Artificial Intelligence and Predictive Modeling in Spinal Oncology: A Narrative Review

Rene Harmen Kuijten, Hester Zijlstra, Olivier Quinten Groot and Joseph Hasbrouck Schwab

Int J Spine Surg 2023, 17 (S1) S45-S56
doi: https://doi.org/10.14444/8500
https://www.ijssurgery.com/content/17/S1/S45

The International Journal of Spine Surgery
2397 Waterbury Circle, Suite 1,
Aurora, IL 60504, Phone: +1-630-375-1432

International Journal of Spine Surgery, Vol. 17, No. S1, 2023, pp. S45–S56
https://doi.org/10.14444/8500

Artificial Intelligence and Predictive Modeling in Spinal Oncology: A Narrative Review

RENE HARMEN KUIJTEN, BSc[1,2]; HESTER ZIJLSTRA, MD[1,2]; OLIVIER QUINTEN GROOT, MD, PhD[1,2]; AND JOSEPH HASBROUCK SCHWAB, MD, MS[1]

[1]*Department of Orthopedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA;* [2]*Department of Orthopedic Surgery, University Medical Center Utrecht, Utrecht University, Heidelberglaan, The Netherlands*

ABSTRACT

Background: Artificial intelligence (AI) tremendously influences our daily lives and the medical field, changing the scope of medicine. One of the fields where AI, and, in particular, predictive modeling, holds great promise is spinal oncology. An accurate patient prognosis is essential to determine the optimal treatment strategy for patients with spinal metastases. Multiple studies demonstrated that the physician's survival predictions are inaccurate, which resulted in the development of numerous predictive models. However, difficulties arise when trying to interpret these models and, more importantly, assess their quality.

Objective: To provide an overview of all stages and challenges in developing predictive models using the Skeletal Oncology Research Group machine learning algorithms as an example.

Methods: A narrative review of all relevant articles known to the authors was conducted.

Results: Building a predictive model consists of 6 stages: preparation, development, internal validation, presentation, external validation, and implementation. During validation, the following measures are essential to assess the model's performance: calibration, discrimination, decision curve analysis, and the Brier score. The structured methodology in developing, validating, and reporting the model is vital when building predictive models. Two principal guidelines are the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis checklist and the prediction model risk of bias assessment. To date, many predictive modeling studies lack the right validation measures or improperly report their methodology.

Conclusions: A new health care age is being ushered in by the rapid advancement of AI and its applications in spinal oncology. A myriad of predictive models are being developed; however, the subsequent stages, quality of validation, transparent reporting, and implementation still need improvement.

Clinical Relevance: Given the rapid rise and use of AI prediction models in patient care, it is valuable to know how to assess their quality and to understand how these models influence clinical practice. This article provides guidance on how to approach this.

Level of Evidence: 4.

Other and Special Categories

Keywords: artificial intelligence, machine learning, orthopedic surgery, prediction tools, clinical decision support, spinal oncology

INTRODUCTION

Artificial intelligence (AI) tremendously influences not only our daily lives but also the medical field, changing the scope of medicine. Improvements in computational power, along with AI-based software platforms, and the availability of more extensive electronic data, have enabled the development of many different applications, such as machine learning (ML)–derived clinical decision support tools, deep learning-based computer vision, and natural language processing.[1] Oosterhoff et al suggested in 2020 that we have reached the peak of inflated expectations in medical AI along with Gartner's hype cycle (Figure 1).[2] Although the promise of AI remains strong, where an individual stands on the hype cycle would depend on their experience and understanding of AI. Individuals new in this field can still be at the peak of inflated expectations, while more experienced individuals might be toiling through the trough of disillusionment as challenges in implementing AI applications are becoming more apparent. The purpose of the present article is to provide a narrative review of AI and predictive modeling in spinal oncology and discuss the potential and limitations of the technology. We present no unpublished data and reference to data from previously published studies.

Spinal Oncology

One of the fields where AI, and, in particular, predictive modeling, has made significant advances is spinal oncology. The spine is the most common location of

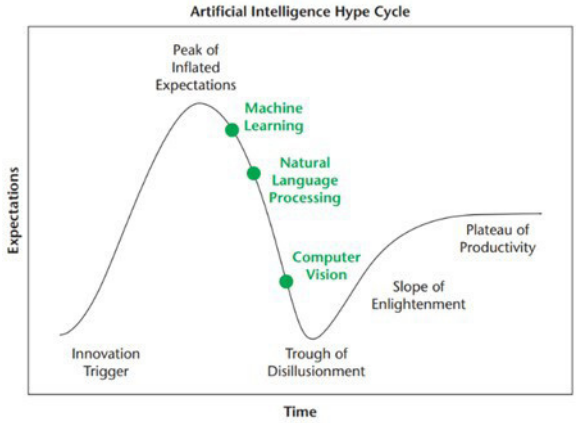

Figure 1. Gartner's hype cycle. Source: Reprinted with permission from Oosterhoff JHF, Doornberg JN. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. EFORT Open Rev. 2020;5(10):593–603. © 2020 Oosterhoff and Doornerg.

metastatic cancer disease,[3–5] and 30% to 90% of patients who die of cancer have spinal metastasis in cadaver studies.[6–9] Up to 50% of spinal metastasis require treatment, and 5% to 10% need surgical management.[8,10,11] Moreover, cancer survival rates are increasing due to earlier detection and improved treatment, and the prevalence of spinal metastasis will also likely increase.[12] In 2005, the landmark article of Patchell et al[13] showed that surgical intervention is efficacious in treating metastatic spinal tumors. Following this, together with the emergence of a myriad of treatments, including personalized systemic therapy and targeted therapy, a systemic decision framework for treating spinal metastases was necessary.[14] In 2015, the neurologic, oncologic, mechanical, and systemic decision framework was developed to determine the optimal therapy for patients with spinal metastases.[15] This framework enabled physicians to apply a systematic approach to treating spinal metastases, resulting in an increased surgery rate.[16] However, spinal surgery is not without risk; surgery complications are a significant source of comorbidity and include wound infections, neurologic impairment, venous thromboembolism, instrumentation failure, and pain.[17–20] Moreover, patients with metastatic spinal disease generally have multiple medical comorbidities and are immunocompromised due to immune suppression.[21] Therefore, treatment goals focus on whether patients will likely recover from the indicated procedure.[22] The appropriate use of surgery for metastatic spinal disease is dependent on the expected risk of surgery and the expected benefit. Accurate expectations for risk and benefit would be valuable to empower informed choice for physicians and patients.

The Emergence of Prediction Tools

Multiple studies have shown that physicians' clinical predictions of the life expectancy of cancer patients are inaccurate.[23,24] In 2005, Nathan et al showed that a better means of prognostication was needed.[25] Consequently, numerous new scoring systems and prognostic calculators were developed.[26–36] Unfortunately, many did not meet the required accuracy, performed inconsistently, or lacked personalized predictions.[26,37] Thirteen survival prediction scores exist, including PATHFx,[38] Skeletal Oncology Research Group ML algorithms (SORG-MLA),[33] Bollen Classification,[39] modified Bauer score,[34] and van der Linden[40] (Supplement 1).[27,41–47] Of these prediction scores, SORG-MLA and PATHFx are the only 2 ML algorithms. Over the past years, SORG-MLA demonstrated its clinical value and promise over other prediction scores such as nomograms or regression models. However, important questions regarding the use of AI in predictive models remain, including the following: (1) How do we interpret prognostic AI models such as SORG-MLA? (2) How do we assess their quality? and (3) How will these models influence clinical practice?

DEVELOPMENT, VALIDATION, AND IMPLEMENTATION OF PREDICTION MODELS

Why Machine Learning?

Statistical models have been widely used to formalize the understanding of data, but since data size and variable inputs increased, these models have become more complex. Fortunately, ML models have become more powerful due to an increase in computational power. According to Bzdok et al,[48] "statistics draws population inference from a sample, and ML finds generalizable predictive patterns." In principle, many methods from statistics and ML can be used for both prediction and inference. However, statistical methods have a long-standing focus on inference, achieved through creating and fitting a project-specific probability model. In contrast, ML concentrates on prediction with general purpose learning algorithms to find patterns in often rich and unwieldy data.[49,50] They are particularly helpful when dealing with "wide data," where the number of input variables exceeds the number of subjects. Thus, where statistical models are generally hypothesis-driven, ML is more exploratory in identifying correlations, and the pattern of correlation is not a causal relationship. This may be recognized as a

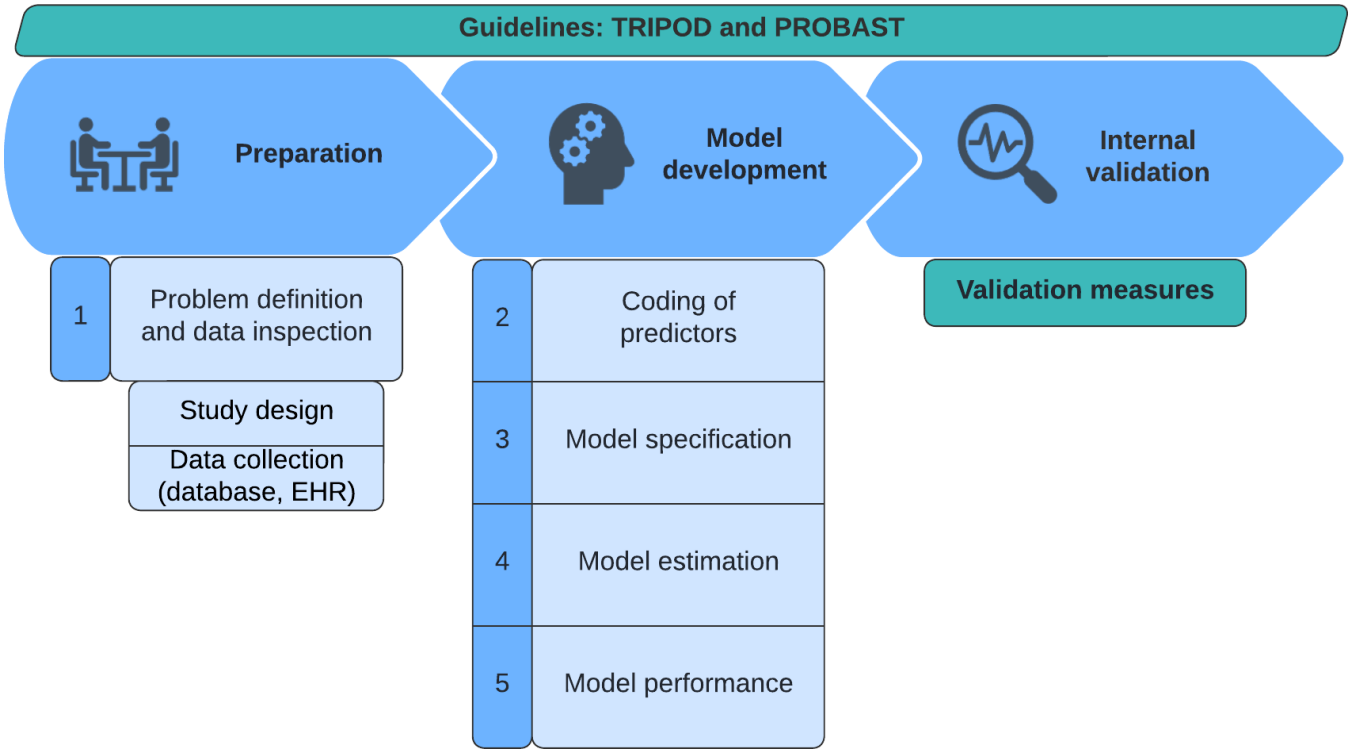

Figure 2. The first 3 stages in model development: preparation, development, and internal validation. TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; PROBAST, prediction model risk of bias assessment tool; EHR, electronic health record.

limitation of ML. However, with the possession of large patient data sets due to electronic health care systems, ML provides the opportunity to find patterns and determine values predictive of the output requested. Therefore, ML offers a more accurate solution for developing prediction models, such as the survival probability of patients with metastatic spine disease, which is complicated and requires multiple aspects to be considered.

Steps in Building Predictive Models

Structured methodology in the development and validation of an ML model is of great importance and is best executed along the ABCD steps of Steyerberg et al.[51] Additionally, 2 important guidelines are important to adhere to: the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)[52] checklist, essential for transparent reporting of a prediction model study, and the prediction model risk of bias assessment (PROBAST),[53] a tool for assessing the risk of bias and applicability of prediction model studies. With the SORG-MLA for 1-year survival, developed and validated multiple times within our research team, as an example, we will go through the steps of model preparation, development, validation, presentation, and implementation (Figures 2–4).[54–57]

The first step is the consideration of the research question and initial data inspection. For the development of the SORG-MLA, the objective was to find predictive variables and develop a predictive algorithm for survival of metastatic spinal disease at intermediate (90-day) and long-term (1-year) time points. Based on the expert knowledge and previous literature, we chose a framework of input variables to consider. Patients were included when they were older than 18 years, had a diagnosis of metastatic spinal disease, and had an initial surgical procedure performed between 1 January 2000, and 31 December 2016. Missing data were imputed with the missForest multiple imputation method, which is currently considered one of the superior imputation methods. Baseline data collection was retrospective, and the definitions of all input variables, generally referred to as predictors, were carefully documented.

The second step is the coding of the predictors. Categorical and continuous predictor variables can be coded in different ways. At the start of model development, coding the variables in a detailed way is preferred so that in a later phase, when relative effects of predictors are known, a user-friendly variable format may be used. For example, when coding the variable of primary tumor histology, we might see that coding the variable in 3

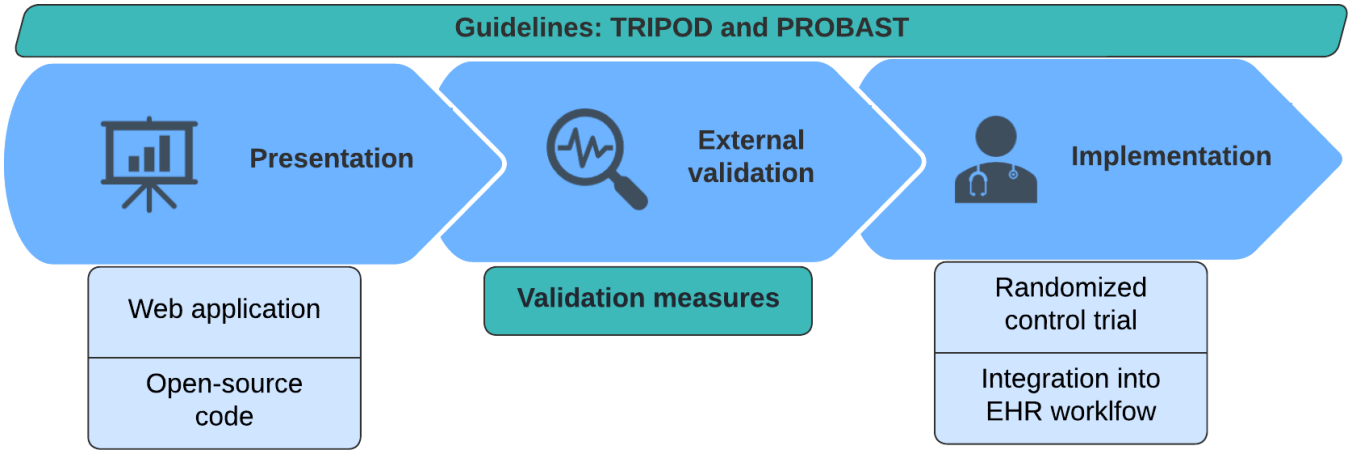

Figure 3. The last 3 stages in model development: presentation, external validation, and implementation. TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; PROBAST, prediction model risk of bias assessment tool; EHR, electronic health record.

groups according to primary tumor instead of coding them all separately would result in similar performance, making the model simpler to use.

The third step is the model specification, where we choose the predictors for inclusion in the prediction model (Figure 5). For SORG-MLA, we used random forest algorithms with 10-fold cross-validation, which enabled us to find the optimal subset of predictors while keeping the variance of the model performance low and avoiding overfitting.

The fourth step is the model estimation: choosing the right ML model (Figure 6). For SORG-MLA, we used 5 different models based on a previous study's method.[58] The data were then divided into a training set (80%) and a holdout validation set (20%). The training set is used to train the models, and the validation set is used to internally validate the model. An independent validation set is essential to test the models on unseen data.

The fifth and sixth steps are the validation and evaluation of model performance, where we determine the quality and performance of the algorithms and alter the algorithm if necessary. Evaluation and validation are ideally performed along the ABCD steps; these will be discussed in the next section.

The seventh, and final, step is the model presentation such that it best addresses the clinical needs. We presented SORG-MLA as an open access web-based application to facilitate accessibility (see https://sorg-apps.

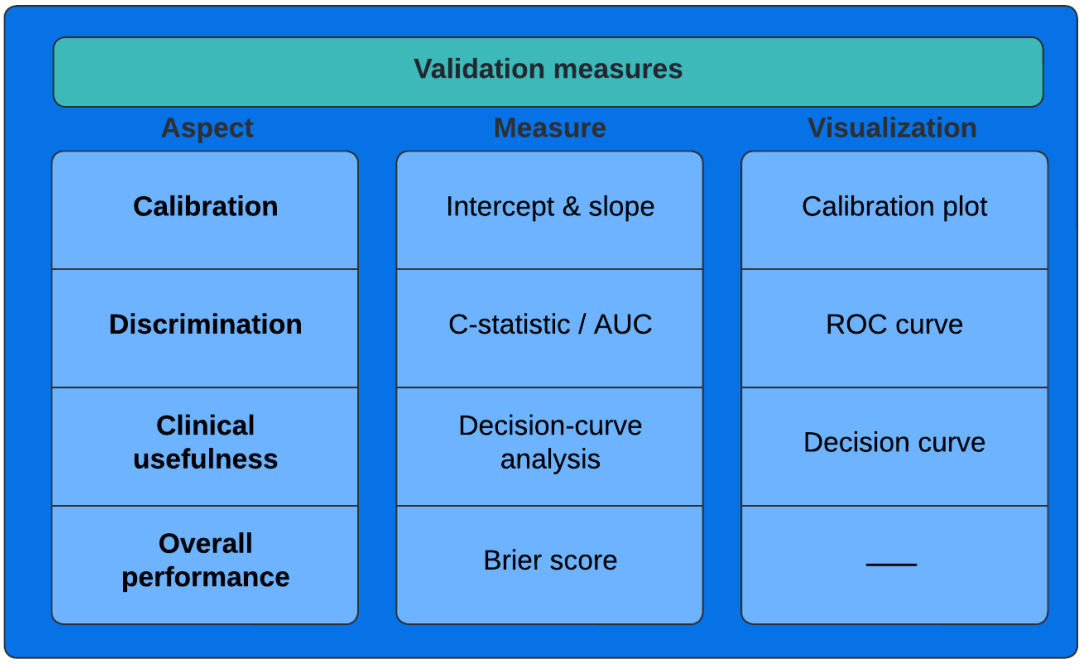

Figure 4. Overview of validation measures. C-statistic, concordance statistic; AUC, area under the curve; ROC, receiver operating characteristic.

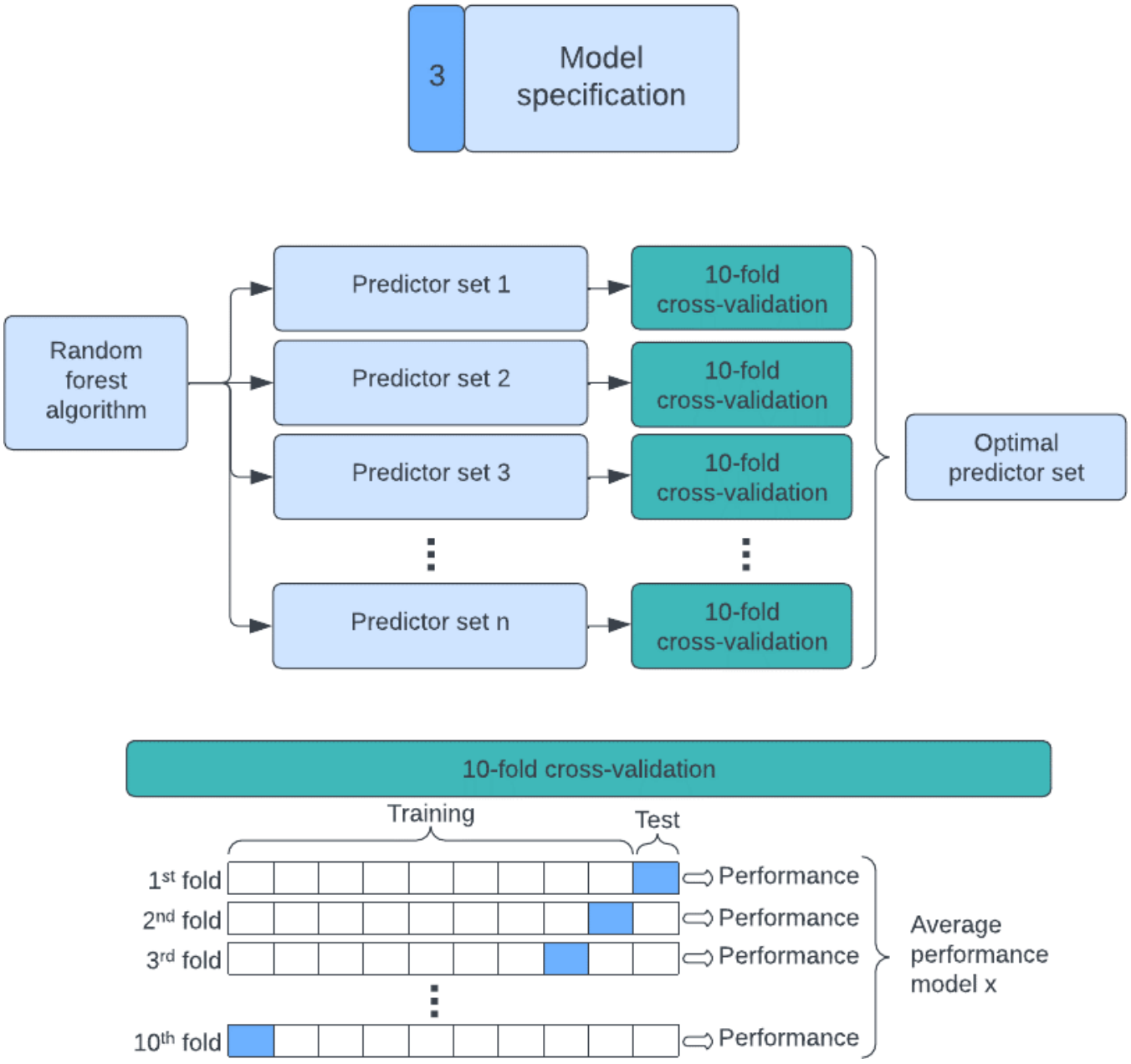

Figure 5. Model specification. With a random forest algorithm, we created many different predictor sets (sets with different input variables) which we tested with 10-fold cross-validation to find the optimal set of predictors. This technique fits the model 10 times, with each fit being performed on a training set of a different 90% of the data with the remaining 10% as a holdout set for validation. Each fit produces a performance metric, and the average of all these fits results in the average performance of a predictor set.

shinyapps.io/spinemetssurvival/). However, ultimately, integration into decision aids and electronic patient records will best support clinical decision-making.[59]

Validation Methods

Model validation is the process by which predictions are compared with independent real-world observations to judge quantitative and qualitative properties of the model. There are 4 important measures based on the "ABCD" steps of Steyerberg et al,[51] which together provide an accurate and well-established validation and evaluation: calibration, discrimination, decision curve analysis, and the Brier score.[38,57,60,61]

Calibration (A and B) refers to the agreement between observed end points and predictions and answers the question: Is the model as reliable when it predicts a 10% probability as when it predicts a 70% probability of mortality?[62] It can be best assessed graphically in a calibration plot with survival predictions on the *x*-axis and real-world observations on the *y*-axis. Perfect calibration

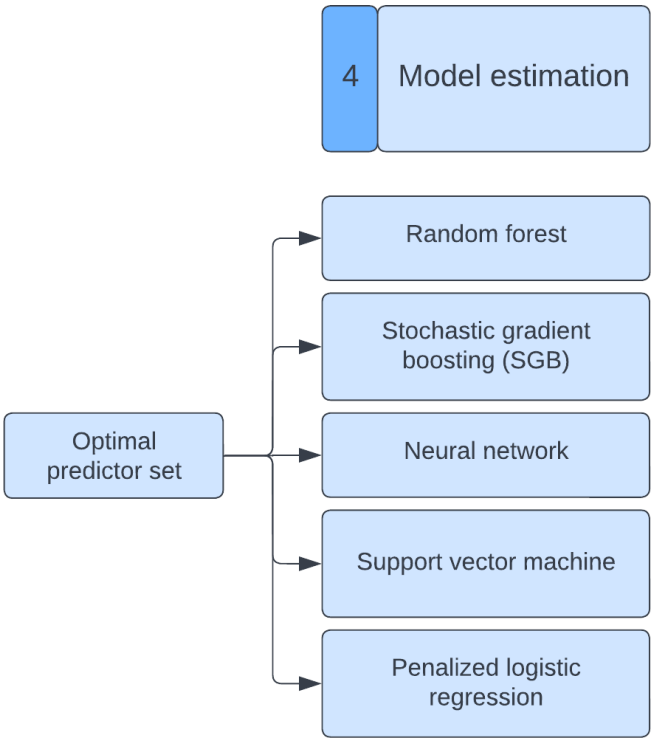

Figure 6. Model estimation. For SORG-MLA, we used 5 different models: random forests, stochastic gradient boosting, neural network, support vector machine, and penalized logistic regression. SORG-MLA, the Skeletal Oncology Research Group machine learning algorithms.

of a model should have a straight line, described with an intercept of 0 and a slope of 1. Imperfect calibration can be observed by deviation from this ideal straight line (Figure 3). This calibration plot helps visualize whether models overestimate or underestimate the outcome. The SORG-MLA achieved an intercept of 0.07 and a slope of 1.26 (Figure 7), showing a near perfect intercept and a slightly higher slope, indicating that there are individuals or subgroups in whom calibration is suboptimal and survival is overestimated.[63]

Discrimination (C) refers to the ability of the model to distinguish the end points, that is, whether a patient is dead or alive at the specified time point. The measure is quantified by the area under the curve of the receiver operating characteristic curve, which represents the probability that the model will be able to differentiate between patients who survived and those who died. Interpretation of this curve can be simplified: 0.51 to 0.69 poor, 0.70 to 0.79 fair, 0.80 to 0.89 good, 0.90 to 0.99 excellent. The SORG-MLA achieved an area under the curve of 0.89.[63]

Even though calibration, discrimination, and the Brier score are essential, these measures do not assess the clinical usefulness or the ability to make better clinical decisions with the model than without. To determine the impact of these models on clinical decisions, it is essential to perform a decision analysis (D). Even though this type of analysis has been around for a significant amount of time, it only recently started gaining popularity as a necessary tool in prediction models.[62,64] Decision curve analysis examines the net benefit of decisions made based on the model predictions. Changing management for all patients and changing management for no patients are the 2 default strategies for decisions without prediction models. Decision curves show whether the clinical prediction model used for management changes offers a greater net benefit than the 2 default strategies. The SORG-MLA showed greater standardized net benefit at all predicted probabilities relative to management decision change based on treating all patients or no patients (Figure 8).[57]

Another important measure, although not recorded in the ABCD steps, is the Brier score: a summary measure that formalizes the performance of predictions. The so-called "null model" of the Brier score corresponds to the scenario where every patient is predicted to have a risk equal to the prevalence of mortality in the whole disease population. The Brier score calculates the error between the prediction and observed outcome for each patient and compares it to the null model. Ideally, zero error between the predictions and outcomes is preferred, resulting in a perfect Brier score of 0. The SORG-MLA achieved a Brier score of 0.13, whereas the null model had a Brier score of 0.25.[63]

Validation of the model can only be adequately assessed when all measures are performed. For example, a model can have excellent discrimination but very poor calibration. Or, a model could have good discrimination and calibration but worse standardized net benefit compared with default changes in management, resulting in a model that harms clinical decision-making. Therefore, assessing and reporting every validation measure mentioned above are essential.

Internal and External Validation

Assessing model validation is executed at 2 stages: internal validation at the end of model development and external validation when the model is already presented. The difference is that internal validation is performed at the institute that develops the model, whereas external validation is done at multiple (different) institutions, assessing the model's generalizability to different patient populations. When validating a prediction model, it is important to not only assess the measures mentioned before but also assess whether the model has been developed correctly. To facilitate this, transparent and complete reporting of the development and

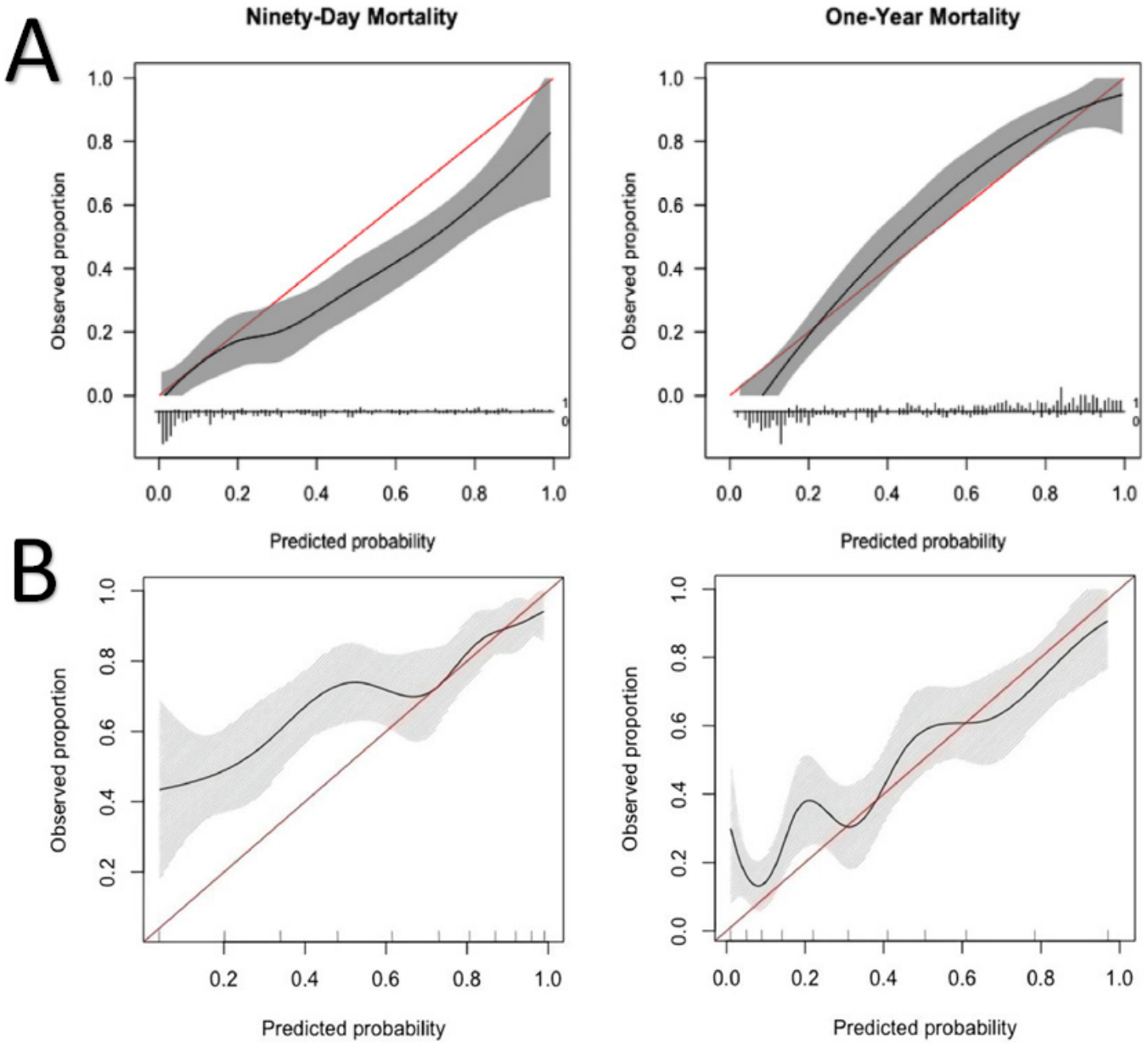

Figure 7. Calibration: calibration plot of SORG-MLA predicting 90-d and 1-y mortality at (A) internal validation and (B) external validation (Taiwan). Comparing these plots demonstrates that SORG-MLA performs differently in other populations, highlighting the importance of external validation. SORG-MLA, the Skeletal Oncology Research Group machine learning algorithms. Source: Reprinted from with permission from *The Spine Journal*, Vol 21, Yang J-J, Chen C-W, Fourman MS, 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, 1670-1678, Copyright 2021, with permission from Elsevier.[56]

validation of a model are required to allow the reader to critically assess the presence of bias, facilitate study replication, and correctly interpret results.[65] External validation of SORG-MLA has been done extensively in the United States and multiple international patient populations (Table).[54–57] However, the overall survival of patients with spinal metastases is improving and will hopefully keep improving due to improved treatments and clinical decision-making.[12] This may result in lower performance of the model in the future. Therefore, it is vital to continuously monitor and validate the performance of ML models so that clinicians and data scientists can identify and assess performance deviations as soon as possible and recalibrate or update models if necessary.

Implementation

Once external validation has been successful, the next step is implementing the model into clinical practice. An essential factor for integrating a model into clinical practice is ensuring clinicians' trust and

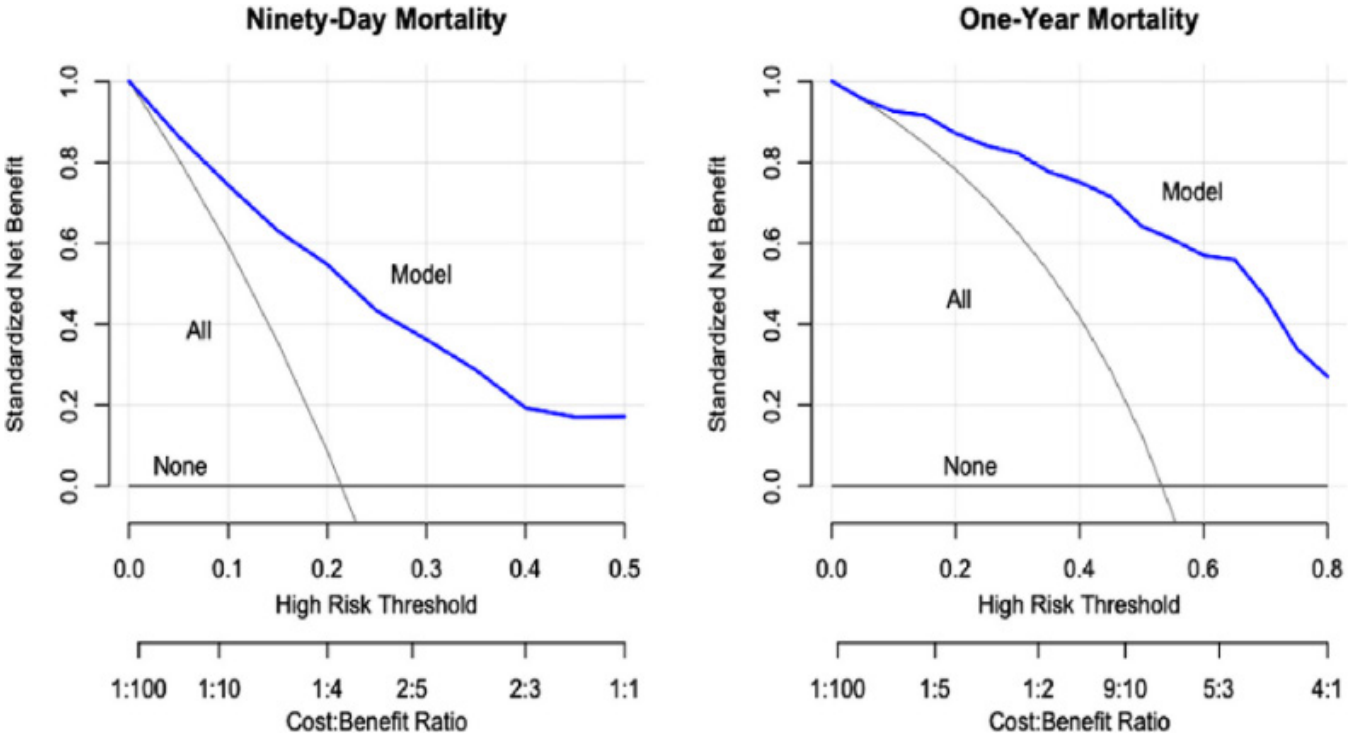

Figure 8. Decision curve analysis: decision curve of SORG-MLA predicting 90-d and 1-y mortality at external validation. SORG-MLA, the Skeletal Oncology Research Group machine learning algorithms. Source: Reprinted from *The Spine Journal*, Vol 21, Shah AA, Karhade AV, Park HY, et al, Updated external validation of the SORG machine learning algorithms for prediction of ninety-day and one-year mortality after surgery for spinal metastasis, 1679–1686, Copyright 2021, with permission from Elsevier.[57]

accurately interpreting the model.[66] To earn this trust, transparent reporting of the model's development, internal, and external validation is essential. Next, we must assess the real-world performance of the model on operational data, thus validating the algorithm on a prospective cohort by comparing the model performance with a surgeon with or without the model. Consequently, the performance of the developed model is ideally assessed with randomized control trials. Guidelines such as CONSORT-AI (Consolidated Standards of Reporting Trials - Artificial Intelligence) and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials - Artificial Intelligence) have been developed to assist in the application for these trials.

To facilitate easy access to SORG-MLA, we presented the model as an open access web application.

Table. External validations of SORG-MLA predicting 90-d and 1-y mortality.

Study	Institution, City, Country	Patients	Calibration: Intercept	Calibration: Slope	Discrimination	Brier Score Model	Brier Score Null	Decision Curve Analysis Performed
Karhade et al, 2020[54]	John Hopkins University, School of Medicine, Baltimore, USA	176						
90-d Mortality			–0.10	0.64	0.75	0.157	0.176	Yes
1-y Mortality			0.43	0.77	0.77	0.199	0.246	Yes
Bongers et al, 2020[55]	Memorial Sloan-Kettering Cancer Center, New York, USA	200						
90-d Mortality			–0.07	0.64	0.81	0.17	0.20	Yes
1-y Mortality			0.57	0.85	0.84	0.16	0.23	Yes
Yang et al, 2021[56]	National Taiwan University Hospital, Taipei, Taiwan	427						
90-d Mortality			0.81	0.51	0.73	0.17	0.19	Yes
1-y Mortality			0.08	0.59	0.74	0.20	0.24	Yes
Shah et al, 2021[57]	David Geffen School of Medicine at UCLA, USA	298						
90-d Mortality			–0.65	0.80	0.84	0.13	0.17	Yes
1-y Mortality			0.08	1.20	0.90	0.13	0.25	Yes

Abbreviation: UCLA, University of California, Los Angeles.

However, a real-time outcome calculator based on the developed ML algorithm and routinely collected data is best established, validated, and integrated within the electronic health record (EHR) systems.[59] This has implications for patient privacy and creates obstacles for implementation.[67] For SORG-MLA, we are currently performing an international, multicenter prospective study to evaluate the survival predictions of surgeons with or without the model. Consequently, if the study shows survival predictions improve significantly, implementation into EHR will follow.

RECOMMENDATIONS AND CHALLENGES

Despite the potential increased benefit of predictive models, there are restrictions and risks associated with ML models. As we have gone through all aspects of model processing, we will highlight several challenges in each stage.

Preparation

The quality of the data from which prediction models are produced determines the quality of those models.[68] Even if the amount of data is large, data inaccuracy and missing data still pose serious problems when EHR data are used and may impact prognostic factors, treatment exposures, and outcome estimation.[69] Because existing ML models are created using small, retrospective cohorts or registries, they frequently lack generalizability. This is particularly problematic in ML algorithms as they tend to amplify the biases and confounds already present in a dataset. Therefore, the PROBAST bias tool is so important. To increase the available data, many institutions are setting up multicenter or international databases or registries. However, these may be constrained by varied terminology affecting data labeling.

Considering the cost and time needed to utilize predictive modeling, spine surgeons, oncologists, and researchers should balance the upfront investment of time and money required to develop and validate predictive models.[70] Predictive ML models can assist clinicians, but if there is no apparent need for more accurate predictions or if simple statistical models suffice, developing these models would not necessarily be advantageous.

Development and Validation

Even though there has been a massive increase in the volume of predictive models, quality and transparent reporting were not performed consistently. Quality of reporting refers to the application and reporting of the established validation measures. Unfortunately, of 18 studies externally validating 10 different ML prediction models in orthopedic surgery, only 39% reported calibration and 50% reported decision curve analysis.[71] Transparent reporting refers to whether an article mentions all required items in development and validation recommended by the TRIPOD checklist and PROBAST tool. A recent study by Groot et al[65] showed that in ML studies in orthopedics, adherence to the TRIPOD guidelines and PROBAST bias tool was limited. They reviewed 59 ML prediction studies published in orthopedic surgery, of which 18 (31%) were in the spine. The overall completeness for the TRIPOD checklist was 53%, and the overall risk of bias was low in 44%, high in 41%, and unclear in 15%.[65]

These results show that many studies incompletely reported their methods and performance measures. This, together with the fact that the relative novelty of this technique is viewed skeptically, makes it harder for clinicians to rely on predictive models. Thus, to enable trust and facilitate implementation, adherence to the guidelines and transparent reporting of these steps are essential. Consequently, TRIPOD-AI and PROBAST-AI were recently proposed for explicit use in AI to further aid in directing the future of this field.[72]

Even so, the aforementioned performance evaluations might not be sufficient to identify harmful or uninformative algorithms.[69] Moreover, recent research has demonstrated that models created using retrospective data may be biased against racial minorities.[73] Last, many AI algorithms are referred to as black boxes: we are unaware of the operations between input and output. Thus, fully interpreting the models becomes difficult. For this reason, the website of SORG-MLA contains explanations for which predictors contradict or support the model, allowing clinicians to interpret and explain the predicted mortality.

Implementation

Aside from challenges in the development and validation, more challenges arise when implementing ML models in clinical practice. As mentioned before, randomized prospective trials are essential to compare the accuracy of the survival prediction of a surgeon with or without the model. However, very few trials have been performed for predictive models in medicine and, to our knowledge, none to date in orthopedics or spine.[61,69,74] Additionally, ethical, legal, political, and administrative barriers must be overcome. Ethical concerns include liability in cases of medical error, doctors' understanding

of how these models produce predictions, and patients' understanding and control of how these models are used in their care.[75] Moreover, issues of privacy, security, and management of patient data are important to consider.

CONCLUSION

A new health care age is being ushered in by the rapid advancement of AI and its applications in spinal oncology. A myriad of new models are being developed, but the subsequent stages, quality of validation, transparent reporting, and implementation still need improvement. Moreover, we must acknowledge that these models are not a single means to an end. When interpreting these algorithms, we must always consider the context of the clinical question regarding the patient. It will be vital as we advance to regularly scan for potential dangers and ensure that patient benefit and safety continue to come first.

REFERENCES

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94–98. doi:10.7861/futurehosp.6-2-94

2. Oosterhoff JHF, Doornberg JN, Machine Learning Consortium. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of gartner's hype cycle. *EFORT Open Rev*. 2020;5(10):593–603. doi:10.1302/2058-5241.5.190092

3. Black P. Spinal metastasis. *Neurosurgery*. 1979;5(6):726–746. doi:10.1227/00006123-197912000-00016

4. Yuh WT, Zachar CK, Barloon TJ, Sato Y, Sickels WJ, Hawes DR. Vertebral compression fractures: distinction between benign and malignant causes with MR imaging. *Radiology*. 1989;172(1):215–218. doi:10.1148/radiology.172.1.2740506

5. Aaron AD. The management of cancer metastatic to bone. *JAMA*. 1994;272(15):1206–1209.

6. Wong DA, Fornasier VL, MacNab I. Spinal metastases: the obvious, the occult, and the impostors. *Spine J*. 1990;15(1):1–4. doi:10.1097/00007632-199001000-00001

7. Lenz M, Freid JR. Metastases to the skeleton, brain and spinal cord from cancer of the breast and the effect of radiotherapy. *Ann Surg*. 1931;93(1):278–293. doi:10.1097/00000658-193101000-00036

8. Cobb CA, Leavens ME, Eckles N. Indications for nonoperative treatment of spinal cord compression due to breast cancer. *J Neurosurg*. 1977;47(5):653–658. doi:10.3171/jns.1977.47.5.0653

9. Sciubba DM, Petteys RJ, Dekutoski MB, et al. Diagnosis and management of metastatic spine disease. A review. *J Neurosurg Spine*. 2010;13(1):94–108. doi:10.3171/2010.3.SPINE09202

10. Bell GR. Surgical treatment of spinal tumors. *Clin Orthop Relat Res*. 1997;(335):54–63.

11. Bilsky MH, Lis E, Raizer J, Lee H, Boland P. The diagnosis and treatment of metastatic spinal tumor. *Oncologist*. 1999;4(6):459–469. doi:10.1634/theoncologist.4-6-459

12. Hsiue PP, Kelley BV, Chen CJ, et al. Surgical treatment of metastatic spine disease: an update on national trends and clinical outcomes from 2010 to 2014. *Spine J*. 2020;20(6):915–924. doi:10.1016/j.spinee.2020.02.010

13. Patchell RA, Tibbs PA, Regine WF, et al. Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: a randomised trial. *Lancet*. 2005;366(9486):643–648. doi:10.1016/S0140-6736(05)66954-1

14. Goodwin CR, Abu-Bonsrah N, Rhines LD, et al. Molecular markers and targeted therapeutics in metastatic tumors of the spine: changing the treatment paradigms. *Spine (Phila Pa 1976)*. 2016;41 Suppl 20:S218–S223. doi:10.1097/BRS.0000000000001833

15. Laufer I, Rubin DG, Lis E, et al. The NOMS framework: approach to the treatment of spinal metastatic tumors. *Oncologist*. 2013;18(6):744–751. doi:10.1634/theoncologist.2012-0293

16. Yoshihara H, Yoneoka D. Trends in the surgical treatment for spinal metastasis and the in-hospital patient outcomes in the United States from 2000 to 2009. *Spine J*. 2014;14(9):1844–1849. doi:10.1016/j.spinee.2013.11.029

17. Sebaaly A, Shedid D, Boubez G, et al. Surgical site infection in spinal metastasis: incidence and risk factors. *Spine J*. 2018;18(8):1382–1387. doi:10.1016/j.spinee.2018.01.002

18. Carl HM, Ahmed AK, Abu-Bonsrah N, et al. Risk factors for wound-related reoperations in patients with metastatic spine tumor. *J Neurosurg Spine*. 2018;28(6):663–668. doi:10.3171/2017.10.SPINE1765

19. Paulino Pereira NR, Ogink PT, Groot OQ, et al. Complications and reoperations after surgery for 647 patients with spine metastatic disease. *Spine J*. 2019;19(1):144–156. doi:10.1016/j.spinee.2018.05.037

20. Groot OQ, Ogink PT, Paulino Pereira NR, et al. High risk of symptomatic venous thromboembolism after surgery for spine metastatic bone lesions: a retrospective study. *Clin Orthop Relat Res*. 2019;477(7):1674–1686. doi:10.1097/CORR.0000000000000733

21. Oostinga D, Steverink JG, van Wijck AJM, Verlaan J-J. An understanding of bone pain: a narrative review. *Bone*. 2020;134. doi:10.1016/j.bone.2020.115272

22. Barzilai O, Fisher CG, Bilsky MH. State of the art treatment of spinal metastatic disease. *Neurosurgery*. 2018;82(6):757–769. doi:10.1093/neuros/nyx567

23. Chow E, Harth T, Hruby G, Finkelstein J, Wu J, Danjoux C. How accurate are physicians' clinical predictions of survival and the available prognostic tools in estimating survival times in terminally ill cancer patients? A systematic review. *Clinical Oncology*. 2001;13(3):209–218. doi:10.1007/s001740170078

24. Viganò A, Dorgan M, Bruera E, Suarez-Almazor ME. The relative accuracy of the clinical estimation of the duration of life for patients with end of life cancer. *Cancer*. 1999;86(1):170–176. doi:10.1002/(SICI)1097-0142(19990701)86:13.0.CO;2-S

25. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: implications for end-of-life orthopedic care. *J Clin Oncol*. 2005;23(25):6072–6082. doi:10.1200/JCO.2005.08.104

26. Ahmed AK, Goodwin CR, Heravi A, et al. Predicting survival for metastatic spine disease: a comparison of nine scoring systems. *Spine J*. 2018;18(10):1804–1814. doi:10.1016/j.spinee.2018.03.011

27. Tomita K, Kawahara N, Kobayashi T, Yoshida A, Murakami H, Akamaru T. Surgical strategy for spinal metastases. *Spine (Phila Pa 1976)*. 2001;26(3):298–306. doi:10.1097/00007632-200102010-00016

28. Tabouret E, Cauvin C, Fuentes S, et al. Reassessment of scoring systems and prognostic factors for metastatic spinal cord compression. *Spine J*. 2015;15(5):944–950. doi:10.1016/j.spinee.2013.06.036

29. Chen H, Xiao J, Yang X, Zhang F, Yuan W. Preoperative scoring systems and prognostic factors for patients with spinal metastases from hepatocellular carcinoma. *Spine (Phila Pa 1976)*. 2010;35(23):E1339–E1346. doi:10.1097/BRS.0b013e3181e574f5

30. Eap C, Tardieux E, Goasgen O, et al. Tokuhashi score and other prognostic factors in 260 patients with surgery for vertebral metastases. *Orthop Traumatol Surg Res*. 2015;101(4):483–488. doi:10.1016/j.otsr.2015.03.007

31. Hernandez-Fernandez A, Vélez R, Lersundi-Artamendi A, Pellisé F. External validity of the tokuhashi score in patients with vertebral metastasis. *J Cancer Res Clin Oncol*. 2012;138(9):1493–1500. doi:10.1007/s00432-012-1222-2

32. Hessler C, Vettorazzi E, Madert J, Bokemeyer C, Panse J. Actual and predicted survival time of patients with spinal metastases of lung cancer: evaluation of the robustness of the Tokuhashi score. *Spine (Phila Pa 1976)*. 2011;36(12):983–989. doi:10.1097/BRS.0b013e3181e8f7f8

33. Katagiri H, Takahashi M, Wakai K, Sugiura H, Kataoka T, Nakanishi K. Prognostic factors and a scoring system for patients with skeletal metastasis. *J Bone Joint Surg Br*. 2005;87(5):698–703. doi:10.1302/0301-620X.87B5.15185

34. Leithner A, Radl R, Gruber G, et al. Predictive value of seven preoperative prognostic scoring systems for spinal metastases. *Eur Spine J*. 2008;17(11):1488–1495. doi:10.1007/s00586-008-0763-1

35. Quraishi NA, Manoharan SR, Arealis G, et al. Accuracy of the revised tokuhashi score in predicting survival in patients with metastatic spinal cord compression (MSCC). *Eur Spine J*. 2013;22(Suppl 1):S21–S26. doi:10.1007/s00586-012-2649-5

36. Rades D, Dunst J, Schild SE. The first score predicting overall survival in patients with metastatic spinal cord compression. *Cancer*. 2008;112(1):157–161. http://doi.wiley.com/10.1002/cncr.v112:1. doi:10.1002/cncr.23150

37. Hibberd CS, Quan GMY. Accuracy of preoperative scoring systems for the prognostication and treatment of patients with spinal metastases. *Int Sch Res Notices*. 2017;2017:1320684. doi:10.1155/2017/1320684

38. Anderson AB, Wedin R, Fabbri N, Boland P, Healey J, Forsberg JA. External validation of pathfx version 3.0 in patients treated surgically and nonsurgically for symptomatic skeletal metastases. *Clin Orthop Relat Res*. 2020;478(4):808–818. doi:10.1097/CORR.0000000000001081

39. Bollen L, van der Linden YM, Pondaag W, et al. Prognostic factors associated with survival in patients with symptomatic spinal bone metastases: a retrospective cohort study of 1 043 patients. *Neuro Oncol*. 2014;16(7):991–998. doi:10.1093/neuonc/not318

40. van der Linden YM, Dijkstra SPDS, Vonk EJA, Marijnen CAM, Leer JWH, Dutch Bone Metastasis Study Group. Prediction of survival in patients with metastases in the spinal column: results based on a randomized trial of radiotherapy. *Cancer*. 2005;103(2):320–328. doi:10.1002/cncr.20756

41. Choi D, Pavlou M, Omar R, et al. A novel risk calculator to predict outcome after surgery for symptomatic spinal metastases; use of a large prospective patient database to personalise surgical management. *Eur J Cancer*. 2019;107:28–36. doi:10.1016/j.ejca.2018.11.011

42. Ghori AK, Leonard DA, Schoenfeld AJ, et al. Modeling 1-year survival after surgery on the metastatic spine. *Spine J*. 2015;15(11):2345–2350. doi:10.1016/j.spinee.2015.06.061

43. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med*. 2014;3(5):1359–1367. doi:10.1002/cam4.292

44. Balain B, Jaiswal A, Trivedi JM, Eisenstein SM, Kuiper JH, Jaffray DC. The oswestry risk index: an aid in the treatment of metastatic disease of the spine. *Bone Joint J*. 2013;95-B(2):210–216. doi:10.1302/0301-620X.95B2.29323

45. Mizumoto M, Harada H, Asakura H, et al. Prognostic factors and a scoring system for survival after radiotherapy for metastases to the spinal column: a review of 544 patients at Shizuoka cancer center Hospital. *Cancer*. 2008;113(10):2816–2822. doi:10.1002/cncr.23888

46. Tokuhashi Y, Matsuzaki H, Oda H, Oshima M, Ryu J. A revised scoring system for preoperative evaluation of metastatic spine tumor prognosis. *Spine (Phila Pa 1976)*. 2005;30(19):2186–2191. doi:10.1097/01.brs.0000180401.06919.a5

47. Sioutos PJ, Arbit E, Meshulam CF, Galicich JH. Spinal metastases from solid tumors. Analysis of factors affecting survival. *Cancer*. 1995;76(8):1453–1459. doi:10.1002/1097-0142(19951015)76:8<1453::aid-cncr2820760824>3.0.co;2-t

48. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods*. 2018;15(4):233–234. doi:10.1038/nmeth.4642

49. Bzdok D. Classical statistics and statistical learning in imaging neuroscience. *Front Neurosci*. 2017;11:543. doi:10.3389/fnins.2017.00543

50. Bzdok D, Krzywinski M, Altman N. Machine learning: a primer. *Nat Methods*. 2017;14(12):1119–1120. doi:10.1038/nmeth.4526

51. 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. doi:10.1093/eurheartj/ehu207

52. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Med. 2015;13(1). 10.1186/s12916-014-0241-z

53. 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. doi:10.7326/M18-1376

54. Karhade AV, Ahmed AK, Pennington Z, et al. External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. *Spine J*. 2020;20(1):14–21. doi:10.1016/j.spinee.2019.09.003

55. Bongers MER, Karhade AV, Villavieja J, et al. Does the SORG algorithm generalize to a contemporary cohort of patients with spinal metastases on external validation? *Spine J*. 2020;20(10):1646–1652. doi:10.1016/j.spinee.2020.05.003

56. Yang J-J, Chen C-W, Fourman MS, 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*. 2021;21(10):1670–1678. doi:10.1016/j.spinee.2021.01.027

57. Shah AA, Karhade AV, Park HY, et al. Updated external validation of the SORG machine learning algorithms for prediction of ninety-day and one-year mortality after surgery for spinal metastasis. *Spine J*. 2021;21(10):1679–1686. doi:10.1016/j.spinee.2021.03.026

58. Wainer J. *Comparison of 14 Different Families of Classification Algorithms on 115 Binary Datasets*. http://arxiv.org/abs/1606.00930. Accessed May 27, 2022.

59. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med*. 2018;6(12):905–914. doi:10.1016/S2213-2600(18)30300-X

60. 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. doi:10.1097/EDE.0b013e3181c30fb2

61. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II external validation, model updating, and impact assessment. *Heart*. 2012;98(9):691–698. doi:10.1136/heartjnl-2011-301247

62. Karhade A, Schwab JH. CORR synthesis: when should we be skeptical of clinical prediction models? *Clin Orthop Relat Res*. 2020;478(12):2722–2728. doi:10.1097/CORR.0000000000001367

63. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Neurosurgery*. 2019;85(4):E671–E681. doi:10.1093/neuros/nyz070

64. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574. doi:10.1177/0272989X06295361

65. Groot OQ, Ogink PT, Lans A, et al. Machine learning prediction models in orthopedic surgery: a systematic review in transparent reporting. *J Orthop Res*. 2022;40(2):475–483. doi:10.1002/jor.25036

66. Verma AA, Murray J, Greiner R, et al. Implementing machine learning in medicine. *CMAJ*. 2021;193(34):E1351–E1357. doi:10.1503/cmaj.202434

67. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019;322(18):1806–1816. doi:10.1001/jama.2019.16489

68. Rocco G. Garbage in, garbage out. *Eur J Cardiothorac Surg*. 2022;61(5):1020–1021. doi:10.1093/ejcts/ezab504

69. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. *JAMA*. 2018;320(1):27–28. doi:10.1001/jama.2018.5602

70. Ghaednia H, Lans A, Sauder N, et al. Deep learning in spine surgery. *Seminars in Spine Surgery*. 2021;33(2):100876. doi:10.1016/j.semss.2021.100876

71. 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 Orthopaedica*. 2021;92(4):385–393. doi:10.1080/17453674.2021.1910448

72. Collins GS, Dhiman P, Andaur Navarro CL, et al. 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):e048008. doi:10.1136/bmjopen-2020-048008

73. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453. doi:10.1126/science.aax2342

74. Poldervaart JM, Reitsma JB, Backus BE, et al. Effect of using the heart score in patients with chest pain in the emergency department: a stepped-wedge, cluster randomized trial. *Ann Intern Med*. 2017;166(10):689–697. doi:10.7326/M16-1600

75. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. 2019;20(5):e262–e273. doi:10.1016/S1470-2045(19)30149-4

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests: The authors report no conflicts of interest in this work.

Disclosures: Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article. Investigation performed at Massachusetts General Hospital, Boston, USA.

Corresponding Author: Rene Harmen Kuijten, Department of Orthopedic Surgery/Orthopedic Oncology Service, Massachusetts General Hospital – Harvard Medical School, 55 Fruit St, Boston, MA 02114, USA; rkuijten@mgh.harvard.edu; rhkuijten@gmail.com

Published 10 May 2023