Original Article

Application of data mining in the provision of in-home medical care for patients with advanced cancer

Chao Yang[1#], Ruihua Yu[1#], Hui Ji[1,2#], Haosheng Jiang[3#], Wanli Yang[1], Feng Jiang[1]

[1]Translational Institute for Cancer Pain, Xinhua Hospital Chongming Branch, Shanghai, China; [2]Department of Anesthesiology, Xinhua Hospital Chongming Branch, Shanghai, China; [3]Department of Oncology, Shanghai International Medical Center, Shanghai, China

Contributions: (I) Conception and design: C Yang, F Jiang; (II) Administrative support: F Jiang; (III) Provision of study materials or patients: R Yu; (IV) Collection and assembly of data: C Yang; (V) Data analysis and interpretation: C Yang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Feng Jiang. Translational Institute for Cancer Pain, Xinhua Hospital Chongming Branch, Shanghai 202150, China. Email: fengjiang@xinhuamed.com.cn.

Background: As the number of patients with cancer rises, home care for patients with advanced disease is becoming increasingly important. To provide guidance for home medical services and hospice care, we investigated the basic information and medical service information of patients with advanced cancer receiving home care by using a data mining algorithm to predict the patients' survival and medical expenses.

Methods: Data from patients with advanced cancer who received home care in Chongming District (Shanghai, China) between 2016 and 2018 were collected. The medical expenses and survival time of the patients were classified and predicted through the use of random forest algorithms, support-vector machine algorithms, and back-propagation (BP) neural network algorithms.

Results: The performances of the 3 algorithms in classifying patient survival and predicting medical expenses were compared. The random forest algorithm, support vector machine, and BP neural network in the classification of patient survival had accuracy of 81.94%±6.12%, 74.61%±7.01%, and 72.90%±8.08%, respectively. The standard mean square errors of the regression model for predicting medical expenses were 0.4194±0.2393, 1.1222±0.0648, and 1.2986±0.1762, respectively.

Conclusions: The random forest algorithm is the most suitable prediction model for predicting medical costs and patient survival with the quantity of data currently available. Further optimization of the random forest algorithm could provide guidance and help medical institutions improve the efficiency and quality of home medical services for patients with advanced cancer.

Keywords: Data mining; patients with advanced cancer; random forest; support vector machine (SVM); back-propagation neural network (BP neural network)

Submitted Apr 20, 2021. Accepted for publication Jun 11, 2021.
doi: 10.21037/tcr-21-896
View this article at: https://dx.doi.org/10.21037/tcr-21-896

Introduction

According to the World Health Organization's Global Cancer Statistics 2018, new cancer cases and cancer deaths worldwide were estimated to tally 18.1 million and 9.6 million, respectively. Globally, 1 in 5 men and 1 in 6 women will develop cancer, and 1 in 8 men and 1 in 11 women will die from it, indicating the high global burden of this disease (1). China has the highest incidence of cancer of any country, and cancer is the 2nd leading cause of death in Chinese people under the age of 70 (2).

Currently, most types of cancer cannot be cured by medical treatment. Patients with advanced cancer often experience anxiety, depression, pain, and other symptoms

(3,4). Hospice care is the main approach to managing these symptoms, and can aid in alleviating patients' suffering and improve their quality of life (QOL) (5). In China, some patients with advanced cancer choose to spend their final days in home care. The Homecare Service Program for advanced cancer patients provides free medical services in patients' homes, and plays an essential role in alleviating patients' suffering and improving their QOL (6). Therefore, it is of considerable significance to understand the pain, QOL, survival, and medical costs of patients with advanced cancer, and to establish survival and medical cost models that will improve medical services and rational use of medical resources and funds for these patients. With the improvement of computing power, data mining has been widely used in medical big data processing field. Machine learning algorithms were used for early disease prediction and prognosis analysis. For example, the random forest algorithm can be used to analyze patient heart rate, diastolic blood pressure, systolic blood pressure, average blood pressure, respiratory rate, blood oxygen saturation. A disease prediction model for early diagnosis of disease and prognosis assessment was constructed by random forest algorithm (7). SVM was reported to predict and judge the occurrence of chronic diseases such as diabetes and medical care management, and provide a basis for patients to choose the best treatment (8). Artificial intelligence learning was also being used in the diagnosis of ischemic stroke (9). In this study, we retrospectively analyzed the home medical case of advanced cancer patients and used three kinds of machine learning algorithms to establish the survival prediction model for advanced cancer patients. We presented the following article in accordance with the MDAR reporting checklist (available at https://dx.doi.org/10.21037/tcr-21-896).

Methods

Patients

Data of 310 patients with pain due to advanced cancer who were treated at Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (Chongming Branch) between January 2016 and December 2018 were collected. The dataset included medical expenses, the duration of home service, address, sex, age, education level, medical insurance, monthly household income, tumor type, previous treatment, pain duration, medical history, previous medication, pain severity [Numeric Rating Scale (NRS) score], self-assessment of health status [Karnofsky Performance Scale (KPS) score], and QOL score.

Statistical analysis

Microsoft Excel software (Microsoft, America) was used to input and review the data, and R3.6.0 statistical software was used to analyze it. The 'randomForest', 'e1071', 'neural network' (nnet), and 'rminer' packages in R were loaded using the random forest algorithm (10). The support vector machine (SVM) (11) and BP neural network (12) algorithms were used to classify and predict the patient's home medical service duration (days) and to perform a regression analysis of drug costs (Chinese Yuan), respectively (13-15). The 10-fold cross-validation method was adopted to verify the 3 classification prediction methods. The dataset was divided into 10 parts. During the experiment, 9 parts were taken as training data, and 1 part was taken as test data in turn. The normalized mean square error (NMSE) of the training sets was calculated (16).

Experiment

The home service case data used in this experiment was obtained from Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (Chongming Branch). The dataset contained 310 cases and 35 attributes, including 29 categorical (*Table 1*) and 6 numeric (*Table 2*) variables.

Developed by Becker, Chambers, and Wilks of Bell Labs, R language is an open-source data analysis software maintained by a large and active global research community. It contains all the algorithms used in this data mining analysis and could efficiently complete the data mining performed for this experiment. The experimental steps were as follows:

(I) Home service data of advanced cancer patients were imported using the read.csv() function.
(II) The nnet, e1071, randomForest, and rminer packages required for this experiment were loaded using the library() function.
(III) The RandomForest(), SVM(), and nnet() functions were used in turn to classify and predict the survival time of patients in the dataset. In the BP neural network, the number of hidden layers (size) was set to 10 and the option to add the jump layer connection (skip) was set as true; and the weight attenuation parameter (decay) was set to 0.1. The 10-fold cross-validation method was used to verify the 3 classification prediction methods.
(IV) The RandomForest(), fit(), and nnet() functions

Table 1 Types of categorical variables

Categorical variables	Attributes	No.
Survival (days)	{<30, 30–90, >90}	3
Address	{Chenjia town, Chengqiao town, …}	17
Sex	{Male or female}	2
Living situation	{Living alone, living in a nursing home, living with family}	3
Household monthly income per capita (Yuan)	{<300, 300–600, >600}	3
Education	{Illiteracy, elementary school, junior high school, high school and above}	4
Primary disease diagnosis	{Liver cancer, lung cancer, stomach cancer...}	21
Transfer status	{Yes, No}	2
Radiotherapy	{Yes, No}	2
Chemotherapy	{Yes, No}	2
Surgery	{Yes, No}	2
Past medical history	{High blood pressure, diabetes, heart disease...}	12
Pain duration	{1 month, 1–6 months, 6 months–1 year, > 1 year}	4
Physical pain	{Yes, No}	2
Visceral pain	{Yes, No}	2
Pain medicine	{NSAIDs, weak opioids, strong opioids, other}	4
Anticonvulsant drug use	{Yes, No}	2
Anti-anxiety drug use	{Yes, No}	2
Glucocorticoid use	{Yes, No}	2
Constipation	{Yes, No}	2
Disgusting vomits	{Yes, No}	2
Vomiting	{Yes, No}	2
Dizziness	{Yes, No}	2
Sweating	{Yes, No}	2
Difficulty urinating	{Yes, No}	2
Drowsiness	{Yes, No}	2
Other negative symptoms	{Yes, No}	2

NSAIDs, non-steroidal anti-inflammatory drugs.

were used in turn to analyze and predict the drug costs of patients in the dataset. The model parameter of the fit() function was set to SVM; in the BP neural network, the number of hidden layer units (size) was set to 10 and the option to add a skip layer connection (skip) was set to true; and the weight attenuation parameter (decay) was set to 0.1. The 10-fold cross-validation method was used to verify the 3 classification prediction methods and calculate the NMSE of the training set.

Ethical approval

The study was approved by the ethics committee of Xinhua Hospital Chongming Branch (No.: CMEC-2021-KT-24) and was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Informed consent was taken from all the patients.

Table 2 Numeric variables

Numeric variables	Mean	Range
Medical expenditure (Yuan)	1,982.03	0–84,384.78
Age	70.58	23–91
Analgesic dosage (mg)	26.64	0–360
NRS	3.78	0–9
QOL	49.84	0–100
KPS	33.54	0–60

NRS, Numeric Rating Scale; QOL, quality of life; KPS, Karnofsky Performance Scale.

Table 3 Performance of 3 algorithms in predicting the survival time of patients

Algorithm	Correct rate (mean ± SD)	Error rate (mean ± SD)
Random forest	81.94%±6.12%	18.06%±6.12%
SVM	74.61%±7.01%	25.39%±7.01%
BP network	72.90%±8.08%	27.10%±8.08%

BP, back-propagation; SVM, support vector machine.

Results

This experiment used case data from 310 patients with pain due to advanced cancer who were treated at Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (Chongming Branch) between January 2016 and December 2018. Because this experimental dataset was a small-scale, multi-attribute dataset, the above 3 classification and regression algorithms could be easily implemented.

The random forest algorithm possesses the advantage of being able to classify and predict data with a small sample size and many variables. With sufficient computing power, the random forest algorithm does not require the deletion of variables; on the contrary, the more variables there are, the higher the classification accuracy.

Our comparison found the accuracy rate of the random forest algorithm for patient survival classification to be 81.94%, which was far higher than the 74.61% of the SVM algorithm or the 72.90% of the BP neural network ($P<0.05$) (*Table 3*). The standard deviation (SD) of the random forest algorithm was also smaller than those of the SVM and BP neural network algorithms. The accuracy and stability of the random forest algorithm for the classification prediction of patient survival time were better than those of the other 2 algorithms (*Figure 1*).

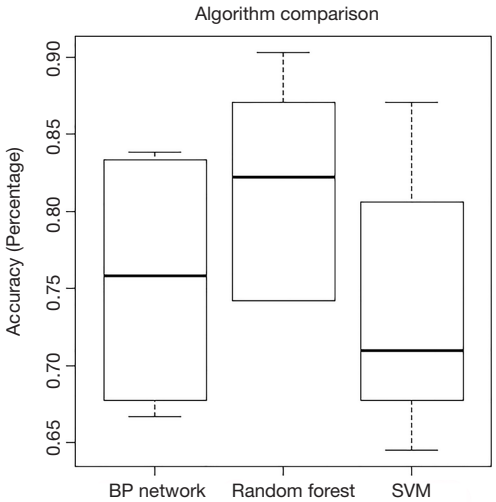

Figure 1 Accuracy of the different algorithms. BP, back-propagation; SVM, support vector machine.

Three data mining algorithms were employed to perform the regression analysis and prediction of patients' medical expenses. The NMSE values of the SVM and BP neural network algorithms in the training set were both above 1 (*Table 4*, *Figure 2*), which indicated that the regression models constructed using these algorithms had extremely poor accuracy for predicting patients' medical expenses. However, the NMSE value of the random forest algorithm in the training set was under 0.5 (*Table 4*, *Figure 2*), indicating that, to a certain degree, the regression model constructed using the random forest algorithm could predict the medical expenses admitted patients would incur.

Discussion

Home-based palliative care has a positive impact on patients. A study published in *The Lancet* showed that early palliative care improves the QOL and satisfaction of patients with cancer (17). Temel *et al.* showed that the combination of palliative care with antitumor therapy not only improves cancer patients' QOL and mood, but also significantly extends their survival (18). However, in China, the overall level of palliative care is significantly lower than the world average. In a global "quality of death" ranking released in 2015, the mainland China was placed 71st. China has a high demand for palliative care, but current medical resources meet only 0.3% of patients' needs (19). In the face of the shortage of home-based health care resources,

Table 4 Results of regression prediction of patient expenses using the 3 algorithms

Algorithm	NMSE (mean ± SD)
Random forest	0.4194±0.2393
SVM	1.1222±0.0648
BP network	1.2986±0.1762

BP, back-propagation; SVM, support vector machine; NMSE, normalized mean square error.

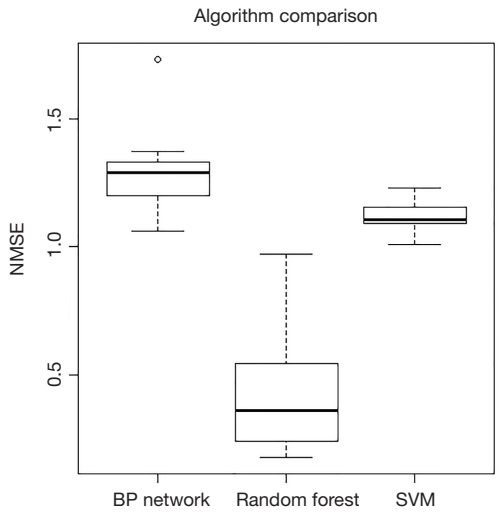

Figure 2 The NMSE of the different algorithms. NMSE, normalized mean square error; BP, back-propagation; SVM, support vector machine.

it is especially important that home-based palliative care resources are applied rationally and economically.

In the present research, we developed the analysis models for predicting survival and health care expenditures based on a dataset of 310 cases through the application of 3 different data mining techniques (random forest, SVM, and neural network). Based on the results of patient survival analysis, developing different strategies for homecare service for different patients, so as to improve the quality and efficiency of in-home medical services. Our medical cost prediction model can provide a reference for the planning of medical funding allocation by home medical service organizations.

The survival prediction of advanced cancer patients is of great significance to home care decision-making and arrangement of patients and relatives. By using data mining methods, researchers find that many biomarkers and clinical symptoms have significant correlation with the survival time of patients with advanced cancer. Cox regression and logistic regression were used to construct survival prediction models, such as: objective palliative prognostic score (OPPS) (20), performance status-based palliative prognostic index (PS-PPI) (21), etc. However, there are still some specific problems in real-world clinical applications. The way data collection methods and the quality of data are the most critical issues. It is difficult to collect clinical laboratory test results in home care service for patients with advanced cancer. Some data that can't be automatically collected by the wearable devices must still be collected manually. Beyond that, the prediction ability of diverse data mining models is variable with different amount of included data and different included factors. Therefore, the establishment of prediction models should be based on application scenarios, and the balance between sensitivity and specificity should be fully considered.

In general, home-based palliative care services can only provide services to patients in a particular geographical area, and differences may exist in the quality of medical services, basic information of patients, and number of medical service points in different regions. Therefore, for the analysis of survival time and medical expenses in home-based palliative care patients, a dataset obtained from a single institution is better than one from multiple institutions. Although single-center datasets generally have a small sample size, the data integrity is usually good.

All of the patients included in this study were enrolled from the same public in-home palliative care organization. Furthermore, their information was obtained through home follow-up, which ensured the standardization, completeness, and accuracy of the data, as well as clear patient outcomes and relevant medical expenses. In this study, there were few cases and many attributes in the dataset. However, the random forest algorithm showed strong adaptability to our dataset, and is able to process both classified data and numerical data without normalization. Meanwhile, it has a good ability to interpret high-dimensional data and is not easy to overfit. Our results showed that the random forest algorithm had an accuracy of more than 80% for predicting patients' survival times, and it also had high accuracy in the prediction of the medical costs of home palliative care, outperforming the other 2 algorithms.

The application of random forest model is helpful to the development of homecare services program and the use of

medical funds by hospice care organizations. In the future, the random forest algorithm needs to be further optimized to improve its accuracy in the classification of patients' survival time and the prediction of medical costs.

Acknowledgments

Funding: This research was supported by Shanghai Health Committee (20184Y0016, 201840131), Chongming Science and Technology Committee (CKY2019-6), Xinhua Hospital (Chongming Branch) (2019YA-03, 2019YA-06) and Fund for Chongming Key Discipline of Anesthesiology.

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at https://dx.doi.org/10.21037/tcr-21-896

Data Sharing Statement: Available at https://dx.doi.org/10.21037/tcr-21-896

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/tcr-21-896). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was approved by ethics committee of Xinhua Hospital Chongming Branch (No.: CMEC-2021-KT-24) and was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Informed consent was taken from all the patients.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Chen W, Sun K, Zheng R, et al. Cancer incidence and mortality in China, 2014. Chin J Cancer Res 2018;30:1-12.
3. Ouyang XY, Yang C, Zhu Z, et al. Cancer pain, a serious threat to patientsmemory. Sheng Li Xue Bao 2019;71:343-9.
4. Ouyang X, Zhu Z, Yang C, et al. Epinephrine increases malignancy of breast cancer through p38 MAPK signaling pathway in depressive disorders. Int J Clin Exp Pathol 2019;12:1932-46.
5. Gotze H, Brahler E, Gansera L, et al. Psychological distress and quality of life of palliative cancer patients and their caring relatives during home care. Support Care Cancer 2014;22:2775-82.
6. Liyan C, Zhihui Y, Ruina L, et al. The Positive Experience of Family Caregivers for Terminal Cancer Patients Receiving Home-based Hospoce Care: A Qualitative Study. Nursing Journal of Chinese People's Liberation Army 2019;36:12-4.
7. Forkan ARM, Khalil I. A clinical decision-making mechanism for context-aware and patient-specific remote monitoring systems using the correlations of multiple vital signs. Comput Methods Programs Biomed 2017;139:1-16.
8. Murray NM, Unberath M, Hager GD, et al. Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review. J Neurointerv Surg 2020;12:156-64.
9. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 2017;15:104-16.
10. Diz J, Marreiros G, Freitas A. Applying Data Mining Techniques to Improve Breast Cancer Diagnosis. J Med Syst 2016;40:203.
11. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34:113-27.
12. Sun ASA, Tan YTY, Zhang DZD. editors. Hybrid prediction model based on BP neural network for lung cancer. IEEE International Symposium on It in Medicine & Education; 2008.
13. Liaw A, Wiener M. Classification and Regression by randomForest. R News; 2002;23.
14. Meyer D. Misc Functions of the Department of Statistics

(e1071), TU Wien; 2012.
15. Cortez P. rminer: Simpler use of data mining methods (e.g. NN and SVM) in classification and regression; 2013.
16. Guojian C, Han Z, Junjie W. Application of data mining in breast cancer recurrence prediction. Intelligent Computer and Applications 2019;9:96-9.
17. Zimmermann C, Swami N, Krzyzanowska M, et al. Early palliative care for patients with advanced cancer: a cluster-randomised controlled trial. Lancet 2014;383:1721-30.
18. Temel JS, Greer JA, Muzikansky A, et al. Early palliative care for patients with metastatic non-small-cell lung cancer. N Engl J Med 2010;363:733-42.
19. (EIU) TEIU. The 2015 Quality of Death Index: Ranking of Palliative Care Across the World (2nd Ed.). 2016.
20. Chen YT, Ho CT, Hsu HS, et al. Objective Palliative Prognostic Score Among Patients With Advanced Cancer. J Pain Symptom Manage 2015;49:690-6.
21. Yamada T, Morita T, Maeda I, et al. A prospective, multicenter cohort study to validate a simple performance status-based survival prediction system for oncologists. Cancer 2017;123:1442-52.

(English Language Editor: J. Reynolds)

Cite this article as: Yang C, Yu R, Ji H, Jiang H, Yang W, Jiang F. Application of data mining in the provision of in-home medical care for patients with advanced cancer. Transl Cancer Res 2021;10(6):3013-3019. doi: 10.21037/tcr-21-896

