scientific reports

Check for updates

OPEN

A machine learning-based predictor for the identification of the recurrence of patients with gastric cancer after operation

Chengmao Zhou[1,2,4,5]✉, Junhong Hu[3,4,5], Ying Wang[1,4], Mu-Huo Ji[1,4], Jianhua Tong[1,4], Jian-Jun Yang[1,2,4]✉ & Hongping Xia[1,2,3,4]✉

To explore the predictive performance of machine learning on the recurrence of patients with gastric cancer after the operation. The available data is divided into two parts. In particular, the first part is used as a training set (such as 80% of the original data), and the second part is used as a test set (the remaining 20% of the data). And we use fivefold cross-validation. The weight of recurrence factors shows the top four factors are BMI, Operation time, WGT and age in order. In training group:among the 5 machine learning models, the accuracy of gbm was 0.891, followed by gbm algorithm was 0.876; The AUC values of the five machine learning algorithms are from high to low as forest (0.962), gbm (0.922), GradientBoosting (0.898), DecisionTree (0.790) and Logistic (0.748). And the precision of the forest is the highest 0.957, followed by the GradientBoosting algorithm (0.878). At the same time, in the test group is as follows: the highest accuracy of Logistic was 0.801, followed by forest algorithm and gbm; the AUC values of the five algorithms are forest (0.795), GradientBoosting (0.774), DecisionTree (0.773), Logistic (0.771) and gbm (0.771), from high to low. Among the five machine learning algorithms, the highest precision rate of Logistic is 1.000, followed by the gbm (0.487). Machine learning can predict the recurrence of gastric cancer patients after an operation. Besides, the first four factors affecting postoperative recurrence of gastric cancer were BMI, Operation time, WGT and age.

The global incidence of gastric cancer is the fourth in malignant tumors and the second in mortality. There are nearly 1 million cases of new gastric cancer worldwide each year, of which nearly 50.0% occur in China. The prognosis of early gastric cancer is good, but the clinical symptoms are atypical and the signs are not obvious. The postoperative recurrence rate (40.0–70.0%) remained high[1]. It has been reported that the average time to recurrence of gastric cancer was 20.5–28.0 months after operation[2]. However, when gastric cancer recurs after the operation, some chemotherapy and immunotherapy can be used in close cooperation to control cancer and reduce necrosis, to create conditions and strive for operation.

In recent years, big data and machine learning have led to innovative changes in many industries. With the development of precision medicine plan, the combination of health and medical big data and machine learning brings people the imagination space of the future big data health cause. The method of machine learning is especially suitable for prediction based on existing data. By capturing complex nonlinear relations in the data, the machine learning algorithm can improve the accuracy of prediction more than the conventional regression model. At present, machine learning can predict the survival of breast cancer patients at an early stage[3]. Using machine learning technique can predict the malignant degree of breast lesions[4]; Machine learning was used to better predict early biochemical recurrence[5]; Machine learning is a branch of artificial intelligence, which has

[1]Department of Anesthesiology, Pain and Perioperative Medicine, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450000, China. [2]School of Medicine, Southeast University, Nanjing 210009, China. [3]Department of Colorectal and Anal Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450000, China. [4]Department of Pathology, School of Basic Medical Sciences, Sir Run Run Hospital, State Key Laboratory of Reproductive Medicine, Key Laboratory of Antibody Technique of National Health Commission, Nanjing Medical University, Nanjing 211166, China. [5]These authors contributed equally: Chengmao Zhou and Junhong Hu. ✉email: zhouchengmao187@foxmail.com; yjyangjj@126.com; xiahongping@njmu.edu.cn

been used in tumor risk assessment, lesion detection, prognosis prediction and treatment response[6]; Deep learning can effectively solve medical problems that were previously thought unsolvable[7].

In conclusion, postoperative recurrence of gastric cancer are key factors affecting the prognosis of gastric cancer, and it is very important to actively explore the adverse factors affecting postoperative recurrence of gastric cancer to detect and evaluate the recurrence or metastasis of postoperative gastric cancer. Therefore, we used a machine learning technique to predict the tumor recurrence of gastric cancer patients after operation.

Materials and method

Data source. Data is available at BioStudies database, accession numbers: S-EPMC4344235. This study included 2012 patients.

Data from the retrospective studies included age, gender, pathological characteristics, treatment-related factors, and the follow-up period related to survival status.

Machine learning algorithm. Logistic regression, a kind of generalized linear regression analysis model, is often used in such fields as automatic disease diagnosis and economic prediction.

The decision tree algorithm belongs to the category of supervisory learning.

In machine learning, a random forest is a kind of classification/regression which contains multiple decision trees (CART tree). The final classification result is decided by each decision tree vote/ average, that is, a few obey the principle of most. The stochastic forest in turn corresponds to the fusion of the model of several decision trees (CART trees).

The GBDT is also called MART. It is an iterative decision tree algorithm. The trees in GDBT are all regression trees. Only the accumulation of the results of regression trees is meaningful, and the addition of the results of classification is not meaningful.

The light GBM (light gradient boosting machine) is a framework to implement GBDT algorithm, which supports efficient parallel training.

Data processing. Data were processed in R (3.5.3) language. $P < 0.05$ was taken as the difference with statistical significance; Multiple imputations were used for missing variables. The machine learning was analyzed by python (3.6.5). The total population was randomly divided into a training group and test group according to the ratio of 8:2. The available data is divided into two parts (sometimes called training-test segmentation). In particular, the first part is used as a training set (such as 80% of the original data), and the second part is used as a test set (the remaining 20% of the data). Then, a prediction model is established by using the training set. To get the best model, manual parameter adjustment and grid search are used. And we use fivefold cross validation[8–10]. Then the trained model is applied to the test set for prediction. Choose the best model according to its performance on the test set[11]. And the data are normalized by us. The parameters of the machine learning model are shown in Supplementary Table 1.

Ethics approval. Because this is only a secondary data analysis study using public databases, there is no need to apply for ethics[25].

Results

Correlation analysis and feature analysis. Comparison of basic indicators of patients in the two groups: there was no statistically significant difference in age and height between the two groups ($P = 0.697$ and $P = 0.982$). See Table 1.

The results of gbm algorithm showed that the first 4 factors were ranked in order: BMI, Operation time, WGT and age, respectively (Figs. 1 and 2).

Training set results. In training group: among the 5 models, the accuracy of gbm was 0.891, followed by gbm algorithm was 0.876; The AUC values of the five algorithms are from high to low as forest (0.962), gbm (0.922), GradientBoosting (0.898), DecisionTree (0.790) and Logistic (0.748). The precision of forest is the highest 0.957, followed by the GradientBoosting algorithm (0.878).The recall rate for forest is up to 0.478, followed by the gbm (0.451) (Table 2 and Fig. 3).

Test set results. In the test group: among the five algorithm models, the highest accuracy of Logistic was 0.801, followed by forest algorithm and gbm; The AUC values of the five algorithms are forest (0.795), GradientBoosting (0.774), DecisionTree (0.773), Logistic (0.771) and gbm (0.771), from high to low. The highest precision rate of Logistic is 1.000, followed by the gbm (0.487).The highest recall rate was 0.309 for the DecisionTree, followed by the gbm algorithm (0.235) (Table 3 and Fig. 4).

Discussion

Currently, the treatment of early gastric cancer mainly involves an open operation. According to the location of gastric cancer and the size of lesions, the proximal or distal subtotal resection or total gastrectomy is selected. Mortality is very high in patients with gastric cancer due to its high morbidity and recurrence rate. Therefore, it is very important to find out the factors affecting recurrence and metastasis to reduce the mortality of gastric cancer. After identifying the influencing factors, correct and effective prevention methods are adopted to treat the patients, so that the postoperative recurrence rate can be effectively reduced and the postoperative quality of life can be greatly improved. The first 4 important factors affecting postoperative recurrence of gastric cancer

RECURR	No	Yes	P-value
N	1607	405	
Age (year)	58.5 ± 11.5	58.5 ± 12.2	0.697
Weight (kg)	61.4 ± 10.2	60.0 ± 10.4	0.011
Height (cm)	162.2 ± 8.4	162.3 ± 8.4	0.982
BMI (kg/m^2)	23.3 ± 3.2	22.8 ± 3.1	0.004
Operation time (min)	168.6 ± 52.0	183.3 ± 55.2	< 0.001
Tumor size (cm)	4.3 ± 2.8	6.7 ± 3.4	< 0.001
Sex			0.980
Male	1110 (69.1%)	280 (69.1%)	
Female	497 (30.9%)	125 (30.9%)	
Location			< 0.001
Upper	234 (14.6%)	65 (16.0%)	
Middle	499 (31.1%)	108 (26.7%)	
Lower	860 (53.5%)	204 (50.4%)	
Whole	14 (0.9%)	28 (6.9%)	
Extent of LN dissection			0.409
D2	131 (8.2%)	28 (6.9%)	
D1 plus	1476 (91.8%)	377 (93.1%)	
Chemotherapy			< 0.001
NO	950 (59.1%)	89 (22.0%)	
YES	657 (40.9%)	316 (78.0%)	

Table 1. Baseline data.

were BMI, Operation time,WGT and age, respectively. And machine learning can predict the recurrence of gastric cancer after the operation.

The risk factors affecting the recurrence of gastric cancer are clinical, pathological and biomolecule. The later the clinical stage, the greater the probability of early recurrence, the shorter the survival time. Studies[12] have shown that elevated early gastric cancer, infiltration depth to the submucosa, and concomitant lymph node metastasis are independent factors affecting postoperative recurrence of early gastric cancer. It was also found that age and macroscopic appearance of the tumor was associated with postoperative recurrence of early gastric cancer, while gender, tumor location, tumor size and histological type were not associated with postoperative recurrence of early gastric cancer[13]. Moriguchi et al.[14] found that regional lymph node metastasis was a risk factor for gastric cancer recurrence. The results of this study were also similar.

Kattan et al.[14] demonstrated a poor prognosis in the upper third of tumors. Our study also showed a correlation between tumor location and tumor recurrence.

Recently, studies have demonstrated the survival benefit of adjuvant chemotherapy after radical resection of gastric cancer[1]. However, in this study, chemotherapy was not negatively associated with tumor recurrence. This negative result may be influenced by the reason and the regimen and indication of adjuvant chemotherapy at each institution.

Bickenbach et al.[15] concluded that high BMI was a postoperative complication of gastric cancer but not of long-term survival. Dhar et al.[16] reported that high BMI was not conducive to the removal of gastric lymph nodes in 787 patients with gastric cancer. The results of Tokunaga et al.[17] showed that the survival rate of patients with gastric cancer with high BMI was higher than that of patients with low BMI. Kruhlikava et al.[18] concluded that BMI did not affect survival in patients with esophagogastric cancer. Migita et al.[19] showed that underweight is a simple and reliable predictor of poor long-term prognosis in patients with gastric cancer. Kulig et al.[20] reported that the median disease-related survival time of patients with high BMI was significantly longer than that of patients with low BMI. The results of this study showed that BMI was negatively correlated with tumor recurrence.

This study is addressed by the classification task. It should be to use accuracy (Ac), sensitivity (Sn), specificity (Sp) Matthews coefficient correlation (MCC) and AUC[21–23]. However, it appears that only 20% of the data correspond to the positive class. The imbalance (1:4) makes ROC, AUC-ROC, and especially accuracy, less useful in assessing the utility of a predictor. So we have adopted the AUC chart with the highest precision. And this imbalance in classification is normal in the application of machine learning-related medicine because the incidence and non-incidence of diseases are unbalanced.

Although machine learning has yielded good results in predicting the postoperative recurrence of gastric cancer, the present study has some limitations. Some patients were excluded due to lack of data, which may lead to selection bias. Besides, due to retrospective data, our study failed to refine the prediction of recurrence in some subgroups of the postoperative gastric cancer population, such as patients with gastric cancer combined with other malignant tumors and patients with gastric cancer with other special medical histories, which may cause some applicability of the study results. Further prospective studies on this aspect are needed in the future.

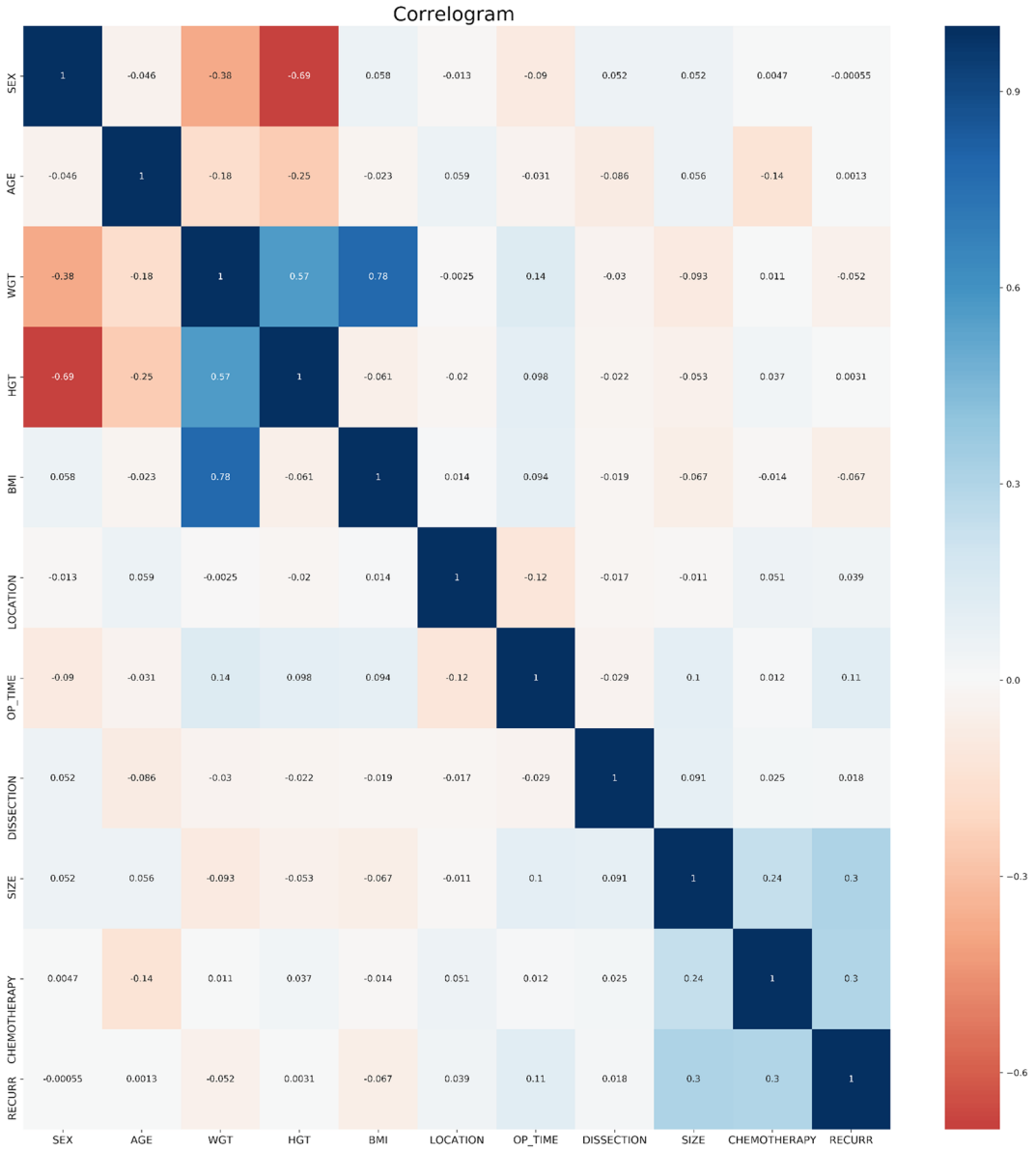

Figure 1. Correlation between variables.

Conclusion

To sum up, machine learning can predict the recurrence of patients with gastric cancer after an operation. Besides, the first four factors affecting postoperative recurrence of gastric cancer were BMI, Operation time, WGT and age.

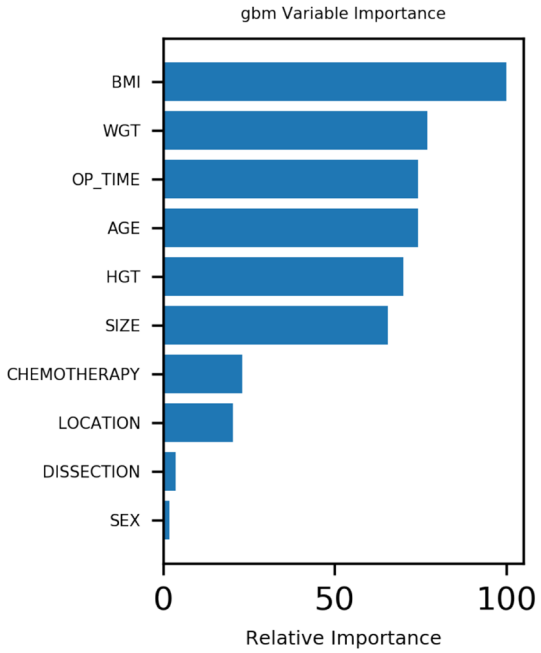

Figure 2. Variable importance of features included in the machine learning algorithm for prediction of recurrence of patients with gastric cancer after operation. *Note*: gbm: LightGBM.

	Accuracy	Precision	Recall	AUC
Logistic	0.799	0.500	0.003	0.748
DecisionTree	0.807	0.529	0.361	0.790
Forest	0.891	0.957	0.478	0.962
GradientBoosting	0.868	0.878	0.401	0.898
gbm	0.876	0.869	0.451	0.922

Table 2. Forecast results for training group.

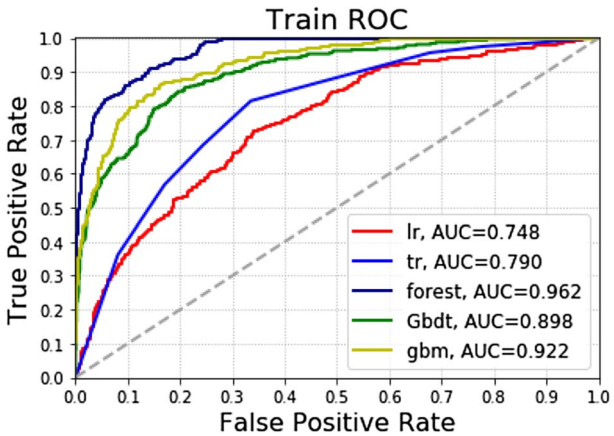

Figure 3. Different machine learning algorithms predict the recurrence of patients with gastric cancer after the operation in the training group.

	Accuracy	Precision	Recall	AUC
Logistic	0.801	1.000	0.012	0.771
DecisionTree	0.782	0.439	0.309	0.773
forest	0.797	0.480	0.148	0.795
GradientBoosting	0.779	0.367	0.136	0.774
gbm	0.797	0.487	0.235	0.771

Table 3. Forecast results for testing group.

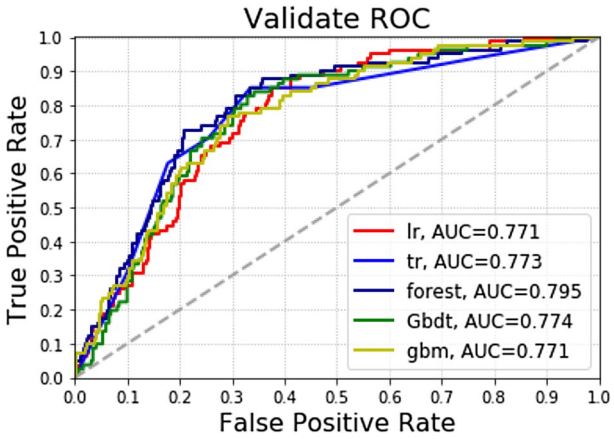

Figure 4. Different machine learning algorithms predict the recurrence of patients with gastric cancer after the operation in the test group.

Data availability

Data is available at BioStudies database, accession numbers: S-EPMC4344235.

Received: 12 June 2020; Accepted: 5 January 2021
Published online: 15 January 2021

References

1. Noh, S. H. *et al.* Adjuvant capecitabine plus oxaliplatin for gastric cancer after D2 gastrectomy (CLASSIC): 5-year follow-up of an open-label, randomised phase 3 trial. *Lancet Oncol.* **15**, 1389–1396. https://doi.org/10.1016/s1470-2045(14)70473-5 (2014).
2. Wu, B., Wu, D., Wang, M. & Wang, G. Recurrence in patients following curative resection of early gastric carcinoma. *J. Surg. Oncol.* **98**, 411–414. https://doi.org/10.1002/jso.21133 (2008).
3. Tahmassebi, A. *et al.* Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. *Investig. Radiol.* **54**, 110–117. https://doi.org/10.1097/rli.0000000000000518 (2019).
4. Uhlig, J. *et al.* Novel breast imaging and machine learning: predicting breast lesion malignancy at cone-beam CT using machine learning techniques. *Am. J. Roentgenol.* **211**, W123–W131. https://doi.org/10.2214/ajr.17.19298 (2018).
5. Wong, N. C., Lam, C., Patterson, L. & Shayegan, B. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. *BJU Int.* **123**, 51–57. https://doi.org/10.1111/bju.14477 (2019).
6. Cuocolo, R., Caruso, M., Perillo, T., Ugga, L. & Petretta, M. Machine learning in oncology: a clinical appraisal. *Cancer Lett.* **481**, 55–62 (2020).
7. Shimizu, H. & Nakayama, K. I. Artificial intelligence in oncology. *Cancer Sci* **111**, 1452 (2020).
8. Charoenkwan, P., Yana, J., Schaduangrat, N., Nantasenamat, C. & Shoombuatong, W. iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics* **112**, 2813–2822 (2020).
9. Charoenkwan, P., Nantasenamat, C., Hasan, M. M. & Shoombuatong, W. iTTCA-Hybrid: improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. *Anal. Biochem.* **599**, 113747 (2020).
10. Charoenkwan, P., Nantasenamat, C., Hasan, M. M. & Shoombuatong, W. Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. *J. Comput. Aided Mol. Des.* **34**, 1105–1116 (2020).
11. Laengsri, V., Shoombuatong, W., Adirojananon, W., Nantasenamat, C. & Nuchnoi, P. ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC Med. Inform. Decis. Mak.* **19**, 212 (2019).
12. Lo, S. S. *et al.* Surgical results of early gastric cancer and proposing a treatment strategy. *Ann. Surg. Oncol.* **14**, 340–347. https://doi.org/10.1245/s10434-006-9077-x (2007).
13. Lai, J. F. *et al.* Prediction of recurrence of early gastric cancer after curative resection. *Ann. Surg. Oncol.* **16**, 1896–1902. https://doi.org/10.1245/s10434-009-0473-x (2009).
14. Moriguchi, S., Maehara, Y., Korenaga, D., Sugimachi, K. & Nose, Y. Risk factors which predict pattern of recurrence after curative surgery for patients with advanced gastric cancer. *Surg. Oncol.* **1**, 341–346. https://doi.org/10.1016/0960-7404(92)90034-i (1992).
15. Bickenbach, K. A. *et al.* Impact of obesity on perioperative complications and long-term survival of patients with gastric cancer. *Ann. Surg. Oncol.* **20**, 780–787. https://doi.org/10.1245/s10434-012-2653-3 (2013).

16. Dhar, D. K. *et al.* Body mass index determines the success of lymph node dissection and predicts the outcome of gastric carcinoma patients. *Oncology* **59**, 18–23. https://doi.org/10.1159/000012131 (2000).
17. Tokunaga, M. *et al.* Better 5-year survival rate following curative gastrectomy in overweight patients. *Ann. Surg. Oncol.* **16**, 3245–3251. https://doi.org/10.1245/s10434-009-0645-8 (2009).
18. Kruhlikava, I., Kirkegård, J., Mortensen, F. V. & Kjær, D. W. Impact of body mass index on complications and survival after surgery for esophageal and gastro-esophageal-junction cancer. *Scand. J. Surg.* **106**, 305–310. https://doi.org/10.1177/1457496916683097 (2017).
19. Migita, K. *et al.* Impact of being underweight on the long-term outcomes of patients with gastric cancer. *Gastric Cancer* **19**, 735–743. https://doi.org/10.1007/s10120-015-0531-y (2016).
20. Kulig, J. *et al.* Implications of overweight in gastric cancer: a multicenter study in a Western patient population. *Eur. J. Surg. Oncol.* **36**, 969–976. https://doi.org/10.1016/j.ejso.2010.07.007 (2010).
21. Shoombuatong, W., Hongjaisee, S., Barin, F., Chaijaruwanich, J. & Samleerat, T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput. Biol. Med.* **42**, 885–889 (2012).
22. Su, W. T., Nalini, S., Virapong, P., Chanin, N. & Watshara, S. PAAP: a web server for predicting antihypertensive activity of peptides. *Future Med. Chem.* **10**, 1749–1767 (2018).
23. Win, T. S. *et al.* HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* **9**, 275–291 (2017).
24. Eom, B. W. *et al.* Survival nomogram for curatively resected Korean gastric cancer patients: multicenter retrospective analysis with external validation. *PLoS ONE* **10**, e0119671. https://doi.org/10.1371/journal.pone.0119671 (2015).
25. Alarcon-Ruiz, C. A., Heredia, P. & Taype-Rondan, A. Association of waiting and consultation time with patient satisfaction: secondary-data analysis of a national survey in Peruvian ambulatory care facilities. *BMC Health Serv. Res.* **19**, 439. https://doi.org/10.1186/s12913-019-4288-6 (2019).

Acknowledgements

We are grateful to the BioStudies database for providing the data[24].

Author contributions

Z.C.M., H.J.H, W.Y., J.M.H., Y.J.J and X.H.P. wrote the main manuscript text and T.J.H. prepared Figs. 1, 2, 3 and 4. All authors reviewed the manuscript.

Funding

This study was supported by grants from the Postgraduate Research&Practice Innovation Program of Jiangsu Province (No. KYCX19_0113) and "Innovative and Entrepreneurial Team" Grant and Southeast University-Nanjing Medical University Cooperative Research Project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-81188-6.

Correspondence and requests for materials should be addressed to C.Z., J.-J.Y. or H.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.