BMJ Open

Serum levels of chemical elements in esophageal squamous cell carcinoma in Anyang, China: a case-control study based on machine learning methods

Tong Lin,[1] Tiebing Liu,[2] Yucheng Lin,[1] Chaoting Zhang,[3] Lailai Yan,[4] Zhongxue Chen,[5] Zhonghu He,[3] Jingyu Wang[4]

To cite: Lin T, Liu T, Lin Y, *et al*. Serum levels of chemical elements in esophageal squamous cell carcinoma in Anyang, China: a case-control study based on machine learning methods. *BMJ Open* 2017;**7**:e015443. doi:10.1136/bmjopen-2016-015443

► Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-015443).

TL and TL contributed equally.

Received 13 December 2016
Revised 31 May 2017
Accepted 20 June 2017

For numbered affiliations see end of article.

Correspondence to
Zhonghu He; zhonghuhe@foxmail.com and Dr Jingyu Wang; wjy@bjmu.edu.cn

ABSTRACT

Objectives Esophageal squamous cell carcinoma (ESCC) is the predominant form of esophageal carcinoma with extremely aggressive nature and low survival rate. The risk factors for ESCC in the high-incidence areas of China remain unclear. We used machine learning methods to investigate whether there was an association between the alterations of serum levels of certain chemical elements and ESCC.

Settings Primary healthcare unit in *Anyang* city, Henan Province of China.

Participants 100 patients with ESCC and 100 healthy controls matched for age, sex and region were included.

Primary and secondary outcome measures Primary outcome was the classification accuracy. Secondary outcome was the p Value of the t-test or rank-sum test.

Methods Both traditional statistical methods of t-test and rank-sum test and fashionable machine learning approaches were employed.

Results Random Forest achieves the best accuracy of 98.38% on the original feature vectors (without dimensionality reduction), and support vector machine outperforms other classifiers by yielding accuracy of 96.56% on embedding spaces (with dimensionality reduction). All six classifiers can achieve accuracies more than 90% based on the single most important element Sr. The other two elements with distinctive difference are S and P, providing accuracies around 80%. More than half of chemical elements were found to be significantly different between patients with ESCC and the controls.

Conclusions These results suggest clear differences between patients with ESCC and controls, implying some potential promising applications in diagnosis, prognosis, pharmacy and nutrition of ESCC. However, the results should be interpreted with caution due to the retrospective design nature, limited sample size and the lack of several potential confounding factors (including obesity, nutritional status, and fruit and vegetable consumption and potential regional carcinogen contacts).

Strengths and limitations of this study

- The classification accuracies achieved by machine learning methods are remarkably higher than most empirical decisions on this small corpus of 100 patients with esophageal squamous cell carcinoma and 100 healthy comparison subjects; in addition, the test error rates provide the quantitative confidence of the prediction results.
- This diagnosis procedure is not expensive and can be conducted in a short period of time, making it possible for clinical use.
- A major limitation of the present work is that the study is retrospective with a relatively small patient cohort. This framework should be evaluated with a larger patient cohort before any real clinical applications are adopted.

INTRODUCTION

Oesophageal cancer (EC) is a cancer with extremely aggressive nature and low survival rate; it has been one of the deadliest cancers worldwide. In 2012, an estimate of 455 800 EC cases and 400 200 deaths occurred in the world.[1 2] Esophageal squamous cell carcinoma (ESCC) is the predominant form of esophageal carcinoma globally; most patients diagnosed as in advanced stages are not amenable to curative treatment. Major risk factors include poor nutritional status, low intake of fruits and vegetables, smoking, alcohol consumption, hot tea drinking, poor oral health, gastro-oesophageal reflux disease, overweight and obesity.[3 4] However, the risk factors for ESCC in the high-incidence areas, such as north-central China, remain unclear.

In addition, most diagnosed patients with ESCC already have had locally advanced EC or distant metastases due to lack of early signs or symptoms.[5] Therefore, the diagnosis test that is practical, non-invasive and can be easily performed is of great interest; new methods like F-fluorodeoxyglucose-positron emission tomography (F-FDG PET) have emerged for the initial staging of patients with EC.[5] Currently, a large body of research in this area aim to identify new biomarker candidates for cancers,

Table 1 Demographic characteristics of normal controls and patients with ESCC from Anyang, China, 2010

Variable	Case (n=100) n (%)	Control (n=100) n (%)	p Value*
Age (years)			
Median (IQR)	56 (55–62)	59 (55–63)	
Gender			
Male	60 (60)	60 (60)	
Female	40 (40)	40 (40)	
History of regular alcohol consumption			
No	82 (82)	81 (81)	0.856
Yes	18 (18)	19 (19)	
History of regular cigarette smoking			
No	54 (54)	57 (57)	0.669
Yes	46 (46)	43 (43)	
Family history of ESCC			
No	71 (71)	83 (83)	0.044
Yes	29 (29)	17 (17)	

*p Values derived from the X^2 test.
ESCC, esophageal squamous cell carcinoma.

such as prostate cancer,[6] breast cancer,[7] lung cancer[8] and gastrointestinal neoplasia.[9]

Chemical elements play essential roles in the biological processes. A number of studies have shown that changes of chemical elements levels might be linked to the risk of some cancers,[10 11] including EC.[12] However, very few relevant studies and only Se, Cu and Zn have been conducted.[12–14] In addition, many chemical elements, such as Mo, Ni, PR, Rb, Sb, Sn, Sr, Th, Ti, Tl, U and V, have not been incorporated. The underlying interactions among these chemical elements can be complex; traditional single variable analysis or correlation analysis may lack the capability to have accurate predictions. Recently, machine learning techniques, such as support vector machines (SVM) and feature selection methods, are gaining popularity in this field for handling high-dimensional input features and yielding better diagnostic accuracies.[15]

In this study, based on recent machine learning techniques and classical statistical methods, a 1:1 matched case–control design was conducted to probe the differences in the serum levels of 38 relatively common chemical elements between patients with ESCC and healthy comparison subjects.

MATERIALS AND METHODS

In the following subsections, we will describe the ESCC serum sample acquisition and preprocessing, evaluation protocol, the main ideas behind dimensionality reduction and classification algorithms, and statistical hypothesis testing. The design and analysis for this study were followed the suggestions from the (strengthening the reporting of observational studies in Epidemiology STROBE guidelines.[16]

Sample collection

At the Cancer Hospital of *Anyang* city, Henan Province of China, 100 patients newly diagnosed with early-stage ESCC were consecutively recruited in 2010. During the same period, 100 age, sex and region-matched healthy comparison subjects were randomly selected from a cohort study[17] about ESCC conducted in Anyang city. Demographic data, personal information and blood samples were obtained from the two groups: patients with ESCC and healthy controls. Specifically, only samples at least 1 week prior to the esophagectomy of patients with ESCC were considered in this case. Patients who suffered from ESCC with some other cancers were excluded from the study. Then each blood sample was centrifuged at 3000 rpm in a 15 min endurance, and separated and stored at −20°C. This study was approved by the Institutional Review Board of Peking University School of Oncology, Beijing. Informed consents were obtained from all participants.

In this study, we defined regular cigarette smoking as a history of smoking at least one cigarette per day for 12 months or 18 packs for 1 year, and regular alcohol consumption was defined as drinking Chinese liquor at least twice per week for 12 months (regular consumption of other beverages such as beer or red wine is very rare in this local area).

Table 2 Classification accuracies (in percentage) and runtime (in seconds) of the patient with ESCC

	Origin	FFS	PCA	FDA	FDAx	LPP	FA
NB	91.70 (0.14)	93.35 (0.12)	90.75 (0.17)	93.90 (0.09)	54.45 (0.18)	89.40 (0.21)	88.95 (0.22)
LR	95.89 (1.70)	94.89 (0.58)	94.05 (0.39)	94.99 (0.47)	94.10 (0.63)	91.31 (0.27)	94.22 (0.53)
NN	97.01 (5.30)	95.05 (12.3)	93.93 (6.40)	94.81 (7.00)	94.33 (6.80)	91.44 (8.50)	94.54 (8.50)
AB	96.05 (76.8)	95.19 (17.3)	88.59 (18.2)	94.26 (1.80)	94.68 (19.4)	71.56 (6.30)	87.84 (12.1)
SVM	97.23 (2.50)	96.56 (1.53)	94.15 (2.30)	94.90 (0.80)	92.15 (1.40)	92.15 (1.50)	94.25 (1.50)
RF	98.38 (16.3)	95.40 (15.0)	91.51 (17.7)	94.88 (15.8)	94.23 (18.8)	89.87 (18.5)	91.94 (17.7)

AB, AdaBoost; ESCC, esophageal squamous cell carcinoma; FA, factor analysis; FDA, Fisher discriminant analysis; FDAx, FDA with its variant; FFS, Fisher feature selection; LPP, locality preserving projection; LR, logistic regression; NB, Naive Bayes; NN, neural network; PCA, principal component analysis; RF, Random Forest; SVM, support vector machine.

Table 3 The projection coefficients w in Fisher discriminant analysis and the Fisher discriminant ratio F used in Fisher feature selection

Feature	w	F	Feature	w	F	Feature	w	F	Feature	w	F
Age	−0.7	0.1	Bi	−1.9	19.7	Se	−0.7	5.5	Rb	−0.9	26.0
Gender	0.0	0.0	Cs	1.0	41.5	Sr	5.0	340.1	Hg	−2.2	37.5
Smoking	−0.1	0.2	Th	0.5	0.2	Li	−1.5	9.9	Pb	−0.7	6.6
Drinking	−0.2	0.0	U	−2.3	72.1	Ni	0.7	1.5	Ca	−2.0	66.4
Family history	−0.1	4.1	La	1.5	5.2	Mo	0.4	1.6	Fe	−0.4	0.3
Be	−0.4	1.1	Ce	2.3	1.1	Ag	0.3	0.0	K	1.3	96.5
B	1.0	68.3	V	−1.4	41.7	Cd	−1.6	3.2	Mg	0.3	96.5
Ai	−0.3	0.0	Cr	−1.5	13.3	Sn	−0.3	14.5	Na	−1.0	43.0
Ti	0.7	47.0	Mn	1.3	5.5	Ba	−0.6	6.8	P	1.7	135.4
Ge	−0.7	17.8	Cu	−1.4	39.4	Pt	0.1	0.0	S	4.7	173.1
As	−1.1	17.8	Zn	1.3	22.1	Ti	−2.1	31.8			

The top two elements with more discriminant information are Sr and S that are marked with bold font. Other important elements (including P, U, Ca, Tl, Bi and Hg) are marked with italic font.

Elements measurement

Each serum sample was put into a quartz tube and 0.3 mL purified HNO_3 (nitric acid) was added. After predigestion at room temperature for 2 hours, 0.5 mL H_2O_2 was added to promote further digestion. The tubes were then placed in a microwave digestion system (Ultrawave, Milestone, Italy) and diluted to 7 mL with deionized water and then diluted to 15 mL with deionised water before analyses. Concentrations of calcium (Ca), magnesium (Mg), potassium (K), phosphorus (P) and sodium (Na) were determined by inductively coupled plasma-atomic emission spectrometry (ICP-AES, American Thermo Electron Corporation iCAP-6300). Also, the levels of other 33 elements, including iron (Fe), selenium (Se), copper (Cu), zinc (Zn), aluminium (Al), manganese (Mn), arsenic (As), molybdenum (Mo), vanadium (V), chromium (Cr), nickel (Ni), lead (Pb), cadmium (Cd), beryllium (Be), boron (B), titanium (Ti), germanium (Ge), strontium (Sr), lithium (Li), silver (Ag), cadmium (Cd), stannum (Sn), barium (Ba), platinum (Pt), thallium (Tl), bismuth (Bi), caesium (Cs), thorium (Th), uranium (U), lanthanum (La), cerium (Ce), rubidium (Rb) and mercury (Hg) were measured by ICP-MS (American PerkinElmer ELAN DRC □). Particularly, the concentrations of six elements are very low (possibly below the detection limit of the spectrometry): Be, Cd, Pt, U, V and Hg. However, we did not directly remove these six 'nuisance' low-concentration elements; instead, these elements were retained to serve as noise for testing the robustness of our algorithm.[18 19]

Quality control

The following issues were considered to ensure the accuracy of the levels of macro and trace elements. (1) All reagents were analytical grade, and water was deionised. (2) All tubes were washed with HNO_3 and rinsed with deionised water. (3) Indium was added into each sample as an internal standard before digestion. (4) We replicated all the blood samples and used several Standard Plasma References, Level I(REF 8883) and II(REF 8884) for quality control. (5) All tubes used were made of polypropylene instead of glass materials to prevent metal contamination. (6) The measurement of element levels was based on the most abundant isotope of each element to avoid interference.

Data normalisation

Each sample contains five demographic characteristics (age, gender, smoking history, drinking history and family history on ESCC), together with concentrations of the aforementioned 38 elements. After data acquisition, preprocessing is performed for later use. The first step is digitisation, with gender (male or female), smoking history, drinking history and family history were represented by 0 or 1. The next step is normalisation: mapping all the concentrations of elements into the interval of (0, 1). The data normalisation procedure is used to avoid numerical difficulties during calculations and to prevent that some variables with greater numeric ranges dominate other variables in smaller numeric ranges, which has been important for practical deployments of neural networks (NN), SVMs and other classifiers.[20] In the same way, ages are linearly transformed from 0 to 100-year olds into (0, 1). After preprocessing, the data from the 100 patients with ESCC and the 100 healthy controls were summarised as a 200×43 matrix, with each row for one subject. Ground-truth labels are used to stand for the case of ESCC or not by +1 or −1, respectively.

Evaluation protocol

To compare the classification performance among different methods, we use the 10 rounds of fivefold cross validation to obtain the average classification accuracies,[21] namely the proportion of correct diagnosis to all of the test subjects. Specifically, the classification accuracy is equal to the ratio between the number of

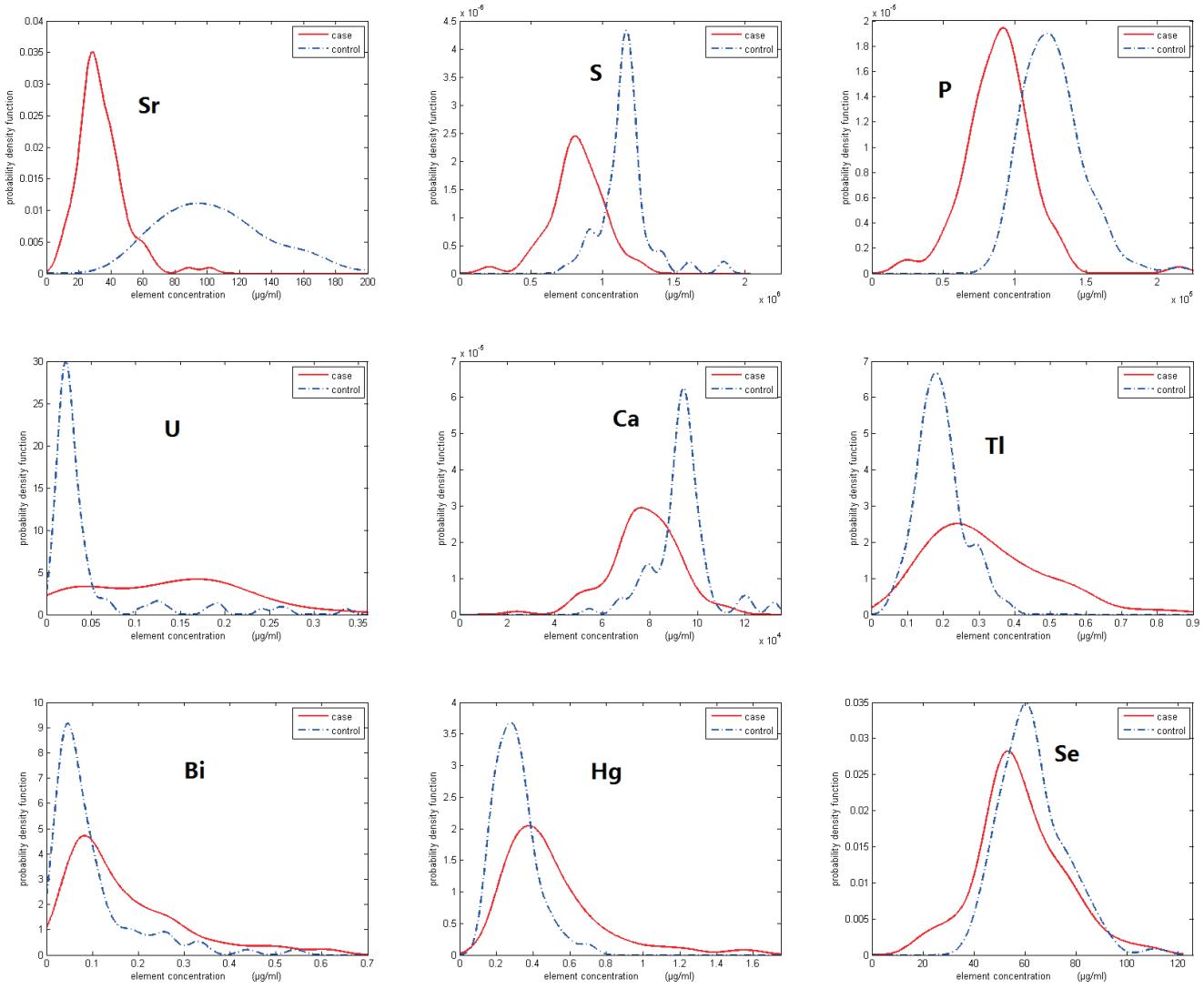

Figure 1 Concentration distributions of eight important elements and one unimportant element (Se) for patients with oesophageal cancer and healthy controls.

correct decision (true positive and true negative for binary problems) and the total number of test subjects. In each round of fivefold cross validation, the labelled examples of the data are randomly partitioned into five chunks (or called folds) of approximately equal size, and then a classification model trained over any four chunks yields a test accuracy on the remaining chunk. For 10 rounds of fivefold cross validation, the test accuracy of each method is computed as the average accuracies over 50(=10×5) chunks. Their means and SD are reported.

For two-class problems, a detailed report of classification results is the confusion matrix consisting of four numbers: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Based on the confusion matrix, one can compute different measures to summarise the results, such as recall=TP/(TP+FN) and precision=TP/(TP+FP). Since our case–control study has an artificial disease prevalence and an unrepresentative disease spectrum, our estimates of precision and recall are not applicable outside of this study. The true disease prevalence should be considered in estimates of recall and precision for a clinical setting. If there is some changeable parameter (threshold) to influence the final decision, one can obtain a visual analysis from computing the receiver operation characteristics (ROC) curve (a sequence of pairs of FP rate and TP rate) by changing the parameter values. Another popular measure is the area under the curve by reducing the ROC curve to a single number,[22] but this measure requires classifiers to change their parameters continuously to yield a function of sensitivity and specificity, which is rather demanding for our developed methods. For detailed description of different approaches to measure classifier performance, one can refer to the Section 19.7 of E. Alpaydin's book.[23] Running time of cross-validation is used to compare the training and prediction time of each

Table 4 Classification accuracies (in percentage) based on single, pair and triple elements

Singles	Sr	Tl	Bi	U	Hg	Ca	P	S
NB	94.41	67.75	59.10	76.65	64.55	74.90	81.50	85.20
LR	94.43	70.45	63.40	77.80	71.55	75.50	81.75	85.80
NN	93.65	70.30	64.50	77.10	71.45	75.85	81.90	86.30
AB	92.13	68.10	62.85	76.95	70.25	73.60	79.75	85.50
SVM	93.86	68.00	57.45	77.00	64.80	74.45	82.40	85.00
RF	91.50	58.10	57.85	66.40	65.20	65.55	73.30	79.55
Pairs	Sr+U	Sr+Ca	Sr+P	Sr+S	U+P	U+S	Ca+S	P+S
NB	96.38	93.15	95.52	93.72	86.82	87.65	83.35	83.25
LR	95.93	94.05	95.15	93.85	87.40	88.20	87.15	85.00
NN	94.25	92.95	95.37	92.43	85.85	86.90	86.80	84.20
AB	93.65	91.15	92.87	91.78	84.30	85.60	84.35	84.80
SVM	96.35	93.05	94.86	94.01	85.70	87.65	83.55	83.10
RF	94.00	91.75	94.45	92.22	82.33	84.90	85.10	82.80
Triples	Sr+U+Ca	Sr+U+P	Sr+U+S	Sr+Ca+P	Sr+Ca+S	Sr+P+S		
NB	94.10	96.90	94.65	93.10	91.90	93.43		
LR	95.80	96.48	96.20	95.20	95.65	94.00		
NN	94.15	95.20	94.95	94.95	94.60	94.00		
Ada	93.35	94.85	93.40	93.85	92.40	92.10		
SVM	95.65	96.98	95.65	94.25	93.75	93.70		
RF	94.00	95.23	95.40	95.15	93.45	94.25		

The best accuracy of each method is marked with bold font in each row.
AB, AdaBoost; LR, logistic regression; NB, Naive Bayes; NN, neural network; RF, Random Forest; SVM, support vector machine.

algorithm. All algorithms were implemented in Matlab on a 2.4 GHz i7-CPU machine with 8 GB memory.

Dimensionality reduction

The main objective of dimensionality reduction is to reduce the computational burden of classifiers and to alleviate the effects of data noise. In the standard paradigm of machine learning and pattern recognition, it is typical to include a stage of feature selection or transformation for dimensionality reduction before the procedure of classification is performed, especially in those applications with hundreds of input features (or variables). In our case, there are only 38 numerical variables for measuring concentrations of the selected chemical elements, but we still apply these dimensionality reduction methods to see if any redundancy can be removed and some improvements in accuracy can be gained. Six methods are used for this objective, including Fisher feature selection (FFS), principal component analysis, Fisher discriminant analysis (FDA) with its variant (FDAx, where the between-class scatter matrix is simply replaced by the 'total' mixture scatter matrix to allow more non-zero eigen values), locality preserving projection and factor analysis.[24–28] Apart from FFS in which a small set of features are chosen directly, other five methods aim at transforming the original features into a low-dimensional embedding space. Note that the Random Forests (RF) method has a built-in mechanism of random feature selection, but it is of interest to see whether the new feature representations from dimensionality reduction approaches can bring some improvements in accuracies for RF. The reader may refer to the review and systematic comparison of different methods for dimensionality reduction.[29]

Classification

A recent comprehensive performance evaluation on the whole UCI classification datasets (available at http://archive.ics.uci.edu/ml) showed that the top rank of the best classifiers includes RF, SVM, NN and boosting. In our EC diagnosis system, we apply six kinds of classifiers. Besides RF,[30] SVM,[31] NN,[32] AdaBoost (AB),[33] two traditional classifiers including Naive Bayes (NB) and logistic regression (LR) are used for further comparisons.[341]

Statistical hypothesis testing

To compare with machine learning approaches, traditional methods of statistical hypothesis testing are also employed. Through the following two hypothesis testing methods, we are able to determine the significant difference in two means of elements' concentrations between patients with EC and healthy comparison subjects. Student's t-test involving two samples is a most widely used testing in biomedical experiments, which is based on the assumption that the data follow the Gaussian (normal) distribution (with equal variances or unequal variances).

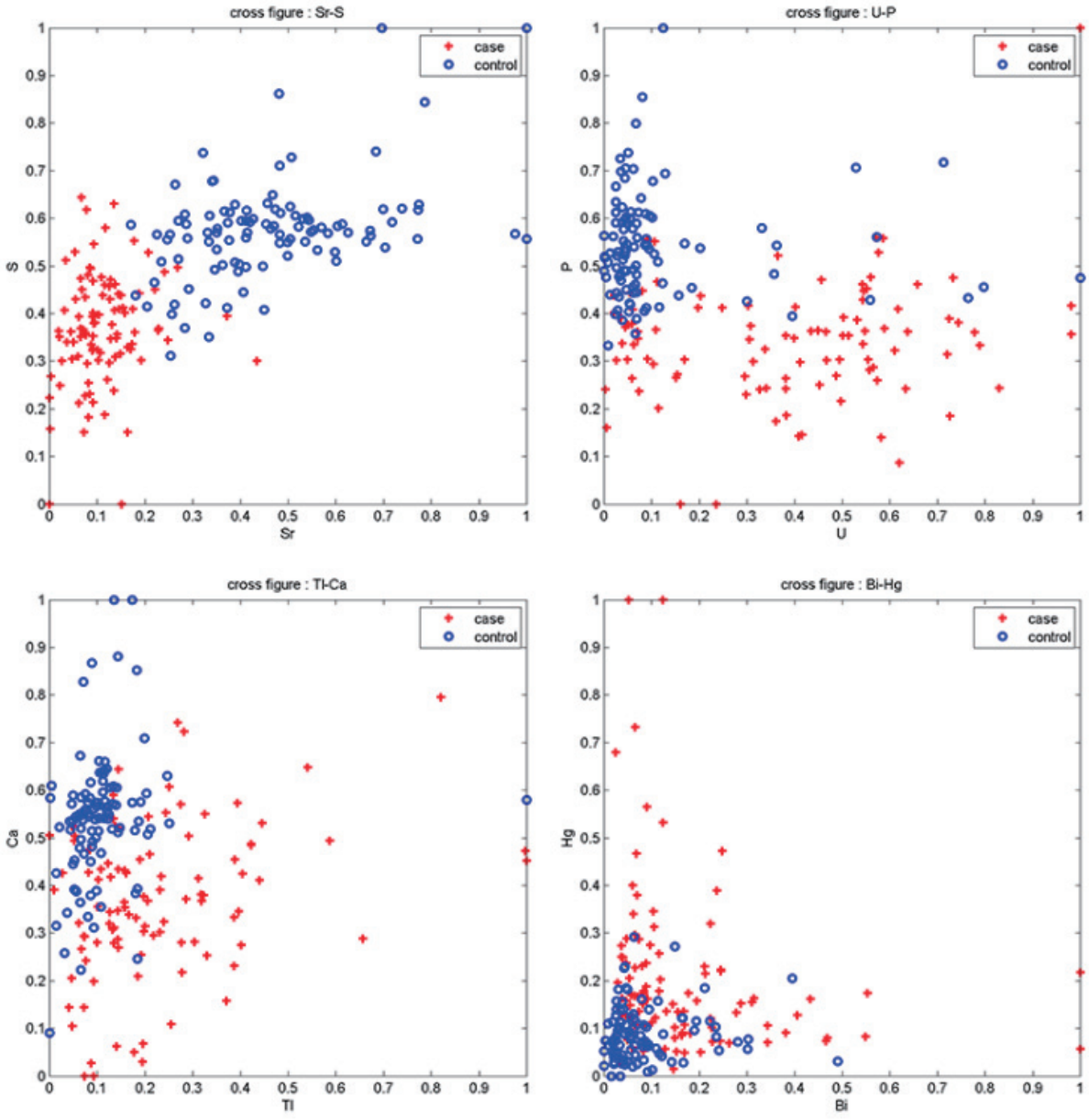

Figure 2 Distributions in normalised concentration for pairs of elements (Sr-S, U-P, Tl-Ca and Bi-Hg).

However, this Gaussian assumption appears to be stringent for most real-world data sets. The counterpart of the t-test is the non-parametric Wilcoxon rank-sum test (also known as the Mann-Whitney U test), which is the most widely used distribution-free hypothesis test. One can refer to Chapter 7 of Thomas Glover and Kevin Mitchell's textbook[35] (Tests of Hypothesis Involving Two Samples) for more details of the two commonly used testing methods.

RESULTS

Table 1 shows the demographic characteristics of the subjects included in this study. A total of 100 patients with ESCC and 100 age, sex and region-controlled normal control subjects were enrolled. In addition, history of regular alcohol consumption and history of regular cigarette smoking were similar between patients with ESCC and controls (p=0.856 and p=0.669, respectively). In contrast, more cases had a family history of ESCC as compared with controls (29.0% vs 17.0%, p<0.05).

The data from patients with ESCC are a 200×43 matrix consisting of 100 patients and 100 healthy subjects with 43 feature variables. We test the aforementioned six classifiers on the original feature space as well as on the embedding spaces by using six dimensionality reduction methods. The dimension of the most embedding spaces is set as 10 for a trade-off between accuracy and complexity, except that the FDA algorithm can only project into 1D because of its inherent limitation for two-class classification problems.

The averaged classification accuracies and running seconds are reported in table 2 based on 10 rounds of

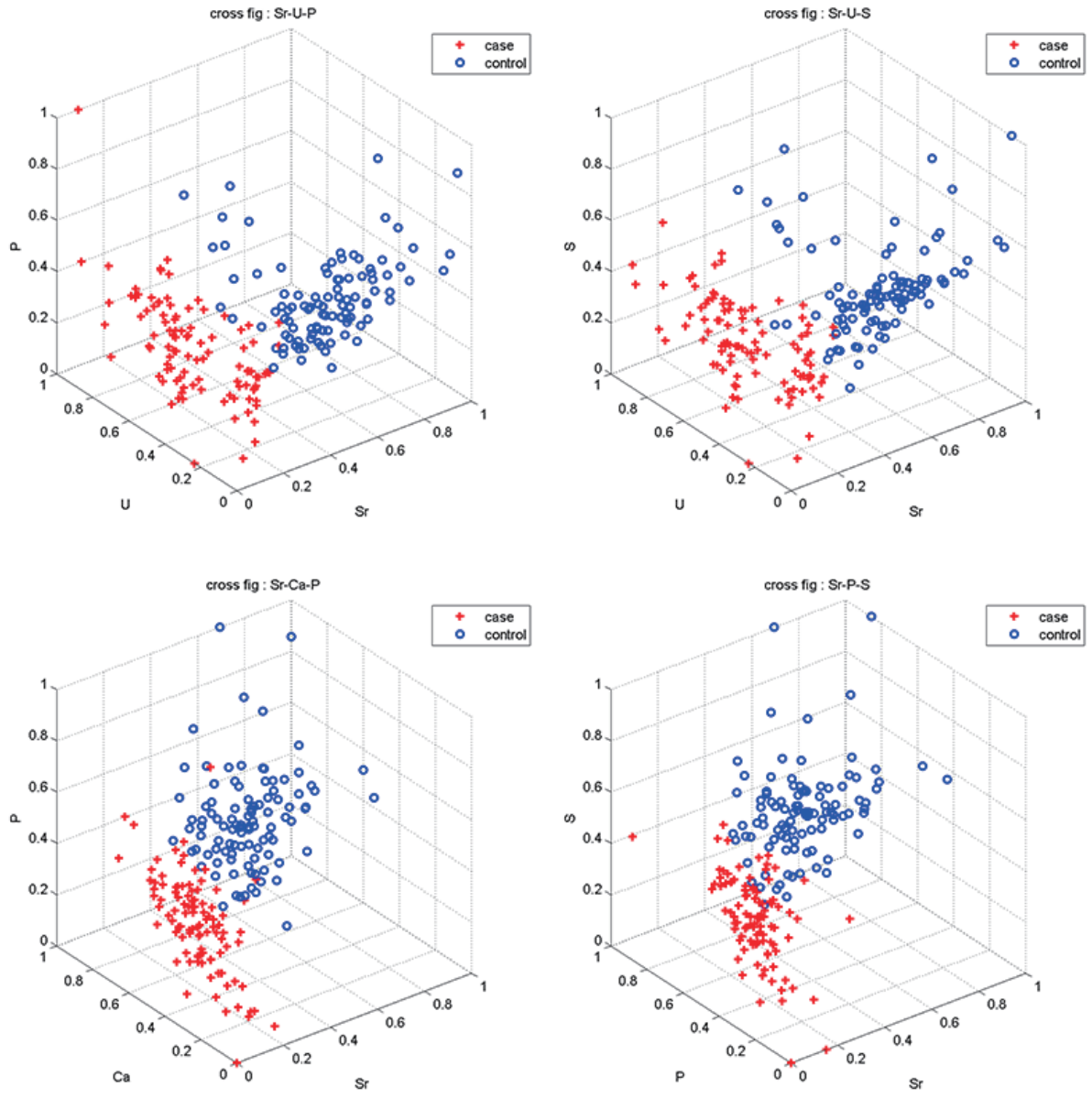

Figure 3 Distributions in normalised concentration for combinations of three elements (Sr-U-P, Sr-U-S, Sr-Ca-P and Sr-P-S).

fivefold cross validation. We can see that RF achieves the best accuracy of 98.38% on the original feature vectors (without dimensionality reduction), and SVM outperforms other classifiers by yielding 96.56% on embedding spaces with dimensionality reduction. In contrast, NB is clearly inferior to other five

Table 5 Classification accuracies (in percentage) on removing a subset of features

Classifiers	Whole	Whole—DemCha	Whole—LowCon	Whole—DemCha—LowCon
NB	92.17	92.23	91.38	91.74
LR	95.92	96.43	95.58	95.88
NN	97.09	97.18	95.58	95.47
AB	96.90	96.89	96.30	96.63
SVM	97.35	97.95	95.71	96.08
RF	98.38	98.31	96.60	96.78

'Whole' means using all available input features in classification. 'DemCha' refers to demographic characteristics including 5 demographic variables: age, gender, smoking history, drinking history and family history. 'LowCon' means the set of six elements with lower concentrations than the detection limit of the spectrometry.
AB, AdaBoost; LR, logistic regression; NB, Naive Bayes; NN, neural network; RF, Random Forest; SVM, support vector machine.

Table 6 The results of hypothesis tests: means and standard deviations of cases and controls, and the p value of t-test and rank-sum test (RS test) (the p values less than alpha=5% are boldfaced)

Feature	Case	Control	t-Test	RS test	Feature	Case	Control	t-Test	RS test
Be	0.10±0.09	0.10±0.10	0.298	0.988	Bi	0.19±0.23	0.10±0.09	<0.001	<0.001
B	33.5±24.1	74.2±45.3	<0.001	<0.001	Cs	0.53±0.20	0.72±0.24	<0.001	<0.001
Al	165±143	366±2111	0.924	0.189	Th	0.11±0.43	0.10±0.24	0.699	0.016
Ti	69.4±16.0	118±291	<0.001	<0.001	U	0.13±0.08	0.05±0.10	<0.001	<0.001
Ge	2.08±0.52	2.37±0.46	<0.001	<0.001	La	0.12±0.11	0.21±0.53	0.024	0.017
As	12.0±4.40	14.1±2.45	<0.001	<0.001	Ce	0.64±0.81	0.80±1.04	0.286	<0.001
Se	58.4±18.0	63.1±12.6	0.020	0.010	V	0.78±0.50	0.83±4.41	<0.001	<0.001
Sr	34.0±14.8	108±37.5	<0.001	<0.001	Cr	10.1±10.5	5.27±7.38	<0.001	<0.001
Li	14.3±15.9	9.09±5.41	0.002	0.691	Mn	8.14±11.4	5.14±10.5	0.020	<0.001
Ni	8.34±5.93	7.32±7.69	0.219	0.008	Cu	9145±254	1122±214	<0.001	<0.001
Mo	4.80±4.60	4.09±3.10	0.204	0.280	Zn	655±166	774±193	<0.001	<0.001
Ag	0.13±0.10	0.54±4.2	0.792	0.100	Rb	150±77.1	181±44.4	<0.001	<0.001
Cd	0.40±0.63	0.28±0.19	0.074	0.306	Hg	0.53±0.35	0.30±0.11	<0.001	<0.001
Sn	3.90±7.90	1.19±3.10	<0.001	<0.001	Pb	7.05±8.14	4.64±4.56	0.010	0.010
Ba	41.2±81.4	18.4±34.4	0.010	0.266	Ca	7.8×10^4 $\pm1.4\times10^4$	9.3×10^4 $\pm1.2\times10^4$	<0.001	<0.001
Pt	18.0±82.7	16.4±73.1	0.867	<0.001	Fe	2124±2204	1966±1100	0.588	0.311
Tl	0.35±0.26	0.20±0.13	<0.001	<0.001	K	1.4×10^5 $\pm4.9\times10^4$	1.6×10^5 $\pm2.7\times10^4$	<0.001	<0.001
P	9.0×10^5 $\pm2.7\times10^4$	1.3×10^5 $\pm2.1\times10^4$	<0.001	<0.001	Mg	1.7×10^4 ±3583	2.2×10^4 ±3155	<0.001	<0.001
S	8.2×10^5 $\pm1.9\times10^5$	1.2×10^6 $\pm1.8\times10^5$	<0.001	<0.001	Na	3.0×10^6 $\pm2.4\times10^5$	3.2×10^6 $\pm2.7\times10^5$	<0.001	<0.001

classifiers possibly because of the 'bogus' conditional independence assumption. Besides, three dimensionality reduction methods based on Fisher discriminant ratios (namely FFS, FDA and FDAx) are often more favourable than other three dimensionality reduction methods without the use of discriminant information. In terms of running time, we found that the dimensionality reduction procedure can evidently speed up the running time for each classifier; on the other hand, three classifiers (namely NB, LR and SVM) are faster than the remaining three classifiers.

Through dimension reduction, the learnt parameters of FDA and FFS can reflect the correlation or significance of the chemical elements to EC. Specifically, the projection vector w of FDA can be regarded as linear weights to the original features; thus, larger weights mean higher contributions. For FFS, the F statistic indicates the discriminant capability of each chemical element: the larger the F is, the more difference there is on that chosen element. As shown in table 3, we can find that strontium (Sr) and sulfur (S) are the top two elements with more discriminant information. Besides, other important elements may include P, U, Ca, Tl, Bi and Hg.

To further examine the discriminant capability, we draw the concentration distributions of these eight important elements and of one unimportant element (Se) for comparison. As shown in figure 1, we can see that the differences between cases and controls on these eight elements are significant, whereas the difference on the element Se is not so much. The top part of table 4 lists the classification accuracies based on each single element.

As we can see, all six classifiers can achieve accuracies more than 90% based on the single most important element Sr. Other two elements with distinctive difference are S and P, providing accuracies around 80%.

We also investigate whether the relation of any two elements is different between cases and controls. Figure 2 shows the distributions in normalised concentration for four selected pairs of elements, including Sr-S, P-U, Ca-Tl and Bi-Hg. We can see that these 2D scatter plots are highly separable, though no straightforward functions are available to describe the pairwise relationship. The middle part of table 4 displays the classification accuracies based on a small set of element pairs. It indicates that the pair of Sr and U achieves the best performance for four classifiers, whereas another pair of Sr and P outperforms other pairs for two classifiers. It appears that Sr and U are most diverse elements and they complement each other, though classification performance on single U is not satisfactory.

When considering the scatter plots of any three important elements, the separability can be improved as

shown in figure 3. Almost any linear classifier can achieve good performance based on these triples of elements. The bottom part of table 4 displays classification results based on a small set of element triples. The triple composed of Sr, U and P achieves the best for five classifiers except RF.

Finally, we attempt to identify the importance of different feature subsets by removing variables from the whole variable set (table 5). Note that this procedure is just a simple way to exclude a specific feature subset, not identical to the complex backward elimination method in the literature of variable selection that will be dependent on the order. When eliminating the five demographic variables (including age, gender, smoking history, drinking history and family history), the classification accuracies are improved for four classifiers compared with the results of using the whole set of original features. For other two classifiers, namely AB and RF, the classification performance only degenerates slightly. It seems that the connections between demographic characteristics and EC are very weak or loose. On the other hand, removing six low-concentration elements becomes detrimental to the classification performance, though the accuracy decreases are not much.

In order to compare to machine learning methods, the traditional hypothesis testing method, including t-test and *rank-sum test*, is applied. As shown in table 6, at the confidence level of 5%, more than half of the features are significantly different between cases and controls in terms of t-test and *rank-sum test*. Also we can notice that the means and SD differ significantly in many features between cases and controls, partly due to singular values or 'outliers' in the data. For example, on some of the elements, the highest concentration is more than thousand times of the average value, which makes a great bias over the hypothesis testing results. By contrast, machine learning methods are more robust when dealing with such outliers in data.

DISCUSSION

In this article, we present a study of chemical elements in serum for patients with non small cell carcinoma (NSCC) based on supervised learning methods. As shown in table 6, at the level of 5%, more than half of chemical elements are shown to be statistically significantly different between cases and controls. To our knowledge, previous relevant studies only focused on Se, Cu and Zn.[12 36 37] Similarly, in our study, we observed lower levels of Se and Zn and higher levels of Cu among patients with ESCC compared with controls. One possible explanation is that Se is a primary component of selenoproteins, of which antioxidant role can regulate the redox status of some molecules and dampen the propagation of free radicals and reactive oxygen species,[38] and Zn has a number of vital functions including cell proliferation, reproduction, immune function and defence against free radicals.[39] Excess Cu has been known to be a potent oxidant causing the generation of ROS in the cells.[40]

However, the values for these studied elements varied significantly among different studies which conducted in different countries, or regions. These inconsistent findings might result from racial factors and geographic variation, and varied sample sizes of relevant studies. Therefore, we used a case–control study matched by age, sex and region in order to make the cases and controls comparable and try to control the potential confounders, such as region, smoking and drinking. In addition, in the studies of chemical elements and health, the means and SD usually differ enormously between the case and control groups over many features. This is mainly because of singular values (or outliers) incurred in measurements to the data. For example, on certain elements, the highest concentration is more than thousand times of the average value, which influenced the test results greatly. However, machine learning methods are much more robust dealing with such 'outliers' problem.

In this study, our analysis based on machine learning methods gives prosperous results. Specifically, RF achieves the best accuracy of 98.38% on the original feature vectors (without dimensionality reduction), and SVM outperforms other classifiers by yielding 96.56% on embedding spaces with dimensionality reduction. All six classifiers can achieve accuracies more than 90% based on the most important single element Sr. The other two elements with distinctive difference are S and P, providing accuracies around 80%.

The contributions of this paper are twofold. First, we provide a principled framework to comprehensively investigate the chemical elements in blood serum of patients with ESCC; this framework can be easily extended to blood serum of other patients with cancer, or even for general diseases. The main impediment of coping with tens of chemical elements (38 in our study) can be efficiently solved by hypothesis testing and machine learning methods nowadays with ordinary computation platforms. Second, we find that great differences exist in element concentrations between patients with ESCC and healthy comparison subjects. Consequently, approaches such as 'crude' diagnosis before gold-standard examination (using biopsy), and early detection of EC, can have potential applications.

There are several merits in our proposed framework. First, the classification accuracies achieved by machine learning methods are remarkably higher than most empirical decisions made by doctors on this small corpus of 100 patients with EC and 100 healthy comparison subjects. Furthermore, the test error rates can tell us the confidence level of the predictions whenever a large data set is available to conduct this chemical element analysis. Second, the expense for this analysis is much lower than that of the gold-standard biopsy for EC and other cancers, though this analysis may only serve as 'crude' diagnosis or prescreening. Third, the time to obtain analysis results in element concentrations by this approach is much shorter than biopsy. Fourth, the blood sample acquisition is lowly invasive, and thus can be easily performed in annual

health examination for early-stage precaution. This point is particularly essential for most patients with cancer, as locally advanced cancer or distant metastases in late stages are associated with high mortality.

There are also some hidden pitfalls in this framework which deserve our cautions. First, possible confounding biases may not be controlled or avoided due to the absence of such factors, including body mass index, dietary intakes and potential regional carcinogen contacts. Moreover, even if these factors were comparable, it remains impossible to eliminate the possibility that diverse genetic features might be associated with ESCC. Second, the present work is only based on a relatively small patient cohort. This framework should be evaluated on a larger patient cohort before any real clinical applications are performed in the future. Third, due to the retrospective design nature of this study, the results showed associations only and gave no cause–effect clues; they may not be generalised to other populations, like Europe or North America, with regard to ESCC. Fourth, among patients with ESCC, lower levels of some elements might be caused by eating difficulty. But on the other hand, higher levels of concentrations in vanadium, manganese and chromium were also observed. Therefore, the differences in element concentrations cannot be simply attributed to eating difficulty. This finding might give some useful hints or ideas for the study of ESCC. Fifth, this study cannot produce clinically relevant estimates of diagnostic accuracy, because clinically relevant estimates would have to come from a study that recruited a clinically relevant patient sample. In addition, case–control studies almost always overestimate sensitivity, specificity and accuracy. Therefore, these classifiers need to be tested in a clinical setting before their use can be recommended. Finally, the contamination of heavy metal elements has become a severe problem in mainland China; however, this effect of heavy metal might be adjusted by using a healthy control group which matches the present patients with ESCC in age, sex and residential areas, as we did in this study. If other physical factors such as age, sex and region are the same or similar, it is safe to attribute the great differences in element concentration between the case group and the control group to the presence of the particular disease, not to other unrelated problem like heavy metal.

The proposed framework may have several new emerging applications. One possibility is to modify the chemical elements in pharmacy for enhancing the levels of certain element concentrations of a patient into a normal interval. Another perspective is to provide a rebalanced diet in nutrition for patients. These new applications will depend on the thorough and deep analysis of the chemical elements in serum among patients and healthy controls.

CONCLUSIONS

These results suggested element profile differences between patients with ESCC and controls, which indicated some potential promising applications in diagnosis, prognosis, pharmacy and nutrition of ESCC. In the future, the results of the analyses will be useful in designs that have larger sample sizes. However, the results should be interpreted with caution due to the retrospective design nature, limited sample size and the lack of several potential confounding factors, such as obesity, nutritional status, and fruit and vegetable consumption and potential regional carcinogen contacts.

Author affiliations
[1]The Key Laboratory of Machine Perception (Ministry of Education), School of EECS, Peking University, Beijing, China
[2]Civil Aviation Medicine Center, Civil Aviation Administration of China, Beijing, China
[3]Key laboratory of Carcinogenesis and Translational Research (Ministry of Education), Laboratory of Genetics, Peking University Cancer Hospital & Institute, Beijing, China
[4]Center of Medical & Health Analysis, School of Public Health, Peking University, Beijing, China
[5]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, Bloomington, Indiana, USA

Contributors JYW, TBL and ZHH proposed and supervised the project. TL and YCL designed and carried out the experiments and analysed the data. LLY, TBL and CTZ contributed to the sample collection, chemical elements measurement, quality control and finding of related literature. ZXC participated in the discussion, helped to improve the proposed methods and copyedited the manuscript. TL, YCL and TBL wrote the manuscript.

Funding This work was supported by the National Natural Science Foundation of China under Grant No. 61375051, 61075119 and 81473033, and the Seeding Grant for Medicine and Information Sciences of Peking University under Grant No. 2014-MI-21.

Competing interests None declared.

Patient consent Obtained.

Ethics approval This study was approved by the Institutional Review Board of Peking University School of Oncology, Beijing.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Additional data are available by emailing wjy@bjmu.edu.cn.

REFERENCES

1. Torre LA, Bray F, Siegel RL, *et al*. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;65:87–108.
2. Torre LA, Siegel RL, Ward EM, *et al*. Global Cancer Incidence and Mortality Rates and Trends–An Update. *Cancer Epidemiol Biomarkers Prev* 2016;25:16–27.
3. Zhang Y. Epidemiology of esophageal cancer. *World J Gastroenterol* 2013;19:5598–606.
4. Stahl M, Budach W, Meyer HJ, *et al*. Esophageal cancer: Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2010;21(Suppl 5):v46–9.
5. Tixier F, Le Rest CC, Hatt M, *et al*. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med* 2011;52:369–78.
6. Salami SS, Schmidt F, Laxman B, *et al*. Combining urinary detection of TMPRSS2:ERG and PCA3 with serum PSA to predict diagnosis of prostate cancer. *Urol Oncol* 2013;31:566–71.

7. Hatse S, Lambrechts D, Verstuyf A, *et al*. Vitamin D status at breast cancer diagnosis: correlation with tumor characteristics, disease outcome, and genetic determinants of vitamin D insufficiency. *Carcinogenesis* 2012;33:1319–26.
8. Yanaihara N, Caplen N, Bowman E, *et al*. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* 2006;9:189–98.
9. Anderson BW, Ahlquist DA. Molecular Detection of Gastrointestinal Neoplasia: Innovations in Early Detection and Screening. *Gastroenterol Clin North Am* 2016;45:529–42.
10. Cunzhi H, Jiexian J, Xianwen Z, *et al*. Serum and tissue levels of six trace elements and copper/zinc ratio in patients with cervical cancer and uterine myoma. *Biol Trace Elem Res* 2003;94:113–22.
11. Milde D, Novák O, Stu ka V, *et al*. Serum levels of selenium, manganese, copper, and iron in colorectal cancer patients. *Biol Trace Elem Res* 2001;79:107–14.
12. Dar NA, Mir MM, Salam I, *et al*. Association between copper excess, zinc deficiency, and TP53 mutations in esophageal squamous cell carcinoma from Kashmir Valley, India–a high risk area. *Nutr Cancer* 2008;60:585–91.
13. Knekt P, Aromaa A, Maatela J, *et al*. Serum vitamin E, serum selenium and the risk of gastrointestinal cancer. *Int J Cancer* 1988;42:846–50.
14. Ma EL, Jiang ZM. Ion-exchange chromatography in simultaneous determination of serum copper and zinc levels in patients with cancer of digestive tract. *Chin Med J* 1993;106:118–21.
15. Abeel T, Helleputte T, Van de Peer Y, *et al*. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 2010;26:392–8.
16. Little J, Higgins JP, Ioannidis JP, *et al*. STrengthening the REporting of Genetic Association Studies (STREGA)–an extension of the STROBE statement. *Genet Epidemiol* 2009;33:581–98.
17. Liu F, Guo F, Zhou Y, *et al*. The Anyang Esophageal Cancer Cohort Study: study design, implementation of fieldwork, and use of computer-aided survey system. *PLoS One* 2012;7:e31602.
18. Liu T, Lu QB, Yan L, *et al*. Comparative Study on Serum Levels of 10 Trace Elements in Schizophrenia. *PLoS One* 2015;10:e0133622.
19. Wang Y, Ou YL, Liu YQ, *et al*. Correlations of trace element levels in the diet, blood, urine, and feces in the Chinese male. *Biol Trace Elem Res* 2012;145:127–35.
20. Hsu CCC CW, Lin CJ. A practical guide to support vector classification. National Taiwan University http://www.csie.ntu.edu.tw/~cjlin (accessed 18 July 2007).
21. Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of machine learning*. Cambridge, UK: The MIT Press 2012.
22. Provost F, Fawcett T. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. *KDD* 1997;97:43–8.
23. Alpaydin E. *Introduction to Machine Learning*. 2nd ed. The MIT Press, 2010.
24. Rice JA. *Mathematical statistics and data analysis*, 2nd ed. New York, USA: Duxbury Press; 1995.
25. Theodoridis S, Koutroumbas K. *Pattern Recognition*. 4th ed. Burlington, Canada: ElsevierL, 2008.
26. He X, Niyogi P, Locality preserving projections. Proceedings of Advances in Neural information processing systems. 2003;16:153–60.
27. *A. N. Factor analysis and lecture notes of machine learning*. Stanford 2008.
28. Harman H. *Modern factor analysis*. Chicago, USA: University of Chicago Press 1976.
29. Maaten L, Postma E, Herik H. Dimensionality reduction: a comparative review. Technical Report, TiCC, Tilburg University 2009.
30. Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
31. Vapnik V. *Statistical Learning Theory*. New York, USA: John Wiley & Sons, 1998.
32. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;323:533–6.
33. Schapire R, Freund Y. *Boosting: Foundations and Algorithms*. Massachusetts, USA: The MIT Press, 2012.
34. Shalev-Shwartz S, Ben-David S. *Understading machine learning*. New York, USA: Cambridge University Press 2014.
35. TGaK M. An Introduction to Biostatistics: the McGraw-Hill Companies. 2001.
36. Mir MM, Dar NA, Salam I, *et al*. Studies on Association Between Copper Excess, Zinc Deficiency and TP53 Mutations in Esophageal Squamous Cell Carcinoma From Kashmir Valley, India-A High Risk Area. *Int J Health Sci* 2007;1:35–42.
37. Wei WQ, Abnet CC, Qiao YL, *et al*. Prospective study of serum selenium concentrations and esophageal and gastric cardia cancer, heart disease, stroke, and total death. *Am J Clin Nutr* 2004;79:80–5.
38. Mistry HD, Broughton Pipkin F, Redman CW, *et al*. Selenium in reproductive health. *Am J Obstet Gynecol* 2012;206:21–30.
39. Ho E, deficiency Z. DNA damage and cancer risk. *J Nutr Biochem* 2004;15:572–8.
40. Gupte A, Mumper RJ. Elevated copper and oxidative stress in cancer cells as a target for cancer treatment. *Cancer Treat Rev* 2009;35:32–46.