Does the SORG algorithm generalize to a contemporary cohort of patients with spinal metastases on external validation?

Bongers, M.E.R.; Karhade, A.V.; Villavieja, J.; Groot, O.Q.; Bilsky, M.H.; Laufer, I.; Schwab, J.H.

The Spine Journal 20 (2020) 1646−1652

Clinical Study

Does the SORG algorithm generalize to a contemporary cohort of patients with spinal metastases on external validation?

Michiel E.R. Bongers, MD[a,*], Aditya V. Karhade, MD, MBA[a], Jemma Villavieja, BS[b], Olivier Q. Groot, MD[a], Mark H. Bilsky, MD[b], Ilya Laufer, MD[b], Joseph H. Schwab, MD, MS[a]

[a] *Department of Orthopedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA*
[b] *Department of Neurosurgery, Memorial Sloan-Kettering Cancer Center, New York, NY, USA*
Received 6 February 2020; revised 5 May 2020; accepted 7 May 2020

Abstract

BACKGROUND CONTEXT: The SORG machine-learning algorithms were previously developed for preoperative prediction of overall survival in spinal metastatic disease. On sub-group analysis of a previous external validation, these algorithms were found to have diminished performance on patients treated after 2010.
PURPOSE: The purpose of this study was to assess the performance of these algorithms on a large contemporary cohort of consecutive spinal metastatic disease patients.
STUDY DESIGN/SETTING: Retrospective study performed at a tertiary care referral center.
PATIENT SAMPLE: Patients of 18 years and older treated with surgery for metastatic spinal disease between 2014 and 2016.
OUTCOME MEASURES: Ninety-day and one-year mortality.
METHODS: Baseline patient and tumor characteristics of the validation cohort were compared to the development cohort using bivariate logistic regression. Performance of the SORG algorithms on external validation in the contemporary cohort was assessed with discrimination (c-statistic and receiver operating curve), calibration (calibration plot, intercept, and slope), overall performance (Brier score compared to the null-model Brier score), and decision curve analysis.
RESULTS: Overall, 200 patients were included with 90-day and 1-year mortality rates of 55 (27.6%) and 124 (62.9%), respectively. The contemporary external validation cohort and the developmental cohort differed significantly on primary tumor histology, presence of visceral metastases, American Spinal Injury Association impairment scale, and preoperative laboratory values. The SORG algorithms for 90-day and 1-year mortality retained good discriminative ability (c-statistic of 0.81 [95% confidence interval [CI], 0.74−0.87] and 0.84 [95% CI, 0.77−0.89]), overall performance, and decision curve analysis. The algorithm for 90-day mortality showed almost perfect calibration reflected in an overall calibration intercept of −0.07 (95% CI: −0.50, 0.35). The 1-year mortality algorithm underestimated mortality mainly for the lowest predicted probabilities with an overall intercept of 0.57 (95% CI: 0.18, 0.96).
CONCLUSIONS: The SORG algorithms for survival in spinal metastatic disease generalized well to a contemporary cohort of consecutively treated patients from an external institutional. Further validation in international cohorts and large, prospective multi-institutional trials is required to

FDA device/drug status: Not applicable.

Author Disclosures: ***MERB:*** Nothing to disclose. ***AVK:*** Nothing to disclose. ***JV:*** Nothing to disclose. ***OQG:*** Nothing to disclose. ***MHB:*** Nothing to disclose. ***IL:*** Nothing to disclose. ***JHS:*** Nothing to disclose.

Funding: The authors report no funding disclosures for this study.

*Corresponding author. Department of Orthopaedic Surgery, Division of Orthopaedic Oncology, Massachusetts General Hospital − Harvard Medical School, Room 3.550, Yawkey Building, 55 Fruit St, Boston, MA 02114, USA.

E-mail address: michielbongers@gmail.com (M.E.R. Bongers).

https://doi.org/10.1016/j.spinee.2020.05.003

confirm or refute the findings presented here. The open-access algorithms are available here: https://sorg-apps.shinyapps.io/spinemetssurvival/.

Keywords: External validation; Machine learning; Mortality; Prediction; Prognostication; Spinal metastases

Introduction

The consideration whether surgical treatment is desirable for patients with spinal metastatic disease, involves balancing the advantages of surgery (avoiding neurological sequelae, increasing comfort, and prolonging survival) against the disadvantages (longer length of stay in the hospital, reoperations and surgical complications with potential increased mortality) [1–7]. Recently, a multicenter, international study by Dea et al. showed that patients with a life expectancy shorter than 3 months but with a good baseline performance status can benefit from surgery in terms of quality of life, opposed to the prerequisite of expected survival longer than three months as a surgical indication [8]. Proper estimation of postoperative survival—in addition to examination of the baseline performance status—can therefore play a decisive role in the preoperative debate concerning different treatment options in these patients [8,9].

The SORG machine-learning (ML) algorithms for prediction of 90-day and 1-year survival in spinal metastatic disease were developed with patients from Massachusetts General Hospital and Brigham and Women's Hospital [10]. Subsequentially, these algorithms were externally validated on an independent patient population from The Johns Hopkins Hospital [11]. However, these algorithms had diminished performance on subanalysis of contemporary patients upon external validation. In patients who underwent surgery before 2010, the 90-day and 1-year algorithms had c-statistic of 0.87 and 0.86 respectively. In comparison, in patients who underwent surgery in 2010 or later, the 90-day and 1-year algorithms had c-statistic of 0.77 and 0.76, respectively. The diminished performance of these algorithms on contemporary cohorts of patients is troubling since the purpose of these tools is application to present-day and future spinal metastatic disease patients.

As such, the primary purpose of this study was to assess the performance of these algorithms on a large contemporary cohort of consecutive spinal metastatic disease patients treated between 2014 and 2016 at the Memorial Sloan Kettering Cancer Center.

Methods

Guidelines

This retrospective external validation study has been performed according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [12]. This study was approved by our institutional review board.

Source of data

Medical records of consecutive patients treated with surgery for metastatic bone disease in the spine at a large tertiary care academic medical center were manually reviewed [JV].

Participants

For this study these inclusion criteria were used: (1) patients of at least 18 years of age or older at the time of surgery, (2) surgical procedure for the resection of metastatic spine lesion performed between 2014 and 2016, (3) pathologic confirmation of primary tumor pathology

Outcome

Survival was established by manual review of the medical records. The last date of review was August 14th, 2019. Ninety-day and 1-year mortality were the primary outcomes of this study, and outcomes were available for 199 (99.5%) and 197 (98.5%) patients, respectively.

Predictors

The variables needed to complete the SORG ML algorithms were collected [10], namely: age (years), gender, body mass index [kilograms per meter squared (kg/m^2)], Eastern Cooperative Oncology Group (ECOG) performance status, histology of primary tumor (groups based on Katagiri et al [13]), visceral metastases (yes/no) [metastases in liver or lung], brain metastases (yes/no), three or more spine metastases (yes/no), previous systemic therapy (yes/no), preoperative presence of any Charlson comorbidity other than metastatic disease (yes/no) [14], American Spinal Injury Association (ASIA) Impairment Scale, and preoperative laboratory characteristics [white blood cell count ($\times 10^3$ per microliter [μL]), hemoglobin (grams per deciliter [g/dL]), platelet count ($\times 10^3/\mu$L), absolute lymphocyte count ($\times 10^3/\mu$L), absolute neutrophil count ($\times 10^3/\mu$L), platelet to absolute lymphocyte ratio, neutrophil to absolute lymphocyte ratio, albumin (g/dL), alkaline phosphatase (international units per liter [IU/L]), calcium (milligrams per deciliter [mg/dL]), creatinine (mg/dL), and international normalized ratio].

Missing data

Data was missing for several variables with the following rates: alkaline phosphatase = 8 (4%), ECOG performance status = 7 (3.5%), albumin = 6 (3%), ASIA = 4 (2%), 1-year survival = 3 (1.5%). Ninety-day survival = 1

(0.5%). Missing values were imputed using the nonparametric MissForest methodology [15].

Statistical analysis and methods

Baseline patient and tumor characteristics of the validation cohort were compared to the development cohort using the chi-square test for categorical variables and the Mann-Whitney U test for continuous variables. The threshold for significance was established a-priori as $p<.05$.

Individual predicted probabilities were calculated for each patient by inputting the variables in the SORG ML algorithms. The same metrics used for the measurement of algorithm performance during internal and initial external validation [10,11]—were used for validation in this study. Which were discrimination, calibration, overall performance, and decision curve analysis. Discrimination was measured using the c-statistic (which is also known as the area under the receiver operating characteristic curve [AUC] for binary classification) and visualized by plotting the receiver operating characteristic curve [16]. The AUC plots sensitivity against 1—specificity for all potential cutoffs for a test and ranges from 0.5 (no better than chance) to 1.0 (perfect discriminative ability). After the AUC was generated, the Optimal Cutpoints method was performed to calculate the Youden index [17], for the maximization of both sensitivity and specificity. Then, The F1-score was computed to calculate the harmonic mean of precision (PPV) and recall (sensitivity) [18]. Additionally assessed discrimination performance measures were sensitivity, specificity, PPV, and negative-predictive value. Calibration—which compares the observed to the predicted proportion of outcomes—was assessed by plotting the calibration plot and subsequently calculating the calibration slope and intercept. An algorithm with perfect calibration has a calibration slope of 1 and an intercept of 0 [16,19,20]. The calibration slope which is lesser than one indicates that the algorithm predicted outcomes are excessive (too high for patients with high predicted probabilities of 90-day or 1-year mortality and too low for patients with low predicted probabilities), a calibration slope greater than one implies the opposite [19,21]. Furthermore, a negative calibration intercept suggests overestimation and a positive intercept suggests underestimation of the outcome [19,21]. Overall performance was measured by assessment and comparison of the Brier score and the null-model Brier score. The Brier score can be assessed by computing the average mean squared difference between the predicted and observed outcomes—the Brier score ranges from 0 (excellent prediction) to 1 (worst prediction) [22]. For correct interpretation of the Brier score a comparison should be performed with the null-model Brier score, which assigns a predicted probability equal to the observed prevalence of the outcome to each patient. Last, decision curve analysis was performed to establish the net benefit (weighted average of true positives and false positives) of the algorithms across a range of different threshold probabilities [23]. Different treatment strategies can be compared using the decision curve analysis. The none line (horizontal) represents the expected net benefit when no management changes are made, whereas the all line represents the net benefit when treatment has been changed for all patients. Different management changes are not distinguished by the decision curve analysis [23].

Statistical software used for data analysis and model validation was: R version 3.5.1 (The R Foundation, Vienna, Austria)

Results

In total, 200 patients were included in this study 90-day and 1-year mortality rates of 55 (27.6%) and 124 (62.9%), respectively. Eighty-nine (44.5%) patients had the female gender, and the median age was 63.4 (interquartile range 54.2–71.0; Table 1).

Baseline characteristics between the developmental cohort and the validation cohort differed significantly ($p<.05$) on primary tumor histology, the presence of visceral metastases, preoperative ASIA-score, albumin value, creatinine value, and 1-year mortality.

The SORG ML algorithm for 90-day mortality prediction in spinal metastatic disease achieved an AUC of 0.81 (95% confidence interval [CI], 0.74–0.87) on external validation (Fig. 1A). At a threshold equal to the Youden index (threshold=0.25), the SORG ML algorithm for 90-day mortality had an F1-score of 0.63 (95% CI, 0.50–0.74). Additional discrimination performance measures are available in Supplemental Table 1. The calibration plot showed good calibration between predicted probability 0.0 and 0.7 in the validation cohort (Fig. 2A). For predicted probabilities of 0.7 and larger the algorithm overestimated the observed proportion of patients with 90-day mortality, reflected in the overall calibration intercept of −0.07 (95% CI, −0.50 −0.35) and calibration slope of 0.64 (95% CI, 0.42–0.86). The Brier score for 90-day mortality was 0.17 compared to the null model Brier score of 0.20. Decision curve analysis showed that the SORG ML algorithm for 90-day mortality prediction resulted in a larger net benefit compared to the default strategies of changing the treatment for all or no patients (Fig. 3A).

The SORG ML algorithm for 1-year mortality prediction in spinal metastatic disease achieved an AUC of 0.84 (95% CI, 0.77–0.89) on external validation (Fig. 1B). At a threshold equal to the Youden index (threshold=0.58), the SORG ML algorithm for 90-day mortality had an F1-score of 0.80 (95% CI, 0.71–0.87) (For additional discrimination performance measures see Supplemental Table 1). The calibration plot showed good calibration between predicted probability 0.25 and 0.7 and predicted probabilities higher than 0.8 (Fig. 2B). With predicted probabilities of 0.25 and lower, and between 0.7 and 0.8 the SORG ML algorithm underestimated 1-year mortality, which was reflected in an

Table 1
Comparison of external validation population to development population

Variable	n (%)	median (IQR)	p value
	Validation cohort(n=200)	Developmental cohort(n=732)	
Age	63.4 (54.2–71.0)	61 (53–69)	.60
Female Sex	89 (44.5)	306 (41.8)	.49
Body mass index (kg/m^2)	26.3 (23.3–29.7)	26.3 (23.1–29.7)	.66
ECOG			.15
0–2	167 (86.5)	440 (81.6)	
3–4	26 (13.5)	99 (18.4)	
Primary Tumor Histology			
Group 1	37 (18.5)	219 (29.9)	**.003**
Group 2	72 (36.0)	254 (34.7)	
Group 3	91 (45.5)	259 (35.4)	
Visceral Metastases	127 (63.5)	252 (34.4)	**<.001**
Brain Metastases	30 (15.0)	81 (11.1)	.17
Three or More Spine Metastases	99 (49.5)	404 (55.2)	.18
Previous Systemic Therapy	111 (55.5)	418 (57.1)	.75
Other Charlson Comorbidity	121 (60.5)	441 (60.7)	1.00
ASIA			**<.001**
Normal (E)	162 (82.7)	379 (52.6)	
Impaired (A-D)	34 (17.3)	342 (47.4)	
Hemoglobin (g/dL)	11.9 (10.1–13.3)	12.1 (10.7–13.3)	.29
Platelet ($10^3/\mu$L)	263.0 (196.8–334.8)	259 (196–337)	.91
Absolute Lymphocyte ($10^3/\mu$L)	0.90 (0.60–1.30)	0.90 (0.58–1.43)	.22
Absolute Neutrophil ($10^3/\mu$L)	7.10 (4.50–10.00)	6.32 (4.48–8.80)	.07
Platelet Lymphocyte Ratio	282.1 (185.9–493.0)	281 (173–461)	.49
Neutrophil Lymphocyte Ratio	8.74 (4.10–13.57)	7.22 (3.64–12.8)	.08
Albumin (g/dL)	4.00 (3.60–4.20)	3.80 (3.40–4.20)	**.002**
Alkaline Phosphatase (IU/L)	102.5 (77.0–141.5)	94.5 (73.0–140)	.26
Creatinine (mg/dL)	0.80 (0.60–0.90)	0.80 (0.69–1.00)	**.005**
INR	1.07 (1.01–1.15)	1.10 (1.00–1.10)	.28
Ninety-Day Mortality	55 (27.6)	181 (25.1)	.53
One-Year Mortality	124 (62.9)	385 (54.3)	**.04**

ASIA, American Spinal Injury Association Impairment Scale; BMI, body mass index; ECOG, Eastern Cooperative Oncology Group performance status; g/dL, grams per deciliter; IU/L, international units per liter; IQR, interquartile range; kg/m^2, kilograms per meter squared; mg/dL, milligrams per deciliter; μL, microliter.

overall calibration intercept of 0.57 (95% CI, 0.18–0.96) and calibration slope of 0.85 (95% CI, 0.59–1.11). The Brier score for 1-year mortality was 0.16 compared to the null model Brier score of 0.23. The decision curve analysis for the SORG ML algorithm to predict 90-day mortality showed that above a high-risk threshold of 0.3 the algorithm resulted in a larger net benefit compared to changing the treatment for all of none patients, below the threshold of 0.3 changing management for all patient provided a greater net benefit than the algorithm (Fig. 3B). An overview of the performance measures is given in Table 2.

Table 2
Overview of the performance measures for 90-day and 1-year mortality on external validation, n=200

Metric	90-day mortality	1-year mortality
Discrimination		
AUC	0.81 (0.74, 0.87)	0.84 (0.77, 0.89)
F1-score*	0.63 (0.50, 0.74)	0.80 (0.71, 0.87)
Calibration		
Intercept	−0.07 (−0.50, 0.35)	0.57 (0.18, 0.96)
Slope	0.64 (0.42, 0.86)	0.64 (0.42, 0.86)
Overall performance		
Brier score	0.17 (0.13, 0.20)	0.16 (0.13, 0.19)
Null-model Brier score	0.20	0.23

AUC, area under the receiver operating curve.
* At the threshold equal to the Youden index (90-d mortality threshold=0.25, 1-year mortality threshold=0.58).

Discussion

Decision guidance for opting between surgery and other treatment modalities for patients with metastatic spine disease by using repeatedly externally validated survival prediction tools, can optimize treatment processes and decrease unnecessary—potentially harmful—treatment of these patients, especially for those with a lesser baseline performance status [8,24,25]. The SORG ML algorithms were developed and externally validated on two independent patient populations from different tertiary care centers in the United States [10,11]. Following the TRIPOD guidelines [12], algorithms should be repeatedly validated for the assessment of possible performance inadequacies

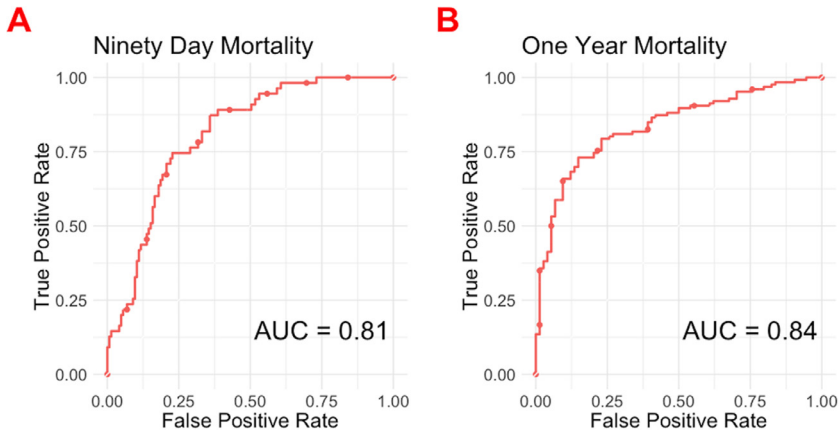

Fig. 1. Area under the receiver operating curve (AUC) for the ML algorithm for (A) 90-day and (B) 1-year mortality on external validation, n=200. AUC, area under the receiver operating characteristic curve; ML, machine learning.

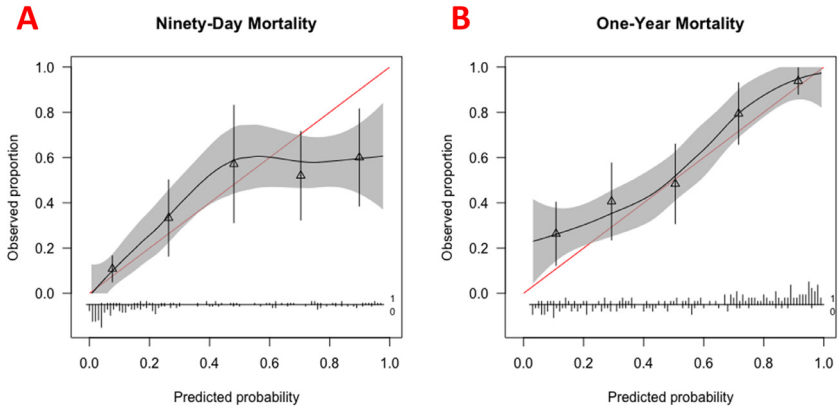

Fig. 2. Calibration plots which indicate the agreement between observed and predicted outcomes for the ML algorithm for 90-day mortality and 1-year mortality on external validation, n=200. Predicted probabilities were subdivided into 5 bins for which the mean observation is plotted against the mean calculated predicted probability, with corresponding 95% confidence interval. ML, machine learning.

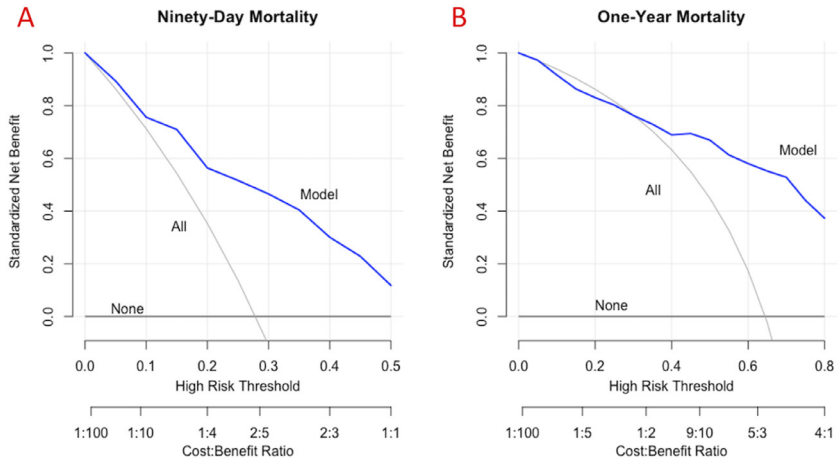

Fig. 3. Decision curve analysis with standardized net benefit by threshold probability for the ML algorithm for 90-day and 1-year mortality on external validation, n=200. ML, machine learning.

among in different independent populations [26,27]. Hence, the purpose of this study was to assess the validity of the SORG ML survival algorithms in an external population of patients with metastatic spinal disease. In this external validation among a contemporary cohort of consecutively treated patients, we found that the SORM ML algorithms retained good discrimination and overall performance. The SORG ML algorithm for 90-day mortality retained

almost perfect overall calibration. For the 1-year mortality algorithm the overall calibration retained moderate and underestimated mainly for the lowest predicted probabilities.

Numerous survival prognostication tools have been developed for patients with spinal metastatic disease. A prior study of nine different scoring systems found that the best tool achieved an AUC of 0.70 for 90-day mortality and an AUC of 0.78 for 1-year mortality [13,28–31]. The SORG ML algorithm for 90-day mortality shows improved discriminative abilities, with an AUC of 0.83 on internal validation, 0.75 on initial external validation, and 0.81 on this secondary external validation. Similarly, the SORG ML algorithm for 1-year mortality also shows improved discriminative abilities, with an AUC of 0.85 on internal validation, 0.77 on initial external validation, and 0.84 on this secondary external validation. A possible explanation for the improved predictability compared to the initial external validation may be that the size of this cohort was larger. Unfortunately, none of the other previously developed algorithms showed results following the TRIPOD guidelines as no calibration results and decision curve analysis were taken into account during development of these tools [12]. A follow-up prospective multi-institutional study assessed calibration of some of the previously developed tools and found that calibration results were poor in all models—where the best model achieved an calibration slope of 0.45 (perfect calibration slope=1) [9]. In contrast, the SORG ML algorithm for 90-day mortality achieved a calibration slope of 0.64 in both the initial as in this secondary external validation; while the SORG ML algorithm for 1-year mortality achieved calibration slopes of 0.77 and 0.85 for the initial and this secondary external validation, respectively [11].

There are several limitations that should be discussed, and this study should thus be interpreted in the context of its design. First, patients were retrospectively included in the validation cohort from a single, though independent, population. Prospective external validation preferably using international multi-institutional collaborations remains to be performed. Second, the baseline characteristics differed between the validation and the developmental cohort on some disease factors. The validation cohort consisted of patients that had a higher proportion of rapidly growing tumors, more patients with visceral metastases, and higher albumin levels [10]. The reason for these differences is unknown but indicate that the algorithms keeps accurate discriminative ability and overall performance with variability in tumor histology. Third, this study suffered from a relatively small sample size. The suggested minimum of 200 events and nonevents for reliable interpretation of calibration results by van Calster et al [19], could not be met as the total cohort consisted of 200 patients. This low number might thus have again caused the calibration incongruities as previous research has shown that calibration plot interpretation could be less reliable in smaller validation cohorts [20,32,33]. Not only reliability of calibration results will improve with larger cohorts, but also subgroup analysis of various patient and treatment characteristics will be more reliable. Fourth, despite that development, internal validation, and primary external validation have been performed in independent datasets, this secondary validation was also performed in a hospital located in the United States; hence, future studies on this topic should seek to validate this algorithm in a non-American population.

Still, these analysis show that the SORG algorithms are—to our knowledge—currently among the best performing externally validated prediction tools for 90-day and 1-year mortality in patients with metastatic spine disease. The algorithms are freely available as a web application at: https://sorg-apps.shinyapps.io/spinemetssurvival/. Health care providers may use this application as an aid to optimize treatment trajectories and as patient education tools in daily clinic. When using the algorithms, clinicians should be aware that the algorithm was developed and validated on surgically treated patients. Thus, applicability to patients treated with different treatment modalities remains to be determined. Future studies can therefore aim to study the performance of these algorithms for non-operatively managed patients.

Conclusions

The SORG algorithms for survival in spinal metastatic disease generalized well to a contemporary cohort of consecutively treated patients from an external institutional. Further validation in international cohorts and large, prospective multi-institutional trials is required to confirm or refute the findings presented here.

Supplementary materials

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.spinee.2020.05.003.

References

[1] Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. Clin Cancer Res 2006;12:6243s–9s. https://doi.org/10.1158/1078-0432.CCR-06-0931.

[2] Kim CH, Chung CK, Sohn S, Lee S, Park SB. Less invasive palliative surgery for spinal metastases. J Surg Oncol 2013;108:499–503. https://doi.org/10.1002/jso.23418.

[3] Nathan SS, Healey JH, Mellano D, Hoang B, Lewis I, Morris CD, et al. Survival in patients operated on for pathologic fracture: implications for end-of-life orthopedic care. J Clin Oncol 2005;23:6072–82. https://doi.org/10.1200/JCO.2005.08.104.

[4] Quinn RH, Randall RL, Benevenia J, Berven SH, Raskin KA. Contemporary management of metastatic bone disease: tips and tools of the trade for general practitioners. J Bone Jt Surg Am 2013;95:1887–95. https://doi.org/10.2106/00004623-201310160-00011.

[5] Prasad D, Schiff D. Malignant spinal-cord compression. Lancet Oncol 2005;6:15–24. https://doi.org/10.1016/S1470-2045(04)01709-7.

[6] Verlaan J-J, Choi D, Versteeg A, Albert T, Arts M, Balabaud L, et al. Characteristics of patients who survived < 3 months or >2 years after surgery for spinal metastases: can we avoid inappropriate patient selection? J Clin Oncol 2016;34:3054–61. https://doi.org/10.1200/JCO.2015.65.1497.

[7] Patchell RA, Tibbs PA, Regine WF, Payne R, Saris S, Kryscio RJ, et al. Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: a randomised trial. Lancet (London, England) 2005;366:643–8. https://doi.org/10.1016/S0140-6736(05)66954-1. [Epub ahead of print].

[8] Dea N, Versteeg AL, Sahgal A, Verlaan J-J, Charest-Morin R, Rhines LD, et al. Metastatic spine disease: should patients with short life expectancy be denied surgical care? An international retrospective cohort study. Neurosurgery 2019. https://doi.org/10.1093/neuros/nyz472.

[9] Nater A, Tetreault LA, Kopjar B, Arnold PM, Dekutoski MB, Finkelstein JA, et al. Predictive factors of survival in a surgical series of metastatic epidural spinal cord compression and complete external validation of 8 multivariate models of survival in a prospective North American multicenter study. Cancer 2018;124:3536–50. https://doi.org/10.1002/cncr.31585.

[10] Karhade AV, Thio QCBS, 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 2019;85:E671–81. https://doi.org/10.1093/neuros/nyz070.

[11] Karhade AV, Ahmed AK, Pennington Z, Chara A, Schilling A, Thio QCBS, et al. External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. Spine J 2020;20:14–21. https://doi.org/10.1016/j.spinee.2019.09.003.

[12] 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. https://doi.org/10.1186/s12916-014-0241-z.

[13] Katagiri H, Okada R, Takagi T, Takahashi M, Murata H, Harada H, et al. New prognostic factors and scoring system for patients with skeletal metastasis. Cancer Med 2014;3:1359–67. https://doi.org/10.1002/cam4.292.

[14] Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P, et al. Practice of epidemiology updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol 2011;173:676–82 https://doi.org/10.1093/aje/kwq433.

[15] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28:112–8. https://doi.org/10.1093/bioinformatics/btr597.

[16] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31. https://doi.org/10.1093/eurheartj/ehu207.

[17] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32–5. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

[18] Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. Lect Notes Comput Sci 2005;3408:345–59. https://doi.org/10.1007/978-3-540-31865-1_25.

[19] van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76. https://doi.org/10.1016/j.jclinepi.2015.12.005.

[20] van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making 2015;35:162–9. https://doi.org/10.1177/0272989X14547233.

[21] van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.

[22] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78:1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

[23] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74. https://doi.org/10.1177/0272989X06295361.

[24] Laufer I, Rubin DG, Lis E, Cox BW, Stubblefield MD, Yamada Y, et al. The NOMS framework: approach to the treatment of spinal metastatic tumors. Oncologist 2013;18:744–51. https://doi.org/10.1634/theoncologist.2012-0293.

[25] Barzilai O, Laufer I, Yamada Y, Higginson DS, Schmitt AM, Lis E, et al. Integrating evidence-based medicine for treatment of spinal metastases into a decision framework: neurologic, oncologic, mechanicals stability, and systemic disease. J Clin Oncol 2017;35:2419–27. https://doi.org/10.1200/JCO.2017.72.7362.

[26] Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. JAMA 2018;320:27–8. https://doi.org/10.1001/jama.2018.5602.

[27] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73. https://doi.org/10.1002/(sici)1097-0258(20000229)19:4<453::aid-sim350>3.0.co;2-5.

[28] Tokuhashi Y, Matsuzaki H, Toriyama S, Kawano H, Ohsaka S. Scoring system for the preoperative evaluation of metastatic spine tumor prognosis. Spine (Phila Pa 1976) 1990;15:1110–3. https://doi.org/10.1097/00007632-199011010-00005.

[29] Bauer HC, Wedin R. Survival after surgery for spinal and extremity metastases. Prognostication in 241 patients. Acta Orthop Scand 1995;66:143–6. https://doi.org/10.3109/17453679508995508.

[30] 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;103:320–8. https://doi.org/10.1002/cncr.20756.

[31] Paulino Pereira NR, Janssen SJ, van Dijk E, Harris MB, Hornicek FJ, Ferrone ML, et al. Development of a prognostic survival algorithm for patients with metastatic spine disease. J Bone Jt Surg Am 2016;98:1767–76. https://doi.org/10.2106/JBJS.15.00975.

[32] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[33] Steyerberg EW, Uno H, Ioannidis JPA, van Calster B, Collaborators. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol 2018;98:133–43. https://doi.org/10.1016/j.jclinepi.2017.11.013.