



Universiteit
Leiden
The Netherlands

External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease

Karhade, A.V.; Ahmed, A.K.; Pennington, Z.; Chara, A.; Schilling, A.; Thio, Q.C.B.S.; ... ; Schwab, J.H.

Citation

Karhade, A. V., Ahmed, A. K., Pennington, Z., Chara, A., Schilling, A., Thio, Q. C. B. S., ... Schwab, J. H. (2020). External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. *The Spine Journal*, 20(1), 14-21.
doi:10.1016/j.spinee.2019.09.003

Version: Publisher's Version

License: [Licensed under Article 25fa Copyright Act/Law \(Amendment Taverne\)](#)

Downloaded from: <https://hdl.handle.net/1887/4302576>

Note: To cite this publication please use the final published version (if applicable).

Clinical Study

External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease

Aditya V. Karhade, BE^a, Ali K. Ahmed, BS^b, Zach Pennington, BS^b,
Alejandro Chara, BS^b, Andrew Schilling, BS^b, Quirina C.B.S. Thio, MD^a,
Paul T. Ogink, MD^a, Daniel M. Sciubba, MD^b, Joseph H. Schwab, MD, MS^{a,*}

^a Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

^b Department of Neurosurgery, The Johns Hopkins University School of Medicine, Baltimore, MD, USA

Received 8 April 2019; revised 3 September 2019; accepted 4 September 2019

Abstract

BACKGROUND CONTEXT: Preoperative survival estimation in spinal metastatic disease helps determine the appropriateness of invasive management. The SORG ML 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease were previously developed in a single institutional sample but remain to be externally validated.

PURPOSE: The purpose of this study was to externally validate these algorithms in an independent population from another institution.

STUDY DESIGN/SETTING: Retrospective study at a large, tertiary care center.

PATIENT SAMPLE: Patients 18 years or older who underwent surgery between 2003 and 2016.

OUTCOME MEASURES: Ninety-day and 1-year mortality.

METHODS: Baseline characteristics of the validation cohort were compared to the developmental cohort for the SORG ML algorithms. Discrimination (c-statistic and receiver operating curve), calibration (calibration slope, intercept, calibration plot, and observed proportions by predicted risk groups), overall performance (Brier score), and decision curve analysis were used to assess the performance of the SORG ML algorithms in the validation cohort.

RESULTS: Overall, 176 patients underwent surgery for spinal metastatic disease, of which 44 (22.7%) experienced 90-day mortality and 99 (56.2%) experienced 1-year mortality. The validation cohort differed significantly from the developmental cohort on primary tumor histology, metastatic tumor burden, previous systemic therapy, overall comorbidity burden, and preoperative laboratory characteristics. Despite these differences, the SORG ML algorithms generalized well to the validation cohort on discrimination (c-statistic 0.75–0.81 for 90-day mortality and 0.77–0.78 for 1-year mortality), calibration, Brier score, and decision curve analysis.

CONCLUSION AND RELEVANCE: Initial results from external validation of the SORG ML 90-day and 1-year algorithms for survival prediction in spinal metastatic disease suggest potential utility of these digital decision aids in clinical practice. Further studies are needed to validate or refute these algorithms in large patient samples from prospective, international, multi-institutional trials. © 2019 Published by Elsevier Inc.

Keywords:

External validation; Machine learning; Neurosurgery; Ninety-day; One year; Orthopedic surgery; Prediction; Prognosis; Spine metastasis; Spine surgery; Survival

FDA device/drug status: Not applicable.

Author disclosures: **AVK:** Nothing to disclose. **AKA:** Nothing to disclose. **ZP:** Nothing to disclose. **AC:** Nothing to disclose. **AS:** Nothing to disclose. **QCBST:** Nothing to disclose. **PTO:** Nothing to disclose. **DMS:** Consulting: Baxter (B), DePuy Synthes (B), Globus (B), Medtronic (C), NuVasive (B), Stryker (C). **JHS:** Consulting or Honorarium: Scientific Advisory Board: Chordoma Foundation (Nonfinancial); Speaking and/or

Teaching Arrangements: AO Spine (Travel Expense Reimbursement), Stryker Spine (B).

* Corresponding author. Department of Orthopedic Surgery, Massachusetts General Hospital, Harvard Medical School, 55 Fruit St, Boston, MA 02114, USA. Tel.: (617) 543-5227; fax: (617) 726-7587.

E-mail address: jhschwab@mgh.harvard.edu (J.H. Schwab).

Background

The spine is the most common site of bony metastasis and metastatic epidural spinal cord compression occurs in up to 10% of cancer patients [1–3]. Surgery is undertaken in spinal metastatic disease to reduce pain, improve quality of life, and preserve ambulatory status [4,5]. Since the randomized controlled trial by Patchell et al., there has been increased enthusiasm for surgical intervention in spinal metastatic disease [6,7]. However, the decision for operative intervention requires preoperative estimation of expected postoperative survival [8].

The SORG machine learning (ML) 90-day and 1-year algorithms were previously developed to preoperatively predict survival in patients with metastatic spinal tumors [9]. Two time-points for estimation of survival were chosen to provide both short- and long-term survival estimates. Individual provider preference on patient counseling may depend on factors such as primary tumor histology and patient functional status. Ninety-day survival estimation may be useful for patients who have relatively poor prognosis such as lung primary tumors. One-year survival estimation may be useful in patients with relatively good prognosis such as prostate primary tumors. Though these algorithms showed promise in the developmental cohort, they were produced from a single institutional sample and remain to be externally validated.

The primary purpose of this study was to externally validate or refute these algorithms in an independent population of patients from another institution. Comparison of the baseline characteristics in the two populations was further undertaken to determine if the algorithms would generalize to the independent population.

Methods

Guidelines

This validation study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [10].

Source of data

Health records review was approved by the institutional review board (NA_00067508) for retrospective chart review on patients who underwent surgery for spinal metastatic disease at a high volume, tertiary care center.

Participants

The inclusion criteria for the study were: (1) adult patients between the ages of 18 and 100 at the time of surgery; (2) surgical resection for metastatic spine lesion between 2003 and 2016; (3) primary tumor histology confirmed by pathology; (4) availability of electronic health records (EHR) with clinical presentation, imaging, and

operative notes; and (5) date of death recorded in the EHR or most recent follow-up available.

Outcome

Survival was ascertained through manual chart review and structured documentation in the EHR. The date of last review was March 8, 2019. The primary outcomes of interest were 90-day and 1-year mortality.

Predictors

The following variables were assessed based on the factors needed as inputs to the SORG ML algorithms [9]: age (years), sex, body mass index (kilograms per meter squared [kg/m^2]), Eastern Cooperative Oncology Group [ECOG] performance status, primary tumor histology (based on groupings by Katagiri et al.) [11], presence of visceral metastases (metastases in liver or lung), presence of brain metastases, three or more spine metastases, previous systemic therapy, preoperative presence of any Charlson comorbidity other than metastatic disease [12], 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 ($\times 10^3/\mu\text{L}$), absolute lymphocyte count ($\times 10^3/\mu\text{L}$), absolute neutrophil count ($\times 10^3/\mu\text{L}$), 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 (INR).

Missing data

Rates of missing data for variables were as follows: body mass index=53 (30.1), alkaline phosphatase=40 (22.7), absolute lymphocyte=47 (26.7), absolute neutrophil=47 (26.7), platelet to lymphocyte ratio=47 (26.7), neutrophil to lymphocyte ratio=47 (26.7), international normalized ratio=15 (8.5). The missForest multiple imputation technique was used for imputation [13].

Statistical analysis and methods

Discrimination was calculated by computing the c-statistic (also known as the area under the receiver operating curve [AUC] for binary classification) and by plotting the receiver operating curve [14]. Models no better than chance have $\text{AUC}=0.5$ and perfect models have $\text{AUC}=1$. Calibration was calculated by computing the calibration slope and intercept and by plotting the calibration plot. Perfect models have calibration intercept=0 and calibration slope=1. The Brier score is the average mean squared difference between the model predictions and the observed outcomes and it captures both discrimination and calibration [15]. The null model Brier score was calculated to compare the incremental benefit of the algorithms to a theoretical null model that

would assign a predicted probability to every patient equal to the observed prevalence of 90-day and 1-year mortality in the validation cohort. Finally, decision curve analysis was undertaken to the net benefit (weighted average of true positives and false positives) of the algorithm in comparison to the default strategies of changing management for all patients or for no patients [16]. Decision curve analysis allows for estimation of the model's expected utility in clinical practice by comparing true positives and false positives generated by the model.

Net benefit is defined as:

$$\text{Net benefit} = \frac{\text{True Positives} - \text{relativeWeight}(\text{False Positives})}{\text{Number of Patients}}$$

Relative weight is defined by the threshold:

$$\text{relative Weight} = \frac{\text{threshold probability}}{1 - \text{threshold probability}}$$

The decision curve analysis offers a visual comparison of various treatment strategies. The point in time the decision curve analysis refers to is the point before surgery when management changes are possible. The decision curve analysis does not prescribe what these management changes may be. The horizontal lines (labeled “none”) refer to the expected net benefit without any changes in management. Without making any changes there are no benefits (net benefit > 0) or harms (net benefit < 0) from this strategy. The slanted vertical line (labeled “all”) refers to the strategy of changing management for all patients. The relative weight varies with the chosen threshold probability and as the threshold probability increases, the relative weight increases. False positives are weighted more and the net benefit becomes negative for the strategy of changing management for all patients.

Importantly, the threshold probability depends on the provider using the algorithms and is not an intrinsic property of the algorithms. Certain providers may have a higher threshold for changing management than others. This may also depend on the management changes suggested. For example, if the management change is increased surveillance and preoperative optimization, a threshold for intervening may be predicted 90-day mortality > 0.25. However, if the management changes are delaying surgery and pursuing nonoperative care, the threshold for 90-day mortality may be > 0.5 or > 0.75. The specific threshold will vary from one provider to the next. The decision curve analysis is useful because once the provider has determined their threshold probability for taking action, the decision curve can show the expected utility (net benefit) of using the algorithms at that specific threshold. Simultaneously, the decision curve analysis allows for the comparison of selectively changing management for patients based on the algorithm's predicted probabilities at this threshold as compared to the default strategies of changing management for all patients or for no patients.

Sensitivity analyses were conducted by measuring the performance of the algorithms across these utilities in both the full population of patients with imputed data as well as the subpopulation of patients with complete data. Additionally, performance of the algorithms was assessed for in subsets of patients who underwent surgery from 2003–2009 relative to 2010–2016.

The baseline characteristics of the validation cohort were compared to those of the developmental cohort with bivariate logistic regression. Significance was defined a priori as $p < .05$.

Data analysis was undertaken with the Anaconda Distribution (Anaconda, Inc., Austin, TX), R version 3.5.1 (The R Foundation, Vienna, Austria), RStudio version 1.0.153 (RStudio, Boston, MA), and Python version 3.6 (Python Software Foundation, Wilmington, DE).

Results

Overall, 176 patients with spinal metastatic disease were included, of which 44 (22.7%) experienced 90-day mortality and 99 (56.2%) experienced 1-year mortality. Seventy-one (40.3%) of the patients were female and the median age was 62 (interquartile range [IQR]=52–69) [Table 1].

In comparison to the developmental cohort for the SORG ML algorithms, the population in this study differed significantly ($p < .05$) on performance status, primary tumor histology, presence of brain metastases, three or more spine metastases, other Charlson comorbidities, preoperative ASIA status, and laboratory characteristics (white blood cell, hemoglobin, platelet).

The SORG ML 90-day algorithm achieved AUC=0.75 in the validation cohort (Fig. 1A). The calibration plot of the algorithm in the validation cohort showed excellent calibration from predicted probability=0 to predicted probability=0.5 (Fig. 2A). With predicted probabilities greater than 0.5, the algorithm underestimated the observed proportion of patients with 90-day mortality. This was reflected in the negative calibration intercept of -0.10 with calibration slope=0.64. For patients with predicted mortality probabilities less than 0.3 ($n=120$), the mean 90-day mortality was 12.5% (95% confidence interval [CI]: 6.5%–18.5%). For patients with predicted probabilities 0.3 to 0.6 ($n=36$), the mean 90-day mortality was 38.9% (95% CI: 22.2%–55.6%). For patients with predicted probabilities greater than 0.6 ($n=20$), the mean 90-day mortality was 55% (95% CI: 31.1%–78.9%). The raw Brier score for 90-day mortality was 0.157 relative to the null model Brier score of 0.176. On decision curve analysis in the validation cohort, the algorithm provided greater net benefit than the default strategies of changing management for all patients or for no patients (Fig. 3A).

The SORG ML 1-year algorithm achieved AUC=0.77 in the validation cohort (Fig. 1B). The calibration plot of the algorithm in the validation cohort showed good calibration for predicted probabilities less than 0.1 and greater than 0.4

Table 1

Baseline characteristics of validation cohort, n=176

Variable	n (%) median (IQR)		p Value
	Validation cohort (n=176)	Developmental cohort (n=732)	
Age	62.0 (52.0–69.0)	61 (53–69)	.87
Female sex	71 (40.3)	306 (41.8)	.72
Body mass index (kg/m ²)	25.9 (22.6–30.7) *	26.3 (23.1–29.7)	.82
ECOG			<.001
0–2	109 (61.9)	440 (81.6) [#]	
3–4	67 (38.1)	99 (18.4) [#]	
Primary tumor histology			<.001
Group 1	61 (34.7)	219 (29.9)	
Group 2	98 (55.7)	254 (34.7)	
Group 3	17 (9.7)	259 (35.4)	
Visceral metastases	49 (27.8)	252 (34.4)	.078
Brain metastases	7 (4.0)	81 (11.1)	.006
Three or more spine metastases	64 (36.4)	404 (55.2)	<.001
Previous systemic therapy	102 (58.0)	418 (57.1)	.84
Other Charlson comorbidity	67 (38.1)	441 (60.7)	<.001
ASIA			.006
Normal (E)	72 (40.9)	379 (52.6) [¶]	
Impaired (A–D)	104 (59.1)	342 (47.4) [¶]	
White blood cell (10 ³ /uL)	9.04 (6.30–12.5)	8.20 (6.06–10.9) **	.009
Hemoglobin (g/dL)	11.7 (10.5–12.8)	12.1 (10.7–13.3) **	.049
Platelet (10 ³ /uL)	225 (169–289)	259 (196–337) **	<.001
Absolute lymphocyte (10 ³ /uL)	0.82 (0.51–1.36) [‡]	0.90 (0.58–1.43) ^{††}	.85
Absolute neutrophil (10 ³ /uL)	7.22 (4.37–9.80) [‡]	6.32 (4.48–8.80) ^{††}	.45
Platelet lymphocyte ratio	247 (150–440) [‡]	281 (173–461) ^{††}	.80
Neutrophil lymphocyte ratio	8.06 (4.11–14.3) [‡]	7.22 (3.64–12.8) ^{††}	.21
Albumin (g/dL)	3.80 (3.40–4.30) [†]	3.80 (3.40–4.20) ^{§§}	.49
Alkaline phosphatase (IU/L)	92.0 (75.8–132.2) [†]	94.5 (73.0–140)	.98
Creatinine (mg/dL)	0.80 (0.60–1.00)	0.80 (0.69–1.00) **	.26
INR	1.00 (1.00–1.10) [§]	1.10 (1.00–1.10) ^{##}	.072
Ninety-day mortality	40 (22.7)	181 (25.1)	.51
One-year mortality	99 (56.2)	385 (54.3)	.99

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/m²), kilograms per meter squared; (mg/dL), milligrams per deciliter; uL, microliter.

Bold values $p < 0.05$ are statistically significant.

* Missing in 53 (30.1).

† Missing in 40 (22.7).

‡ Missing in 47 (26.7).

§ Missing in 15 (8.5).

|| Missing in 90 (12.3).

¶ Missing in 193 (26.4).

Missing in 11 (11.5).

** Missing in 89 (12.2).

†† Missing in 218 (29.8).

‡‡ Missing in 215 (29.3).

§§ Missing in 185 (25.3).

||| Missing in 194 (26.5).

Missing in 89 (12.2).

(Fig. 2B). Overall, the algorithm had higher predicted probabilities than the observed proportions in the validation cohort and this was reflected in the calibration intercept with 0.43 and calibration slope=0.77. For patients with predicted mortality probabilities less than 0.3 ($n=60$), the mean 1-year mortality was 28.3% (95% CI: 16.6%–40.1%). For patients with predicted probabilities 0.3 to 0.6 ($n=41$), the mean 1-year mortality was 53.7% (95% CI: 37.7%–70.0%). For patients with predicted probabilities greater than 0.6

($n=75$), the mean 1-year mortality was 55% (95% CI: 31.1%–78.9%). The Brier score for 1-year mortality was 0.199 relative to the null model Brier score of 0.246. On decision curve analysis in the validation cohort, the algorithm provided greater net benefit than the default strategies of changing management for all patients or for no patients (Fig. 3B).

The discrimination of the SORG ML 90-day algorithm in subpopulations of patients with lung, breast, prostate,

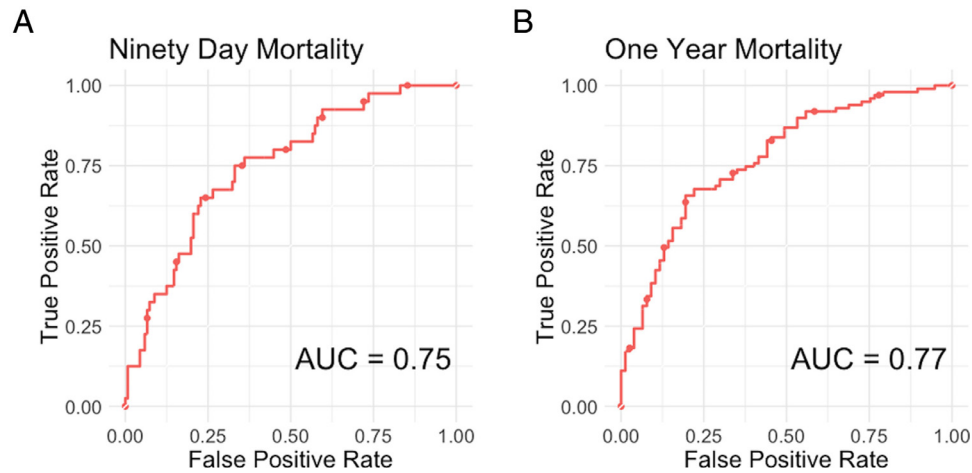


Fig. 1. (A-B) Receiver operating curve for 90-day (A) and 1-year mortality (B) in validation cohort, n=176. AUC, area under receiver operating curve.

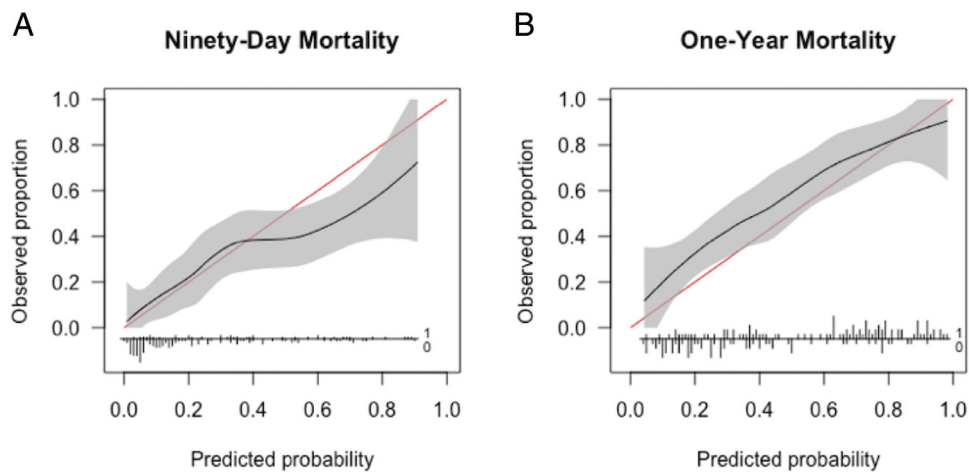


Fig. 2. (A-B) Calibration plots for 90-day (A) and 1-year (B) mortality in validation cohort, n=176.

and renal primaries was 0.70, 0.87, 0.60, and 0.76, respectively (Table 2). The discrimination of the SORG ML 1-year algorithm in subpopulations of patients with lung, breast, prostate, and renal primaries was 0.92, 0.76, 0.75, and 0.78, respectively (Table 2).

On sensitivity analysis in the subpopulation of patients (n=91) with complete data available for the SORG ML algorithms without requiring multiple imputation, the AUC of the SORG 90-day algorithm in the validation cohort was 0.81 and the AUC of the 1-year algorithm was 0.78

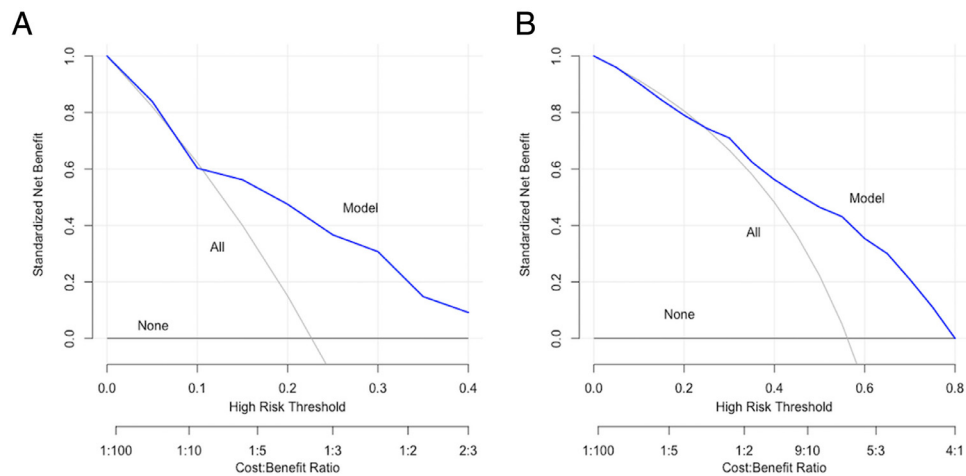


Fig. 3. (A-B) Decision curve analysis with net benefit by threshold probability for 90-day (A) and 1-year (B) mortality in validation cohort, n=176.

Table 2

Discrimination (c-statistic) of algorithms by primary tumor histology in validation cohort

Validation cohort	Ninety-day	One-year
Overall (n=176)	0.75 (0.66–0.83)	0.77 (0.70–0.83)
Lung (n=33)	0.70 (0.49–0.85)	0.92 (0.68–0.99)
Breast (n=37)	0.87 (0.56–0.97)	0.76 (0.51–0.91)
Prostate (n=27)	0.60 (0.36–0.80)	0.75 (0.49–0.90)
Renal (n=36)	0.76 (0.53–0.90)	0.78 (0.58–0.90)

Table 3

Discrimination (c-statistic) of SORG ML algorithms by primary tumor histology in subset of validation cohort with complete data, n=91

Validation cohort	Ninety-day	One-year
Overall (n=91)	0.81 (0.70–0.89)	0.78 (0.67–0.87)
Lung (n=20)	0.85 (0.58–0.96)	0.97 (0.67–0.99)
Breast (n=22)	0.83 (0.41–0.97)	0.67 (0.36–0.88)
Prostate (n=11)	0.67 (0.31–0.90)	0.89 (0.42–0.99)
Renal (n=13)	0.77 (0.41–0.94)	0.97 (0.62–0.99)

(Table 3). For patients with complete date, the discrimination of the SORG ML 90-day algorithm in patients with lung, breast, prostate, and renal primaries was 0.85, 0.83, 0.67, and 0.77, respectively (Table 3). For patients with complete date, the discrimination of the SORG ML 1-year algorithm in patients with lung, breast, prostate, and renal primaries was 0.97, 0.67, 0.89, and 0.97, respectively (Table 3).

In patients with complete data who underwent surgery from 2003 to 2009 (n=31), the AUC of the SORG 90-day algorithm was 0.87 and the AUC of the SORG 1-year algorithm was 0.86. In patients with complete data who underwent surgery from 2010 to 2016 (n=60), the AUC of the SORG 90-day algorithm was 0.77 and the AUC of the SORG 1-year algorithm was 0.76.

Discussion

Determining the appropriateness of surgical intervention in spinal metastatic disease requires reliable, externally validated scoring systems to preoperatively predict the survival at key postoperative time-points such as 90-day and 1-year [1,4,17–19]. The SORG ML algorithms were previously developed for this purpose but were not externally validated [9]. In this study, these algorithms were externally validated in an independent population from a high volume, tertiary care center.

The population of patients used in this study to externally validate the SORG ML algorithms was previously used to compare nine scoring systems in spinal metastatic disease (Tokuhashi, revised Tokuhashi, Tomita, Bauer, modified Bauer, Katagiri, van der Linden, SORG Classic, SORG nomogram) [20–27]. In the prior study, the SORG nomogram achieved the best performance for 90-day with

AUC=0.70 and the Tokuhashi, SORG nomogram, and Katagiri scoring systems achieved the best performance for 1-year mortality with AUC=0.78. Relative to these benchmarks, the SORG ML algorithms improved on 90-day mortality (AUC=0.75–0.81) and performed similarly on 1-year mortality (AUC=0.77–0.78). However, the prior study examined discrimination of scoring systems alone as the primary mode of determining the optimal preoperative tool. As prescribed by the TRIPOD guidelines for prediction, calibration and decision curve analysis are critical [8,10,14] and this study also demonstrated fair calibration and good decision curve analysis of the SORG ML algorithms. Nater et al. recently completed the first calibration study of eight multivariate models (Tokuhashi, revised Tokuhashi, Tomita, van der Linden, Oswestry Spinal Risk Index, Bauer, modified Bauer, Bollen) of survival in spinal metastatic disease in a prospective multicenter study and found that all had poor calibration. In the study, the Bartels model achieved the best calibration with calibration slope=0.45. Relative to these benchmarks [28], the SORG ML algorithms achieved good performance with calibration slope=0.64 for 90 days and 0.77 for 1 year.

The discrepancies in calibration may be due to two factors. The first factor is the difference in baseline characteristics (performance status, primary tumor histology, metastatic tumor burden, overall comorbidity burden, preoperative motor deficit, and laboratory characteristics) between the validation cohort and developmental cohort [29]. The second factor is the relatively small sample size of the validation cohort. Previous studies have shown that at smaller sample sizes for validation cohorts, interpretations of the calibration plot can be unreliable and Van Calster et al. have recommended a minimum of 200 events (90-day and 1-year mortality) and 200 nonevents (90-day and 1-year survival) to reliably assess calibration plots [30–32]. In the context of spinal metastatic disease, achieving a sample size of 200 patients who experienced 90-day mortality, given the observed 90-day mortality of 22.7%, would require a total sample size of 881. The 176 patients in this study spanned 13 years at a high-volume tertiary care center. Without multi-institutional collaboration, it would be very difficult to collect the minimum number of patients recommended for reliable assessment of calibration plots. Although not as compelling as perfect calibration plots, the observed 90-day and 1-year mortality for the risk strata predicted probabilities less than 0.3, 0.3 to 0.6, and greater than 0.6 show that the models do predict probabilities that approximate the observed proportions of events.

There are several limitations that must be acknowledged to appropriately contextualize the findings of this study. Though external validation of the SORG ML algorithms was achieved in an independent population, the study design was retrospective, and the patients came from a different but single institution. The sample size of the study was relatively small, particularly in analyses that examined subpopulations of patients with single primary tumor

histology. Future studies are needed that will assess the predictive performance of the SORG ML algorithms in large, prospective, multicenter samples. In addition, the validation cohort originated from the same nation as the developmental cohort. International validation in multiple countries remains to be undertaken. The sample size of the present study also posed limitations potentially on optimal calibration of the models in the developmental cohort as well as the subpopulation of patients with complete data not requiring multiple imputations. In addition, as a web-based tool, the SORG ML algorithms require reliable access to internet connectivity which may not be available in all settings. Traditional scores and printable paper-based nomograms remain useful for resource-constrained settings. Furthermore, the SORG ML algorithms provide survival estimates at specific time-points (90 days and 1 year) as opposed to prior scores.

Despite these limitations, this study is the first external validation of the SORG ML algorithms in an independent population. The SORG ML algorithms achieved the best performance for 90-day mortality and achieved similar performance for 1-year mortality. The nature of machine learning algorithms allows them to be updated easily with new data over time. Another avenue of future research for the SORG ML algorithms is to retrain them by combining the patients from both institutions (developmental and validation) and externally validating them on a third population, potentially from multiple institutions. Such work is ongoing and continuous updating of these models will help improve prognostication in spinal metastatic disease.

Conclusions

Initial results from external validation of the SORG ML 90-day and 1-year algorithms for survival prediction in spinal metastatic disease suggest potential utility of these digital decision aids in clinical practice. Further studies are needed to validate or refute these algorithms in large patient samples from prospective, international, multi-institutional trials.

References

- [1] Lawton AJ, Lee KA, Cheville AL, Ferrone ML, Rades D, Balboni TA, et al. Assessment and management of patients with metastatic spinal cord compression: a multidisciplinary review. *J Clin Oncol* 2018;37:61–71.
- [2] Loblaw DA, Laperriere NJ, Mackillop WJ. A population-based study of malignant spinal cord compression in Ontario. *Clin Oncol* 2003;15:211–7.
- [3] Klimo Jr. P, Schmidt MH. Surgical management of spinal metastases. *Oncologist* 2004;9:188–96.
- [4] Verlaan JJ, 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.
- [5] Prasad D, Schiff D. Malignant spinal-cord compression. *Lancet Oncol* 2005;6:15–24.
- [6] 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* 2005;366:643–8.
- [7] Kelly ML, Kshetty VR, Rosenbaum BP, Seicean A, Weil RJ. Effect of a randomized controlled trial on the surgical treatment of spinal metastasis, 2000 through 2010: a population-based cohort study. *Cancer* 2014;120:901–8.
- [8] 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.
- [9] 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.
- [10] 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* 2015;13:1.
- [11] 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.
- [12] Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P, et al. 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.
- [13] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2011;28:112–8.
- [14] 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.
- [15] Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 1950;78:1–3.
- [16] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565–74.
- [17] Harel R, Angelov L. Spine metastases: current treatments and future directions. *Eur J Cancer* 2010;46:2696–707.
- [18] 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, mechanical stability, and systemic disease. *J Clin Oncol* 2017;35:2419–27.
- [19] Barzilai O, Fisher CG, Bilsky MH. State of the art treatment of spinal metastatic disease. *Neurosurgery* 2018;82:757–69.
- [20] Ahmed AK, Goodwin CR, Heravi A, Kim R, Abu-Bonsrah N, Sankey E, et al. Predicting survival for metastatic spine disease: a comparison of nine scoring systems. *Spine J* 2018;18:1804–14.
- [21] Tokuhashi Y, Matsuzaki H, Toriyama S, Kawano H, Ohsaka S. Scoring system for the preoperative evaluation of metastatic spine tumor prognosis. *Spine* 1990;15:1110–3.
- [22] Tokuhashi Y, Matsuzaki H, Oda H, Oshima M, Ryu J. A revised scoring system for preoperative evaluation of metastatic spine tumor prognosis. *Spine* 2005;30:2186–91.
- [23] Tomita K, Kawahara N, Kobayashi T, Yoshida A, Murakami H, Akamaru T. Surgical strategy for spinal metastases. *Spine* 2001;26:298–306.
- [24] Bauer HC, Wedin R. Survival after surgery for spinal and extremity metastases: prognostication in 241 patients. *Acta Orthop Scand* 1995;66:143–6.
- [25] 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 Vol* 2005;87:698–703.
- [26] van der Linden YM, Dijkstra SP, Vonk EJ, Marijnen CA, Leer JW. Prediction of survival in patients with metastases in the spinal column: results based on a randomized trial of radiotherapy. *Cancer* 2005;103:320–8.
- [27] 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 Joint Surg Am* Vol 2016;98:1767–76.
- [28] 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* 2010;21:128–38.
- [29] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68:279–89.
- [30] Steyerberg EW, Harrell Jr. FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245–7.
- [31] Steyerberg EW, Uno H, Ioannidis JPA, van Calster B. Poor performance of clinical prediction models: the harm of commonly applied methods. *J Clin Epidemiol* 2018;98:133–43.
- [32] 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.