Journal of Surgical Oncology

WILEY

RESEARCH ARTICLE OPEN ACCESS

Development of Machine Learning Models for Predicting the 1-Year Risk of Reoperation After Lower Limb Oncological Resection and Endoprosthetic Reconstruction Based on Data From the PARITY Trial

Jiawen Deng[1] | Myron Moskalyk[2] | Matthew Shammas-Toma[1] | Ahmed Aoude[3] | Michelle Ghert[4,5] | Sahir Bhatnagar[6] | Anthony Bozzo[3]

[1]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada | [2]Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada | [3]Division of Orthopaedic Surgery, McGill University, Montréal, Québec, Canada | [4]Division of Orthopaedic Surgery, McMaster University, Hamilton, Ontario, Canada | [5]Department of Orthopaedics, University of Maryland School of Medicine, University of Maryland, Baltimore, Maryland, USA | [6]Department of Epidemiology and Biostatistics, McGill University, Montréal, Québec, Canada

Correspondence: Anthony Bozzo (anthony.bozzo.med@ssss.gouv.qc.ca)

Received: 3 June 2024 | **Revised:** 5 August 2024 | **Accepted:** 18 August 2024

Funding: The authors received no specific funding for this work.

Keywords: endoprosthetic reconstruction | machine learning | malignant bone tumor | prognostic modeling

ABSTRACT

Background: Oncological resection and reconstruction involving the lower extremities commonly lead to reoperations that impact patient outcomes and healthcare resources. This study aimed to develop a machine learning (ML) model to predict this reoperation risk.

Methods: This study was conducted according to TRIPOD + AI. Data from the PARITY trial was used to develop ML models to predict the 1-year reoperation risk following lower extremity oncological resection and reconstruction. Six ML algorithms were tuned and calibrated based on fivefold cross-validation. The best-performing model was identified using classification and calibration metrics.

Results: The polynomial support vector machine (SVM) model was chosen as the best-performing model. During internal validation, the SVM exhibited an AUC-ROC of 0.73 and a Brier score of 0.17. Using an optimal threshold that balances all quadrants of the confusion matrix, the SVM exhibited a sensitivity of 0.45 and a specificity of 0.81. Using a high-sensitivity threshold, the SVM exhibited a sensitivity of 0.68 and a specificity of 0.68. Total operative time was the most important feature for reoperation risk prediction.

Conclusion: The models may facilitate reoperation risk stratification, allowing for better patient counseling and for physicians to implement measures that reduce surgical risks.

1 | Introduction

Surgical management of malignant lower extremity bone tumors and soft tissue sarcomas typically consists of amputation or limb salvage surgery. Limb salvage surgery, which includes oncological resection followed by extensive endoprosthetic reconstruction, has emerged as the preferred approach in 70%–85% of cases [1, 2] due to its superior functional outcomes [3, 4] and similar oncological results [5, 6] compared to amputations. While the success rate of oncological limb salvage surgeries has greatly improved

https://doi.org/10.1002/jso.27854

due to advancements in surgical techniques, neoadjuvant therapies, and endoprosthetic hardware, it is still associated with a high risk of postoperative complications such as surgical site infections and prosthetic failures, which often necessitate revision surgeries [7]. For instance, a previous analysis of pediatric patients undergoing lower extremity endoprosthetic reconstruction for bone tumors demonstrated a 30% reoperation rate in the first year after surgery [8]. These revision surgeries can severely impact patient outcomes and quality of life, increase patients' healthcare expenses, and exacerbate disease burdens placed on healthcare systems [9–11].

Thus, a patient's reoperation risk is an important factor to consider and discuss with patients when planning limb salvage surgeries. Nevertheless, it can be difficult to predict and quantify reoperation risk based on surgeon experience alone. Machine learning (ML) has been shown to be an effective approach to predict postoperative clinical outcomes [12, 13]. Having accurate, ML-powered predictions on reoperation risk can help set realistic patient expectations during preoperative and postoperative counseling. Patients predicted to have a high reoperation risk can also receive tailored treatments aimed at reducing their risk such as prehabilitation programs to improve their functional status [14] the use of silver-coated implants to prevent infection [15, 16], the use of negative pressure wound therapy [17], or alternative surgical options such as amputation [18]. In the postoperative phase, patients' reoperation riskscan be reduced using strategic suction drain placements, vigilant monitoring and early intervention for common postoperative complications such as hematomas, and early mobilization protocols [19–21].

To our knowledge, there are currently no predictive ML models available to provide individualized reoperation risk stratification in the context of lower extremity oncological resection and reconstruction surgeries. This largely stems from the rarity of bone cancers [22] and the accompanying lack of high-quality training and validation data. However, the recent dissemination of results from the Prophylactic Antibiotic Regimens in Tumor Surgery (PARITY) trial [23] provides the high-quality and multicentered data needed to develop such a model. Therefore, the purpose of this study is to develop and internally validate ML models that can provide individualized predictions of 1-year reoperation risk following oncological resection and endoprosthetic reconstruction of the lower extremities.

2 | Materials and Methods

This study was conducted and reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI (TRIPOD + AI) Checklist for Prediction Model Development (Supporting Information S1: Table S1) [24].

2.1 | Data Sources

The PARITY trial [23] is a multicentered, blinded, parallel-group randomized controlled trial that compared a 5-day versus 1-day prophylactic antibiotics regimen on reducing postoperative surgical site infections in patients with a primary bone tumor or soft tissue sarcoma that underwent lower extremity oncologic resection and complex endoprosthetic reconstruction. It was performed across 48 clinical sites in 12 countries from January 1, 2013, to October 29, 2019. Detailed eligibility criteria and study protocol has been described previously [25]. Baseline and 1-year follow up data from all 604 patients analyzed in the trial was included in the current study.

2.2 | Outcome of Interest

The outcome of interest for our predictive models was unplanned additional surgeries within 1-year following the original lower extremity resection/reconstruction surgery. This outcome was recorded as a secondary outcome of the PARITY trial.

2.3 | Feature Selection

Candidate features were first selected based on availability (≤ 30% of missing data) and expert domain (see Supporting Information S1: Table S2 for the list of candidate features). Dimension of the feature set was further reduced using Least Absolute Shrinkage and Selection Operator (LASSO) penalized logistic regression [26] and Boruta [27]. A union of features selected by the two methods was chosen as the final feature set (see Table 1) used for predictive modeling. Given the "one in ten" rule commonly used in predictive modeling, which states that 10 events are needed for every predictor included in the model [28, 29], we aimed to include around 10 features following feature selection.

2.4 | Data Preprocessing

The PARITY data set was randomly split into an 80% derivation and 20% internal validation cohort, stratified by outcome classification. The derivation cohort was further split into different training and testing subsets during cross-validation for model tuning and evaluation. A data preprocessing pipeline was constructed to identify and impute missing data, correct for class imbalance, and improve compatibility with ML algorithms before data is fed into the predictive models (Figure 1). These preprocessing steps were applied after cross-validation splits and after the derivation-validation split to avoid potential data leakage [30].

2.4.1 | Categorical Data Encoding

Ordinal categorical features, including tumor grade and the amount of preoperative chemotherapy received, were encoded as ordinal integer values. Categorical features with no clear ordinality, including ethnicity and tumor type, were transformed via one-hot encoding [31].

2.4.2 | Missing Data Imputation

Amount of missing data for each feature column within the PARITY data set is described as footnotes in Table 1 and

TABLE 1 | Summary of baseline patient features used for ML predictions.

Features	No additional operations ($N = 449$)	Required additional operations ($N = 155$)	Total ($N = 604$)	SMD (95% CI)
Age	41.07 (21.36)	41.60 (23.31)	41.21 (21.86)	0.02 (−0.16 to 0.21)
Ethnicity[a]				
Asian	99 (22.1%)	14 (9.0%)	113 (18.8%)	0.37 (0.18 to 0.55)
Other[b]	348 (77.9%)	141 (91.0%)	489 (81.2%)	—
AJCC stage[c]				0.37 (0.18 to 0.56)
Grade I	56 (14.0%)	7 (4.8%)	63 (11.6%)	
Grade II	89 (22.3%)	25 (17.1%)	114 (20.9%)	
Grade III	175 (43.9%)	79 (54.1%)	254 (46.6%)	
Grade IV	79 (19.8%)	35 (24.0%)	114 (20.9%)	
Tumor type				
Bone tumor	356 (79.3%)	130 (83.9%)	486 (80.5%)	0.12 (−0.06 to 0.30)
Oligometastatic bone disease	49 (10.9%)	7 (4.5%)	56 (9.3%)	0.24 (0.06 to 0.43)
Other[d]	44 (9.8%)	18 (11.6%)	62 (10.3%)	—
Malignant tumor	405 (90.2%)	151 (97.4%)	556 (92.1%)	0.30 (0.12 to 0.49)
Tumor located in femur	373 (83.1%)	125 (80.6%)	498 (82.5%)	0.06 (−0.12 to 0.25)
Tumor located in tibia	78 (17.4%)	31 (20.0%)	109 (18.0%)	0.07 (−0.12 to 0.25)
Preoperative chemotherapy[e]				0.40 (0.21 to 0.58)
None received	233 (52.0%)	81 (52.3%)	314 (52.1%)	
1 cycle	3 (0.7%)	5 (3.2%)	8 (1.3%)	
2 cycles	55 (12.3%)	30 (19.4%)	85 (14.1%)	
3 cycles	46 (10.3%)	14 (9.0%)	60 (10.0%)	
4 cycles	34 (7.6%)	7 (4.5%)	41 (6.8%)	
5 cycles	24 (5.4%)	3 (1.9%)	27 (4.5%)	
6 cycles	26 (5.8%)	8 (5.2%)	34 (5.6%)	
7 cycles	8 (1.8%)	0 (0.0%)	8 (1.3%)	
> 7 cycles	19 (4.2%)	7 (4.5%)	26 (4.3%)	
Total operative time (hours)	4.78 (2.14)	5.97 (2.77)	5.09 (2.37)	0.48 (0.30 to 0.66)
Use of negative pressure wound therapy[f]	49 (10.9%)	34 (22.1%)	83 (13.8%)	0.30 (0.12 to 0.49)
Surgical drain duration (days)[g]	3.49 (2.98)	4.67 (7.12)	3.79 (4.45)	0.22 (0.03 to 0.40)

Abbreviations: AJCC, American Joint Committee on Cancer; CI, confidence interval; SMD, standardized mean difference.

[a] 2 (0.4%) patients have missing ethnicity data in the no additional operations group.

[b] Additional ethnicity categories were reported, but they are omitted here because they were not selected by statistical feature selection methods following one-hot encoding.

[c] 50 (11.1%) patients have missing AJCC stage data in the no additional operations group, and 9 (5.8%) patients have missing AJCC stage data in the required additional operations group.

[d] Additional tumor type categories (i.e., soft tissue sarcoma) were reported, but they are omitted here because they were not selected by statistical feature selection methods following one-hot encoding.

[e] 1 (0.2%) patients have missing preoperative chemotherapy data in the no additional operations group.

[f] 1 (0.6%) patients have missing negative pressure wound therapy data in the required additional operations group.

[g] 4 (0.9%) patients have missing surgical drain duration data in the no additional operations group, and 2 (1.3%) patients have missing surgical drain duration data in the required additional operations group.

Supporting Information S1: Table S3. Five feature columns had missing data requiring imputation. Four of these columns have less than 1% of missing data. Tumor stage has the highest percentage of missing data, with 59 patients (9.8%) requiring imputation.

Because the amount of missing data was low for a majority of the feature columns, and because missing tumor stage can be adequately inferred based on other features such as operation time, tumor type, and the number of adjuvant chemotherapy cycles, we assumed the data to be missing at random. Missing data was imputed using multivariate imputation by chained equations (MICE) over 10 iterations [32]. The training and testing subsets were imputed separately to avoid data leakage [30]. The use of MICE was treated as a hyperparameter when tuning ML algorithms with native missing data handling methods, and we trialed both MICE and the algorithms' native approachesduring model tuning.

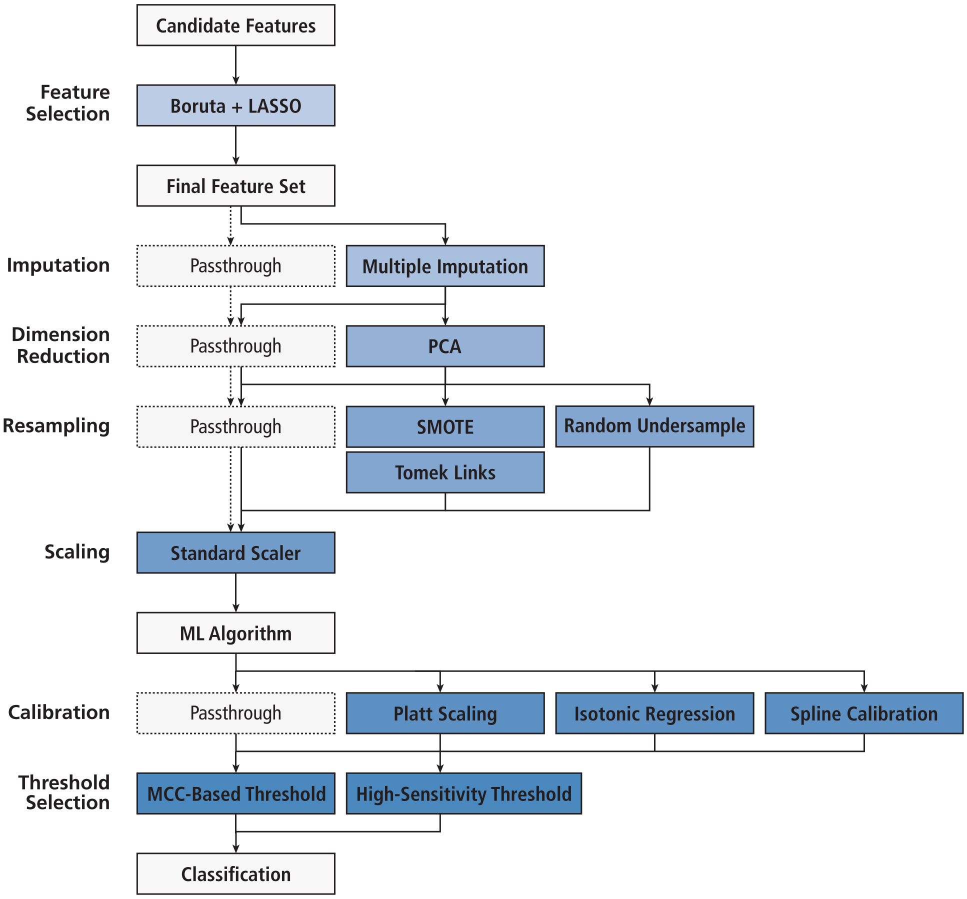

FIGURE 1 | Schematic diagram showing the layout of the ML pipeline used in this study. The arrows show the flow of data within the pipeline during model fitting and predictions. Dimension reduction and resampling techniques cannot be used if there is missing data in the data set, thus they are always skipped if the imputation step is set to passthrough. This is shown by arrows with the dotted line. LASSO, Least Absolute Shrinkage and Selection Operator (Penalized Logistic Regression); MCC, Matthew's correlation coefficient; ML, machine learning; PCA, principal component analysis; SMOTE, Synthetic Minority Oversampling Technique.

2.4.3 | Addressing Multicollinearity

As assessments for potential multicollinearity using variance inflation factors (VIFs) [33] conducted before model development identified potential multicollinearity within features selected via expert domain, we trialed the use of principal component analysis [34] as a hyperparameter in the data preprocessing pipeline.

2.4.4 | Resampling

Given that the data set was imbalanced, resampling was trialed using a combination of Synthetic Minority Oversampling Technique (SMOTE) [35] and Tomek Links [36] as well as random undersampling. ML algorithms' sample weight scaling factors were also tuned to correct for the imbalance during model tuning.

2.4.5 | Scaling

Each feature within the data set was scaled according to its variance, and the data distribution was re-centered around the mean [37].

2.5 | Hyperparameter Tuning

We fitted and tuned six commonly-used ML classification algorithms on the data set to select for the best-performing model: (1) Penalized Logistic Regression, (2) Support Vector Machine (SVM), (3) Random Forest, (4) Light Gradient-Boosting Machines (LightGBM) [38], (5) eXtreme Gradient-Boosting (XGBoost) [39], and (6) Neural Networks.

Each algorithm was tuned to minimize cross-entropy loss across stratified fivefold cross-validation. The optimal hyperparameters were selected using Bayesian Optimization [40]. Bayesian Optimization starts with random parameter searches to gather data points for building a probabilistic model that predicts the performance of different hyperparameter settings. An acquisition function then uses this model to identify the most promising parameters for the next round of evaluations. The results are used to update the performance model, and the process is repeated until a pre-established performance budget is exhausted [41]. Bayesian Optimization is empirically considered to be superior compared to traditional random search techniques, and both are generally more effective than grid-search methods [42]. As random search, which is a worse-performing hyperparameter tuning method, can reliably identify hyperparameters from the top 5% of the most performant combinations with 60 iterations [40], we aimed to perform Bayesian Optimization at or beyond this performance budget. Ultimately, we set the budget for Bayesian Optimization to at least 500 iterations per model (Supporting Information S1: Table S3).

For neural networks, the maximum number of layers in the network was capped at 3. Dropout was used to prevent overfitting [43], and performance-based learning rate decay was trialed to improve network performance.

2.6 | Calibration

After hyperparameter tuning, all models underwent calibration. Three calibration approaches were trialed: (1) Platt/sigmoidal scaling [44, 45], (2) isotonic regression [44] and (3) spline-based calibration [46]. Results from the different calibration approaches (including no additional calibration) were compared, and the calibration approach with the lowest average cross-entropy loss on fivefold cross-validation was selected.

2.7 | Threshold Selection

Following calibration, each model underwent threshold tuning to maximize their classification performance. Matthew's correlation coefficient (MCC) [47] was calculated for every classification threshold between 0.01 and 0.99 at an interval of 0.01, and the threshold with the highest average MCC across fivefold cross-validation was selected as the optimal threshold. This approach creates a threshold that finds a balance between all four quadrants of the confusion matrix.

Given that clinicians may prefer more conservative risk classifications with higher sensitivities during patient counseling or treatment planning, we also created a high-sensitivity threshold. To create the high-sensitivity threshold, we calculated the sensitivity for every classification threshold between 0.01 and 0.99 at an interval of 0.01, and the threshold that produces an average sensitivity that is the nearest and higher than 0.70 during fivefold cross-validation was selected as the high-sensitivity threshold.

2.8 | Model Evaluation

To select the most optimal ML model, we followed a previously published framework proposed for the evaluation of clinical prediction models [48]. The classification performance of each tuned and calibrated model was assessed using area under the receiver operating characteristic curve (AUC-ROC), MCC, sensitivity, and specificity. MCC ranges from −1 to 1 (with higher values indicating better classification performance) and serves as a good representation of all four categories within the confusion matrix (true positives, false negatives, true negatives, and false positives). It is generally considered to be a more informative and robust measure of classification performance compared to raw accuracy or F_1 scores [47]. Calibration performance was assessed using Brier score, calibration curve slope and calibration curve intercept. All metrics were averaged across fivefold cross-validation, and 95% confidence intervals (CIs) were used to assess performance variance between folds.

2.9 | Internal Validation

The best performing model was internally validated using the internal validation cohort based on the same classification and calibration metrics assessed during model evaluation and selection. Feature importance for the best-performing model was assessed using the permutation importance method [49].

2.10 | Statistical Analysis and Software

Continuous baseline patient features used for ML predictions were summarized as mean and standard deviation. Categorical baseline patient features were summarized as frequency and percentages. Distribution of feature data between the control patients and patients who required additional operations was compared using standardized mean differences with corresponding 95% CIs. Descriptive statistics were generated using *arsenal* and *stddiff* in *R*.

LASSO-based feature selection was conducted using *scikit-learn* in Python, and Boruta-based feature selection was conducted using *boruta* in *R*. Data preprocessing was conducted using *scikit-learn* and *imbalanced-learn* in Python. ML models were fitted, tuned, and calibrated using *scikit-learn*, *lightgbm*, *xgboost*, *tensorflow*, *keras*, *scikit-optimize*, and *ml-insights* in Python. Feature importance was assessed using *eli5* in Python.

3 | Results

The total PARITY data set included 604 patients, 449 (74.3%) of whom did not require additional operations and 155 (25.7%) patients who required additional unplanned operations. Patient characteristics, summarized based on the final feature set chosen, are tabulated in Table 1. Data from all trial participants were included. The final feature set included in ML modeling is also tabulated in Supporting Information S1: Table S3.

3.1 | Model Performance on Cross-Validation

Classification performance and calibration metrics of the tuned and calibrated models during fivefold cross-validation are listed in Table 2. Average AUC-ROC ranged from 0.65 to 0.69. Average Brier score was 0.18 for all models. Average calibration curve slope ranged from 0.68 to 1.54 and average calibration curve intercept ranged from −0.12 to 0.07. Using the optimal threshold derived from maximizing MCC during cross-validation, the average MCC ranged from 0.25 to 0.28, the average sensitivity ranged from 0.33 to 0.60, and the average specificity ranged from 0.70 to 0.88. Using the high-sensitivity threshold which aimed to keep sensitivity above 0.70 during cross-validation, the average MCC ranged from 0.14 to 0.23, the average sensitivity ranged from 0.70 to 0.74, and the average

TABLE 2 | Average performance of machine-learning models during fivefold cross-validation.

Model	Threshold-independent metrics				Threshold-dependent metrics (balanced threshold)			Threshold-dependent metrics (high sensitivity threshold)		
	AUC-ROC	Brier score	Calibration curve slope	Calibration curve intercept	Matthew's correlation coefficient	Sensitivity	Specificity	Matthew's correlation coefficient	Sensitivity	Specificity
Penalized Logistic Regression (Ridge)	0.67 (0.59 to 0.75)	0.18 (0.17 to 0.19)	0.68 (0.47 to 0.89)	0.07 (0.04 to 0.09)	0.28 (0.12 to 0.43)	0.60 (0.48 to 0.71)	0.70 (0.61 to 0.79)	0.21 (0.07 to 0.35)	0.72 (0.63 to 0.81)	0.52 (0.41 to 0.63)
Support Vector Machine (Polynomial)	0.68 (0.60 to 0.76)	0.18 (0.16 to 0.19)	0.92 (0.62 to 1.22)	−0.01 (−0.07 to 0.06)	0.27 (0.12 to 0.41)	0.49 (0.37 to 0.62)	0.78 (0.37 to 0.62)	0.22 (0.07 to 0.37)	0.71 (0.61 to 0.81)	0.54 (0.42 to 0.66)
Random Forest	0.69 (0.62 to 0.76)	0.18 (0.17 to 0.19)	1.50 (0.88 to 2.13)	−0.12 (−0.25 to 0.00)	0.27 (0.13 to 0.41)	0.49 (0.40 to 0.58)	0.79 (0.73 to 0.84)	0.23 (0.16 to 0.30)	0.73 (0.66 to 0.81)	0.53 (0.48 to 0.58)
LightGBM	0.65 (0.57 to 0.73)	0.18 (0.17 to 0.19)	1.54 (0.11 to 2.97)	−0.12 (−0.46 to 0.22)	0.27 (0.16 to 0.37)	0.48 (0.41 to 0.56)	0.79 (0.73 to 0.85)	0.14 (0.03 to 0.24)	0.74 (0.63 to 0.85)	0.41 (0.38 to 0.44)
XGBoost	0.66 (0.61 to 0.70)	0.18 (0.17 to 0.19)	1.40 (0.67 to 2.13)	−0.10 (−0.28 to 0.09)	0.25 (0.15 to 0.35)	0.33 (0.27 to 0.39)	0.88 (0.85 to 0.91)	0.19 (0.11 to 0.28)	0.70 (0.64 to 0.76)	0.52 (0.46 to 0.58)
Neural Network	0.66 (0.59 to 0.74)	0.18 (0.17 to 0.20)	0.76 (0.58 to 0.93)	0.06 (0.03 to 0.10)	0.26 (0.14 to 0.38)	0.41 (0.30 to 0.52)	0.83 (0.75 to 0.91)	0.18 (0.07 to 0.28)	0.71 (0.61 to 0.81)	0.49 (0.41 to 0.57)

Note: The performance metrics are tabulated as mean (95% confidence interval).

Abbreviations: AUC-ROC, area under the receiver operating characteristic curve; LightGBM, light gradient-boosting machine; XGBoost, eXtreme gradient-boosting.

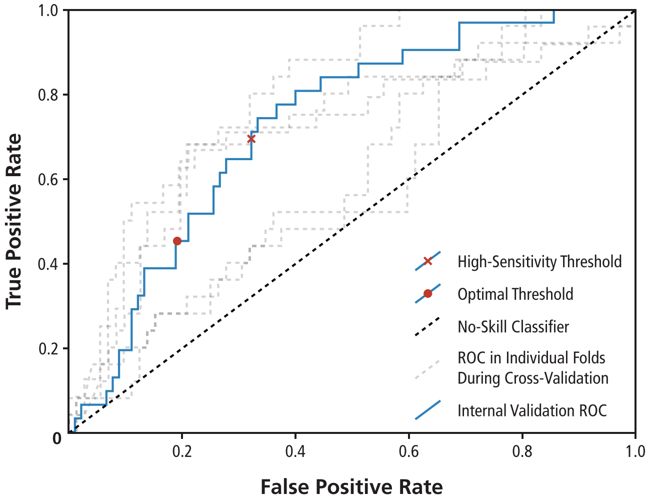

FIGURE 2 | ROC curves for cross-validation and internal validation of the polynomial support vector machine model. The area under the curve was 0.73. ROC, receiver operating characteristic (curve).

specificity ranged from 0.41 to 0.54. The optimal hyperparameter combinations identified during cross-validation for each model are tabulated in Supporting Information S1: Table S4.

Overall, the cross-validation performances of different models using the optimal threshold were very similar. However, both the Random Forest and polynomial SVM model exhibited better classification performances compared to the remaining models. The Random Forest model yielded an average AUC-ROC of 0.69 (95% CI: 0.62 to 0.76), while the polynomial SVM model yielded an average AUC-ROC of 0.68 (95% CI: 0.60 to 0.76). Both models also had the highest MCCs under both the optimal threshold and the high-sensitivity threshold, and both models exhibited similar sensitivity and specificity under the optimal threshold and the high-sensitivity threshold (see Table 2). The polynomial SVM was designated as the best performing model overall because it exhibited better calibration metrics during cross-validation as measured by the average calibration curve slope and intercept compared to the Random Forest model (calibration curve slope, 0.92 [95% CI: 0.62 to 1.22] vs. 1.50 [95% CI: 0.88 to 2.13]; calibration curve intercept, −0.01 [95% CI: −0.07 to 0.06] vs. −0.12 [95% CI: −0.25 to 0.00]).

3.2 | Model Performance During Internal Validation

On internal validation with the validation cohort, the polynomial SVM model exhibited an AUC-ROC of 0.73 (Figure 2), Brier score of 0.17, calibration slope of 0.75, and calibration intercept of 0.04 (the calibration plot is appended as Supporting Information S1: Figure S1). Using the optimal threshold identified during cross-validation, the MCC was 0.26, the sensitivity was 0.45, and the specificity was 0.81. Using the high-sensitivity threshold identified during cross-validation, the MCC was 0.31, the sensitivity was 0.68, and the specificity was 0.68. The performance metrics of the SVM model on the internal validation cohort are similar if not better than metrics observed during cross-validation, with the exception of a worse calibration slope observed during internal validation (Figure 3).

3.3 | Feature Importance and Model Deployment

The polynomial SVM model pipeline was re-trained on the entire PARITY data set and incorporated into an online reoperation risk calculator. The feature importance of the production model is shown in Figure 4. Total operative time was the most important feature. Other ML algorithms were also retrained and incorporated into the calculator for demonstration purposes. The calculator is available at https://parity-ml.shinyapps.io/reop-estimator/.

4 | Discussions

Need for additional unplanned operations is a common disposition for patients undergoing oncological resection and endoprosthetic reconstruction of the lower limbs [50]. In this study, we developed ML models to predict the risk of reoperations in this patient population based on easily accessible patient characteristics and operative parameters. The developed models have the potential to enable early postoperative risk stratification, which allows patients at higher risks for reoperation to receive prophylactic measures and to be followed more closely. Orthopedic surgeons can also test the impact of changing potential surgical parameters in the ML model to help inform treatment decisions and provide their patients with individualized reoperation risk predictions.

An important consideration for the clinical deployment of our ML model is the selection of decision thresholds. In our study, we produced both an optimal threshold aimed at maximizing the model's MCC metrics as well as a high-sensitivity threshold aimed at increasing the model's true positive rate. The use of decision thresholds will always represent a trade-off between sensitivity and specificity [51]. In many clinical scenarios, such as deciding when to implement vigilant monitoring and early postoperative ambulation protocols or providing surgical counseling to patients, the high-sensitivity threshold may be preferred to ensure that high-risk patients receive proper prophylactic measures and conservative prognoses. In scenarios where the model is used to make drastic changes to surgical planning, such as converting a limb-salvage surgery into an amputation, the optimal threshold may be preferred to ensure that both specificity and sensitivity are accounted for. Clinicians should tailor the threshold selection to the specific clinical context and balance the need for accuracy with the potential consequences of misclassifying false positive and negative cases.

This study has several notable strengths. First, the PARITY data set is relatively complete with low amounts of missing data. Combined with multiple imputation techniques, we were able to train the models on credible data that do not rely heavily on missingness assumptions [52]. The data set also originated from an international, multicenter trial, which enables the models to be generalizable across most, if not all, oncological patients undergoing lower extremity resections with endoprosthetic reconstructions. Secondly, we tested a wide range of ML algorithms, data preprocessing techniques, and calibration methods. We also used Bayesian Optimization with high numbers of iterations for hyper parameter tuning, which allowed us to

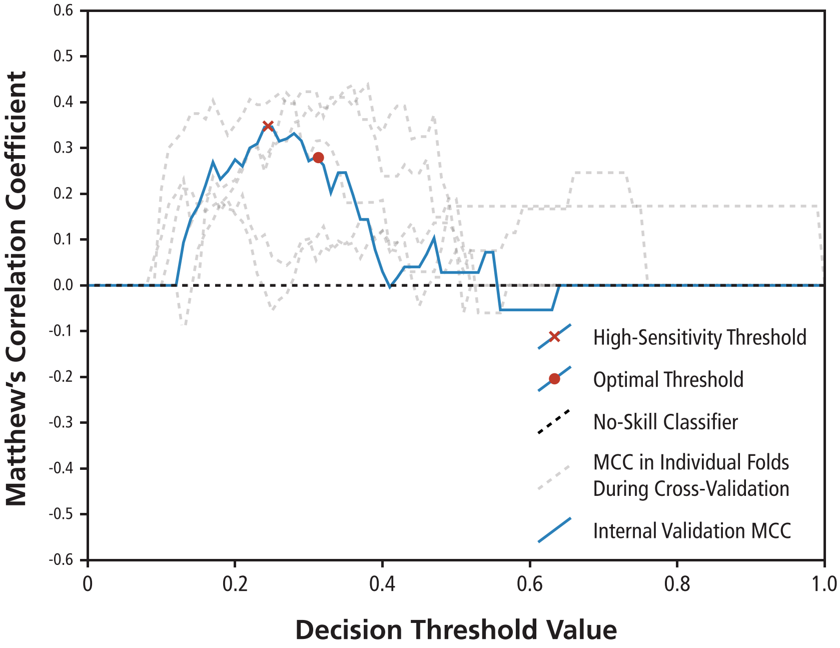

FIGURE 3 | Line plot showing the MCC of the polynomial support vector machine model at every possible decision threshold values for cross-validation and internal validation. MCC, Matthew's correlation coefficient.

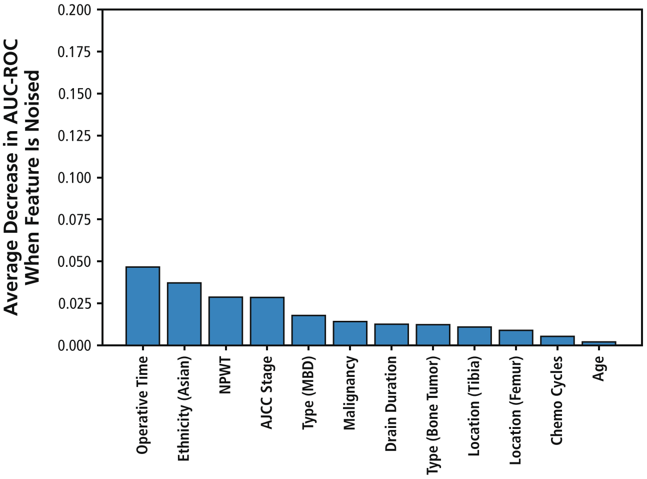

FIGURE 4 | Feature importance plot for the finalized polynomial support vector machine model, showing the change in the model's AUC-ROC when individual features are replaced with noise. AJCC, American Joint Committee on Cancer; AUC-ROC, area under the receiver operating characteristic curve; NPWT, negative-pressure wound therapy; MBD, metastatic bone disease.

identify the most optimal ML model configurations for our data set. Lastly, we employed fivefold cross-validation during our hyperparameter tuning process, which reduces the potential for the models to overfit [53]. Based on our assessment of feature importance, the features prioritized by the ML models generally correspond with previous investigations using more traditional statistical approaches. For instance, operative time was the most important predictor of reoperation risk in both our best performing ML model and also in previous PARITY publications using univariate and multivariate analyses [54]. Conversely, while tumor type was shown to be a top three predictor of reoperations risk in the previous multivariate analysis, our SVM model placed one-hot encoded tumor type features behind other features that were considered to be less important by previous analyses such as ethnicity and tumor stage [54]. Prior work has indeed shown that variable importance can differ when both regression and ML techniques, or both linear and nonlinear models, are used to analyze the same data set. Our findings highlight the value of using different methods to fully understand which factors truly influence clinical outcomes [55].

A key limitation of the current study is that our feature selection process relied on data availability in the PARITY data set. With this approach, our model may not have captured all the relevant information needed to predict reoperation risks accurately, such as surgeon experience or surgeon age. Another limitation of the study is its small sample size due to the rarity of relevant patient cases. Soft tissue sarcomas, for instance, account for only 1% of all adult malignancies [56]. Osteosarcoma, which is the most common type of primary bone tumor, has an incidence rate of 4–5 per year per million persons [57]. The PARITY trial generated one of the largest and most comprehensive prospective data set on malignant bone tumors and soft tissue sarcomas [58]; however, our sample size was still lower than the 50+ events per feature threshold needed to obtain optimal predictions from modern ML algorithms and deep-learning neural networks, as demonstrated in previous simulation studies [59].

A potential solution to this study's limitations is incorporating multimodal data, which is readily available in the preoperative setting. For instance, magnetic resonance imaging (MRI) of the tumor is routinely acquired in this patient population [60]. Incorporating imaging data could potentially allow ML models to identify additional features associated with reoperation, such as the presence of critical neurovascular structures or extensive perilesional edema. These may entail close or positive margins, which are known to be associated with higher local recurrence rates and reoperation rates [61]. Similarly, contextual information about the tissues near the tumor, the predicted dead space following the resection, and other factors related to infection risk, may be extracted from the MRI by multimodal models. Similar multimodal approaches in other clinical investigations have yielded increased ML model performance. For instance, a prior systematic review found that multimodal ML models generally yielded higher accuracy and AUC-ROC metrics compared to single modality models when applied to clinical diagnostic and prognostic tasks [62].

5 | Conclusion

Predicting postsurgical outcomes in orthopedic oncology is complex. This current study serves as a proof-of-concept for using ML models to predict risk of reoperation in patients with lower extremity bone tumors and soft tissue sarcomas that require oncological resection and endoprosthetic reconstruction. External validation and incorporation of multimodal data will improve the generalizability and accuracy of the predictions.

Acknowledgments

We would like to thank The Prophylactic Antibiotic Regimens in Tumor Surgery (PARITY) investigators for providing the anonymized trial data needed for model development.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Supporting data is not directly available from the authors of this study. To request access to deidentified PARITY trial data, please contact the corresponding author(s) of the original trial publication [23]. Statistical codes used in this study are available upon reasonable request if the requesting individual has access to the PARITY data set.

Ethics Statement

This work is exempt from ethics review under The Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2) as it only involves secondary analysis of anonymized data.

References

1. R. Veth, R. van Hoesel, M. Pruszczynski, J. Hoogenhout, B. Schreuder, and T. Wobbes, “Limb Salvage in Musculoskeletal Oncology,” *Lancet Oncology* 4 (2003): 343–350.

2. G. E. Mason, L. Aung, S. Gall, et al., “Quality of Life Following Amputation or Limb Preservation in Patients With Lower Extremity Bone Sarcoma,” *Frontiers in Oncology* 3 (2013): 210.

3. B. T. Rougraff, M. A. Simon, J. S. Kneisl, D. B. Greenberg, and H. J. Mankin, “Limb Salvage Compared With Amputation for Osteosarcoma of the Distal End of the Femur. A Long-Term Oncological, Functional, and Quality-of-Life Study,” *Journal of Bone and Joint Surgery. American volume* 76 (1994): 649–656.

4. A. M. Davis, M. Devlin, A. M. Griffin, J. S. Wunder, and R. S. Bell, “Functional Outcome in Amputation Versus Limb Sparing of Patients With Lower Extremity Sarcoma: A Matched Case-Control Study,” *Archives of Physical Medicine and Rehabilitation* 80 (1999): 615–618.

5. M. A. Ghert, A. Abudu, N. Driver, et al., “The Indications for and the Prognostic Significance of Amputation as the Primary Surgical Procedure for Localized Soft Tissue Sarcoma of the Extremity,” *Annals of Surgical Oncology* 12 (2005): 10–17.

6. R. Niimi, A. Matsumine, K. Kusuzaki, et al., “Usefulness of Limb Salvage Surgery for Bone and Soft Tissue Sarcomas of the Distal Lower Leg,” *Journal of Cancer Research and Clinical Oncology* 134 (2008): 1087–1095.

7. E. R. Henderson, J. S. Groundland, E. Pala, et al., “Failure Mode Classification for Tumor Endoprostheses: Retrospective Review of Five Institutions and a Literature Review,” *Journal of Bone and Joint Surgery. American Volume* 93 (2011): 418–429.

8. A. Bozzo, C. M. Yeung, M. Van De Sande, M. Ghert, J. H. Healey, and I. Parity, “Operative Treatment and Outcomes of Pediatric Patients with an Extremity Bone Tumor: A Secondary Analysis of the PARITY Trial Data,” *Journal of Bone and Joint Surgery. American Volume* 105 (2023): 65–72.

9. Z. Butte, K. Tanaka, V. Andaya, et al., “Risk of Endoprosthetic Infection and Impact of Health-Related Quality of Life in Patients With Osteosarcoma and Giant Cell Tumor of Bone; A Retrospective Case-Control Study,” *Annals of Joint* 5 (2020): 27.

10. M. Allami, A. Yavari, A. Karimi, M. Masoumi, M. Soroush, and E. Faraji, “Health-Related Quality of Life and the Ability to Perform Activities of Daily Living: A Cross-Sectional Study on 1079 War Veterans With Ankle-Foot Disorders,” *Military Medical Research* 4 (2017): 37.

11. J. Woodfield, P. Deo, A. Davidson, T. Y. Chen, and A. van Rij, “Patient Reporting of Complications After Surgery: What Impact Does Documenting Postoperative Problems From the Perspective of the Patient Using Telephone Interview and Postal Questionnaires Have on the Identification of Complications after Surgery?” *BMJ Open* 9 (2019): e028561.

12. T. Davenport and R. Kalakota, “The Potential for Artificial Intelligence in Healthcare,” *Future Healthcare Journal* 6 (2019): 94–98.

13. F. Shamout, T. Zhu, and D. A. Clifton, “Machine Learning for Clinical Outcome Prediction,” *IEEE Reviews in Biomedical Engineering* 14 (2021): 116–126.

14. C. E. Guerra-Londono, J. P. Cata, K. Nowak, and V. Gottumukkala, “Prehabilitation in Adults Undergoing Cancer Surgery: A Comprehensive Review on Rationale, Methodology, and Measures of Effectiveness,” *Current Oncology* 31 (2024): 2185–2200.

15. J. Hardes, M. P. Henrichs, G. Hauschild, M. Nottrott, W. Guder, and A. Streitbuerger, “Silver-Coated Megaprosthesis of the Proximal Tibia in Patients With Sarcoma,” *Journal of Arthroplasty* 32 (2017): 2208–2213.

16. A. Streitbuerger, M. P. Henrichs, G. Hauschild, M. Nottrott, W. Guder, and J. Hardes, “Silver-Coated Megaprostheses in the Proximal Femur in Patients with Sarcoma,” *European journal of orthopaedic surgery & traumatology: orthopedie traumatologie* 29 (2019): 79–85.

17. C. Gusho, R. Phillips, J. Cook, and A. Evenski, “A Systematic Review and Meta-Analysis of Negative Wound Pressure Therapy Use in Soft Tissue Sarcoma Resection,” *Iowa orthopaedic journal* 43 (2023): 52–59.

18. M. Kirilova, A. Klein, L. H. Lindner, et al., “Amputation for Extremity Sarcoma: Indications and Outcomes,” *Cancers* 13 (2021): 5125, https://doi.org/10.3390/cancers13205125.

19. C. Radtke, M. Panzica, K. Dastagir, C. Krettek, and P. M. Vogt, “Soft Tissue Coverage of the Lower Limb Following Oncological Surgery,” *Frontiers in oncology* 5 (2015): 303.

20. A. Puri, “Limb Salvage in Musculoskeletal Oncology: Recent Advances,” *Indian Journal of Plastic Surgery: Official Publication of the Association of Plastic Surgeons of India* 47 (2014): 175–184.

21. A. Shehadeh, M. El Dahleh, A. Salem, et al., “Standardization of Rehabilitation After Limb Salvage Surgery for Sarcomas Improves

Patients' Outcome," *Hematology/Oncology and Stem Cell Therapy* 6 (2013): 105–111.

22. Y. Xu, F. Shi, Y. Zhang, et al., "Twenty-Year Outcome of Prevalence, Incidence, Mortality and Survival Rate in Patients With Malignant Bone Tumors," *International Journal of Cancer* 154 (2024): 226–240.

23. Prophylactic Antibiotic Regimens in Tumor Surgery (PARITY) Investigators, M. Ghert, P. Schneider, et al., "Comparison of Prophylactic Intravenous Antibiotic Regimens After Endoprosthetic Reconstruction for Lower Extremity Bone Tumors: A Randomized Clinical Trial," *JAMA Oncology* 8 (2022): 345–353.

24. G. S. Collins, K. G. M. Moons, P. Dhiman, et al., "Tripod+Ai Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods," *BMJ (Clinical Research Ed.)* 385 (2024): e078378.

25. M. Ghert, B. Deheshi, G. Holt, et al., "Prophylactic Antibiotic Regimens in Tumour Surgery (Parity): Protocol for a Multicentre Randomised Controlled Study," *BMJ Open* 2 (2012): e002197, https://doi.org/10.1136/bmjopen-2012-002197.

26. R. Muthukrishnan, and R. Rohini. "LASSO: A Feature Selection Technique in Predictive Modeling for Machine Learning," in *IEEE International Conference on Advances in Computer Applications (ICACA)*, IEEE, 2016.

27. M. B. Kursa and W. R. Rudnicki, "Feature Selection With the Boruta Package," *Journal of Statistical Software* 36 (2010): 1–13, https://doi.org/10.18637/jss.v036.i11.

28. F. E. Harrell, Jr., K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati, "Regression Modelling Strategies for Improved Prognostic Prediction," *Statistics in Medicine* 3 (1984): 143–152.

29. Jr Harrell FE, Jr., K. L. Lee, and D. B. Mark, "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Statistics in Medicine* 15 (1996): 361–387.

30. A. Apicella, F. Isgrò, and R. Prevete. "Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning." Published ahead of print, February 21, 2024, https://doi.org/10.2139/ssrn.4733889.

31. S. K. Ashenden, A. Bartosik, P.-M. Agapow, et al., "Chapter 2—Introduction to Artificial Intelligence and Machine Learning," in *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, eds. S. K. Ashenden. (Academic Press, 2021), 15–26.

32. S. van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software* 45 (2011): 1–67.

33. G. James, D. Witten, T. Hastie, et al., *An Introduction to Statistical Learning: With Applications in R* (Springer Science & Business Media, 2013).

34. I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374 (2016): 20150202.

35. N. V. Chawla, K. W. Bowyer, L. O. Hall, et al., "Smote: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research* 16 (2002): 321–357.

36. I. Tomek, "Two Modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics* 6 (1976): 769–772.

37. L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The Choice of Scaling Technique Matters for Classification Performance," *Applied Soft Computing* 133 (2023): 109924.

38. G. Ke, Q. Meng and T. Finley, "LightGBM: A highly efficient gradient boosting decision tree," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, eds. U. von Luxburg, I. Guyon and S. Bengio, et al. Curran Associates Inc, 2017), 3149–3157.

39. T. Chen, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016).

40. D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization* 13 (1998): 455–492.

41. J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems 25.*

42. R. Turner, D. Eriksson, M. McCourt, et al., "Bayesian Optimization is Superior to Random Search For Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020," in NeurIPS 2020 Competition and Demonstration Track, (PMLR), 3–26.

43. N. Srivastava, G. Hinton, A. Krizhevsky, et al., "Dropout: A Simple Way to Prevent Neural Networks From Overfitting," *Journal of Machine Learning Research* 15 (2014): 1929–1958.

44. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning—ICML'05.* (ACM Press, 2005).

45. J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers* 10 (1999): 61–74.

46. B. Lucena, "Spline-Based Probability Calibration," Published ahead of print, September 20, 2018, https://doi.org/10.48550/arXiv.1809.07751.

47. D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation," *BMC Genomics* 21 (2020): 6.

48. E. W. Steyerberg, A. J. Vickers, N. R. Cook, et al., "Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures," *Epidemiology* 21 (2010): 128–138.

49. L. Breiman, "Random Forests," *Machine Learning* 45 (2001): 5–32.

50. T. Morii, H. Yabe, H. Morioka, et al., "Postoperative Deep Infection in Tumor Endoprosthesis Reconstruction Around the Knee," *Journal of Orthopaedic Science: Official Journal of the Japanese Orthopaedic Association* 15 (2010): 331–339.

51. J. Chubak, G. Pocobelli, and N. S. Weiss, "Tradeoffs between Accuracy Measures for Electronic Health Care Data Algorithms," *Journal of Clinical Epidemiology* 65 (2012): 343–349.

52. K. J. Lee, J. B. Carlin, J. A. Simpson, and M. Moreno-Betancur, "Assumptions and Analysis Planning in Studies With Missing Data in Multiple Variables: Moving Beyond the Mcar/Mar/Mnar Classification," *International Journal of Epidemiology* 52 (2023): 1268–1275.

53. S. S. Al-Zaiti, A. A. Alghwiri, X. Hu, et al., "A Clinician's Guide to Understanding and Critically Appraising Machine Learning Studies: A Checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML)," *European Heart Journal. Digital Health* 3 (2022): 125–140.

54. J. K. Kendal, D. Slawaska-Eng, A. Gazendam, et al., "Risk Factors for All-Cause Early Reoperation Following Tumor Resection and Endoprosthetic Reconstruction: A Secondary Analysis from the PARITY Trial," *The Journal of Bone and Joint Surgery. American Volume* 105 (2023): 4–9.

55. M. Saarela and S. Jauhiainen, "Comparison of Feature Importance Measures As Explanations for Classification Models," *SN Applied Sciences* 3 (2021): 272.

56. A. C. Gamboa, A. Gronchi, and K. Cardona, "Soft-Tissue Sarcoma in Adults: An Update on the Current State of Histiotype-Specific

Management in an Era of Personalized Medicine," *CA: A Cancer Journal for Clinicians* 70 (2020): 200–229.

57. G. Ottaviani and N. Jaffe, "The Epidemiology of Osteosarcoma," *Cancer Treatment and Research* 152 (2009): 3–13.

58. M. Ghert, "Selected Secondary Analyses From the PARITY Trial: Introduction," *Journal of Bone and Joint Surgery. American Volume* 105 (2023): 2–3.

59. T. van der Ploeg, P. C. Austin, and E. W. Steyerberg, "Modern Modelling Techniques Are Data Hungry: A Simulation Study for Predicting Dichotomous Endpoints," *BMC Medical Research Methodology* 14 (2014): 137.

60. C. H. Lohmann, S. Rampal, M. Lohrengel, and G. Singh, "Imaging in Peri-Prosthetic Assessment: An Orthopaedic Perspective," *EFORT Open Reviews* 2 (2017): 117–125.

61. P. W. O'Donnell, A. M. Griffin, W. C. Eward, et al., "The Effect of the Setting of a Positive Surgical Margin in Soft Tissue Sarcoma," *Cancer* 120 (2014): 2866–2875.

62. S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines," *NPJ Digital Medicine* 3 (2020): 136.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.