens het promotiefeest de 21^{e}! Hoop career-limiting uurtjes die jullie ongeschonden moeten doorlopen.

Dear Patrick, Chip and Conny Lane

Thank you for your hospitality in Marco Island. You represent what I liked best in America: positive, fun and enjoying life. Chip, I just had to borrow that one-liner you told us after partying that last day.

Beste Obelixers, beste vrienden

Het spijt me maar ik ga jullie niet allemaal opnoemen want daarvoor zijn er veel teveel van jullie. Het zwaarste aan het onderzoeksleven was natuurlijk de tijd zonder geel-groen. Het afscheid naar Amerika de avond voor de vlucht was op z'n zachtst gezegd legendarisch en het gesigneerde shirt hing jarenlang op m’n kamer. Blij jullie nu wel weer in levende lijven te kunnen zien en dank voor alles de afgelopen 18 jaar.

Beste Elsien en Brent, Olav en Augusta, Helena en Laurens
De meeste van jullie kunnen dit nog niet lezen, maar later wel. Jullie zijn het zonnetje in m’n leven en jullie hebben geen idee hoe blij ik er van word jullie te zien.

Beste Richard en Kitty, Stephanie en Tom, Vincent en Ilse
Dank voor alle steun die ik als nakomertje van kinds af aan al van jullie heb gekregen. Zonder jullie was ik nooit geworden wie ik nu ben. M’n talenknobbel is al vroeg gestimuleerd dankzij het tot 10 tellen in al die verschillende talen.

Beste pap en mam
Elk gebouw heeft een stevig fundament nodig en het fundament wat jullie me hebben gegeven is rock solid. Jullie zijn er altijd voor me geweest en hebben me het fundament gegeven voor alle avonturen van de afgelopen jaren. Soms blik ik terug op de afgelopen jaren van Luzern naar Amerika naar Utrecht en vraag ik me af hoe het allemaal zo is gelopen. Hoe de ideeën in me opkwamen weet ik soms niet meer, maar waar ik altijd van op aan kon was jullie steun en vertrouwen. Dit boek is in de allereerste plaats aan jullie opgedragen. Vanuit het diepste van m’n hart: dank voor alles.

Curriculum Vitae Auctoris

Paul Theodor Ogink was born on November 4th 1988 in Roosendaal, the Netherlands. He grew up in Roosendaal and graduated from Gertrudiscollege top of his class in 2007. He subsequently started medical school at Radboud University, Nijmegen. As a student he joined the student rugby club NSRV Obelix which he headed as president in 2011. Before graduation he did a rotation abroad at Luzerner Kantonspital, Switzerland. After graduating he moved to Boston to join Dr. David Ring's research group at the Hand and Upper Extremity department at Mass General Hospital. After six months he returned to Boston to spend the next 2 years with Dr. Joseph Schwab's Skeletal Oncology Research Group. He coordinated a prospective study investigating the association between single nucleotide polymorphisms and lumbar disc herniation. During this time he became interested in machine learning based prediction models, which became the focus of his research. Currently he has 44 peer-reviewed articles and has presented his work at several international conferences. In 2019 he was nominated for the van Rensprijs at the Dutch orthopedic society's annual conference.

Paul currently lives in Utrecht, the Netherlands and is in his residency orthopedic surgery in this city.