Prediction of comorbid anxiety and depression using machine learning models in cancer survivors

Rui Yan
Fudan University https://orcid.org/0000-0002-5330-7861
Jiwei Wang
Fudan University
Xiaoguang Yang
Fudan University
Jinming Yu (jmy@fudan.edu.cn)
https://orcid.org/0000-0002-8537-839X

Research article

Keywords: anxiety, depression, cancer survivor, machine learning, prediction model

Posted Date: June 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-32449/v1

Abstract

Background

We aimed to investigate the prevalence and associated factors of comorbid anxiety and depression (CAD) in Chinese cancer survivors (CS), and develop novel prediction model using traditional logistic regression and machine-learning algorithms to predict CS' CAD, which would be helpful to accurately distinguish individuals with high risk of psychological problems.

Methods

A cross-sectional study of 1546 CSs were conducted in Shanghai China. Self-reported structured questionnaire was used to collect information about basic socio-demographic factors, socioeconomic status, life behavior, health conditions, fatigue, social support, anxiety (Zung self-rating anxiety scale) and depression (Zung self - rating depression scale). Stepwise logistic regression and three machine learning algorithms (support vector machine, decision tree, and random forest) were used to construct the prediction model of CS' CAD.

Results

18.24% CSs reported CAD. The AUC of models to predict the high-risk CAD in the training set was best in that of RF (0.839), followed by LR (0.811), SVM (0.808), and DT (0.794). Social support and several comorbid chronic diseases were important top ranking contributing factors of CAD in all the predictive models.

Conclusion

CAD was a common psychological problems in cancer survivors (CSs).The machine leaning techniques and logistic regression showed similar performance to classify Chinese CSs at risk of CAD using questionnaire-based survey. Effective treatment to control the coexisting disease, more social support should be provided to CSs to improve psychological health. Psychological care was recommend to be incorporated into regular health care and management of cancer survivors.

Introduction

Cancer has become one of the major public health problems threatening individuals' health, and it is estimated that one in five global cancer patients live in China(1). An individual is considered as a cancer survivor (CS) since cancer diagnosis, throughout the balance of his or her life (2). Although the better primary health care and continuing medical advances allowed CS could live longer after cancer diagnosis (1), CS still face a range of challenges from physical, social, professional and economic factors. These

challenges may also case psychological issues that might be long-term existed problems throughout CS' entire survival(3).

Anxiety and depression are common psychological problems in CSs(4). Anxiety and depression often co-occur in clinical practice(5). Among individuals with depression, 67% also had a current comorbid anxiety. And among persons with anxiety, 63% had a current comorbid depressive disorder(5). Anxiety or depression was associated with the development and progression of cancer. An individual is considered as a comorbid anxiety and depression (CAD) when she/he has both current anxiety and depression disorder. The contribution of CAD on functional impairment and cancer progression may substantially exceeded the contribution of their independent parts(6, 7). Since the presence of CAD represent one of great psychological burden and distress in CSs, identifying available risk factors of CAD and building prediction models more accurately may be beneficial to early prevention strategies for CSs with high risk of CAD in the most high-risk population.

Screening for anxiety and depression is resource-intensive and depends on access to medical resources. Walker et al. found that 73 percent of cancer patients with depression did not receive effective treatment, and only 5 percent of cancer patients saw a psychologist(8). However, in China, the supply of psychiatrists and psychologist is insufficient, and physician have little time to actively consult patients about their emotional and psychological problems. Physicians were insensitive to patients' emotions and non-verbal information, screening psychological distress in cancer patients through anxiety and depression scale was often regarded as a burden for physicians. So, building an accurate CAD model would help the physicians based on simple indicators to identify those with high-risk of CAD, and those with high CAD risk would be given appropriate advice and transfer treatment.

In recent years, machine learning algorithms had been widely used to predict outcomes and guide decisions in clinical practice. Several studies evaluate individuals' mental health by using common machine learning techniques, such as support vector machine (SVM) and decision trees (DT) and random forest (RF) (9-11), which shown that machine learning algorithms improved the accuracy of prediction. Furthermore, the recent advancement of machine learning puts more importance in explaining/interpreting machine learning trying to open the "black box" of machine learning algorithms(12), particularly probe into the model-based variable importance evaluation to visualize the effect of features on outcome. However, few application of machine learning was found to build prediction models and to classify cancer survivors with high-risk of CAD.

The purpose of this study was to examine the prevalence of CAD basing on a cross-sectional study. We also investigated the associated factors with CAD, and establish prediction models for CAD in Chinese CSs using traditional logistic regression (LR) and several machine learning algorithms, including random forest (RF), decision trees (DT), and support vector machines (SVM). Additionally, by comparing the four predictive models, we also aimed to find several top contributing variables that more associated with the CAD, which may be helpful to develop reference tools assisting doctors to predict mental disorders and support patient care.

Methods

Study population

Data were acquired from a cross-sectional study of 1546 CSs, in Shanghai Cancer Rehabilitation Club (SCRC), from June to September 2018. All the recruited participants were new members registered to SCRC in 2018, pathologic diagnosed with cancer, able to independently participate in the activities of SCRC. Participants were asked to finish a self-reported structured questionnaire including a range of questions about basic socio-demographic factors, socioeconomic status, life behavior, health conditions, social support, anxiety and depression. Informed consent was obtained from each study participant. Our study was approved by the Medical Research Ethics Committee of the school of public health, Fudan University (The international registry NO. IRB00002408 &FWA00002399).

Target variable: Comorbid anxiety and depression (CAD)

As a dependent variable, CAD was considered as CSs who were both anxiety and depression. Anxiety and Depression was assessed by using the Zung self-rating anxiety scale (SAS)(13) and Zung self-rating depression scale (SDS)(14), respectively. Both SAS and SDS were a 20-item self-administrated scale, and each question is scored on a scale of 1 to 4 (rarely, sometimes, frequently, and always). The total score of each scale ranges between 20 and 80, and were then multiplied by 1.25 to obtain a standard scale. SAS standard scores ≥50 indicated anxiety and SDS standard scores ≥53 indicated depression. The respondents who experienced both anxiety and depression were categorized within the CAD group.

Contributing features

Contributing features included basic socio-demographic factors (age, gender and marital status), socioeconomic status (education level, working status and income), life behavior (smoking, drinking, dietary intake frequency of vegetables, fruits, fish, shrimp/crab/shell, eggs, milk, bean products and nuts), health conditions (BMI, comorbid chronic disease, cancer treatment, time since cancer diagnosis, recurrence and metastasis), and social support.

Marital status was divided as married and divorced/widowed/ separated/single. Education level was categorized as less than senior high school, senior high school and above senior high school. Income was categorized as <2000 yuan/month, 2000-4000 yuan/month and ≥4000 yuan/month. BMI was categorized as <18.5 kg/m^2, 18.5–22.9 kg/m^2, 23.0–27.4 kg/m^2 and ≥27.5 kg/m^2 according to the World Health Organization (WHO) recommendation for Asians. Surgery, radiotherapy, chemotherapy, traditional Chinese medicine, biotherapy, recurrence and metastasis were divided as yes and no. Time since cancer diagnosis was categorized as <1 years, 1~3 years, 3~5 years and ≥5 years.

Questionnaire included questions about a list of comorbid chronic diseases (CCD), including hypertension, hyperlipidemia, hyperuricemia, diabetes mellitus, heart and cardiovascular diseases, stroke, respiratory diseases, digestive diseases, and musculoskeletal diseases. And each type of CCD was

categorized as “yes” or “no”. All these CCD must be clinical diagnosed by physician from secondary or tertiary hospitals in China.

Smoking was categorized as never smoked, former smoker and current smoker. Drinking frequency was categorized as no, occasionally and usually. Dietary intake frequency of each food items (vegetables, fruits, eggs, fish and nuts) were obtained through a food frequency questionnaire (FFQ) that included 4 frequency categories for each kind of food (<1 times/week, 1-2 times/week, 3-4 times/week and ≥5 times/week).

The level of physical activity was measured by the long form of the International Physical Activity Questionnaire (IPAQ)(15) , and according to the IPAQ scoring guideline, physical activity level were then categorized into three groups: high, moderate, and low.

Social Support was assessed by the Multidimensional Scale of Perceived Social Support (MSPSS) (16). MSPSS is composed of 12 items to measure perceived social support form family, friends and a significant other. Respondents use a 7-point Likert-type scale (from “very strongly disagree” to “very strongly agree”) with each item. The total MSPSS score was calculated by adding all the item scores together and then dividing by 12, and higher total scores represent higher social support.

Machine learning algorithms to predict CAD

Three machine learning algorithms were used to train models to predict CAD: Support Vector Machine (SVM)(17), Decision Tree (DT)(18) and Random Forest (RF)(19). The data set (n=1546) was randomly divided as a training set (n=1160, 75%) to train prediction models and a testing set (n=386, 25%) to evaluate the real performance of the prediction methods. Since using models with feature selection was more efficient than that searching routine for all external predictors into the model, we applied the feature selection by filter to the entire training set with cross-validate by the R package *caret*. And 13 features (gender, cancer site, hypertension, hyperlipidemia, heart disease, stroke, respiratory diseases, digestive diseases, musculoskeletal diseases, smoking, fish intake frequency, egg intake frequency, and social support) were finally selected using simple univariate statistical methods and was then used to build machine learning models.

A 10-fold cross validation was implemented to tune hyper-parameters and to prevent performance overfitting. This means that, the training dataset was split in 10 equally-sized random folds, at each time, a random subsample containing 90% of the training data was used to train a prediction model, and the remaining 10% part of the training data was used as validation. The above process was repeated 10 times until all folds had served as the test set. Via the 10-fold cross validation, the optimal hyper-parameters were searched through grid search for each machine learning prediction methods. The area under the receiver operating characteristics (ROC) curve (AUC) was used to assess performance during parameter selection. The optimal parameters of each machine learning algorithms selected by grid search were：SVM (cost=3, gamma=0.005); DT (maxdepth=5, minbucket=5, cp=0.005, xval=5); RF (maxnodes=13; mtry=6, ntree=500). These final optimal hyper-parameters were then passed to the

machine learning models and applied to the testing set to evaluate the performance of the prediction methods on new data. Sensitivity, specificity, accuracy and AUC were used as evaluation measures to predict CSs with CAD. For the machine learning models with optimal hyper-parameters, model-based variable importance evaluation was conducted to quantifying the importance of each feature.

Statistical Analysis

Means and standard deviations were calculated for continuous variables, and numbers and percentages were computed for categorical variables. The distribution of CAD among different socio-demographic factors, socioeconomic status, life behavior, health conditions were compared using Chi-square test or Students't-test. Multivariate logistic regression was used to identify factors associated with CAD using the odds ratio (OR) and its corresponding 95% confidence interval (95%CI), adjusted for all other confounders. A stepwise logistic regression(20) was used to build the final model for predicting CAD. The machine learning algorithm was developed using R 3.6.1 with the package *mlr* 2.15. All statistical analyses were performed by R version 3.6.1. A two-sided P value < 0.05 was considered significant.

Results

In our study, the top five cancer site were breast cancer (n=491), lung cancer (n=241), colorectal cancer (n=179), thyroid cancer (n=123) and gastric cancer (n=120). Combining all cancer sites, the prevalence of CAD in CS was 18.24%, and the prevalence of CAD (9.76%-25.71%) varied by cancer site. (Table 1).

The characteristics for a total of 1546 CSs are summarized in Table 2. After adjusted for potential confounding variables, when compared with male CSs, female CSs had higher risk of CAD (OR = 2.12, 95% CI: 1.51-2.97) (Table 2). 50 to 59 years old CSs had higher levels of CAD than those below 50 years of age (OR = 1.80, 95% CI: 1.06-3.07). CSs with underweight had higher risk of CAD than those with normal weight (OR =1.81, 95% CI: 1.01-3.23). CSs underwent chemotherapy had higher risk of CAD than those without chemotherapy. CS with diagnoses of hyperlipidemia(OR = 1.49, 95% CI: 1.13-1.96), Diabetes(OR = 1.47, 95% CI: 1.06-2.05), heart and cardiovascular diseases(OR = 1.93, 95% CI: 1.40-2.66), stroke(OR = 2.07, 95% CI: 1.20-3.56), respiratory diseases (OR =2.29, 95% CI: 1.61-3.27), Digestive diseases (OR =1.38, 95% CI: 1.07-1.79) and musculoskeletal diseases (OR = 2.41, 95% CI: 1.79-3.24) were significantly more likely to report CAD. CS with frequent intake of fish, Shrimp/crab/shell, milk and nuts ≥5 times/week had lower risk of CAD than those with intake of these less than 1 times/week. Decreased risk of CAD was found among those with higher scores of social support (OR =0.97, 95% CI: 0.96-0.99).

Transitional logistic regression and three machine learning methods were used to predict CAD for CS. The predictive performance ROC plots of training set and testing set were shown in Table 3. In the training set, the ROC curve shown that the diagnostic performance of the four logistic regression and machine learning algorithms to predict CAD ranked from high to low were RF, LR, SVM, and DT. In the testing set, RF classifier achieved the best prediction performance with 0.771, 0.701, 0.68 and 0.684 in AUC, accuracy (Acc.), and sensitivity (Sens.) and specificity (Spec.). Similar prediction performance was found in logistic

regression with a sensitivity of 0.675, specificity of 0.693, classification accuracy of 0.689, and AUC of 0.769.

Feature importance

Table 4 shown the feature importance ranking for predicting CAD in logistic regression and the three machine learning models. Seven of the thirteen contributing factors (gender, fish intake frequency, hypertension, digestive diseases, musculoskeletal diseases, stroke and social support) were identified in both logistic regression and machine learning models. Additionally, the RF model also put emphasis on age, marital status, working status, time since diagnosis and radiotherapy.

Discussion

This study examined the association of comorbid anxiety and depression with various factors among and used different machine learning algorithms to predict CAD among 1546 Chinese cancer survivors. Our study indicated that the prevalence of CAD was 18.2%, and random forest (RF) and logistic regression achieved similar predictive performance with 0.771 and 0.769 in AUC in testing set, respectively. Social support and comorbid chronic diseases were important top ranking associated factors with CAD in all the predictive models.

Our results indicated that nearly one in five cancer survivors reported CAD which indicated that CAD had become a severe psychological distress and burden that faced by numerous CSs. As two negative psychological problems, anxiety and depression may decline body's immunity, and be associated with the development and progression of cancer. Anxiety and depression might also lead to lower adherence to treatment, higher suicide risk, and additional health expenditures (21, 22). Compared with CS with simple anxiety or depression, CAD was associated with more complex symptom, and greater negative impact on the cancer treatment effect and CS' quality of life. It is important to found risk factors of CAD, reduce the prevalence of CAD, conduct psychological health promotion, and finally improve CS' psychological health.

Machine learning algorithms have several advantages beyond traditional statistical approaches in building predictive model (23): no linear requirement for data distribution, discovery and utilization of the complex interaction and non-linear relationships among related factors automatically, and high tolerance to outliers and missing values. These advantages allowed machine learning techniques has high application value in predictive classification. In this study, we applied the several machine learning algorithms to develop predictive models to determine CSs at risk of CAD. Among the three machine learning algorithms, in external validation, RF shown the best prediction performance in the testing dataset with an accuracy of 68.4% (AUC = 0.771, 95%CI=0.714, 0.829). However, the performance of RF was similar to the performance of logistic regression (AUC = 0.769, 95%CI=0.711, 0.862). Our study indicated that although machine learning algorithms have the potential to identify cancer survivors at risk of CAD using questionnaire-based survey, however the difference in AUC between LR and random forest was not statistically significant. A previous systematic review indicated that clinical prediction models

based on machine learning would not lead to better AUC than that based on LR(24). Logistic regression is a powerful tool that easy to understand and implement, especially in epidemiologic studies. Future studies would possibly show a better performance if a single machine learning algorithm is combined with various algorithms rather than a single algorithm model.

Machine learning were considered as “black-box”, and had limitations to reveal how features are interacting and what effect they have on the outcome (15, 25). However, the feature importance provided information evaluate the importance ranking of each factor in the classification decision. Six of thirteen selected features were comorbid chronic diseases. Our previous study had shown that comorbid chronic diseases was associated with greater probability of anxiety and depression(26), which indicating that physical problem might had great influence on CSs’ psychological health, and effectively monitoring and management of CCDs should be provided for CSs to address the psychological and psychical problems. Social support was another important predictor of CAD. Social support refers to the material and spiritual support that provided from family, friends and others. The association between social support and anxiety, depression among cancer patients had been proven in previous studies (27, 28). Cancer may had negative influence on individuals’ social relationships and the support the patient receives after cancer diagnosis. Social support may act as a coping resource that could deal with the side effect of cancer treatments and mitigate the adverse psychological issues.

Several limitations of our study should be acknowledged. First, due to the good reliability and validity of SAS and SDS in Chinese population (29, 30), we used questionnaire-based scales to screen anxiety and depression rather than a clinical diagnosis of either. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM -5) should be as gold standard diagnostic instruments in further studies. Second, several important clinical indicators that may associated with psychological health, such as cancer stage and pathological type, were not collected in our study. The performance of our predictive model might be improved if we composed of more detailed basic questions. Third, larger sample size of CSs from multicenter should be included in the future study to verify and improve the stability and reliability of the current machine learning models. At last, the relationship between CAD and several factors, such as physical activity and comorbid chronic diseases, might be bidirectional. Given the cross-sectional nature of current study, we cannot deduce the causal relationships between associated factors and CAD.

Several clinical implication of our study should also be emphasized. A high percentage of Chinese CSs suffered from both anxiety and depression, and it could be beneficial to screen these psychological distress and improve CS’ quality of life. However, in China, the cancer management is more focused on cancer treatment outcome, instead of psychological health. Psychological problems was infrequently discussed or treated, and remain on the periphery of cancer care. Both logistic regression and machine learning models indicated several intervention targets to improve CSs’ psychological health, such as frequently intake of fish and egg, and effective treatment to control the coexisting disease, and provide more social support. Psychological care was strongly recommend to be incorporated into regular health

care and management of cancer survivors. Additionally, the cooperation of oncologist, physician , psychologist, family and others were needed to provide supportive collaborative care.

In conclusion, our study shown that machine learning techniques have the potential to identify Chinese cancer survivors at risk of CAD using questionnaire-based survey of socio-demographic factors, socioeconomic status, life behaviors, health conditions and social support. Random forest model and logistic regression have the similarly good model performance, and would facilitate early screening of high risk of CAD and changing modifiable risk factors.

Declarations

Acknowledgements: Shanghai Cancer Rehabilitation Club provided invaluable resources for field investigation. We are grateful to all involved cancer survivors to participate in this study. We also thank all workers and volunteers involved in the acquisition of data.

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

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflict of Interest: The authors declare that they have no conflict of interest.

Research involving human participants：The study was approved by the Medical Research Ethics Committee of the school of public health, Fudan University (The international registry NO. IRB00002408 &FWA00002399).

Informed consent: Informed consent was obtained from all individual participants included in the study.

Data availability: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Authors ' contributions

Rui Yan and Xiaoguang Yang proposed and designed the study. Rui Yan and Xiaoguang Yang built models and wrote the paper. Rui Yan and Xiaoguang Yang wrote the paper. Rui Yan and Jiwei Wang did ground work of collecting original data and revised the paper. Jinming Yu supervised the study and revised the paper. All authors have read and approved the final manuscript.

Reference

1. Zeng H, Chen W, Zheng R, Zhang S, Ji JS, Zou X, et al. Changing cancer survival in China during 2003-15: a pooled analysis of 17 population-based cancer registries. The Lancet Global health. 2018;6(5):e555-e67.
2. Denlinger CS, Carlson RW, Are M, Baker KS, Davis E, Edge SB, et al. Survivorship: introduction and definition. Clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network : JNCCN. 2014;12(1):34-45.
3. Khan NF, Evans J, Rose PW. A qualitative study of unmet needs and interactions with primary care among cancer survivors. Brit J Cancer. 2011;105:S46-S51.
4. Mitchell AJ, Ferguson DW, Gill J, Paul J, Symonds P. Depression and anxiety in long-term cancer survivors compared with spouses and healthy controls: a systematic review and meta-analysis. Lancet Oncology. 2013;14(8):721-32.
5. Lamers F, van Oppen P, Comijs HC, Smit JH, Spinhoven P, van Balkom AJ, et al. Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the Netherlands Study of Depression and Anxiety (NESDA). J Clin Psychiatry. 2011;72(3):341-8.
6. Dong L, Freedman VA, Mendes de Leon CF. The association of comorbid depression and anxiety symptoms with disability onset in older adults. Psychosom Med. 2019.
7. Lowe B, Spitzer RL, Williams JB, Mussell M, Schellberg D, Kroenke K. Depression, anxiety and somatization in primary care: syndrome overlap and functional impairment. Gen Hosp Psychiatry. 2008;30(3):191-9.
8. Walker J, Hansen CH, Martin P, Symeonides S, Ramessur R, Murray G, et al. Prevalence, associations, and adequacy of treatment of major depression in patients with cancer: a cross-sectional analysis of routinely collected clinical data. Lancet Psychiatry. 2014;1(5):343-50.
9. Omurca SI, Ekinci E. An Alternative Evaluation of Post Traumatic Stress Disorder with Machine Learning Methods. 2015 International Symposium on Innovations in Intelligent Systems and Applications (Inista) Proceedings. 2015:231-7.
10. Chi MY, Guo SW, Ning YP, Li J, Qi HC, Gao MJ, et al. Using Support Vector Machine to Identify Imaging Biomarkers of Major Depressive Disorder and Anxious Depression. Comm Com Inf Sc. 2014;472:63-7.
11. Wade BS, Joshi SH, Pirnia T, Leaver AM, Woods RP, Thompson PM, et al. Random Forest Classification of Depression Status Based On Subcortical Brain Morphometry Following Electroconvulsive Therapy. Proc IEEE Int Symp Biomed Imaging. 2015;2015:92-6.
12. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A. 2019;116(44):22071-80.
13. Zung WW. A rating instrument for anxiety disorders. Psychosomatics. 1971;12(6):371-9.
14. Zung WW. A Self-Rating Depression Scale. Arch Gen Psychiatry. 1965;12:63-70.
15. Hagstromer M, Oja P, Sjostrom M. The International Physical Activity Questionnaire (IPAQ): a study of concurrent and construct validity. Public Health Nutr. 2006;9(6):755-62.

16. Dahlem NW, Zimet GD, Walker RR. The Multidimensional Scale of Perceived Social Support: a confirmation study. J Clin Psychol. 1991;47(6):756-61.
17. Cristianini N, Scholkopf B. Support vector machines and kernel methods - The new generation of learning machines. Ai Mag. 2002;23(3):31-41.
18. Quinlan JR. C4.5: programs for machine learning: Morgan Kaufmann Publishers Inc.; 1993. 302 p.
19. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
20. Gortmaker SL. Applied Logistic Regression. Contemporary Sociology. 1994;23(1):159-.
21. Dauchy S, Dolbeault S, Reich M. Depression in cancer patients. EJC supplements : EJC : official journal of EORTC, European Organization for Research and Treatment of Cancer [et al]. 2013;11(2):205-15.
22. Misono S, Weiss NS, Fann JR, Redman M, Yueh B. Incidence of suicide in persons with cancer. J Clin Oncol. 2008;26(29):4731-8.
23. Boulesteix AL, Schmid M. Machine learning versus statistical modeling. Biom J. 2014;56(4):588-93.
24. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.
25. Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. PloS one. 2019;14(3).
26. Yan R, Xia J, Yang RR, Lv BH, Wu P, Chen WL, et al. Association between anxiety, depression, and comorbid chronic diseases among cancer survivors. Psycho-Oncology. 2019;28(6):1269-77.
27. de Tejada MGZ, Bilbao A, Bare M, Briones E, Sarasqueta C, Quintana JM, et al. Association between social support, functional status, and change in health-related quality of life and changes in anxiety and depression in colorectal cancer patients. Psycho-Oncology. 2017;26(9):1263-9.
28. Manne SL, Kashy DA, Virtue S, Criswell KR, Kissane DW, Ozga M, et al. Acceptance, social support, benefit-finding, and depression in women with gynecological cancer. Qual Life Res. 2018;27(11):2991-3002.
29. Tanakamatsumi J, Kameoka VA. Reliabilities and Concurrent Validities of Popular Self-Report Measures of Depression, Anxiety, and Social Desirability. J Consult Clin Psych. 1986;54(3):328-33.
30. Peng H ZY, Gi Y, Tang W, Li Q, Yan X. Analysis of reliability and validity of Chinese version SDS scale in women of rural area. Shanghai Med Pharm J. 2013;34(14):20-3.

Tables

Table 1 Prevalence of anxiety and depression among CSs by tumor site

Cancer Site	Number	CAD	
		N	%
All sites	1546	282	18.24%
Breast	491	103	20.98%
Lung	241	34	14.41%
Colorectum	179	32	17.88%
Thyroid	123	28	22.76%
Stomach	120	19	15.83%
Cervix	45	9	20.00%
Liver	41	4	9.76%
Uterus	35	9	25.71%
Ovary	34	8	23.53%
Testis	30	5	16.67%
All others [a]	207	31	14.98%

Abbreviations: CAD, comorbid anxiety and depression

[a] All other cancer sites include kidney (N = 28), nasopharynx (N = 22), prostate (N = 21), bladder (N = 20), oral cavity (N = 12), esophagus (N = 10), pancreas (N = 10), gallbladder (N = 8), leukemia (N = 8), melanoma of skin (N = 7), larynx (N = 6), bone (N = 4), brain (N = 2), and other cancer sites (N = 28).

Table 2 Prevalence of CAD in CS among various basic characters

Characteristics	N(%)	CAD		χ^2/t	P	OR(95%CI)[a]
		No N(%)	Yes N(%)			
Gender						
Male	415(26.84)	369(88.92)	46(11.08)			ref
Female	1131(73.16)	895(79.13)	236(20.87)	19.48	**<.001**	**2.12(1.51,2.97)**
Age (years)						
<50	130(8.41)	112(86.15)	18(13.85)			ref
50-59	584(37.77)	453(77.57)	131(22.43)			**1.80(1.06,3.07)**
60-69	733(47.41)	611(83.36)	122(16.64)			1.24(0.73,2.12)
≥70	99(6.40)	88(88.89)	11(11.11)	13.19	**0.004**	0.78(0.35,1.73)
Marital status						
Married	1374(88.87)	1115(81.15)	259(18.85)			ref
Unmarried/widowed/divorced	172(11.13)	149(86.63)	23(13.37)	3.08	0.080	0.67(0.42,1.05)
Education						
<senior high school	865(55.95)	700(80.92)	165(19.08)			ref
senior high school	504(32.6)	411(81.55)	93(18.45)			0.96(0.72,1.27)
>senior high school	177(11.45)	153(86.44)	24(13.56)	3.02	0.221	0.67(0.42,1.06)
Household per capita income (yuan/month)						
<2000	278(17.98)	221(79.50)	57(20.50)			ref
2000~	677(43.79)	561(82.87)	116(17.13)			0.80(0.56,1.14)
≥4000	591(38.23)	482(81.56)	109(18.44)	1.53	0.466	0.88(0.61,1.25)
Working status						
On-the-job	58(3.75)	50(86.21)	8(13.79)			ref
Sick leave	49(3.17)	41(83.67)	8(16.33)			1.22(0.42,3.53)
Unemployed/retired	1439(93.08)	1173(81.51)	266(18.49)	0.95	0.623	1.42(0.66,3.02)
Living alone						
No	1468(94.95)	1199(81.68)	269(18.32)			ref
Yes	78(5.05)	65(83.33)	13(16.67)	0.14	0.712	0.89(0.48,1.64)
BMI (kg/m^2)						
Underweight (<18.5)	63(4.08)	46(73.02)	17(26.98)			**1.81(1.01,3.23)**
Normal weight (18.5–24.9)	1020(65.98)	847(83.04)	173(16.96)			ref
Pre-obese (25.0–29.9)	401(25.94)	322(80.30)	79(19.70)			1.20(0.89,1.61)
Obese (≥30)	62(4.01)	49(79.03)	13(20.97)	5.23	**0.156**	1.30(0.69,2.45)
Time since diagnosis (years)						
<1	86(5.56)	69(80.23)	17(19.77)			ref
1~	604(39.07)	505(83.61)	99(16.39)			0.8(0.45,1.41)
3~	526(34.02)	415(78.9)	111(21.1)			1.09(0.61,1.92)
≥5	330(21.35)	275(83.33)	55(16.67)	4.96	0.175	0.81(0.44,1.49)
Cancer Treatment						
Surgery						
No	139(8.99)	117(84.17)	22(15.83)			ref
Yes	1407(91.01)	1147(81.52)	260(18.48)	0.60	0.440	1.21(0.75,1.94)
Radiotherapy						
No	1206(78.01)	990(82.09)	216(17.91)			ref
Yes	340(21.99)	274(80.59)	66(19.41)	0.40	0.527	1.1(0.81,1.5)
Chemotherapy						
No	549(35.51)	467(85.06)	82(14.94)			ref

Yes	997(64.49)	797(79.94)	200(20.06)	6.23	**0.013**	**1.43(1.08,1.89)**
TMC						
No	686(44.37)	565(82.36)	121(17.64)			ref
Yes	860(55.63)	699(81.28)	161(18.72)	0.20	0.657	1.08(0.83,1.4)
Biotherapy						
No	1453(93.98)	1192(82.04)	261(17.96)			ref
Yes	93(6.02)	72(77.42)	21(22.58)	1.25	0.264	1.33(0.81,2.21)
Recurrance						
No	1454(94.05)	1191(81.91)	263(18.09)			ref
Yes	92(5.95)	73(79.35)	19(20.65)	0.38	0.573	1.18(0.7,1.99)
Metastasis						
No	1463(94.63)	1200(82.02)	263(17.98)			ref
Yes	83(5.37)	64(77.11)	19(22.89)	1.27	0.259	1.36(0.8,2.3)
Comorbid chronic diseases						
Hypertension						
No	1028(66.49)	846(82.3)	182(17.7)			ref
Yes	518(33.51)	418(80.69)	100(19.31)	0.59	0.442	1.11(0.85,1.46)
Hyperlipidemia						
No	1110(71.8)	927(83.51)	183(16.49)			ref
Yes	436(28.2)	337(77.29)	99(22.71)	8.12	**0.004**	**1.49(1.13,1.96)**
Hypeluricemia						
No	1451(93.86)	1194(82.29)	257(17.71)			ref
Yes	95(6.14)	70(73.68)	25(26.32)	4.43	**0.03**	**1.66(1.03,2.67)**
Diabetes						
No	1308(84.61)	1082(82.72)	226(17.28)			ref
Yes	238(15.39)	182(76.47)	56(23.53)	5.28	**0.022**	**1.47(1.06,2.05)**
Heart and cardiovascular						
No	1311(84.8)	1094(83.45)	217(16.55)			ref
Yes	235(15.2)	170(72.34)	65(27.66)	16.49	***0*.001**	**1.93(1.40,2.66)**
Stroke						
No	1481(95.8)	1219(82.31)	262(17.69)			ref
Yes	65(4.2)	45(69.23)	20(30.77)	7.14	**0.008**	**2.07(1.20,3.56)**
Respiratory diseases						
No	1377(89.07)	1148(83.37)	229(16.63)			ref
Yes	169(10.93)	116(68.64)	53(31.36)	21.90	**<0.001**	**2.29(1.61,3.27)**
Digestive diseases						
No	793(51.29)	667(84.11)	126(15.89)			ref
Yes	753(48.71)	597(79.28)	156(20.72)	6.04	**0.014**	**1.38(1.07,1.79)**
Musculoskeletal diseases						
No	1269(82.08)	1072(84.48)	197(15.52)			ref
Yes	277(17.92)	192(69.31)	85(30.69)	35.05	**<0.001**	**2.41(1.79,3.24)**
Number of comorbidity						
0	388(25.1)	335(86.34)	53(13.66)			ref
1-2	694(44.89)	588(84.73)	106(15.27)			1.14(0.8,1.63)
≥3	464(30.01)	341(73.49)	123(26.51)	30.82	**<0.001**	**2.28(1.60,3.25)**
Smoking						
Never	1259(81.44)	1010(80.22)	249(19.78)			ref
Ever	229(14.81)	200(87.34)	29(12.66)			**0.59(0.39,0.89)**
Current	58(3.75)	54(93.1)	4(6.9)	11.77	**0.003**	**0.3(0.11,0.84)**

Drinking						
Never	1177(76.13)	958(81.39)	219(18.61)			ref
Ever	333(21.54)	273(81.98)	60(18.02)			0.96(0.7,1.32)
Current	36(2.33)	33(91.67)	3(8.33)	2.49	0.289	0.4(0.12,1.31)
Physical activity						
Low	240(15.52)	190(79.17)	50(20.83)			ref
Moderate	404(26.13)	323(79.95)	81(20.05)			0.95(0.64,1.42)
High	902(58.34)	751(83.26)	151(16.74)	3.33	0.189	0.76(0.54,1.09)
Sedentary time						
<4 h/day	855(55.30)	713(83.39)	142(16.61)			ref
4-6 h/day	491(31.76)	390(79.43)	101(20.57)			1.30(0.98,1.73)
>6 h/day	200(12.94)	161(80.5)	39(19.5)	3.53	0.171	1.22(0.82,1.8)
Food frequency						
Vegetables						
<1 times/week	44(2.85)	31(70.45)	13(29.55)			ref
1-2 times/week	50(3.23)	42(84)	8(16)			0.45(0.17,1.23)
3-4 times/week	91(5.89)	75(82.42)	16(17.58)			0.51(0.22,1.18)
≥5 times/week	1361(88.03)	1116(82)	245(18)	4.02	0.260	0.52(0.27,1.02)
Fruits						
<1 times/week	80(5.17)	59(73.75)	21(26.25)			ref
1-2 times/week	126(8.15)	98(77.78)	28(22.22)			0.8(0.42,1.54)
3-4 times/week	195(12.61)	170(87.18)	25(12.82)			**0.41(0.22,0.79)**
≥5 times/week	1145(74.06)	937(81.83)	208(18.17)	8.63	**0.035**	0.62(0.37,1.05)
Eggs						
<1 times/week	110(7.12)	84(76.36)	26(23.64)			ref
1-2 times/week	218(14.1)	176(80.73)	42(19.27)			0.77(0.44,1.34)
3-4 times/week	316(20.44)	259(81.96)	57(18.04)			0.71(0.42,1.20)
≥5 times/week	902(58.34)	745(82.59)	157(17.41)	2.73	0.435	0.68(0.42,1.09)
Fish						
<1 times/week	178(11.51)	134(75.28)	44(24.72)			ref
1-2 times/week	549(35.51)	455(82.88)	94(17.12)			**0.63(0.42,0.95)**
3-4 times/week	447(28.91)	368(82.33)	79(17.67)			**0.65(0.43,0.99)**
≥5 times/week	372(24.06)	307(82.53)	65(17.47)	5.71	0.126	**0.65(0.42,0.99)**
Shrimp/crab/shell						
<1 times/week	568(36.74)	438(77.11)	130(22.89)			ref
1-2 times/week	574(37.13)	492(85.71)	82(14.29)			**0.56(0.41,0.76)**
3-4 times/week	230(14.88)	187(81.3)	43(18.7)			0.78(0.53,1.14)
≥5 times/week	174(11.25)	147(84.48)	27(15.52)	15.14	**0.002**	**0.62(0.39,0.98)**
Nuts						
<1 times/week	604(39.07)	472(78.15)	132(21.85)			ref
1-2 times/week	385(24.9)	320(83.12)	65(16.88)			0.73(0.52,1.01)
3-4 times/week	226(14.62)	183(80.97)	43(19.03)			0.84(0.57,1.23)
≥5 times/week	331(21.41)	289(87.31)	42(12.69)	12.70	**0.005**	**0.52(0.36,0.76)**
Milk						
<1 times/week	299(19.34)	231(77.26)	68(22.74)			ref
1-2 times/week	213(13.78)	178(83.57)	35(16.43)			0.67(0.43,1.05)
3-4 times/week	227(14.68)	186(81.94)	41(18.06)			0.75(0.49,1.15)
≥5 times/week	807(52.2)	669(82.9)	138(17.1)	5.24	0.155	**0.70(0.51,0.97)**

Bean						
<1 times/week	320(20.7)	257(80.31)	63(19.69)			ref
1-2 times/week	518(33.51)	430(83.01)	88(16.99)			0.84(0.58,1.20)
3-4 times/week	396(25.61)	323(81.57)	73(18.43)			0.92(0.63,1.34)
≥5 times/week	312(20.18)	254(81.41)	58(18.59)	1.03	0.794	0.93(0.63,1.39)
Social support score		66.18±10.20	62.64±11.35		**<.0001**	**0.97(0.96,0.99)**

Abbreviations: CAD, comorbid anxiety and depression; BMI, body mass index; TMC, traditional Chinese medicine

[a] adjusted for all other factors

Table 3 The predictive performance of each model

	Training set				Testing set			
	Sens	Spec	Acc	AUC(95%CI)	Sens	Spec	Acc	AUC(95%CI)
LR	0.761	0.703	0.715	0.811(0.781-0.841)	0.675	0.693	0.689	0.769(0.711-0.826)
SVM	0.709	0.765	0.753	0.808(0.777-0.838)	0.688	0.605	0.622	0.747(0.693-0.801)
DT	0.725	0.746	0.741	0.794(0.760-0.827)	0.701	0.621	0.637	0.682(0.618-0.747)
RF	0.789	0.735	0.747	0.839(0.812-0.869)	0.701	0.68	0.684	0.771(0.714-0.829)

Abbreviations: Sens, sensitivity; Spec, specificity; Acc, accuracy; AUC, the area under the receiver operating characteristics curve; LR, logistic regression; RF, Random Forest; DT, Decision Tree; SVM, Support Vector Machine

Table 4 Feature importance ranking for predicting CAD in training set

Feature Importance Ranking	LR	RF	DT	SVM
1	Musculoskeletal diseases	Cancer site	Musculoskeletal diseases	Hypertension
2	Fish intake frequency	Heart disease	Cancer site	Hyperlipidemia
3	Gender	Musculoskeletal diseases	Gender	Gender
4	Social Support	Social Support	Smoking	Musculoskeletal diseases
5	Stroke	Fish intake frequency	Hyperlipidemia	Fish intake frequency
6	Marital status	Smoking	Fish intake frequency	Smoking
7	Age	Gender	Stroke	Egg intake frequency
8	work	Digestive diseases	Respiratory diseases	Cancer site
9	Time since diagnosis	Stroke	Social Support	Heart disease
10	Fruit intake frequency	Egg intake frequency	Hypertension	Respiratory diseases
11	Hypertension	Hypertension	Heart disease	Stroke
12	Digestive diseases	Respiratory diseases	Egg intake frequency	Digestive diseases
13	Radiotherapy	Hyperlipidemia	Digestive diseases	Social Support

Abbreviations: LR, logistic regression; RF, Random Forest; DT, Decision Tree; SVM, Support Vector Machine