Surgical Oncology 36 (2021) 23–27

Contents lists available at ScienceDirect

Surgical Oncology

journal homepage: http://www.elsevier.com/locate/suronc

Deep learning enabled prediction of 5-year survival in pediatric genitourinary rhabdomyosarcoma

Hriday P. Bhambhvani [a], Alvaro Zamora [a], Kyla Velaer [a], Daniel R. Greenberg [b], Kunj R. Sheth [a,*]

[a] Department of Urology, Stanford University Medical Center, Stanford, CA, USA
[b] Department of Urology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

ARTICLE INFO

Keywords:
Urogenital
Sarcoma
Prognostication
Machine learning
Artificial intelligence

ABSTRACT

Background: Genitourinary rhabdomyosarcoma (GU-RMS) is a rare, pediatric malignancy originating from embryonic mesenchyme. Current approaches to prognostication rely upon conventional statistical methods such as Cox proportional hazards (CPH) models and have suboptimal predictive ability. Given the success of deep learning approaches in other specialties, we sought to develop and compare deep learning models with CPH models for the prediction of 5-year survival in pediatric GU-RMS patients.

Methods: Patients less than 20 years of age with GU-RMS were identified within the Surveillance, Epidemiology, and End Results (SEER) database (1998–2011). Deep neural networks (DNN) were trained and tested on an 80/20 split of the dataset in a 5-fold cross-validated fashion. Multivariable CPH models were developed in parallel. The primary outcomes were 5-year overall survival (OS) and disease-specific survival (DSS). Variables used for prediction were age, sex, race, primary site, histology, degree of tumor extension, tumor size, receipt of surgery, and receipt of radiation. Receiver operating characteristic curve analysis was conducted, and DNN models were tested for calibration.

Results: 277 patients were included. The area under the curve (AUC) for the DNN models was 0.93 for OS and 0.91 for DSS. AUC for the CPH models was 0.82 for OS and 0.84 for DSS. The DNN models were well-calibrated: OS model (slope = 1.02, intercept = −0.06) and DSS model (slope = 0.79, intercept = 0.21).

Conclusions: A deep learning-based model demonstrated excellent performance, superior to that of CPH models, in the prediction of pediatric GU-RMS survival. Deep learning approaches may enable improved prognostication for patients with rare cancers.

1. Introduction

Soft-tissue sarcomas are the most common extracranial tumors in children, and 40% of these tumors are rhabdomyosarcomas [1]. Roughly 375 cases of rhabdomyosarcoma are diagnosed in the United States annually, and 15%–20% of these arise from sites along the genitourinary tract including the bladder, prostate and uterus [1,2]. Historically associated with poor survival, patients with low and/or intermediate-risk genitourinary rhabdomyosarcoma (GU-RMS) have seen a drastic improvement in prognosis over the past few decades owing to the combined effort of clinical trials led by the International Society for Pediatric Oncology (SIOP), the Children's Oncology Group (COG), and other professional organizations [2]. Indeed, patients with low-risk disease have attained 3-year survival rates of greater than 95% with modern treatment paradigms [3]. However, for patients with metastatic and/or relapsed disease, survival rates remain poor, with a 3-year overall survival ranging from 34% to 42% [4,5]. GU-RMS patients proceed along a risk-stratified treatment regimen, with patients classified as low-risk, intermediate-risk, or high-risk based on site of disease, histology, and TNM staging. Accurate prognostic assessment of these patients informs the shared decision-making process and patient counseling.

There is currently no clinical decision-making support tool to aid in survival prediction or prognostication for GU-RMS. There is one report of a survival prediction nomogram for all patients with RMS, but this tool does not focus on genitourinary disease [6]. Furthermore, the authors used a multivariable Cox proportional hazards (CPH) model and achieved only moderate predictive ability with an area under the curve (AUC) of a receiver operating characteristic (ROC) graph of 0.74 [6]. Machine learning (ML) algorithms, and in particular deep learning

* Corresponding author. Department of Urology, Stanford University School of Medicine, 300 Pasteur Drive, Palo Alto, CA, 94305, USA.
E-mail addresses: hriday@stanford.edu (H.P. Bhambhvani), shethk@stanford.edu (K.R. Sheth).

https://doi.org/10.1016/j.suronc.2020.11.002
Received 29 September 2020; Accepted 15 November 2020
Available online 20 November 2020

approaches, have been recognized in recent years as robust predictors of oncologic outcomes [7–9]. These techniques can improve prediction over conventional statistical methods, like CPH models, through the detection of high dimensional, non-linear associations between variables.

In this context, we sought to use a national, multi-institution database with longitudinal follow-up to identify pediatric patients with GU-RMS and subsequently develop deep neural networks (DNNs) to predict both 5-year disease-specific and overall survival. We hypothesized that the DNNs would perform better than a multivariable Cox proportional hazards model in predicting GU-RMS survival.

2. Methods

2.1. Data source

The National Cancer Institute's Surveillance Epidemiology and End Results (SEER) database covers data from 18 population-based registries and represents approximately 28% of the population [10]. The SEER*-Stat software (Version 8.3.6) case listing session was used to identify patients from the SEER 18 (November 2018 submission) database. This study was conducted using de-identified data in compliance with the Health Insurance Portability and Accountability Act and the 1964 Helsinki Declaration and its amendments; thus, formal informed consent was not obtained, and institutional review board approval was not required.

2.2. Inclusion criteria

We included patients less than 20 years of age diagnosed with rhabdomyosarcoma of the genitourinary tract based on the International Classification of Disease for Oncology Version 3 (ICD-O-3) coding system between 1998 and 2011. Data was only collected through 2011 to allow for complete 5-year follow-up before the cut-off date of December 31, 2016 in the November 2018 SEER submission. We extracted patients with site codes C51 (vulva), C52 (vagina), C53 (cervix), C54/55 (uterus), C61 (prostate), C62 (testis) C63, (other male genital organs), and C67 (urinary bladder) as well as histology codes 8900–8905, 8910, 8912, 8920, and 8991, all of which represent rhabdomyosarcoma. Patients without histological confirmation of diagnosis and those with more than one primary cancer were excluded.

2.3. Model development and deployment

We chose to develop deep neural network models on the basis of previous studies that have suggested they are superior to traditional statistical models in survival analysis [8,11]. 80% of the study cohort was used to develop the model (i.e. the training set), while the remaining 20% was used to evaluate the developed models (i.e. the testing set). Model inputs spanned a variety of demographic, clinical, and therapeutic variables and were selected based on prior work in RMS that has identified them as predictive of survival [6,12]. These covariates included in the analysis were age at diagnosis, sex, race, tumor extension (localized to primary site, regional invasion, or distant metastasis), tumor histology (embryonal, alveolar, or other), tumor size (in centimeters), specific primary site, radiotherapy treatment (yes/no), and surgical excision (yes/no). The final endpoints of this study were 5-year overall survival (OS) and 5-year disease specific survival (DSS).

The final DNN models (Final_DSS.h5 and Final_OS.5), as well as the source code used to develop them, are publicly accessible at https://github.com/alvarozamora/UroRhabdo. Briefly, we constructed fully-connected networks with 8 layers and 256 nodes in each layer. Five such networks, one for each fold of validation, were constructed in an ensemble approach, as previously described [13]. Training of the model was performed using the adaptive moment estimation (Adam) optimization technique with final parameters: learning rate = 0.0003, $\beta 1 = 0.9$, $\beta 2 = 0.999$, epsilon = 10^{-8}, and weight decay = 10^{-4}. Binary cross-entropy was used as the loss function. Deep learning models were trained with the assistance of the Stanford University Sherlock high performance computing cluster. We adhered to the established Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [14].

Multivariable Cox proportional hazards were constructed in parallel for comparison using the same covariates. Code was implemented using Python version 3.6 (Python Software Foundation, Wilmington, Delaware, USA) and the deep learning framework PyTorch, version 1.4. CPH models were built using R version 3.5.3 (The R Foundation for Statistical Computing). To validate our dataset, univariate hazards analysis was conducted to compare survival rates between the more recent patient cohort diagnosed between 2005 and 2011 and the earlier patient cohort diagnosed between 1998 and 2004.

2.4. Model assessment

All model assessments were conducted in a 5-fold cross validated fashion. Model discrimination was tested for with ROC curve analysis. The AUC was calculated. For a perfect model with no false positives and only true positives, the AUC is equal to 1. For a model that is no better than chance, the AUC is equal to 0.5. We compared the AUC of the DNN models to that of the CPH models to assess which model had superior discriminatory performance. We additionally assessed the DNN models for calibration, which refers to the agreement between subgroups of predicted probabilities and their observed frequencies in a given population [15]. We then developed calibration plots for each model and calculated calibration intercept and slope. A perfect model has a calibration intercept of 0 and a calibration slope of 1.

3. Results

In total, 277 patients with GU-RMS met our inclusion criteria; patient and disease characteristics are summarized in Table 1. Notably, most patients were male (n = 204, 73.6%), and the most common primary site was the testis (n = 101, 36.5%). The most common histological subtype was embryonal (n = 206, 74.4%), and tumors were most frequently between 5 cm and 10 cm in size (n = 91, 32.9%). 5-year OS and DSS were 79.4% and 80.9%, respectively.

Multivariable CPH models achieved an AUC of 0.82 for OS and 0.84 for DSS. Factors associated with all-cause mortality are summarized in Table 2. Notably, RMS originating in the ovary (HR = 26.41, 95% CI 4.31–161.89, p = 0.0004), prostate (HR = 5.99, 95% CI 2.21–16.22, p = 0.0004) and bladder (HR = 3.65, 95% CI 1.18–11.26, p = 0.02) were associated with poor survival, compared to RMS originating in the testis. Increasing age was associated with worse survival (HR = 1.10, 95% CI 1.04–1.15, p = 0.0003). Finally, as compared to localized RMS, disease with regional invasion (HR = 4.34, 95% CI 1.45–12.97, p = 0.009) or distant metastasis (HR = 15.46, 95% CI 5.11–46.70, p < 0.0001) was associated with worse survival. In univariate hazards analysis, patients diagnosed between 2005 and 2011 had similar survival (HR = 0.74, 95% CI 0.45–1.22, p = 0.24) as compared to patients diagnosed between 1998 and 2004.

The 5-year deep learning-based DSS model achieved a mean AUC of 0.91, with a standard deviation of 0.05, across all folds (Fig. 1). ROC plots from each fold for the DSS model are reported in Supplemental Fig. 1. The calibration slope of the DSS model was 0.79, and the calibration intercept was 0.21 (Fig. 2).

The 5-year deep learning-based OS model achieved a mean AUC of 0.93, also with a standard deviation of 0.05 across all folds (Fig. 1). ROC plots from each fold for the OS model are reported in Supplemental Fig. 2. The calibration slope of the OS model was 1.02, and the calibration intercept was −0.06 (Fig. 2).

Table 1
Baseline patient and disease characteristics.

Characteristic	N (%)
Age, mean (std dev)	8.52 (6.53)
Sex	
Female	73 (26.4)
Male	204 (73.6)
Race	
White	207 (74.7)
Black	54 (19.5)
Asian	11 (4.0)
American Indian/Alaska Native	2 (0.7)
Unknown	3 (1.1)
Primary Site	
Cervix	16 (5.8)
Ovary	4 (1.4)
Uterus	2 (0.7)
Vagina	25 (9.0)
Vulva	7 (2.5)
Bladder	46 (16.6)
Prostate	40 (14.4)
Testis	101 (36.5)
Other male genital organs	36 (13.0)
Tumor size	
0 cm–5 cm	77 (27.8)
5 cm–10 cm	91 (32.9)
Greater than 10 cm	53 (19.1)
Unknown	56 (20.2)
Tumor histology	
Embryonal	206 (74.4)
Alveolar	26 (9.4)
Other	28 (10.1)
Unknown	17 (6.1)
Tumor extension	
Localized	89 (32.1)
Regional invasion	62 (22.4)
Distant metastasis	56 (20.2)
Unknown	70 (25.3)
Received surgery (yes)	217 (78.3)
Received radiotherapy (yes)	132 (47.7)
5-year overall surviving	220 (79.4)
5-year disease specific surviving	224 (80.9)

Table 2
Factors associated with all-cause mortality: Multivariable cox proportional hazards.

Characteristic	HR	95% CI	p-value
Age	1.10	1.04–1.15	0.0003
Sex			
Male	Ref		
Female	1.68	0.49–5.78	0.41
Race			
White	Ref		
Black	1.49	0.74–3.00	0.27
Asian	0.57	0.08–4.29	0.59
American Indian/Alaska Native	5.00	0.93–26.82	0.06
Primary Site			
Testis	Ref		
Cervix/Uterus	0.46	0.04–5.34	0.53
Ovary	26.4	4.31–161.89	0.0004
Vagina	2.50	0.18–34.61	0.49
Vulva	2.21	0.22–22.58	0.50
Bladder	3.65	1.18–11.26	0.02
Prostate	5.99	2.21–16.22	0.0004
Other male genital organs	2.09	0.23–18.80	0.51
Tumor size (cm)	1.00	1.00–1.00	0.19
Tumor histology			
Embryonal	Ref		
Alveolar	1.15	0.53–2.51	0.72
Other	1.45	0.54–3.85	0.46
Unknown	0.15	0.02–1.17	0.07
Tumor extension			
Localized	Ref		
Regional invasion	4.34	1.45–12.97	0.009
Distant metastasis	15.50	5.11–46.7	<0.0001
Unknown	3.03	0.33–28.11	0.33
Surgery			
No	Ref		
Yes	1.28	0.59–2.75	0.53
Radiotherapy			
No	Ref		
Yes	1.03	0.57–1.88	0.92

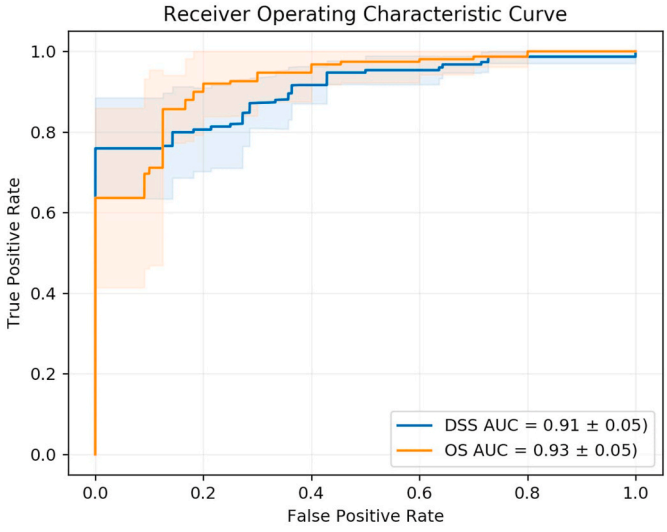

Fig. 1. Receiver operating characteristic curve of deep neural network models.

4. Discussion

To the best of our knowledge, this is the first study to apply machine learning to prognostication of rhabdomyosarcoma. We showed that deep neural networks were robust predictors of 5-year overall and disease-specific survival, and superior to traditional Cox proportional hazards models. CPH models identified increasing age, primary sites of the bladder, prostate, and ovary, as well as regional invasion/distant metastasis to be independently associated with increased all-cause mortality.

Our DNN models achieved robust discriminative performance, with an AUC of 0.93 for overall survival and 0.91 for disease-specific survival, both of which were notably superior to the CPH models developed in parallel. The DNN models were also well-calibrated, with the OS model in particular achieving a near perfect calibration slope of 1.02 and intercept of −0.06. In combination, high performance in both discrimination and calibration suggests that the DNN models are accurate, reliable predictors of 5-year survival in GU-RMS. The primary advantage of deep learning models over traditional statistical methods is ability to detect nonlinear relationships between model inputs (i.e. clinical, demographic, and treatment data) and model outputs (i.e. survival data) [16]. This likely underlies the superior performance of our DNN models as compared to our study's CPH models. As aforementioned, there is only one report of a survival prediction nomogram, which only achieved an AUC of 0.74, for children with RMS of any primary site [6]. Again, the ability to detect nonlinear relationships likely underlies the superior performance of our DNN models compared to Yang and colleague's nomogram. Additionally, in the generalized RMS model, the authors categorized primary site as either favorable or unfavorable, a dichotomization that inevitably results in loss of granular prognostic information derived from knowledge of a unique primary site. In comparison, knowledge of the specific primary site may have contributed to the improved performance of both our DNN and CPH models.

Though deep learning models have high predictive power, an inherent drawback to them is the lack of explanatory power. In this regard, CPH models are more informative in allowing for interpretation

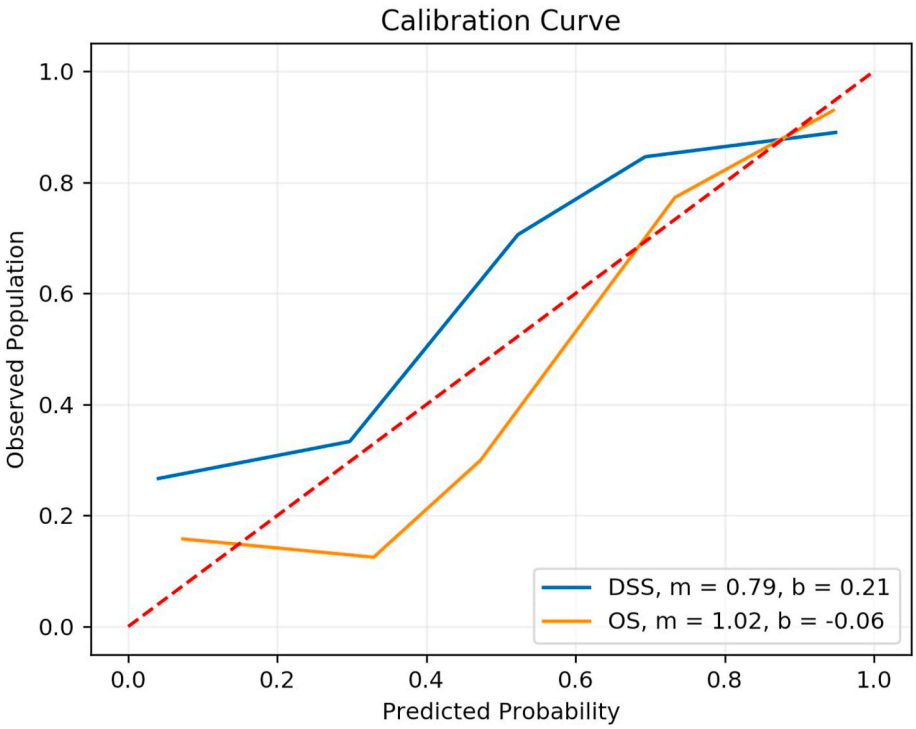

Fig. 2. Calibration curve of deep neural network models.

of what covariates drive differences in survival. For example, with our CPH model, we quantitatively (via hazard ratios) identified increasing age, primary bladder/prostate/ovarian disease, and regional invasion/distant metastasis as independently statistically significantly associated with all-cause mortality, in an analysis adjusted for race, sex, tumor size, and tumor histology. As of yet, DNN models are not able to offer this degree of explanatory power.

Recent years have seen a renaissance in deep learning, owing largely to increases in computational power and the availability of large datasets [17]. In the realm of urology, a variety of studies have demonstrated the utility of deep learning approaches, ranging from automated detection of bladder cancer on cystoscopy to prediction of kidney stone composition from digital photographs [18,19]. Though all of these studies fall in the realm of medical image analysis, deep learning models have also shown promise in the prediction of oncologic events, including risk stratification and survival prediction based on clinical and disease attributes [20–22]. A common problem among these studies and many others utilizing machine learning methods is the lack of deployment of the developed models or availability of source code. While these "proof of concept" studies demonstrate incredible performance of deep learning models in an internal, investigator-curated cohort, they fail to make the developed models publicly available. In this context, we made our final models as well as our source code publicly available; interested investigators may download our models, or even generate their own.

Due to the rarity of GU-RMS, most published studies of this disease are retrospective analyses of clinical studies or small, single-institution studies. These single-institution and retrospective studies often fail to rigorously identify associations between survival outcomes and risk factors due to small sample sizes and a lack of generalizability. SEER data presents an opportunity to access information on GU-RMS patients at the population level, leading to both the development of generalizable models as well as the identification of true risk factors of poor survival. Furthermore, SEER data includes patients receiving both guideline-concordant and guideline-divergent therapy, and the results reported herein therefore represent the full spectrum of GU-RMS in the United States.

This study should be considered in the context of its limitations. First, the algorithms developed in the study have not undergone external validation, and rigorous external validation is necessary to determine if a proposed model is applicable outside of the study cohort. Though we present models that are not externally validated, the use of a nationally curated database that collects data from 28% of the United States should improve generalizability. Additionally, by making our models and code publicly available, we offer investigators the opportunity to externally validate our models or generate their own. Second, the source of our data is the SEER program and our study is, therefore, subject to the limitations common to all studies of administrative datasets. These include missing data, lack of granularity for specific predictive factors, and reliance on the accurate coding of diagnoses, procedures, and disease characteristics. However, the fact that our models were able to robustly predict 5-year survival despite these limitations is promising and suggests that machine learning techniques may allow for optimal interpretation of administrative data. Finally, an inherent limitation in the study of GU-RMS is the rarity of the disease, reflected in the relatively small size of our cohort. To train our models on as many patients as possible, we included patients whose index diagnosis was as early as 1998. While this long period of enrollment bolsters our sample size, it also leads to the inclusion of non-contemporary patients whose survival might not be reflective of GU-RMS today. To address this, we compared survival in the more recent half of our cohort (diagnosed between 2005 and 2011) to the earlier half (diagnosed between 1998 and 2004), finding no difference in survival in a univariate CPH analysis. Nevertheless, given modern treatment paradigms, it is possible survival estimates from our models could slightly underestimate present-day survival of GU-RMS patients.

5. Conclusion

In summary, we report the development of deep neural network models that are robust predictors of 5-year overall and disease-specific survival in genitourinary rhabdomyosarcoma. Furthermore, these models performed superior to traditional multivariable Cox proportional hazards models. These results suggest deep learning approaches may enable optimal prognostication of patients with rare cancers.

Source of funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Funding

None.

Declaration of competing interest

The authors have no conflicts of interest to report.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.suronc.2020.11.002.

References

[1] D.A. Siegel, J. Li, S.J. Henley, et al., Geographic variation in pediatric cancer incidence - United States, 2003-2014, MMWR Morb. Mortal. Wkly. Rep. 67 (25) (2018) 707–713.
[2] M. Harel, F.A. Ferrer, L.H. Shapiro, J.H. Makari, Future directions in risk stratification and therapy for advanced pediatric genitourinary rhabdomyosarcoma, Urol. Oncol. 34 (2) (2016) 103–115.
[3] D.O. Walterhouse, A.S. Pappo, J.L. Meza, et al., Shorter-duration therapy using vincristine, dactinomycin, and lower-dose cyclophosphamide with or without radiotherapy for patients with newly diagnosed low-risk rhabdomyosarcoma: a report from the Soft Tissue Sarcoma Committee of the Children's Oncology Group, J. Clin. Oncol. 32 (31) (2014) 3547–3552.
[4] O. Oberlin, A. Rey, E. Lyden, et al., Prognostic factors in metastatic rhabdomyosarcomas: results of a pooled analysis from United States and European cooperative groups, J. Clin. Oncol. 26 (14) (2008) 2384–2389.
[5] G. Bisogno, A. Ferrari, A. Prete, et al., Sequential high-dose chemotherapy for children with metastatic rhabdomyosarcoma, Eur. J. Canc. 45 (17) (2009) 3035–3041.
[6] L. Yang, T. Takimoto, J. Fujimoto, Prognostic model for predicting overall survival in children and adolescents with rhabdomyosarcoma, BMC Canc. 14 (2014) 654.
[7] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.
[8] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Med. Res. Methodol. 18 (1) (2018) 24.
[9] D.W. Kim, S. Lee, S. Kwon, W. Nam, I.H. Cha, H.J. Kim, Deep learning-based survival prediction of oral cancer patients, Sci. Rep. 9 (1) (2019) 6994.
[10] Surveillance Epidemiology and End results program (www.seer.cancer.gov). Database: Incidence: SEER 18, November 2018 submission.
[11] F.E. Ahmed, Artificial neural networks for diagnosis and survival prediction in colon cancer, Mol. Canc. 4 (2005) 29.
[12] J.C. Chisholm, J. Marandet, A. Rey, et al., Prognostic factors after relapse in nonmetastatic rhabdomyosarcoma: a nomogram to better define patients who can be salvaged with further therapy, J. Clin. Oncol. 29 (10) (2011) 1319–1325.
[13] B. Liu, Q. Cui, T. Jiang, S. Ma, A combinational feature selection and ensemble neural network method for classification of gene expression data, BMC Bioinf. 5 (2004) 136.
[14] W. Luo, D. Phung, T. Tran, et al., Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view, J. Med. Internet Res. 18 (12) (2016) e323.
[15] F.J.W.M. Dankers, A. Traverso, L. Wee, et al., Prediction modeling methodology. 2018 dec 22 (Chapter 8). Available from:, in: P. Kubben (Ed.), Fundamentals of Clinical Data Science [Internet], Cham (CH): Springer, 2019, https://doi.org/10.1007/978-3-319-99713-1_8 https://www.ncbi.nlm.nih.gov/books/NBK543534/.
[16] J.H. Chen, S.M. Asch, Machine learning and prediction in medicine - beyond the peak of inflated expectations, N. Engl. J. Med. 376 (26) (2017) 2507–2509.
[17] A. Esteva, A. Robicquet, B. Ramsundar, et al., A guide to deep learning in healthcare, Nat. Med. 25 (1) (2019) 24–29.
[18] K.M. Black, H. Law, A. Aldoukhi, J. Deng, K.R. Ghani, Deep learning computer vision algorithm for detecting kidney stone composition, BJU Int. 125 (6) (2020) 920–924.
[19] E. Shkolyar, X. Jia, T.C. Chang, et al., Augmented bladder tumor detection using deep learning, Eur. Urol. 76 (6) (2019) 714–718.
[20] S. Jeong, Y. Ge, J. Chen, et al., Latent risk intrahepatic cholangiocarcinoma susceptible to adjuvant treatment after resection: a clinical deep learning approach, Front. Oncol. 10 (2020) 143.
[21] D.W. Kim, S. Lee, S. Kwon, W. Nam, I.H. Cha, H.J. Kim, Deep learning-based survival prediction of oral cancer patients, Sci. Rep. 9 (1) (2019) 6994.
[22] B. Lee, S.H. Chun, J.H. Hong, et al., DeepBTS: prediction of recurrence-free survival of non-small cell lung cancer using a time-binned deep neural network, Sci. Rep. 10 (1) (2020) 1952.