Can a Bayesian belief network for survival prediction in patients with extremity metastases (PATHFx) be externally validated in an Asian cohort of 356 surgically treated patients?

Hsiang-Chieh HSIEH [1a], Yi-Hsiang LAI [2a], Chia-Che LEE [3], Hung-Kuan YEN [1,4], Ting-En TSENG [2,3], Jiun-Jen YANG [2], Shin-Yiing LIN [3], Ming-Hsiao HU [3], Chun-Han HOU [3], Rong-Sen YANG [3], Rikard WEDIN [5], Jonathan A FORSBERG [6], and Wei-Hsin LIN [3]

[1] Department of Orthopaedic Surgery, National Taiwan University Hospital, Hsin-Chu branch, Hsin-Chu City, Taiwan; [2] Department of Medical Education, National Taiwan University Hospital, Taipei City, Taiwan; [3] Department of Orthopaedic Surgery, National Taiwan University Hospital, Taipei City, Taiwan; [4] Department of Medical Education, National Taiwan University Hospital, Hsin-Chu branch, Hsin-Chu City, Taiwan; [5] Department of Trauma and Reparative Medicine, Karolinska University Hospital, and Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; [6] Department of Orthopaedic Surgery, Johns Hopkins University, Baltimore, MD, USA
[a] Shared first authorship.
Correspondence: W-H Lin: oweihsin@gmail.com
Submitted 2022-05-01. Accepted 2022-07-27.

Background and purpose — Predicted survival may influence the treatment decision for patients with skeletal extremity metastasis, and PATHFx was designed to predict the likelihood of a patient dying in the next 24 months. However, the performance of prediction models could have ethnogeographical variations. We asked if PATHFx generalized well to our Taiwanese cohort consisting of 356 surgically treated patients with extremity metastasis.

Patients and methods — We included 356 patients who underwent surgery for skeletal extremity metastasis in a tertiary center in Taiwan between 2014 and 2019 to validate PATHFx's survival predictions at 6 different time points. Model performance was assessed by concordance index (c-index), calibration analysis, decision curve analysis (DCA), Brier score, and model consistency (MC).

Results — The c-indexes for the 1-, 3-, 6-, 12-, 18-, and 24-month survival estimations were 0.71, 0.66, 0.65, 0.69, 0.68, and 0.67, respectively. The calibration analysis demonstrated positive calibration intercepts for survival predictions at all 6 timepoints, indicating PATHFx tended to underestimate the actual survival. The Brier scores for the 6 models were all less than their respective null model's. DCA demonstrated that only the 6-, 12-, 18-, and 24-month predictions appeared useful for clinical decision-making across a wide range of threshold probabilities. The MC was < 0.9 when the 6- and 12-month models were compared with the 12-month and 18-month models, respectively.

Interpretation — In this Asian cohort, PATHFx's performance was not as encouraging as those of prior validation studies. Clinicians should be cognizant of the potential decline in validity of any tools designed using data outside their particular patient population. Developers of survival prediction tools such as PATHFx might refine their algorithms using data from diverse, contemporary patients that is more reflective of the world's population.

Survival estimation is important for the management of skeletal metastasis. Patients with short remaining life might not benefit from a major operation, and patients with longer survival could face revision surgery if not initially provided with more durable reconstruction. Predicting survival, however, is difficult. Several preoperative scoring systems (PSSs) have thus been developed for this purpose (1). Among them, PATHFx (https://www.pathfx.org/), developed in 2011, is a modern machine learning-based algorithm using data from 189 patients who underwent surgery for skeletal metastases (2). PATHFx showed good performance in several external cohorts from developed regions such as the North America, Italy, the Scandinavian peninsula, Australia, and Japan (2-7). It has also been recently updated to the 3rd version (3) to provide predictions for patients treated both operatively and with radiation only. However, some studies suggested PSSs could

DOI 10.2340/17453674.2022.4545

perform differently between ethnogeographically distinct cohorts (4,5,8-12), and they should have been validated before being applied onto a specific population. The PATHFx offers survival predictions at 6 different time points: 1, 3, 6, 12, 18, and 24 months. In the absence of sudden adverse events, a patient's predicted survival probability at short term should consistently be higher than that at a longer term because survival typically follows the law of attrition by time. If a survival prediction model frequently made paradoxical calculations, its clinical utility might be questioned. In the literature, however, few studies have assessed the performance of a PSS based on model consistency.

We asked in this study: (1) Does PATHFx v3.0 generalize well to a Taiwanese cohort predominantly composed of patients of Han Chinese descent?; (2) Does PATHFx v3.0 show good model consistency across its predictions at various time points?

Patients and methods

Study design and setting

This was a retrospective study based on 356 patients aged ≥ 18 years undergoing surgery for long-bone carcinoma metastases between 2014 and 2019 at a tertiary center in Taiwan (Figure 1, see Supplementary data). In general, the indications for surgery were patients with an ASA classification ≤ IV or patients considered fit for surgery based on a multidisciplinary assessment by a medical oncologist, anesthesiologist, and orthopedic oncologist, and the presence of a complete pathologic fracture or an impending pathologic fracture deemed unlikely to heal with nonoperative treatment alone. An impending fracture was diagnosed if the lesion in question had a Mirels score ≥ 9 and caused pain or weakness in the limbs involved (13). We excluded patients who received their first surgery at an outside institution.

Participants' baseline characteristics

98% (349/356) of the patients were of Han Chinese descent based on their self-reported ethnicity. The median age was 61 years (25–95) and 48% of the patients were male (Table 1). 44% of the patients had a Kitagiri Group 1 cancer; 20% had a Group 2; 36% had a Group 3. The most common primary tumors were non-small-cell lung cancer (23 %) and breast cancer (16%). A pathologic fracture occurred in 55% of the patients; visceral metastases were present in 51%; lymph node metastases were found in 43%; other bone metastases were identified in 72%. 21% of the patients had an Eastern Cooperative Oncology Group (ECOG) score of 3–4, and 79% had an ECOG of 0–2. Follow-up was censored at patients' death or 2 years after the first surgery. 5%, 19%, 34%, 55%, 65%, and 72% patients died within 1, 3, 6, 12, 18, and 24 months of surgery, respectively (Figure 2). 6 patients were lost to follow-up within 90 days; 30 were lost to follow-up within 1 year.

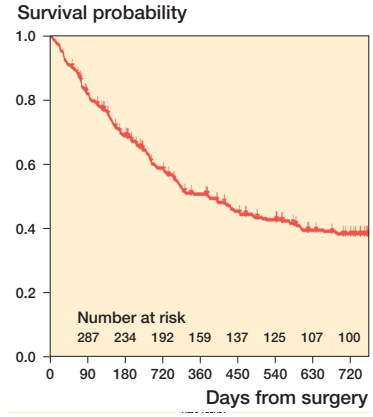

Figure 2. Kaplan–Meier curve of 356 patients in this study.

Baseline characteristics in the validation cohort differed from those in the PATHFx v3.0 development cohort in several regards, such as the makeup of primary oncologic diagnoses, the presence of visceral metastases, and ECOG score (all $p < 0.001$). Patients' age, sex, presence of lymph node metastases, number of bone metastases, and survival rates at different time points were similar between the validation and the development cohorts.

Treatment

In general, radiosensitive tumors such as breast, prostate, lung cancer, and hematologic malignancies were stabilized with a nail or plate-and-screws construct and given adjuvant radiotherapy postoperatively. Radioresistant tumors such as renal cell carcinoma and hepatocellular carcinoma were typically treated with extended curettage, cement augmentation, and nail/plate fixation, or resected and replaced with a prosthesis. Megaprosthesis arthroplasty was considered for patients with an unsalvageable joint or extensive metaphyseal bone loss if they had a reasonably long survival and for those who had oligometastasis and may benefit from wide excision of metastatic tumor. 59% of patients underwent intramedullary nailing, 23% were treated with plate-and-screws fixation, and 18% received endoprosthetic reconstruction.

Prognostic variables and outcome

The following preoperative data were extracted: age at the time of surgery, sex, preoperative hemoglobin concentration (g/dL), absolute lymphocyte count (k/uL), the presence of visceral and lymph node metastases, impending or completed pathologic fracture, number of bone metastases, the ECOG score, and the patient's primary tumor type (7,8). The surgeon's estimation of survival was omitted since this was not recorded in the electronic medical records. The primary outcomes were 1-, 3-, 6-, 12-, 18-, and 24-month mortality, which were defined as the time between the patient's first surgery for an extremity metastasis and death of any cause. The 3-month and 12-month survival have historically been used for reporting oncologic outcomes. These 2 time points are also

Table 1. Comparison of external validation population with development population. Values are n (%) unless otherwise specified

Variables	Training set Developmental cohort (n = 397)	Validation set Scandinavia (n = 815)	Italy (n = 287)	Japan (n = 261)	Taiwan (n = 356)	p-value versus developmental cohort	Scandinavia	Italy	Japan
Age at surgery, mean (SD)	62.4 (13.7)	66.4 (12.7)	63.1 (11.7)	61.8 (12.3)	61.2 (12.8)	0.2	< 0.01	0.1	0.6
Sex						0.1	0.3	0.1	0.2
Male	170 (43)	369 (45)	120 (42)	139 (53)	172 (48)				
Female	227 (57)	446 (55)	167 (58)	122 (47)	184 (52)				
Oncologic diagnosis group [a]						< 0.001	< 0.001	< 0.001	< 0.001
1	108 (27)	173 (21)	63 (22)	60 (23)	158 (44)				
2	72 (18)	74 (9)	44 (15)	75 (29)	71 (20)				
3	211 (53)	567 (69)	173 (60)	126 (48)	127 (36)				
Missing	6 (2)	1 (<1)	7 (2)	0 (0)	0 (0)				
Visceral metastases						< 0.001	0.01	< 0.001	0.1
Yes	247 (62)	325 (40)	91 (32)	114 (44)	180 (51)				
No	148 (37)	441 (54)	161 (56)	147 (56)	176 (49)				
Missing	2 (<1)	49 (6)	35 (12)	0 (0)	0 (0)				
Lymph node metastases						0.2	0.004	0.4	< 0.001
Yes	152 (38)	169 (21)	96 (33)	71 (27)	153 (43)				
No	245 (62)	143 (18)	146 (51)	190 (73)	203 (57)				
Missing	0 (0)	503 (62)	45 (16)	0 (0)	0 (0)				
Skeletal metastases						1.0	< 0.001	< 0.001	< 0.001
Solitary	112 (28)	123 (15)	139 (48)	112 (43)	100 (28)				
Multiple	285 (72)	666 (81)	144 (50)	149 (57)	256 (72)				
Missing	0 (0)	26 (3)	4 (1)	0 (0)	0 (0)				
Eastern Cooperative Oncology Group performance status score						< 0.001	< 0.001	< 0.001	< 0.001
0–2	222 (56)	558 (69)	123 (42)	166 (64)	283 (79)				
3–4	164 (41)	257 (31)	106 (37)	95 (36)	73 (21)				
Missing	11 (3)	0 (0)	58 (20)	0 (0)	0 (0)				
Survival > 1 month						0.9	< 0.001	–	0.1
Yes	379 (96)	707 (87)	–	240 (92)	339 (95)				
No	18 (5)	108 (13)	–	21 (8)	17 (5)				
Missing	0 (0)	0 (0)	–	0 (0)	0 (0)				
Survival > 3 months						0.3	< 0.001	< 0.001	0.6
Yes	309 (78)	557 (68)	267 (93)	218 (84)	287 (81)				
No	88 (22)	258 (32)	20 (7)	43 (17)	63 (18)				
Missing	0 (0)	0 (0)	0 (0)	0 (0)	6 (2)				
Survival > 6 months						0.4	< 0.001	–	1.0
Yes	248 (63)	372 (46)	–	179 (69)	234 (66)				
No	149 (38)	443 (54)	–	82 (31)	108 (30)				
Missing	0 (0)	0 (0)	–	0 (0)	14 (4)				
Survival > 12 months						0.4	< 0.001	< 0.001	0.02
Yes	189 (48)	241 (30)	181 (63)	152 (58)	159 (45)				
No	208 (52)	574 (70)	106 (37)	109 (42)	167 (47)				
Missing	0 (0)	0 (0)	0 (0)	0 (0)	30 (8)				
Survival > 18 months						0.8	< 0.001	–	0.3
Yes	134 (34)	156 (19)	–	113 (43)	123 (35)				
No	263 (66)	659 (81)	–	148 (57)	191 (54)				
Missing	0 (0)	0 (0)	–	0 (0)	42 (12)				
Survival > 24 months						0.7	< 0.001	–	0.6
Yes	105 (26)	117 (14)	–	79 (30)	98 (28)				
No	292 (74)	698 (86)	–	182 (70)	204 (72)				
Missing	0 (0)	0 (0)	–	0 (0)	54 (15)				
Hemoglobin concentration (mg/dL)	11 (10–13) [b]	11.5 (3.5)	11.5 (1.4)	12.4 (6.0)	11 (2.0)	0.2	0.002	< 0.001	< 0.001
Absolute lymphocyte count (K/uL)	1.2 (1.3)	1.2 (0.7)	1.3 (0.5)	1.3 (0.9)	1.3 (0.9)	0.2	0.06	1	1
Surgeon's estimate of survival (months)	9 (6–18) [b]	11.8 (17.2)	11.2 (7.0)	8.9 (4.6)	Not available	–	–	–	–
Pathologic fracture status						0.02	< 0.001	0.5	< 0.00
Completed	84 (44)	614 (75)	143 (50)	105 (40)	195 (55)				
Impending	105 (56)	196 (24)	131 (46)	156 (60)	161 (45)				
Missing	all RT cases	5 (1)	3 (5)	0 (0)	0 (0)				
Preoperative systemic therapy						–	–	–	–
Chemotherapy	–	–	–	–	227 (64)				
Target therapy	–	–	–	–	121 (34)				
Hormone therapy	–	–	–	–	59 (17)				
Immunotherapy	–	–	–	–	24 (7)				
Preoperative radiotherapy	–	–	–	–	215 (60)	–	–	–	–
Surgical treatment						–	–	–	–
Intramedullary nail	–	–	–	–	210 (59)				
Plate-and-screws fixation	–	–	–	–	81 (23)				
Endoprosthetic reconstruction	–	–	–	–	65 (18)				

[a] Cancer types of pulmonary, gastric, and hepatoma and melanoma were assigned to Group 1; sarcomas and other carcinomas carcinomas were assigned to Group 2; breast, prostate, thyroid cancer, renal cell carcinoma, multiple myeloma, and malignant lymphoma were assigned to Group 3.

[b] The original article provided only IQR without SD.

meaningful for surgeons when they need to determine whether surgery would be beneficial (3-month survival probability) and whether they should pursue a more durable reconstruction (12-month survival probability). The 1-month prediction may help clinicians and patients decide if they should seek palliative treatment. With the advances in surgical implants and techniques, patients with favorable survival estimates at 1–6 months may be candidates for procedures such as percutaneous cementoplasty and minimally invasive nail or plate stabilization. The 18–24 months predictions allow surgeons to better assess the need for more aggressive surgical strategies such as tumor resection and prosthetic replacement. Patients with unknown final survival status due to loss to follow-up were excluded from analyses of model performance and calculation of actual survival rates.

Missing data

The lymphocyte count was missing in 8 patients (2.2%). The missForest methods were used to impute missing values. Loss to follow-up occurred in 0% of the patients at 1 month, 2% at 3 months, 4% at 6 months, 8% at 12 months, 12% at 18 months, and 15% at 24 months.

Statistics and assessment of model performance

We manually retrieved survival predictions at the 6 different time points from the PATHFx application (https://www.pathfx.org/). The model's performance was evaluated by discrimination (i.e., concordance index; c-index), calibration (i.e., calibration slope and intercept), overall performance (i.e., Brier score [BS]), decision curve analysis (DCA), and model consistency (i.e., whether the shorter-term survival prediction probability was higher than the longer-term survival prediction probability in paired comparisons).

The c-index measures goodness of fit of a model, and typically ranges from 0.5 to 1.0 (**14**). A c-index = 0.5 indicates random guessing and 1.0 a perfect prediction. In general, a c-index ≥ 0.7 indicates a model has good discriminatory ability, and a c-index ≥ 0.8 indicates excellent discrimination. Calibration evaluates the agreement between the predicted outcomes and the actual outcomes by plotting a calibration curve and measuring its slope and intercept. A perfect calibration has a slope of 1 and an intercept of 0 (**14**). Calibration analysis may detect whether a model overestimates or underestimates the examined outcome when there is a negative or positive intercept, respectively.

The BS is the average mean squared difference between the model predictions and the observed outcomes. A BS of 0 suggests the perfect model and a BS of 1 signifies the worst possible model (**14**). However, the prevalence of the outcome (in this case, the actual survival rate) must be considered. The BS of the null model was calculated by assigning a probability equal to the prevalence of the outcome to each patient. If a prediction model's BS is lower than the null model's, the model is deemed as having good performance.

The DCA identifies whether a treatment decision would do more good than harm by appraising the cost-to-benefit ratio. The clinician can choose an applicable threshold probability for a certain treatment, assess the corresponding net benefit on the decision curve, and determine if the treatment is clinically beneficial (**14**).

We devised a metric called model consistency (MC) to assess whether PATHFx provided consistent predictions at the 6 different timepoints. A consistent result is defined as having an intuitively reasonable prediction pair, in which the survival probability at a shorter term is higher than that at a longer term (e.g., a 3-month and 18-month survival probability of 90% and 20%, respectively). An inconsistent result refers to a paradoxical pair of predictions (e.g., a 6-month and 12-month survival probability of 30% and 40%, respectively). MC is the ratio of consistent prediction pairs divided by all prediction pairs. An MC = 1 indicates the best consistency and an MC = 0 signifies the worst.

Baseline clinical and demographic data and survival rates at different timepoints of the development and external validation cohorts were compared by either Student's t-tests for continuous variables or chi-square tests for categorical ones. 95% confidence interval are represented as CI for applicable statistics. The significance level was set at 0.05. We used R for Mac (v4.0.4) (R Foundation for Statistical Computing, Vienna, Austria) for statistical analyses.

Ethics, funding, and potential conflicts of interests

This study was designed following the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guideline (**15**) and approved by our Research Ethics Committee (201912022RIND, Treatment effect on patients with osteolytic tumors: a retrospective study). No funding was received for this study. The authors involved in the original PATHFx studies, JF and RW, were not part of data extraction and analysis and did not have access to our original Taiwanese dataset. JF and RW are shareholders in Prognostix AB, Sweden.

Results

Assessment of model performance

The c-indexes of PATHFx's 1-, 3-, 6-, 12-, 18-, and 24-month survival predictions were 0.71 (CI 0.58–0.84), 0.66 (CI 0.59–0.73), 0.65 (CI 0.59–0.71), 0.69 (CI 0.64–0.75), 0.68 (CI 0.62–0.75), and 0.67 (CI 0.60–0.74), respectively (Table 2, see Supplementary data), indicating an acceptable but not great discriminatory ability of the model in our Taiwanese cohort. The calibration intercept was 1.00 (CI 0.50–1.51) for 1-month survival prediction; 1.60 (CI 1.31–1.90) for 3-month; 1.39 (CI 1.13–1.64) for 6-month; 0.61 (CI 0.36–0.85) for 12-month; 0.42 (CI 0.16–0.68) for 18-month; and 0.52 (CI 0.24–0.80) for 24-month survival prediction (Figure 3 and Figure 4, see Sup-

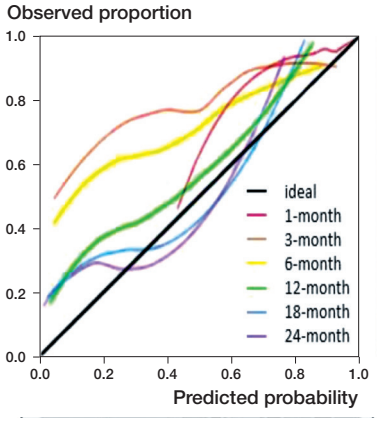

Figure 3. Calibration plots of predictions by PATHFx are shown for 1-month, 3-month, 6-month, 12-month, 18-month, and 24-month survival. The calibration plot visualizes how accurate the predictions are over different probabilities. The diagonal line represents the optimal calibration; the closer the line of the model, the more accurate the prediction. The calibration intercepts were 1.00 (CI 0.50–1.51) for 1-month, 1.60 (CI 1.31–1.90) for 3-month, 1.39 (CI 1.13–1.64) for 6-month, 0.61 (CI 0.36–0.85) for 12-month, 0.42 (CI 0.16–0.68) for 18-month, and 0.52 (CI 0.24–0.80) for 24-month survival prediction. The calibration slopes were 0.75 (CI 0.28–1.21) for 1-month, 0.51 (CI 0.25–0.76) for 3-month, 0.52 (CI 0.29–0.75) for 6-month, 0.71 (CI 0.48–0.94) for 12-month, 0.59 (CI 0.38–0.79) for 18-month, and 0.51 (CI 0.31–0.71) for 24-month survival prediction.

plementary data). These positive intercepts indicated PATHFx tended to underestimate the survival of Taiwanese patients, especially with its shorter-term (1-, 3-, and 6-month) predictions. The BSs of the 6 models were all < 0.25 and less than that of their respective null model (Table 2, see Supplementary data), suggesting a generally adequate fit of these models to the actual outcome of the patients.

The DCA demonstrated that only the 6-, 12-, 18-, and 24-month predictions made by PATHFx could provide clinical benefit across a wide range of threshold probabilities. The 1- and 3-month predictions provided minimal benefits that were seen only when the threshold possibilities were very high (0.9–1.0 and 0.7–0.9, respectively [Figure 5]). In other words, the users would likely only find PATHFx helpful to decision-making when the risk-to-benefit ratio of the proposed treatment is very high at these 2 time points.

Model consistency (MC) of PATHFx

When we extracted survival estimates from PATHFx, we noticed some patients were given predictions against the law of attrition by time. We defined such counterintuitive predictions as having model inconsistency. In our cohort, model inconsistency rarely occurred with the PATHFx predictions at 1, 3, and 24 months, which all had an MC close to 1.0 (Table 3). However, the 6-month predictions had a less than optimal MC of 0.86 when compared with the 12-month predictions; the 12-month predictions had an MC of 0.88 when compared with the 18-month predictions. We could not compare this study's MC with those of the development and external validation studies because none of them mentioned similar observations and reported this metric.

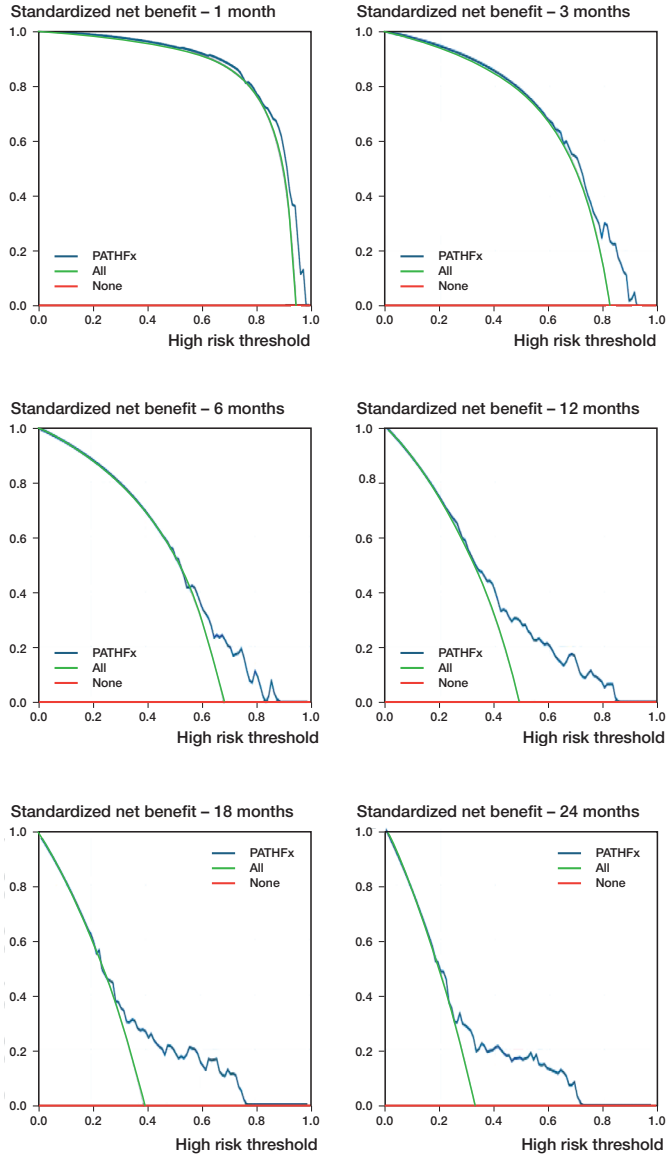

Figure 5. Decision-curve analysis plots of predictions by PATHFx are shown for 1-month, 3-month, 6-month, 12-month, 18-month, and 24-month survival. The range of threshold probabilities are 0.9~1.0 (1-month), 0.7~0.9 (3-month), 0.5~0.9 (6-month), 0.3~0.8 (12-month), 0.3~0.8 (18-month), and 0.2~0.7 (24-month), respectively.

Discussion

What is already known

Management of patients with skeletal metastasis will benefit from an accurate survival prediction tool because overtreatment or undertreatment might thus be avoided. Machine-learning algorithms such as the PATHFx have shown great promise in survival estimation. However, several studies dem-

Table 3. Model consistency of PATHFx algorithm

Time	1-month	3-month	6-month	12-month	18-month	24-month
1-month	–	1	1	1	1	1
3-month	1	–	0.97 (0.95–0.98)	0.96 (0.94–0.98)	0.97 (0.95–0.99)	0.99 (0.98–1.00)
6-month	1	0.97 (0.95–0.98)	–	0.86 (0.83–0.90)	0.92 (0.90–0.95)	0.98 (0.97–1.00)
12-month	1	0.96 (0.94–0.98)	0.86 (0.83–0.90)	–	0.88 (0.85–0.92)	0.99 (0.97–1.00)
18-month	1	0.97 (0.95–0.99)	0.92 (0.90–0.95)	0.88 (0.85–0.92)	–	1
24-month	1	0.99 (0.98–1.00)	0.98 (0.97–1.00)	0.99 (0.97–1.00)	1	–

onstrated that a machine-learning model's performance could vary in ethnogeographically distinct populations (**8,16**) and repeated validation in different cohorts is needed (**9,15**).

Novel insights

We found that PATHFx's discriminatory ability was not optimal in our Taiwanese cohort. In addition, the ubiquitously positive calibration intercepts across the model's 6 prediction time points indicated PATHFx tended to underestimate the survival of our patients, especially in the shorter term (1, 3, and 6 months). These results suggest that PATHFx might need further fine-tuning if it is to be used in regions with distinct clinico-demographic compositions or healthcare systems.

Generalizability of PATHFx v3.0

Developed in 2011, PATHFx was one of the earliest models in orthopedics that employed machine-learning techniques (**2**). It was updated in 2018 with more contemporary data, and now provides survival predictions at 6 time points. Validation studies performed in the US, Italy, Japan, Australia, and Scandinavian countries demonstrated that PATHFx retained excellent discriminatory ability in these developed regions, with c-indexes ranging from 0.70 to 0.85 (**2-6**) (Table 4, see Supplementary data). However, these Western cohorts are clearly different from the Han Chinese-predominant population in Taiwan. Even in Japan, a country also considered "Asian," the Han Chinese do not constitute a large ethnic group. Therefore, our study is meaningful as the target population is an untested one. In this external validation cohort, only the c-index for 1-month prediction was slightly greater than 0.70. The c-indexes of predictions at the other five time points were all below 0.70. These results indicate PATHFx had less than optimal discrimination in our Taiwanese cohort. In comparison, the c-indexes for survival predictions at the 6 time points were noticeably lower than their counterparts in the development and other external validation studies (Table 4, see Supplementary data). Not surprisingly, the demographic data of our cohort differed from those of other patient populations. For example, a substantially higher percentage of our patients had an oncologic group 1 cancer (Table 1). As the primary cancer type was a prognosticator of patient survival (**17,18**), it was not inconceivable that discrepancies in the distribution of primary cancer diagnoses among different cohorts could lead to variations in model performance. In addition, our Taiwanese cohort, when compared with the other cohorts, had a higher proportion of patients with an ECOG score of 0–2 (Table 1). The ECOG score has repeatedly been reported as a prognostic factor for patients with cancer (**19,20**), and its distinct distribution in our cohort might contribute to the decreased discriminatory ability of PATHFx in this study.

We also observed uniformly positive calibration intercepts across the 6 time points, suggesting that PATHFx in general underestimates postoperative survival in our cohort (Figure 3 and Figure 4, see Supplementary data). Although the exact cause(s) of this underestimation is/are hard to determine, one potential contributor is the unique Taiwanese National Health Insurance (NHI), a government-run program that provides universal coverage to all Taiwanese citizens at an affordable premium. Patients in Taiwan are typically less financially constrained to receive critical cancer treatment that impacts survival, such as molecular targeted therapy for lung cancer with EGFR mutation. On decision curve analysis, we found that using PATHFx 1- and 3-month survival predictions as an aid to make treatment decisions provided only marginal clinical benefit when the threshold probability, i.e., the risk-to-benefit ratio, of the proposed operation was very high (Figure 5). PATHFx's 6-, 12-, 18-, and 24-month survival predictions, on the other hand, could be clinically beneficial when the tentative treatment strategy had a moderate risk-to-benefit ratio (Figure 5). Considering these findings, we felt PATHFx might be more useful as a decision-making aid with its mid- to longer-term predictions.

Model consistency of PATHFx

As PATHFx makes survival predictions at several different time points, clinicians might feel perplexed or find it difficult to convey the prediction results to their patients if the estimations appear counterintuitive. We devised model consistency (MC) as a metric to evaluate how often such mismatches occurred. In this study, the 1-, 3-, and 24-month survival predictions were rather consistent, as their MCs were all close to 1.0. A slight decline in MC was observed when the 6-month and 12-month predictions were compared with the 12-month (MC = 0.86) and 18-month predictions (MC = 0.88), respectively. Since inconsistencies could happen with any survival prediction models, it is important to interpret their estimations

in the context of the complete survival trajectory. Nevertheless, future studies may consider reporting this metric so that readers can make a more comprehensive assessment of the model in question.

A model that does not always produce consistent predictions may deter clinicians from adopting it into their practice. The performance of PATHFx might be improved if future updates are done by incorporating more diverse and international datasets in model retraining. Another way to increase model performance may be creating region-specific versions of PATHFx or adding ethnogeographic modifiers to the algorithms. Doing so would allow users in different parts of the world to choose a version that is most applicable to their clinical setting. Furthermore, as advances in medical therapies will undoubtedly alter the survival of patients with cancer metastasis, we believe PATHFx needs to be continually updated with newer data that is more reflective of modern cancer treatment. PATHFx developers might also explore whether adding granular details, such as tumor molecular subtypes and response to systemic and local therapies, into the algorithms would enhance the application's predictive accuracy and consistency.

Limitations

Several limitations should be kept in mind when interpreting our results. First, the Taiwanese healthcare system is a universal one that covers every citizen at an affordable premium (21). Taiwanese patients might therefore have more ready access to certain advanced chemotherapeutics such as molecular targeted therapy and immunotherapy. Second, in an age of globalization and international migration, the racial composition and genetic pool of a population can become mixed and change over time. A tool developed and validated years ago may then suffer from performance declines when applied to more modern patients. This highlights the importance of continual updates and external validations for survival prediction models as time goes by. Third, survival estimation is only one aspect in the treatment decision-making process. The priority for patients with incurable metastatic cancer is perhaps maintenance of quality of life. In the case of femoral pathologic fractures, pain relief and ambulatory function can be difficult to obtain with nonoperative means. The current prediction models are not comprehensive enough for clinicians to base the treatment decision solely on their outputs. Fourth, the PATHFx model does not include the type and extent of surgery in its current algorithm. In clinical practice, however, one must also consider potential complications associated with the proposed operation. A well-meaning, large-extent surgery performed based on good survival prediction could actually lead to complications that negatively impact the patient's function and survival. Future studies should try to develop algorithms to predict major complications and important functional outcomes after surgical intervention. These predictions could help clinicians make a more comprehensive assessment of the proposed treatment, and better inform their patients during the shared decision-making process. Lastly, survival time of patients with metastatic cancer ultimately hinges upon the availability of effective medical treatment. When breakthroughs in cancer therapies occur, PATHFx, or any survival prediction models, would likely need to be retrained to stay up to date.

Conclusion

Prediction models developed in one part of the world should be externally validated before they are applied to patients in other regions. PATHFx did not demonstrate great discrimination in our Taiwanese cohort composed mostly of patients of Han Chinese descent, and showed a general tendency to underestimate the actual survival in this population. These findings are not as encouraging as results from prior validation studies in other patient populations. If developers of PATHFx intend the application to gain wider acceptance, they should consider retraining and refining the PATHFx models using data from diverse, contemporary patients that is more reflective of the world's population.

All authors have contributed to the research design and/or interpretation of data, and the drafting and revising of the manuscript.

The authors would like to thank all healthcare professionals from various departments of National Taiwan University Hospital for their contribution in providing multidisciplinary care for their patients. They also thank the staff at the Department of Medical Research for gathering data from the institutional integrative medical database.

Acta thanks Olivier Quinten Groot and other anonymous reviewers for help with peer review of this study.

1. **Ogink P T, Groot O Q, Karhade A V, Bongers M E R, Oner F C, Verlaan J J, et al.** Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review. Acta Orthop 2021; 92: 526-31. doi: 10.1080/17453674.2021.1932928,
2. **Forsberg J A, Eberhardt J, Boland P J, Wedin R, Healey J H.** Estimating survival in patients with operable skeletal metastases: an application of a Bayesian belief network. PLoS One 2011; 6: e19956. doi: 10.1371/journal.pone.0019956.
3. **Anderson A B, Wedin R, Fabbri N, Boland P, Healey J, Forsberg J A.** External validation of PATHFx Version 3.0 in patients treated surgically and nonsurgically for symptomatic skeletal metastases. Clin Orthop Relat Res 2020; 478: 808-18. doi: 10.1097/CORR.0000000000001081.
4. **Ogura K, Gokita T, Shinoda Y, Kawano H, Takagi T, Ae K, et al.** Can a multivariate model for survival estimation in skeletal metastases (PATHFx) be externally validated using Japanese patients? Clin Orthop Relat Res 2017; 475: 2263-70. doi: 10.1007/s11999-017-5389-3.
5. **Piccioli A, Spinelli M S, Forsberg J A, Wedin R, Healey J H, Ippolito V, et al.** How do we estimate survival? External validation of a tool for survival estimation in patients with metastatic bone disease—decision analysis and comparison of three international patient populations. BMC Cancer 2015; 15: 1-8. doi: 10.1186/s12885-015-1396-5.
6. **Meares C, Badran A, Dewar D.** Prediction of survival after surgical management of femoral metastatic bone disease: a comparison of prognostic models. J Bone Oncol 2019; 15: 100225. doi: 10.1016/j.jbo.2019.100225.

7. **Overmann A L, Clark D M, Tsagkozis P, Wedin R, Forsberg J A.** Validation of PATHFx 2.0: an open–source tool for estimating survival in patients undergoing pathologic fracture fixation. J Orthop Res 2020; 38: 2149-56. doi: 10.1002/jor.24763.
8. **Yang J J, Chen C W, Fourman M S, Bongers M E R, Karhade A V, Groot O Q, et al.** International external validation of the SORG machine learning algorithms for predicting 90-day and 1-year survival of patients with spine metastases using a Taiwanese cohort. Spine J 2021; 10.1016/j.spinee.2021.01.027. doi: 10.1016/j.spinee.2021.01.027.
9. **Groot O Q, Bindels B J, Ogink P T, Kapoor N D, Twining P K, Collins A K, et al.** Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. Acta Orthop 2021; 92: 385-93. doi: 10.1080/17453674.2021.1910448.
10. **Hu M-H, Yen H-K, Chen I-H, Wu C-H, Chen C-W, Yang J-J, et al.** Decreased psoas muscle area is a prognosticator for 90-day and 1-year survival in patients undergoing surgical treatment for spinal metastasis. Clin Nutr 2022. doi: 10.1016/j.clnu.2022.01.011.
11. **Yen H-K, Ogink PT, Huang C-C, Groot OQ, Su C-C, Chen S-F, et al.** A machine learning algorithm for predicting prolonged postoperative opioid prescription after lumbar disc herniation surgery: an external validation study using 1,316 patients from a Taiwanese cohort. Spine J 2022. doi: 10.1016/j.spinee.2022.02.009.
12. **Tseng T E, Lee C C, Yen H K, Groot O Q, Hou C H, Lin S Y, et al.** International validation of the SORG machine-learning algorithm for predicting the survival of patients with extremity metastases undergoing surgical treatment. Clin Orthop Relat Res 2021. doi: 10.1097/CORR.0000000000001969.
13. **Mirels H.** Metastatic disease in long bones: a proposed scoring system for diagnosing impending pathologic fractures. Clin Orthop Relat Res 1989: 256-64.
14. **Karhade A V, Schwab J H.** CORR synthesis: when should we be skeptical of clinical prediction models? Clin Orthop Relat Res 2020; 478: 2722. doi: 10.1097/CORR.0000000000001367.
15. **Collins G S, Reitsma J B, Altman D G, Moons K G.** Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015; 102: 148-58. doi: 10.1002/bjs.9736.
16. **Karhade A V, Thio Q, Ogink P T, Bono C M, Ferrone M L, Oh K S, et al.** Predicting 90-Day and 1-year mortality in spinal metastatic disease: development and internal validation. Neurosurg 2019; 85: E671-E681. doi: 10.1093/neuros/nyz070.
17. **Chen C-H, Tzai T-S, Huang S-P, Wu H-C, Tai H-C, Chang Y-H, et al.** Clinical outcome of Taiwanese men with metastatic prostate cancer compared with other ethnic groups. Urology 2008; 72: 1287-92. doi: 10.1016/j.urology.2008.01.026.
18. **Lin Y-J, Lin C-N, Sedghi T, Hsu S H, Gross C P, Wang J-D, et al.** Treatment patterns and survival in hepatocellular carcinoma in the United States and Taiwan. PloS one 2020; 15: e0240542.
19. **Sørensen M S, Gerds T A, Hindsø K, Petersen M M.** External validation and optimization of the SPRING model for prediction of survival after surgical treatment of bone metastases of the extremities. Clin Orthop Relat Res 2018; 476: 1591. doi: 10.1097/01.blo.0000534678.44152.ee.
20. **Willeumier J, Van der Linden Y, Van der Wal C, Jutte P, van der Velden J, Smolle M, et al.** An easy-to-use prognostic model for survival estimation for patients with symptomatic long bone metastases. J Bone Jt Surg 2018; 100: 196-204. doi: 10.2106/JBJS.16.01514.
21. **Wu T-Y, Majeed A, Kuo K N.** An overview of the healthcare system in Taiwan. Lond J Prim Care 2010; 3: 115-19. doi: 10.1080/17571472.2010.11493315.

Supplementary data

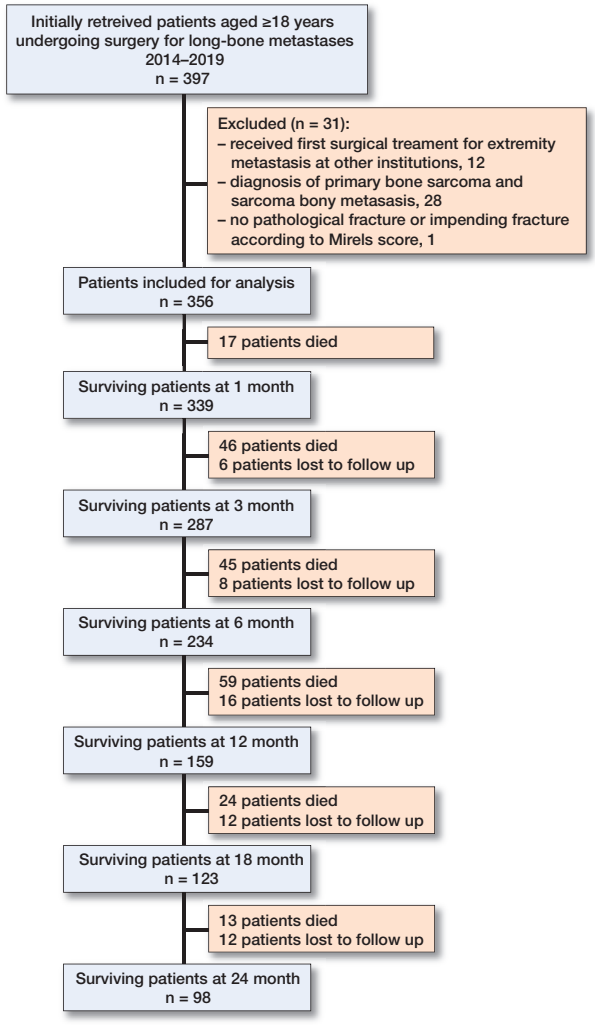

Figure 1. Flow diagram showing the enrolled patients.

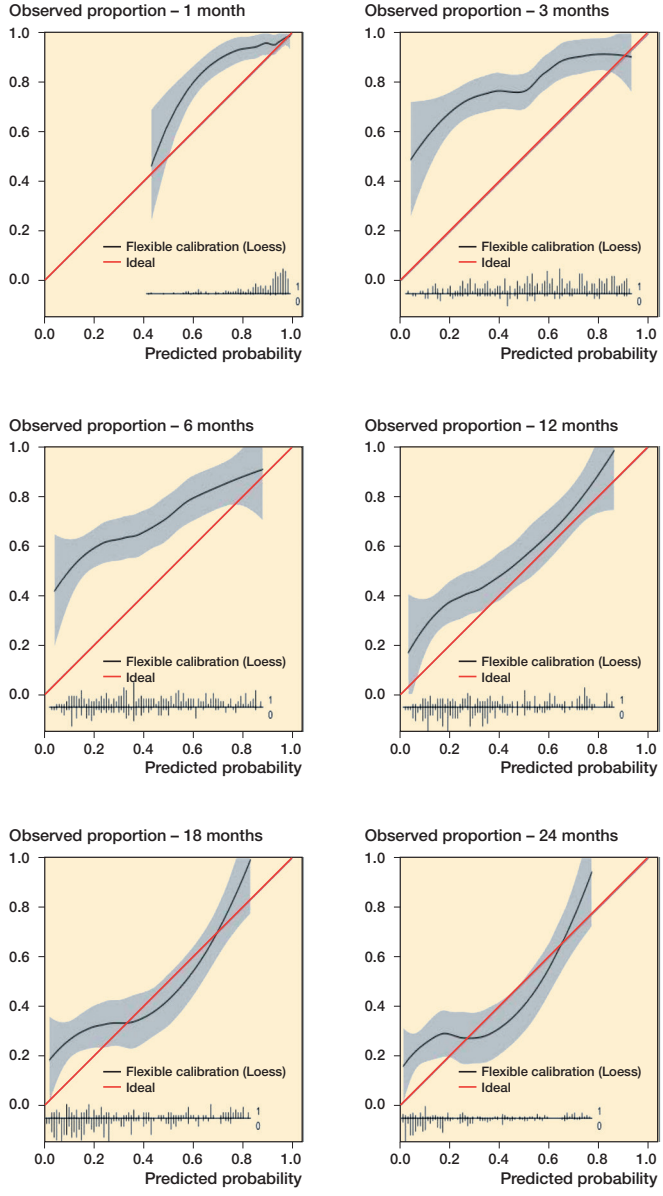

Figure 4. Calibration plots of predictions by PATHFx are shown for 1-month, 3-month, 6-month, 12-month, 18-month, and 24-month survival. The calibration plot visualizes how accurate the predictions are over different probabilities. The diagonal line represents the optimal calibration; the closer the line of the model, the more accurate the prediction. The calibration intercepts were 1.00 (CI 0.50–1.51) for 1-month, 1.60 (CI 1.31–1.90) for 3-month, 1.39 (CI 1.13–1.64) for 6-month, 0.61 (CI 0.36–0.85) for 12-month, 0.42 (CI 0.16–0.68) for 18-month, and 0.52 (CI 0.24–0.80) for 24-month survival prediction. The calibration slopes were 0.75 (CI 0.28–1.21) for 1-month, 0.51 (CI 0.25–0.76) for 3-month, 0.52 (CI 0.29–0.75) for 6-month, 0.71 (CI 0.48–0.94) for 12-month, 0.59 (CI 0.38–0.79) for 18-month, and 0.51 (CI 0.31–0.71) for 24-month survival prediction.

Table 2. Concordance indices (95% CI) (C-index) and Brier scores with the null model in parentheses of PATHFx 3.0 algorithms by primary tumor histology in the validation cohort (n = 356)

Factor	n	C-index	Brier score
1-month survival prediction			
Overall	356	0.71 (0.58–0.84)	0.04 (0.05)
Solid organ	333	0.71 (0.58–0.84)	0.05 (0.05)
Lung	116	0.78 (0.61–0.94)	0.05 (0.05)
Breast	58	0.76 (NA)	0.03 (0.03)
Liver	37	NA	NA
Hematologic malignancies	23	NA	NA
Kidney	21	0.89 (0.67–1.00)	0.06 (0.09)
Prostate	19	0.81 (0.42–1.00)	0.07 (0.09)
3-month survival prediction			
Overall	356	0.66 (0.59–0.73)	0.14 (0.16)
Solid organ	333	0.65 (0.58–0.73)	0.15 (0.16)
Lung	116	0.71 (0.57–0.85)	0.13 (0.16)
Breast	58	0.57 (0.37–0.75)	0.07 (0.07)
Liver	37	0.58 (0.33–0.83)	0.14 (0.14)
Hematologic malignancies	23	0.65 (NA)	0.05 (0.05)
Kidney	21	0.89 (0.75–1.00)	0.13 (0.19)
Prostate	19	0.63 (0.35–0.93)	0.20 (0.22)
6-month survival prediction			
Overall	356	0.65 (0.59–0.71)	0.20 (0.22)
Solid organ	333	0.64 (0.58–0.71)	0.21 (0.22)
Lung	116	0.66 (0.56–0.77)	0.21 (0.22)
Breast	58	0.69 (0.58–0.80)	0.11 (0.11)
Liver	37	0.57 (0.34–0.81)	0.16 (0.16)
Hematologic malignancies	23	0.59 (0.26–0.91)	0.16 (0.16)
Kidney	21	0.87 (0.71–1.00)	0.15 (0.24)
Prostate	19	0.56 (0.27–0.84)	0.24 (0.24)
12-month survival prediction			
Overall	356	0.69 (0.64–0.75)	0.22 (0.25)
Solid organ	333	0.68 (0.62–0.74)	0.22 (0.25)
Lung	116	0.69 (0.59–0.79)	0.22 (0.25)
Breast	58	0.59 (0.43–0.75)	0.17 (0.17)
Liver	37	0.73 (0.56–0.91)	0.21 (0.25)
Hematologic malignancies	23	0.67 (0.43–0.91)	0.19 (0.20)
Kidney	21	0.76 (0.53–0.99)	0.19 (0.25)
Prostate	19	0.53 (0.22–0.85)	0.23 (0.24)
18-month survival prediction			
Overall	356	0.68 (0.62–0.75)	0.21 (0.24)
Solid organ	333	0.67 (0.61–0.74)	0.21 (0.24)
Lung	116	0.65 (0.54–0.77)	0.21 (0.23)
Breast	58	0.68 (0.49–0.87)	0.23 (0.24)
Liver	37	0.79 (0.63–0.95)	0.19 (0.24)
Hematologic malignancies	23	0.63 (0.37–0.89)	0.21 (0.21)
Kidney	21	0.70 (0.45–0.96)	0.20 (0.23)
Prostate	19	0.65 (0.36–0.95)	0.20 (0.23)
24-month survival prediction			
Overall	356	0.67 (0.60–0.74)	0.19 (0.22)
Solid organ	333	0.66 (0.59–0.73)	0.20 (0.21)
Lung	116	0.62 (0.49–0.74)	0.20 (0.21)
Breast	58	0.71 (0.52–0.90)	0.23 (0.25)
Liver	37	0.76 (0.56–0.96)	0.17 (0.21)
Hematologic malignancies	23	0.63 (0.39–0.90)	0.21 (0.22)
Kidney	21	0.72 (0.47–0.97)	0.20 (0.23)
Prostate	19	0.75 (0.51–0.99)	0.19 (0.23)

Solid-organ malignancies indicated all kinds of malignancies but excluded hematopoietic malignancies. Some of the c-indices and their 95% confidence intervals were not available as no or only 1 patient died at the time point.
NA = not available.

Table 4. Summary of included studies

Character of studies						Character of patients			
Author, year	State	Study period	Institution /cohort	Proportion of Han Chinese	Type of studY	Sample size	Period (months)	Mortality n (%)	AUC (95% CI)
PATHFx, 2022	Taiwan	2014–2019	NTUH	98%	Validation	356	1	17 (5)	0.71 (0.58–0.84)
							3	63 (18)	0.66 (0.59–0.73)
							6	108 (30)	0.65 (0.59–0.71)
							12	167 (47)	0.69 (0.64–0.75)
							18	191 (54)	0.68 (0.62–0.75)
							24	204 (72)	0.67 (0.60–0.74)
Forsberg, 2011	USA	1999–2003	MSKCC	1.5%	Develop-mental	189	3	60 (32)	0.85 (0.80–0.93)
							12	110 (58)	0.83 (0.77–0.90)
Ashley, 2019 [a]	USA	2012–2016	MDR	1.5%	Validation	192	1	6 (3)	0.82 (0.68–0.95)
							3	35 (18)	0.83 (0.77–0.90)
							6	65 (34)	0.79 (0.73–0.86)
							12	105 (55)	0.79 (0.73–0.86)
							18	129 (67)	0.79 (0.72–0.86)
							24	137 (71)	0.76 (0.69–0.84)
Ashley, 2019 [a]	USA	2016–2018	IBMR	1.5%	Validation	197	1	17 (7)	0.70 (0.58–0.82)
							3	71 (36)	0.77 (0.70–0.84)
							6	102 (52)	0.77 (0.70–0.83)
							12	129 (66)	0.78 (0.71–0.85)
							18	151 (77)	0.79 (0.71–0.86)
							24	160 (81)	0.82 (0.75–0.90)
Forsberg, 2012	Scandi-navia	1999–2009	SSMR	0.40%	Validation	815	3	258 (32)	0.79 (0.76–0.82)
							12	574 (70)	0.76 (0.72–0.80)
Piccioli, 2015	Italy	2010–2013	OORC	0.53%	Validation	287	3	20 (7)	0.80 (0.72–0.88)
							12	106 (37)	0.77 (0.72–0.82)
Ogura, 2017	Japan	2009–2015	NCCH, CIH, UTH, TUH, JUH	0.78%	Validation	261	1	21 (8)	0.77 (0.63–0.86)
							3	43 (17)	0.80 (0.72–0.87)
							6	82 (31)	0.83 (0.77–0.89)
							12	109 (42)	0.80 (0.75–0.86)
Meares, 2019	Australia	2003–2014	RNC, JHH	5.6%	Validation	114	3	38 (33)	0.70 (0.69–0.70)
							6	56 (49)	0.70 (0.69–0.70)
							12	79 (69)	0.71 (0.70–0.71)
							24	95 (83)	0.75 (0.74–0.75)

[a] These 2 cohorts came from the same study.

Abbreviations: NTUH = National Taiwan University Hospital; USA = United States of America; MSKCC = Memorial Sloan-Kettering Cancer Center; MDR = Military Health System Data Repository; IMBR = International Bone Metastasis Registry; SSMR = Scandinavian Skeletal Metastasis Registry; OORC = 13 orthopedic oncology referral centers; NCCH = National Cancer Center Hospital; CIH = Cancer Institute Hospital; UTH = University of Tokyo Hospital; TUH = Teikyo University Hospital; JUH = Juntendo University Hospital; RNC = Royal Newcastle Centre; JHH = John Hunter Hospital; AUC = area under the receiver operating characteristics curve; CI = confidence interval.