

Optimal Machine Learning Models for Developing Prognostic Predictions in Patients With Advanced Cancer

Review began 11/21/2024

Review ended 12/14/2024

Published 12/22/2024

© Copyright 2024

Hamano et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.76227

Jun Hamano ¹, Ayano Takeuchi ², Tomoya Keyaki ³, Hidemasa Nose ⁴, Kenichi Hayashi ⁵

¹. Palliative and Supportive Care, University of Tsukuba, Tsukuba, JPN ². Science and Engineering, Chuo University, Tokyo, JPN ³. Mathematics, Graduate School of Science and Technology, Keio University, Yokohama, JPN ⁴. Science and Engineering, Graduate School of Science and Engineering, Chuo University, Tokyo, JPN ⁵. Mathematics, Keio University, Yokohama, JPN

Corresponding author: Jun Hamano, junhamano@md.tsukuba.ac.jp

Abstract

Context: Accurate prognosis prediction for cancer patients in palliative care is critical for clinical decision-making and personalized care. Traditional statistical models have been complemented by machine learning approaches; however, their comparative effectiveness remains underexplored.

Objectives: To assess the prognostic accuracy of statistical and machine learning models in predicting 30-day survival in patients with advanced cancer using objective data, such as the result of the blood test.

Methods: A secondary analysis of the Japan-Prognostic Assessment Tools Validation (J-ProVal) study was performed from September 2012 to April 2014. We used data from 58 palliative care services in Japan and enrolled 915 patients. Four models, fractional polynomial (FP) regression, Kernel Fisher discriminant analysis (KFDA), Kernel support vector machine (KSVM), and XGBoost, were compared using 17 objective clinical characteristics. Models were evaluated with the area under the receiver operating characteristic curve (AUC) as the primary metric.

Results: The KSVM model demonstrated the highest predictive accuracy (AUC: 0.834), outperforming the FP model (AUC: 0.799). XGBoost showed comparatively lower performance; however, it was likely limited by the size of the dataset.

Conclusions: Machine learning, particularly KSVM, has high predictive accuracy in palliative care when sufficient data are available. However, our findings suggest that traditional statistical models offer advantages in stability and interpretability, underscoring the importance of tailored model selection based on data characteristics.

Categories: Palliative Care

Keywords: advanced cancer patients, machine learning models, palliative and end-of-life care, prognostic prediction, traditional statistical models

Introduction

Accurate prognosis prediction is essential for informed clinical decision-making, resource allocation, and patient-centered treatment options and goals of care discussion [1-4]. However, predicting the prognosis of cancer patients is challenging due to a complex interplay of various factors. Physician-based prognostication is often inaccurate, with a tendency to overestimate survival [5].

Prognostic models for cancer patients in palliative care have evolved over the years. Since the 1990s, subjective data, such as physical symptoms, have been incorporated into these models [6]. Recently, there has been a shift toward using objective data, including laboratory test results. Previous studies reported that objective data-based models offer superior performance in palliative care settings [7,8].

Fractional polynomials (FP) modeling is a method that is recommended by the Guidelines for Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis [9]. Our previous study reported that FP modeling had better accuracy than existing prediction models using only objective data in palliative care [7].

Traditional statistical methods, such as FP, have been the primary approach for developing objective data-based prognostic models in palliative care. However, machine learning techniques for developing prediction models have emerged as a promising alternative method in recent years [10,11]. The FP model is an extension of traditional regression in statistics that expresses the relationships between the outcome and the patient characteristics. In this study, we apply three machine learning techniques. Kernel Fisher

How to cite this article

Hamano J, Takeuchi A, Keyaki T, et al. (December 22, 2024) Optimal Machine Learning Models for Developing Prognostic Predictions in Patients With Advanced Cancer. Cureus 16(12): e76227. DOI 10.7759/cureus.76227

discriminant analysis (KFDA) and Kernel support vector machines (KSVM) are methods that use Kernel techniques to handle complex, nonlinear patterns in data. XGBoost is a tree-based method that combines multiple decision trees to improve prediction accuracy, offering features like nonlinear prediction and efficient computation. Statistical models enable us to understand relationships within data, emphasizing the evaluation of statistical parameters and their inference by confidence intervals and hypothesis testing. They prioritize interpretability, allowing for clear insights into the influence of each variable. In contrast, machine learning techniques aim to learn patterns from data to enhance prediction and classification accuracy. They often prioritize predictive performance over interpretability, seeking to generalize to future data.

Although there is a burgeoning interest in using machine learning for prognostication, the comparative performance of statistical and machine learning models in palliative care remains to be elucidated. We aimed to compare the prognostic accuracy of statistical and machine-learning models in predicting the survival of cancer patients using objective data in palliative care.

Materials And Methods

We performed a secondary analysis of the Japan-prognostic assessment tools validation (J-ProVal) study. The J-ProVal was a multicenter prospective cohort study that was performed to investigate the feasibility and accuracy of existing prognostic tools. The methodology has been reported in detail elsewhere [12-14].

We compared the prediction performances of various statistical methods or machine learning methods in predicting the prognosis of patients with advanced cancer. Data were obtained from 58 palliative care services in Japan from September 2012 through April 2014. The participating services included 19 hospital palliative care teams, 16 palliative care units, and 23 home-based palliative care services. The primary physician of each patient performed an evaluation and recorded the demographic and clinical characteristics of the participants. Data from laboratory tests were only obtained if the patient underwent blood tests as a clinical necessity within one week after enrolment.

The J-ProVal study was conducted in accordance with the ethical standards of the Declaration of Helsinki and the ethical guidelines for research presented by the Ministry of Health, Labour, and Welfare of Japan. The Institutional Review Boards of all participating services approved this study and the use of the existing J-ProVal data for secondary analysis. The authors did not have access to any information that could identify individual participants during or after data collection.

Patients

Eligible patients were enrolled consecutively as they were referred to the participating services throughout the study period. All services were asked to evaluate and collect data on a specific number of patients, such as 20, 40, 60, 80, or 100, based on the size of the service. Patients were eligible for the study if they were adults (aged 20 years or older); had locally advanced or metastatic cancer (including hematopoietic neoplasms); and had been admitted to a palliative care unit, referred to a hospital palliative care team, or were receiving home-based palliative care.

Measurements and variables

Survival time was defined as the period from the day of referral to the date of death. If patients survived for <6 months after enrollment, survival time was defined as 180 days and they were censored at that time.

The physician assessed variables on the day of admission, including the patient's age, sex, site of the primary tumor and metastases, anticancer treatment during the one-month period before assessment (such as chemotherapy, hormone therapy, or radiotherapy), symptoms and general condition, vital signs (heart rate and respiratory rate), and the results of blood tests performed during the one-week period before assessment.

We also recorded the demographic and clinical characteristics of the participants, including the age, gender, site of primary cancer, metastatic site, Eastern Cooperative Oncology Group performance status, and anticancer therapy (including chemotherapy, hormone therapy, and radiotherapy).

Statistical analysis

We collected the following patient characteristics: heart rate, respiratory rate, leukocyte count ($10^9/L$), neutrophil count ($10^9/L$), lymphocyte count ($10^9/L$), platelet count ($10^9/L$), urea (mg/dL), creatinine (mg/dL), alanine transaminase (U/L), alkaline phosphatase (U/L), total bilirubin (mg/dL), lactate dehydrogenase (U/L), albumin (g/dL), and C-reactive protein (mg/dL). Using patient characteristics as features, we predicted whether patients would survive 30 days after enrollment.

We also included additional features calculated from these 14 patient characteristics: neutrophil to lymphocyte ratio [15], prognostic nutritional index (prognostic indicator derived from serum albumin levels

and total lymphocyte count) [16], and platelet to lymphocyte ratio [17]. Thus, we used a total of 17 features in the analysis, which is the same as the analyses conducted in our previous study [7]. Four models were used for prediction: FP regression [18], KFDA[19], KSVM[20], and XGBoost [21].

FP model

In this study, we apply the logistic regression model as a statistical method. The logistic regression model expresses (the logarithms of) the odds of the outcomes y by a weighted sum such as $\beta_0 + \beta_1 x_1 + \dots + \beta_d x_d$, where x_j s are patient characteristics (explanatory variables). FP model is an extension of the logistic regression model by using the exponent of explanatory variables, hence the odds are expressed by $\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_d x_d)$. In our analysis, we use the same value of p_j s in Hamano et al. (2018). This model is called a linear model even when x_j^{-1} or x_j^2 is used in the context of statistical models because the outcome is expressed by the sum of them (linear combination). While this model is less flexible than the machine learning techniques that will be described below, it has advantages in the sense that we can understand the influence of each variable x_j to y by the magnitude of β_j , evaluate its stability by the confidence intervals, and test its significance.

KFDA and KSVM

KFDA and KSVM are commonly used discriminative models. KFDA discriminates classes of observations with a d -dimensional feature vector $\mathbf{x}$ using the function $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$, where $\mathbf{a}$ is the d -dimensional coefficient vector. In KFDA, $\mathbf{a}$ is determined by maximizing the ratio of the between-group variance to the within-group variance [19].

KSVM is a method that discriminates sets by maximizing the distance (margin) between two parallel hyperplanes $g(\mathbf{x}) = 1$ and $g(\mathbf{x}) = -1$ that separate the two classes of data. KSVM discriminates classes of observations with the d -dimensional feature vector $\mathbf{x}$ using the function $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} - b$, where $\mathbf{w}$ is a d -dimensional coefficient vector and b is a constant. In KSVM, sets are discriminated such that $y \cdot g(\mathbf{x}) \geq 1$ for a feature $\mathbf{x}$ and an objective function $y \in \{-1, 1\}$. However, this condition only applies in linearly discriminative cases; thus, the constraints are relaxed and set $y \cdot g(\mathbf{x}) \geq 1 - \xi$, where ξ is the slack variable and represents the degree of violation between two hyperplanes $g(\mathbf{x}) = 1$ and $g(\mathbf{x}) = -1$. In KSVM, the sum of slack variables increases as the margin increases. KSVM discriminates sets by increasing the margin while keeping the sum of slack variables as small as possible [19].

In practice, linear discriminant methods, such as KFDA and KSVM, sometimes fail to discriminate data. In such cases, discriminant performance can be improved by applying nonlinear transformations to the data. For example, Figure 1 below shows data that were difficult to discriminate linearly, but by applying the transformation $(x_1, x_2) \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$, linear discrimination became possible (Figure 2).

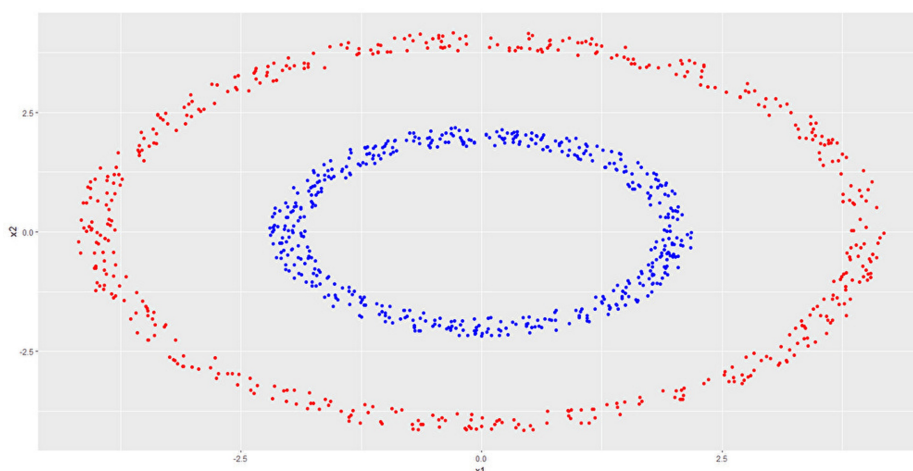


FIGURE 1: Example of nonlinear transformation (colors indicate groups)

Data that cannot be discriminated linearly

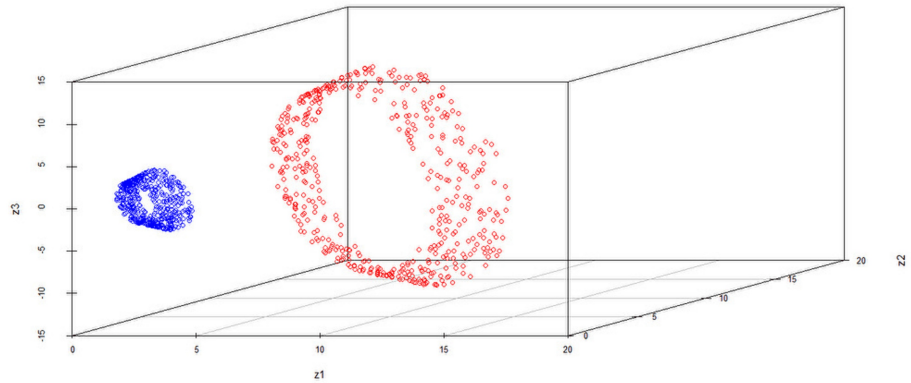


FIGURE 2: Example of nonlinear transformation (colors indicate groups)

Data after transformation to allow for linear discrimination

The kernel method is an efficient method for computing the inner product of data with such nonlinear transformations. Using the kernel method, the inner product of the transformed values $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ by a nonlinear function ϕ is represented by the kernel function $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. In many analytical methods, it is not necessary to explicitly express the nonlinear transformation; rather, it is sufficient to perform an inner product calculation [22]. This technique, known as the kernel trick, achieves calculations in complex nonlinear spaces. Thus, KFDA and KSVM are methods that apply kernel methods to KFDA and KSVM, respectively [19,20]. Both of these methods use the kernel method for flexible nonlinear discrimination. On the other hand, they differ in terms of dimensionality compression, robustness against outliers, and computational speed. KFDA performs dimensionality reduction to avoid the "curse of dimensionality," in which accuracy is reduced when the number of features is large. Compared to KFDA, KSVM is less affected by outliers and can reduce the number of calculations by using only part of the data [20,23]. For the purpose of the study, a kernel function called Gaussian kernel $(\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp \{-||\mathbf{x}_i - \mathbf{x}_j||^2 / (2\sigma^2)\}, \sigma > 0)$ was used.

XGBoost

One of the methods for discrimination is based on decision trees. A decision tree is a method that repeatedly divides the data according to the conditions related to the feature values and discriminates the observed values by the assigned leaf nodes. While decision trees are easy to interpret, they often exhibit poor generalization ability for unknown data [10]. One solution to this issue is ensemble learning. Among the various ensemble techniques, boosting is a method that improves predictive performance by sequentially adding weak classifiers, thereby reducing the discrepancy between predicted and observed values. For the present study, we used XGBoost, a boosting method using decision trees. XGBoost is characterized by the regularization of the objective function, an efficient split-finding algorithm, the ability to handle sparsity caused by missing values and dummy variables, and fast execution speed. XGBoost has demonstrated excellent results in data analysis [21].

Evaluation

In the analysis, we used the area under the receiver operating characteristic curve (AUC) for evaluation according to our previous study [7]. To evaluate generalization performance for unknown data, five-fold cross-validation was conducted for four methods, and the test AUC values were compared. Training AUC values were used. For the FP model, the variables and degrees determined in our previous study were used. For statistical analyses, KFDA, kernlab, and XGBoost packages in R were used.

Results

From the total of 2426 patients, those who had missing values for the 17 features or whose date of death was missing were excluded. Therefore, 915 patients were included in the analysis. Table 1 shows the descriptive statistics of each feature for all patients and for the patients that were included in the analysis. The results of the prediction of 30-day survival for each method are shown in Table 2. Based on the AUC values, the FP model, which is an extension of the polynomial model, outperformed KFDA and XGBoost, which are more flexible methods. On the other hand, the prediction performance of KSVM, which is similar to KFDA in terms of the use of the kernel method, exceeded that of the FP model.

Characteristic	N	%
Age (mean, SD)	68.3±12.9	
Gender		
Male	562	61.4
Site of primary cancer		
Lung	222	24.7
Gastrointestinal	244	27.2
Gynecological	47	5.2
Urogenital	44	4.9
Breast	47	5.2
Other	293	32.7
Metastatic site		
Anywhere	767	84.0
Liver	378	41.5
Bone	256	28.1
Lung	314	34.4
Central nervous system	113	12.4
ECOG performance status		
0-1	63	6.9
2	163	17.9
3	332	36.4
4	355	38.9
Anticancer therapy		
Chemotherapy	191	20.9
Hormone therapy	7	0.8
Radiotherapy	43	4.7

TABLE 1: Patient characteristics

Methods	FP model	KFDA	KSVM	XGBoost
ROC-AUC (95%CI)	0.799 (0.734-0.864)	0.765 (0.696-0.835)	0.834 (0.807-0.861)	0.784 (0.713-0.848)

TABLE 2: Predictive accuracy of 30-day survival

ROC-AUC, receiver operating characteristic-area under the curve; FP, fractional polynomial; KFDA, Kernel Fisher discriminant analysis; KSVM, Kernel support vector machine

Discussion

To the best of our knowledge, this is the first survey to compare the prognostic accuracy of statistical and machine-learning models in predicting the survival of cancer patients using only objective data in palliative care. The most important finding of our study was that KSVM, which is one of the machine learning models,

had better prognostic accuracy for 30-day survival among cancer patients in palliative care. The superior predictive ability of KSVM compared to FP may be attributed to the nonlinear nature of the function, which can flexibly define discriminant boundaries even when the data are distributed in a complex manner using the kernel method [24,25]. However, because the confidence intervals largely overlap with those of FP, it is reasonable to consider that FP possesses some predictive capability, despite being a relatively simple statistical model, rather than viewing KSVM as inherently superior. It is important to compare the accuracy of the statistical models, such as FP, and machine learning techniques, such as KSVM, because the performance of predictive accuracy depends on the size of training data and intrinsic characteristics of the data itself [26]. It is expected that machine learning techniques will perform better than statistical methods as the size of the dataset increases; however, when the results of the two are comparable, the latter has beneficial aspects thanks to the stability and interpretability of the results [27-29]. Notably, XGBoost showed inferior predictive ability compared to predictive models developed using statistical methods. One possible interpretation is that the decision tree used in XGBoost may not effectively approximate linear models, whereas the FP model approximates a linear model [30]. Another possible interpretation is that XGBoost does not perform well on small data sets [31-33]; therefore, XGBoost may not be the optimal method for developing prognostic models in palliative care when working with limited data.

In this study, we analyzed the data by dichotomizing survival time as an outcome, thereby placing it in a more tractable supervised learning setting. The development of predictive models with survival time itself as the outcome remains a methodological challenge. Future investigations may explore the use of statistical methods based on the proportional hazards model and its implementation in machine learning techniques [34-36]. Such survival-time response regression models for survival time response often focus on the estimation of hazard ratios, which requires the potentially unrealistic assumption of proportional hazards. Alternative approaches, such as methods that use pseudo-values to directly estimate the survival function, or those that focus on the restricted mean survival time as the primary outcome [37,38], could offer promising solutions. These statistical and machine learning-based techniques are still in development, and their performance on datasets of the size we analyzed needs to be thoroughly validated in future studies.

Our study had several limitations. The first limitation was that this validation study was conducted only in patients who required blood tests in ordinary clinical settings. In the J-ProVal study, 1263 patients (52.1 %) had no laboratory data and were excluded from the present analysis. For this reason, 1106 (46.8%) patients were excluded. The possible reason with no laboratory data could be the physician's practice policy, the patient's refusal, and the patients with too serious conditions. A second limitation is that the patients with missing values for the 17 characteristics or missing date of death were excluded. As a result, only 915 patients (37.7%) could be included in the analysis. Therefore, the results of this study must be interpreted with caution because of selection bias. A third limitation is that our study only included Japanese participants, so caution is needed when applying our results to other ethnic groups.

Conclusions

Our findings suggest that when a large amount of data is available on cancer patients receiving palliative care, a machine learning method, such as KSVM, could be a viable option for developing a prognostic prediction model using only objective data. However, further studies are needed to compare the accuracy of the KSVM and FP models, as the accuracy of KSVM may vary depending on the size of the training data, whereas the FP model demonstrates stable and consistent accuracy.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Jun Hamano, Ayano Takeuchi, Hidemasa Nose

Acquisition, analysis, or interpretation of data: Jun Hamano, Ayano Takeuchi, Tomoya Keyaki, Hidemasa Nose, Kenichi Hayashi

Drafting of the manuscript: Jun Hamano, Ayano Takeuchi, Tomoya Keyaki, Hidemasa Nose, Kenichi Hayashi

Critical review of the manuscript for important intellectual content: Jun Hamano, Ayano Takeuchi, Tomoya Keyaki, Hidemasa Nose, Kenichi Hayashi

Disclosures

Human subjects: Consent for treatment and open access publication was obtained or waived by all participants in this study. Institutional Review Board of Seirei Mikatahara General Hospital issued approval 201210. This study was conducted in accordance with the ethical standards of the Declaration of Helsinki

and the ethical guidelines for research presented by the Ministry of Health, Labour, and Welfare of Japan.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** This work was supported in part by JSPS KAKENHI Grant Number 22H03305. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

- Hui D, Paiva CE, Del Fabbro EG, et al.: Prognostication in advanced cancer: update and directions for future research. *Support Care Cancer*. 2019, 27:1973-84. [10.1007/s00520-019-04727-y](#)
- Degner LF, Kristjanson LJ, Bowman D, et al.: Information needs and decisional preferences in women with breast cancer. *J Am Med Assoc*. 1997, 277:1485-92. Accessed: August 28, 2017: <http://www.ncbi.nlm.nih.gov/pubmed/9145723>.
- Kirk P, Kirk I, Kristjanson LJ: What do patients receiving palliative care for cancer and their families want to be told? A Canadian and Australian qualitative study. *BMJ*. 2004, 328:1343. [10.1136/bmj.38103.423576.55](#)
- Steinhauser KE, Christakis NA, Clipp EC, McNeilly M, McIntyre L, Tulsky JA: Factors considered important at the end of life by patients, family, physicians, and other care providers. *J Am Med Assoc*. 2000, 284:2476-82. [10.1001/jama.284.19.2476](#)
- Glare P, Virik K, Jones M, Hudson M, Eychmuller S, Simes J, Christakis N: A systematic review of physicians' survival predictions in terminally ill cancer patients. *BMJ*. 2003, 327:195-8. [10.1136/bmj.327.7408.195](#)
- Maltoni M, Caraceni A, Brunelli C, et al.: Prognostic factors in advanced cancer patients: evidence-based clinical recommendations--a study by the Steering Committee of the European Association for Palliative Care. *J Clin Oncol*. 2005, 23:6240-8. [10.1200/JCO.2005.06.866](#)
- Hamano J, Takeuchi A, Yamaguchi T, et al.: A combination of routine laboratory findings and vital signs can predict survival of advanced cancer patients without physician evaluation: a fractional polynomial model. *Eur J Cancer*. 2018, 105:50-60. [10.1016/j.ejca.2018.09.037](#)
- Simmons CP, McMillan DC, McWilliams K, et al.: Prognostic tools in patients with advanced cancer: a systematic review. *J Pain Symptom Manage*. 2017, 53:962-70.e10. [10.1016/j.jpainsymman.2016.12.330](#)
- Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent reporting of a multivariable prediction Model for individual prognosis or diagnosis (Tripod): the Tripod statement. *Ann Intern Med*. 2015, 162:55-63. [10.7326/M14-0697](#)
- Zhang B, Shi H, Wang H: Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach. *J Multidiscip Healthc*. 2023, 16:1779-91. [10.2147/JMDH.S410301](#)
- Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI: Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015, 13:8-17. [10.1016/j.csbj.2014.11.005](#)
- Baba M, Maeda I, Morita T, et al.: Survival prediction for advanced cancer patients in the real world: a comparison of the Palliative Prognostic Score, Delirium-Palliative Prognostic Score, Palliative Prognostic Index and modified Prognosis in Palliative Care Study predictor model. *Eur J Cancer*. 2015, 51:1618-29. [10.1016/j.ejca.2015.04.025](#)
- Baba M, Maeda I, Morita T, et al.: Independent validation of the modified prognosis palliative care study predictor models in three palliative care settings. *J Pain Symptom Manage*. 2015, 49:853-60. [10.1016/j.jpainsymman.2014.10.010](#)
- Hamano J, Morita T, Ozawa T, et al.: Validation of the simplified palliative prognostic index using a single item from the communication capacity scale. *J Pain Symptom Manage*. 2015, 50:542-7.e4. [10.1016/j.jpainsymman.2015.04.021](#)
- Templeton AJ, McNamara MG, Šeruga B, et al.: Prognostic role of neutrophil-to-lymphocyte ratio in solid tumors: a systematic review and meta-analysis. *J Natl Cancer Inst*. 2014, 106:dju124. [10.1093/jnci/dju124](#)
- Zhang C, Wang H, Ning Z, Xu L, Zhuang L, Wang P, Meng Z: Prognostic nutritional index serves as a predictive marker of survival and associates with systemic inflammatory response in metastatic intrahepatic cholangiocarcinoma. *Oncotargets Ther*. 2016, 9:6417-23. [10.2147/OTT.S112501](#)
- Templeton AJ, Ace O, McNamara MG, et al.: Prognostic role of platelet to lymphocyte ratio in solid tumors: a systematic review and meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2014, 23:1204-12. [10.1158/1055-9965.EPI-14-0146](#)
- Royston P, Altman DG: Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *J R Stat Soc Ser C*. 1994, 43:429-67. [10.2307/2986270](#)
- Yang J, Frangi AF, Yang JY, Zhang D, Jin Z: KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE Trans Pattern Anal Mach Intell*. 2005, 27:230-44. [10.1109/TPAMI.2005.33](#)
- Soman KP, Loganathan R, Ajay V: *Machine Learning with SVM and Other Kernel Methods*. PHI Learning Pvt Ltd, 2024.
- Tianqi Chen, Tong He MB: Xgboost: extreme gradient boosting. *R package version 0.4-2*. 2015, 4:1-4.
- Hofmann T, Schölkopf B, Smola AJ: Kernel methods in machine learning. *Ann Stat*. 2008, 36:1171-220. [10.1214/009053607000000677](#)
- Jp STA: Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *J Mach Learn Res*. 2007, 1027-61.
- Cortes C, Vapnik V, Saitta L: Support-vector networks. *Mach Learn*. 1995, 203:273-97. [10.1007/BF00994018](#)
- Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*. Springer, New York; 2009. [10.1007/978-0-387-84858-7](#)
- Bennett M, Hayes K, Kleczyk EJ, Mehta R: Similarities and differences between machine learning and traditional advanced statistical modeling in healthcare analytics. *Arxiv*. 2024,

27. Domingos P: A few useful things to know about machine learning . Commun ACM. 2012, 55:78-87. [10.1145/2347736.2347755](#)
28. Bzdok D, Altman N, Krzywinski M: Statistics versus machine learning . Nat Methods. 2018, 154-3.
29. Breiman L: Statistical modeling: the two cultures . Statist Sci. 2001, 16:199-231. [10.1214/SS/1009213726](#)
30. Raymaekers J, Rousseeuw PJ, Verdonck T, Yao R: Fast linear model trees by PILOT. Mach Learn. 2024, 113:6561-610. [10.1007/S10994-024-06590-3/FIGURES/12](#)
31. Li J, Liu H, Yang Z, Han L: A credit risk model with small sample data based on G-XGBoost . Appl Artif Intell. 2021, 35:1550-66. [10.1080/08839514.2021.1987707](#)
32. Zou M, Jiang WG, Qin QH, Liu YC, Li ML: Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting. Materials (Basel). 2022, 15:5298. [10.3390/ma15155298](#)
33. Silvey S, Liu J: Empirical sample size determination for popular classification algorithms in clinical research. medRxiv. 2024, 2024. [10.1101/2024.05.03.24306846](#)
34. Kleinbaum DG KM: Survival Analysis: A Self-Learning Text (Statistics in the Health Sciences) . Springer, 1996.
35. Binder H, Schumacher M: Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics. 2008, 9:14. [10.1186/1471-2105-9-14](#)
36. Sanz H, Reverter F, Valim C: Enhancing SVM for survival data using local invariances and weighting . BMC Bioinformatics. 2020, 21:193. [10.1186/s12859-020-3481-2](#)
37. Schenk A, Berger M, Schmid M: Pseudo-value regression trees. Lifetime Data Anal. 2024, 30:439-71. [10.1007/s10985-024-09618-x](#)
38. Tian L, Zhao L, Wei LJ: Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics. 2014, 15:222-33. [10.1093/biostatistics/kxt050](#)