

# SCIENTIFIC REPORTS



OPEN

## Identification of risk factors in epidemiologic study based on ROC curve and network

Jiao Jin<sup>1,\*</sup>, Shixin Zhou<sup>2,\*</sup>, Qiujin Xu<sup>3</sup> & Jinbing An<sup>4</sup>

Received: 26 October 2016

Accepted: 28 March 2017

Published: 24 April 2017

This article proposes a new non-parametric approach for identification of risk factors and their correlations in epidemiologic study, in which investigation data may have high variations because of individual differences or correlated risk factors. First, based on classification information of high or low disease incidence, we estimate Receiver Operating Characteristic (ROC) curve of each risk factor. Then, through the difference between ROC curve of each factor and diagonal, we evaluate and screen for the important risk factors. In addition, based on the difference of ROC curves corresponding to any pair of factors, we define a new type of correlation matrix to measure their correlations with disease, and then use this matrix as adjacency matrix to construct a network as a visualization tool for exploring the structure among factors, which can be used to direct further studies. Finally, these methods are applied to analysis on water pollutants and gastrointestinal tumor, and analysis on gene expression data in tumor and normal colon tissue samples.

Identification of possible risk factors of specific diseases in epidemiologic studies is helpful in guiding diagnosis, therapy or disease control. This process is usually considered as a problem of variable selection in mathematics. However, due to individual differences or complicated interaction of risk factors, the epidemiologic investigation data often have serious variation and the relationship between response variable and explanatory variables can not be appropriately expressed by specific mathematical models, which may reduce the reliability of classical methods for variable selection. Therefore, it is desirable to develop appropriate analysis methods suitable for the epidemiologic data.

The conventional methods for variable selection include steps to construct some evaluation functions based on specific parametric models and identify significant risk factors through optimization process<sup>1,2</sup>. These methods usually have severe limitations on the distribution of random errors and mathematical forms of models, such as linear model<sup>3</sup>, Cox model<sup>4,5</sup> and logistic model<sup>6</sup>. However, besides influence of large variation of observations, the bias of selected mathematical model may lead to inappropriate conclusions<sup>7,8</sup>. For example, some important variables may be rejected by selected model mistakenly, or inconsistent conclusions may be obtained after use of different models.

In contrast to parametric methods, random forest is often used to select variables through change of certain measurement on prediction accuracy when selected variables are eliminated<sup>9–11</sup>. In addition, methods based on some probability function<sup>12,13</sup> or network<sup>14,15</sup> are also effective choices to evaluate specific genes or tissues in studies of biomedical science. These methods are non-parametric methods without severe limitations on models or data, and therefore more suitable for the problems with high variation data and unknown factor structure in epidemiologic studies.

Noting the binary feature of high and low disease incidences in epidemiologic investigation data, and two components of true positive rate (TPR) and false positive rate (FPR) in ROC curve<sup>16,17</sup>, we select ROC curve to describe the relationship between risk factors and disease incidence, and screen for the candidate important risk factors. ROC curve has a well-established theoretical basis<sup>18,19</sup>, and is widely used for many problems<sup>20,21</sup>. Furthermore, we define a new type of correlation matrix based on distance of ROC curves corresponding to any

<sup>1</sup>School of Statistics, Beijing Normal University, Beijing, 100875, China. <sup>2</sup>Department of Cell Biology, School of Basic Medicine, Peking University Health Science Center, Beijing, 100191, China. <sup>3</sup>Chinese Research Academy of Environmental Science, Beijing, 100012, China. <sup>4</sup>Faculty of Foundational Education, Peking University Health Science Center, Beijing, 100191, China. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.A. (email: termison@bjmu.edu.cn)

pair of factors, and then use it to evaluate the correlated effect of risk factors on disease and to construct a network as a visualization tool for exploring the structure among factors.

### Screening of risk factors based on ROC curve

Suppose that  $k$ -dimensional random vector  $\mathbf{F} = (F_1, \dots, F_k)$  denotes the risk factors, where each  $F \in \mathbf{F}$  has nonempty support set  $S_F \subseteq \mathbb{R}$ , and random variable  $D$  denotes the state of disease, where  $D = 1$  represents diseased population, and  $D = 0$  represents healthy population. To study the impact of  $\mathbf{F}$  on disease incidence  $\pi$ , we can investigate observation  $\mathbf{u} = (u_1, \dots, u_n)$  from diseased population, and  $\mathbf{v} = (v_1, \dots, v_m)$  from healthy population, where each vector  $u_i$  or  $v_j$  denotes observations of factors  $(F_1, \dots, F_k)$ .

For any factor  $F \in \mathbf{F}$ , ROC curve is defined as a graph of true positive rate (TPR) in  $y$ -axis versus false positive rate (FPR) in  $x$ -axis. For the sake of simplicity, ROC can be expressed by a series of  $(R_X(f), R_Y(f))$  in coordinate system  $X \times Y$  for various values of  $f \in S_F$ , where

$$R_Y(f) = P(F > f | D = 1), \quad R_X(f) = P(F > f | D = 0), \quad (1)$$

and their values can be estimated by  $\mathbf{u}$  and  $\mathbf{v}$ , respectively.

Because both  $R_X(f)$  and  $R_Y(f)$  are monotone functions with range  $(0, 1)$  and connected by common value  $f \in S_F$ , if we define  $R_X^{-1}(t) = \sup\{f : R_X(f) \geq t; 0 < t < 1\}$ , then ROC curve can also be expressed as a graph of  $(t, R(t))$  with only one parameter  $t$  on  $(0, 1)$ :

$$R(t) = R_Y[R_X^{-1}(1 - t)], \quad 0 \leq t \leq 1. \quad (2)$$

Now, suppose the larger value of the variable  $F$  increases the disease incidence  $\pi$ , that is for  $\forall f \in S_F$ ,  $P(D = 1 | F > f) > P(D = 1)$ , then according to definition of conditional probability, we can reach the conclusion:

$$P(D = 1 | F > f) > P(D = 1) \Rightarrow R_Y(f) \geq R_X(f). \quad (3)$$

Similarly, for  $\forall f \in S_F$ , we can also obtain

$$R_Y(f) \geq R_X(f) \Rightarrow P(D = 1 | F > f) > P(D = 1). \quad (4)$$

Because of the equivalence property of  $(R_X(f), R_Y(f))$  and  $(t, R(t))$ , the conclusions above suggest that if the larger value of the variable  $F$  increases the disease incidence  $\pi$ , then the ROC curve  $R(t)$  is above the diagonal of bounded region  $\mathcal{A} = (0, 1) \times (0, 1)$  constantly. Similarly, if  $R(t)$  is above the diagonal of region  $\mathcal{A}$  constantly, then the larger value of the variable  $F$  may result in the larger disease incidence  $\pi$ .

Based on this fact, we can evaluate whether  $F$  plays an important role in influencing disease incidence  $\pi$  through hypothesis testing with null-hypothesis of independence between variable  $F$  and  $D$ , that is  $H_0: D \perp F$ . To construct appropriate test statistic, suppose consistent estimation of  $R(t)$  is  $\hat{R}(t)$ , then using the conclusion on ROC curve<sup>22</sup>, as  $\min(n, m) \rightarrow \infty, n/m \rightarrow \lambda$ , we have

$$\delta(t) = \sqrt{n}[\hat{R}(t) - R(t)] = \lambda^{1/2}B_1(1 - t) + B_2(1 - t) + o_p(1), \quad (5)$$

where  $B_1(t)$  and  $B_2(t)$  are two identical independent Brownian bridges.

Suppose the null hypothesis of  $H_0$  is true, we should have  $R_X(f) = R_Y(f)$  for  $\forall f \in S_F$ , which means that  $R(t) = t$ . In this situation, if we define the symbol  $\delta_0(t) = \sqrt{n}[\hat{R}(t) - t]$ , then we can obtain its asymptotic distribution based on Brownian bridge. Because the integral on  $\delta_0(t)$  is connected with Area under ROC Curve (AUC), which is well known in epidemiologic study, we can construct test statistic based on AUC:

$$S_A = \left| \int_0^1 \delta_0(t) dt \right| = \left| \int_0^1 \{\sqrt{n}[\hat{R}(t) - t]\} dt \right| = \sqrt{n}|AUC - 0.5|. \quad (6)$$

If the value  $S_A$  is larger than a certain critical value, we can reject the null hypothesis of  $H_0$ , which means that the impact of variable  $F$  on disease incidence  $\pi$  can not be explained as random fluctuations.

Because the asymptotic distribution of  $\delta_0(t)$  is expressed as a linear combination of two Brownian bridge processes, we can obtain the empirical distribution of test statistic  $S_A$  through method of asymptotic simulation, and judge whether the ROC curve is significantly deviated from the diagonal. Specifically, we can simulate two independent Brownian bridges  $B_1(t)$  and  $B_2(t)$  using relationship between Brownian bridge and Brownian motion, and construct stochastic process of  $\delta_0(t)$  by equation (5) and  $S_A$  by equation (6). Repeat this process for  $n$  times, and we can obtain  $n$  simulated observations of  $S_A$ , from which we can obtain the empirical distribution of test statistic  $S_A$  together with the hypothesis threshold as the null hypothesis  $H_0$  is true, and then we can complete the hypothesis test to judge whether the ROC curve has significant deviation from the diagonal, which can be used to screen for the variable  $F \in \mathbf{F}$  with important impact on disease.

### Construction of network based on correlation matrix

Although ROC curve can express the effects of any variable  $F \in \mathbf{F}$  on disease  $D$ , we can only evaluate their effects one by one. In fact, the variables often have correlation with one another, therefore it is necessary to analyze the interaction among risk factors. Similarly, the measurements of correlation, such as Pearson coefficient of correlation, Kendall coefficient or other measurements, also have such disadvantages, which can not express the correlation among many variables.

Considering the unknown structure of correlation among many variables, we select network to evaluate their correlation. Network is constructed by many knots with connections between certain pair of variables and therefore can describe the complex interaction among interesting variables<sup>23,24</sup>.

Provided that two factors  $F_i$  and  $F_j$  have synergistic effect, they should have similar ROC curves and small value of difference between their AUC. Based on the assumption above, we define distance  $d_{ij}$  between any pair of variables of  $F_i$  and  $F_j$  to evaluate the correlation of  $F_i$  and  $F_j$  on disease  $D$ :

$$d_{ij} = \left| \int_0^1 \{R_i(t) - R_j(t)\} dt \right| = |AUC_i - AUC_j|, \quad (7)$$

whose value can be estimated by  $\hat{R}_i(t)$  and  $\hat{R}_j(t)$ , and we can denote its estimation as  $\hat{d}_{ij}$ . Then, we select the expression below to estimate the correlation of  $F_i$  and  $F_j$ :

$$\hat{r}_{ij} = \exp\{-\sqrt{n} * \hat{d}_{ij}\}. \quad (8)$$

Here, the value of  $\hat{r}_{ij}$  is always confined into interval of (0, 1), and the larger value of  $\hat{r}_{ij}$  is, the stronger correlation of  $F_i$  and  $F_j$  in effects on disease  $D$  should be.

If we obtain all the estimations of correlations among factors  $\hat{r}_{ij}$ , then we can use the matrix of  $(\hat{r}_{ij})_{k \times k}$  as adjacency matrix to construct network to analyze the structure of correlation among these factors. However, when the value of  $k \times k$  is too large, some connections may be noisy ones, whose values should be converted into 0 through some criteria to avoid interference of them on the analysis.

In fact, based on equation (5), for any pair of  $F_i$  and  $F_j$ , we have

$$\begin{aligned} \hat{R}_i(t) - \hat{R}_j(t) &= [R_i(t) - R_j(t)] + \lambda^{1/2} [B_{i1}(1-t) - B_{j1}(1-t)] \\ &\quad + [B_{i2}(1-t) + B_{j2}(1-t)] + o_p(1). \end{aligned} \quad (9)$$

Based on this result, suppose  $F_i$  and  $F_j$  have strong connection and similar ROC curves, and express it as null hypothesis  $R_i(t) = R_j(t)$ , then the value of  $\hat{r}_{ij}$ , as a function of  $\sqrt{n} * \hat{d}_{ij}$ , can also be evaluated by the aid of distribution of Brownian bridge. Thus, it is a good choice to make decision on whether certain connections  $\hat{r}_{ij}$  should take values 0 through approximated distribution of  $\hat{r}_{ij}$ , just as some statistical methods do in testing whether some parameters should take values 0.

Now, similar to the process of simulation mentioned above, we can simulate Brownian bridges  $B_{i1}$ ,  $B_{i2}$  and  $B_{j1}$ ,  $B_{j2}$ , from which we can obtain simulated values of  $\hat{r}_{ij}$  by equations (7) and (8). Repeat this procedure for  $n$  times, and we can obtain  $n$  simulated observations of  $\hat{r}_{ij}$  and obtain its empirical distribution. Then, for a given level  $\beta$ , we take the quantile of empirical distribution as threshold and transform the ones lower than threshold to 0. Here, the smaller the value of  $\beta$  is, the fewer nodes in the network there are.

After rearrangement of each element, the matrix  $\hat{r}_{ij}$  should have more explicit information. Therefore, we can take the amendatory matrix  $\hat{r}_{ij}$  to construct a network to explore the relationship among these factors, such as individual groups and their central nodes, which can be clues for further experimental or theoretical studies.

The methods above are completed by R software, and the programs are available in appendix, through which the readers can update the programs based on their new methodology.

## Examples and Results

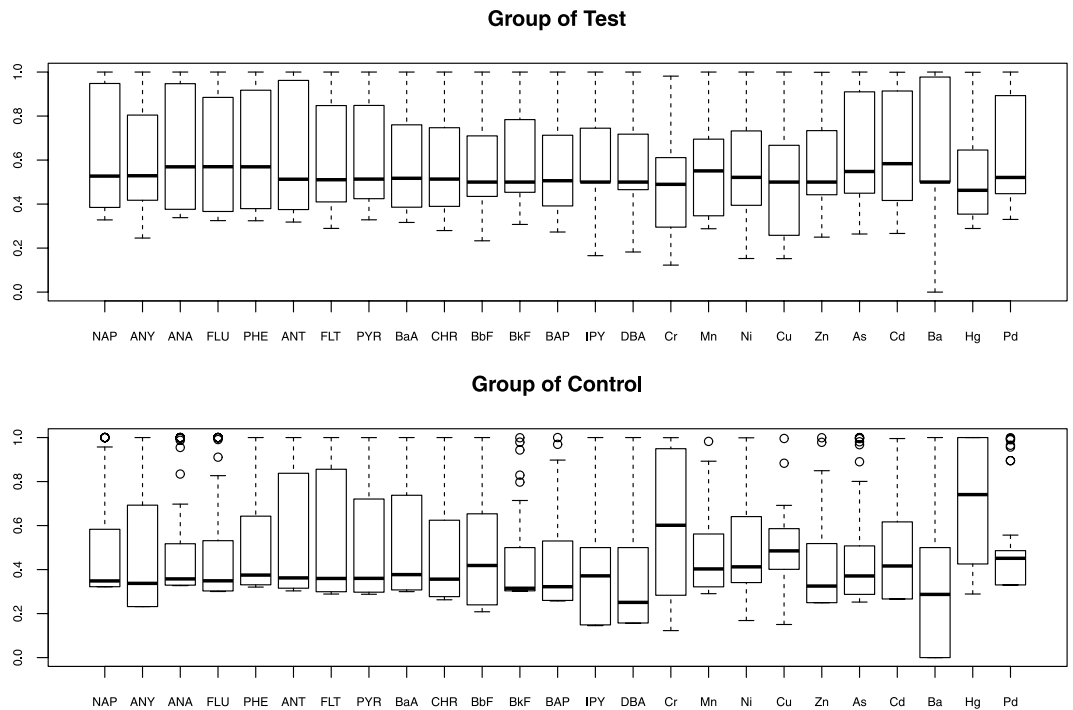
**Example 1.** In this part, we apply the introduced methods to problem of correlation between gastrointestinal (GI) tumors and pollutants in local drinking water, particularly polycyclic aromatic hydrocarbons (PAHs) and heavy metals. Some reports have suggested that high levels of PAHs in the air may be associated with cancer<sup>25-27</sup>. However, few studies have assessed the presence of both PAHs and heavy metals in sources of drinking water, which may have stronger influence in GI tumors.

In the current study, Huai'an region, located in the middle of Jiangsu Province, has been one of the surveillance spots with high cancer incidence for 30 years in China. Furthermore, Huai'an has the highest incidence of GI cancers in Jiangsu Province, and patients suffering from GI cancers (mainly esophageal, stomach, and liver cancers) account for more than two-thirds of all cancer patients in Huai'an<sup>28</sup>. Therefore, on the basis of the cancer surveillance data for incidence and mortality, three counties in Huai'an (Xuyi, Jinhu, and Chuzhou) with a high cancer incidence were selected as test group, and the Tongshan district of Xuzhou city, which has a low cancer incidence, was selected as the control group.

To study the important risk factors which may affect the disease incidence  $\pi$ , based on related literature and other information source about candidate risk factors of GI tumors, we select and measure 25 risk factors for each sample of water, including 15 PAHs and 10 heavy metals, and the  $j$ th factor is denoted as  $F_j$ , whose observations are denoted as  $f_{ij}$ .

Firstly, we give the basic information of all the 25 risk factors by two box graphs, where one is corresponding to test group and the other to control group. However, considering that observations correspond to different substances, which suggest that the values may not be comparable to each other, and each variable may have high variability, we make data transformation on raw data by monotone function below

$$z_{ij} = \frac{1}{1 + \exp[-(f_{ij} - \alpha_j)/\beta_j]}, \quad (10)$$



**Figure 1. Box graphs of 25 candidate risk factors.**

NAP(V1) 0.000	ANY(V2) 0.000	ANA(V3) 0.000	FLU(V4) 0.000	PHE(V5) 0.000
ANT(V6) 0.000	FLT(V7) 0.0040	PYR(V8) 0.000	BaA(V9) 0.004	CHR(V10) 0.004
BbF(V11) 0.004	BkF(V12) 0.000	BAP(V13) 0.000	IPY(V14) 0.000	DBA(V15) 0.000
Cr(V16) 0.012	Mn(V17) 0.048	Ni(V18) 0.104	Cu(V19) 0.832	Zn(V20) 0.000
As(V21) 0.000	Cd(V22) 0.004	Ba(V23) 0.000	Hg(V24) 0.000	Pd(V25) 0.000

**Table 1. P-Values corresponding to different water pollutant risk factors.**

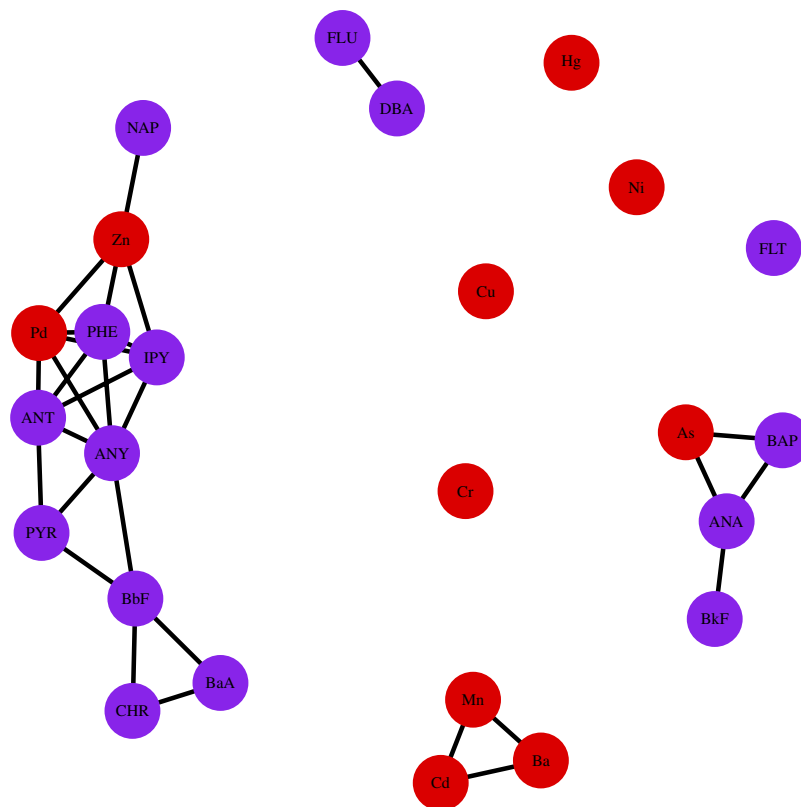
where the parameters of  $\alpha_j$  are median of  $f_{ij}$ , and  $\beta_j$  are median absolute deviation of  $f_{ij}$ . Through data transformation on raw data, all the observations can be confined into interval of (0, 1), which can ensure that the data of each variable is comparable in a single box graph. In fact, this transformation is not necessary for analysis in this article, because monotone transformation will not change the values of  $R_X(t)$  and  $R_Y(t)$ , therefore the ROC should be same to former ones.

The box graphs on transformation  $z_{ij}$  are shown in Fig. 1. According to this figure, the values of many factors have high variation, which means that it is not very reliable to perform conventional statistical analysis based on such investigation data. For example, we perform variable selection by logistic model for each variable at one time, and only 5 variables are selected at level  $\alpha = 0.05$ : BkF (V12), Cr (V16), Zn (V20), As (V21) and Ba (V23). Through probit model, we obtain similar results.

Secondly, we classify the samples with high and low cancer incidences as the test group and the control group, which are denoted as  $D = 1$  and  $D = 0$ , respectively. And then, for each  $F_j$  from 25 risk factors, based on the information of  $D \in \{1, 0\}$ , we calculate its ROC curve  $R_j(t)$  and obtain the value of test statistic  $S_{A_j}$  based on equation (6). Then, through process of simulation for  $n = 1,000$  times, we obtain the empirical distribution of test statistic  $S_A$  under null hypothesis  $H_0$ , and then make hypothesis testing and give p-value of each observation of  $S_{A_j}$ . The p-values are shown in Table 1, and Cu (V19) is excluded from candidates at level  $\alpha = 0.05$ .

Furthermore, we also carry out variable selection by random forest based on the R package ‘randomForest’ for comparison. The parameter of ‘ntree’ is 1,000, and the measurement of importance for variables is ‘Accuracy’. We also provide the top 10 variables: BkF (V12), ANY (V2), NAP (V1), Hg (V24), IPY (V14), FLU (V4), BAP (V13), PYR (V8), ANT (V6) and DBA (V15). However, if the measurement of importance for variables is changed to ‘Gini’, then the top 10 variables are: BkF (V12), ANY (V2), NAP (V1), FLU (V4), FLT (V7), Cr (V16), ANA (V3), PYR (V8), BAP (V13) and CHR (V10). These results show that methods based on ROC and random forest, as nonparametric methods, give close conclusions, and the results are accorded with experimental study, which implies good performance of nonparametric methods.

Finally, we obtain measurement  $\hat{r}_{ij}$  between any pair of variables  $F_i$  and  $F_j$  based on equation (8) and obtain matrix of  $(\hat{r}_{ij})_{25 \times 25}$ , wherein the values lower than threshold at given level  $\beta$  are converted into 0. Then, based on the R package ‘igraph’, we take  $(\hat{r}_{ij})_{25 \times 25}$  as adjacency matrix to construct network, where each node  $V_k$  corre-



**Figure 2.** Network of water pollutant risk factors based on  $(\hat{r}_{ij})$  as  $\beta = 0.01$ .

sponds to certain factor  $F_k$ , and the purple dots and red dots denote the PAHs and heavy metals, respectively. The networks are shown in Figs 2 and 3, where the level  $\beta$  take values of 0.01 and 0.02, respectively.

Based on analysis of network, we can see that almost all the PAHs act as a group, and these results match the studies of PAHs and heavy metals for environmental pollution, such as air pollution, and cancer development<sup>25,29–31</sup>. In particular, we find that heavy metal As (V21) has strong connection with most PAHs. This finding of connection between As (V21) and PAHs may imply the existence of PAHs-arsenic co-contaminated sites<sup>32</sup>, because many PAHs-arsenic co-contaminated sites, such as wood preservation sites, coking or chemical industry sites, and mining or metallurgy industry sites, are common around our survey locations. This finding may indicate the importance of remediation technologies for PAHs-arsenic combined pollution in the future, such as microbial degradation methods<sup>33–35</sup>.

For comparison, we also use Pearson sample correlation coefficient matrix  $(\hat{P}_{ij})_{25 \times 25}$  as adjacency matrix to construct network. Similar to the process on matrix  $(\hat{r}_{ij})_{25 \times 25}$ , the ones lower than the threshold through test hypothesis on coefficient are converted into 0, and the network based on  $(\hat{P}_{ij})$  as  $\beta = 0.01$  is shown in Fig. 4. We can find that this network can hardly give more information, and this phenomenon may be resulted from the sensitivity of  $P_{ij}$  on outliers of observation and the nonlinear relationship between some pairs of variables of  $F_i$  and  $F_j$ .

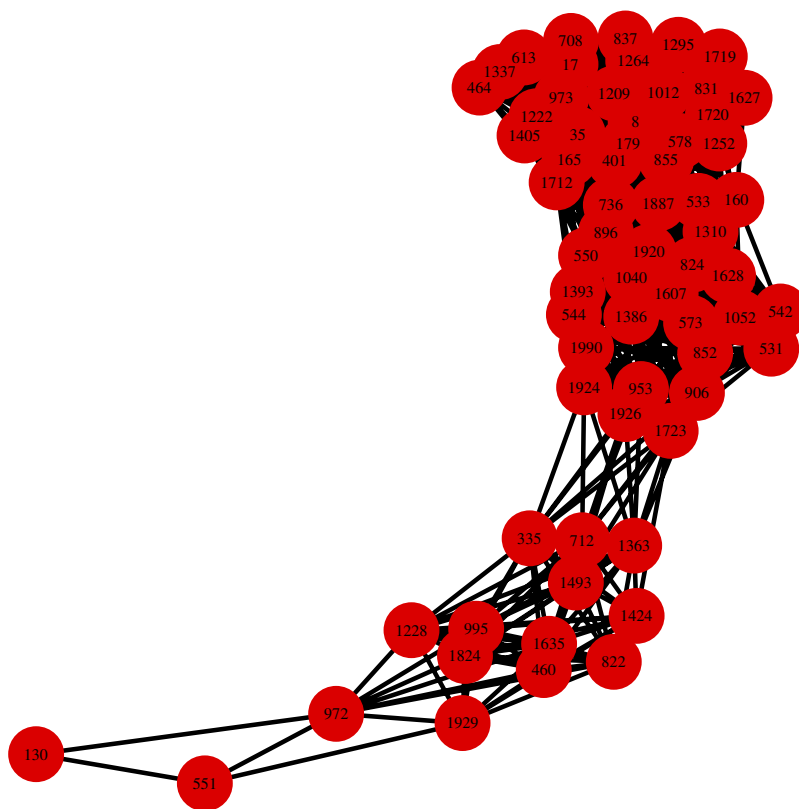
**Example 2.** To show more application of this method, we also use it to analysis of gene expression data in colon tissues, where the data is produced by U. Alon (1999). In this data set, the gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes, and two thousand out of around 6,500 genes were selected based on the confidence in the measured expression levels<sup>36</sup>.

In this example, we consider the genes in this data set as risk factors, and obtain about 100 genes as  $\beta = 0.01$ . Through the annotations of these candidate genes, we note that there are some genes having function connected with tumor of colon. For example, cadherins are the principal components of Adhesion Junctions (AJs) and cluster at sites of cell-cell contact in most solid tissues. These cell adhesion molecules play a significant role in the development of colorectal cancer and mediate the metastases of this common malignancy. Loss or downregulation of E-cadherin expression is a significant feature for colorectal cancer progression or the development of metastases<sup>37,38</sup>. Furthermore, besides E-cadherin, some other genes involved in the signaling pathway of Adhesion Junctions (AJs), including LAR protein, DEP1 (Protein Tyrosine Phosphatase), alpha-catenin, alpha-actinin and actin, also appear in this candidate gene set. These genes together with their annotations are shown in Table 2. The fact that quite a few genes in Adhesion Junctions coexist in the filtered gene set indicates this method can be used to screen for the genes related to colorectal cancer.



ID	Name	Description
481	R09468	PROTEIN-TYROSINE PHOSPHATASE PTP-S (Rattus norvegicus)
806	Z13009	H. sapiens mRNA for E-cadherin
1337	R70016	Human F-actin capping protein beta subunit mRNA, complete cds
1393	X55187	Human mRNA for alpha-actinin, partial cds
1720	U03100	Human alpha2(E)-catenin mRNA, complete cds
1929	Y00815	Human mRNA for LCA-homolog, LAR protein (leukocyte antigen related)

**Table 2.** Annotations for some genes in signaling pathway of Adhesion Junctions.



**Figure 5.** Network of genes in colon data based on  $(\hat{r}_{ij})$  as  $\beta = 0.01$ .

We also construct network based on  $(\hat{r}_{ij})$  corresponding to these candidate genes as level  $\beta = 0.01$  and the result is shown in Fig. 5, through which we find that these genes can be roughly divided into two groups, where the six genes except LAR protein coexist in one group, and LAR protein is in the other group. It suggests that our method may give clues to connections among genes.

The data sets used in Example 1 are presented in files of “Supplementary Data 1.csv” and “Supplementary Data 2.csv”, which are the observations and classification information of samples, respectively. The data set used in Example 2 is produced by U. Alon (1999) and is available on the web at <http://www.molbio.princeton.edu/colon-data>. The programs for data analysis in Example 1 and Example 2 are presented in “Supplementary RiskFactor.R”.

## Discussion

In epidemiologic studies, because of high variability, complex structure among correlative factors, and individual differences of data, it is unreasonable to construct specific mathematical models directly to study the influence of risk factors on disease, while the proposed methods, as non-parametric statistical methods without severe mathematical conditions, such as normality or linear style as in classical statistical methods, are appropriate to explore the relationship between various risk factors and disease incidence. Specifically, ROC curve is only related with probability functions  $R_X(t)$  and  $R_Y(t)$ , and can be estimated directly by quantiles, thus the statistic  $S_A$  or  $\hat{r}_{ij}$  based on ROC curve is not sensitive to outliers, variability of data or individual differences, and can give more reliable conclusions.



Furthermore, although there is no explicit formulation between risk factors and disease incidence, according to equations (3) and (4), ROC curves imply the dose-effect relationship between selected risk factors and disease incidence, similar to classical linear model, which should help to evaluate and screen for the important candidate risk factors. Admittedly, because the larger values of the variable  $F$  do not necessarily increase or decrease the disease incidence  $\pi$  directly, this method may miss some factors without clear dose-effect relationship between  $F$  and  $\pi$ , therefore the users should pay attention to such limitation in real work.

In addition, as factors have complex correlation with each other, network analysis is a desirable choice to explore the complex interaction among different factors through many pairs of factors. The proposed method gives a nice visualization of the network based on correlation matrix  $r_{ij}$  among all risk factors. It is worth noting that the definition of  $r_{ij}$  is constructed by  $R_i(t)$  and  $R_j(t)$ , which uses the information both from risk factors and disease status, while the traditional correlation matrix only uses the information from risk factors. Thus, this method can give more important information in exploring complicated relationship between risk factors and the disease in epidemiologic studies, and is helpful for directing further experimental analyses.

Finally, as shown in the two examples, the proposed method may provide useful tools in other biomedical problems with similar data structure. We can screen risk factors and filter for certain connections  $r_{ij}$  in network by relatively objective criterion, namely quantile of distribution, which can be approximated by some function of Brownian bridge. Thus, it is desirable in real studies, especially for the problems with big data, where some criteria, such as number of selected objects or proportion of total candidates, may be inconvenient for further studies. Incidentally, because the obtained networks may be too complex to efficiently interpret, it is still necessary to improve the proposed method to simplify the networks more efficiently and reliably in the future.

## References

1. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360 (2001).
2. Zhang, C. Nearly unbiased of variable selection under minmax concave penalty. *The Annals of Statistics* **38**, 894–942 (2010).
3. Johnson, B. A. Variable selection semiparametric linear regression with censored data. *Journal of the Royal Statistical Society. Series B* **70**, 351–370 (2008).
4. Tibshirani, R. The lasso method for variable selection in cox model. *Statistics in Medicine* **16**, 385–395 (1997).
5. Fan, J. & Li, R. Variable selection for cox's proportional hazard models and frailty model. *The Annals of Statistics* **30**, 74–99 (2002).
6. Austin, P. C. & Tu, J. V. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology* **57**, 1138–1146 (2004).
7. Candolo, C., Davison, A. & Demtrio, C. A note on model uncertainty in linear regression. *Journal of the Royal Statistical Society. Series D* **52**, 165–177 (2003).
8. Clyde, M. & George, E. I. Model uncertainty. *Statistical Science* **19**, 81–94 (2004).
9. Archer, K. J. & Kimes, R. V. Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis* **52**, 2249–2260 (2008).
10. Genuer, R., Poggi, J. M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognition Letters* **31**, 2225–2236 (2010).
11. Kursu, M. B. Robustness of random forest-based gene selection methods. *BMC Bioinformatics* **15**, 8 (2014).
12. Schug, J., Schuller, W. P. *et al.* Promoter features related to tissue specificity as measured by shannon entropy. *Genome Biology* **6**, R33 (2005).
13. Sundaramurthy, G. & Eghbalnia, H. R. A probabilistic approach for automated discovery of perturbed genes using expression data from micorarray or rna-seq. *Computers in Biology and Medicine* **67**, 29–40 (2015).
14. Chen, X. O. & Blanchette, M. Prediction of tissue-specific cis-regulatory modules using bayesian networks and regression trees. *BMC Bioinformatics* **8**, S2 (2007).
15. Deng, S. G., Qi, J. C. & *et al.* Network-based identification of reliable bio-markers for cancers. *Journal of Theoretical Biology* **383**, 022–027 (2015).
16. Fawcett, T. An introduction to roc analysis. *Pattern Recognition Letters* **27**, 861–874 (2006).
17. Lloyd, C. J. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* **93**, 1356–1364 (1998).
18. Horváth, L., Horváth, Z. *et al.* Confidence bands for roc curves. *Journal of Statistical Planning and Inference* **138**, 1894–1904 (2008).
19. Bradley, A. P. Roc curve equivalence using the kolmogorov-smirnov test. *Pattern Recognition Letters* **34**, 470–475 (2013).
20. Baker, S. G. The central role of receiver operating characteristic (roc) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute* **95**, 511–515 (2003).
21. Rodríguez-álvarez, M. X., Tahoces, P. G. & *et al.* Comparative study of roc regression techniques-applications for the computer-aided diagnostic system in breast cancer detection. *Computational Statistics and Data Analysis* **55**, 888–902 (2011).
22. Hsieh, F. & Turnbull, B. W. Non-parametric and semi-parametric estimation of the receiver operating characteristic curve. *The Annals of Statistics* **24**, 25–40 (1996).
23. Liu, K. Q., Liu, Z. P. & *et al.* Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics* **13**, 126 (2012).
24. Lu, X. & Deng, E. A. Y. A co-expression modules based gene selection for cancer recognition. *Journal of Theoretical Biology* **362**, 75–82 (2014).
25. Callen, M. S., Lopez, J. M. *et al.* Nature and sources of particle associated polycyclic aromatic hydrocarbons (pah) in the atmospheric environment of an urban area. *Environmental Pollution* **183**, 166–174 (2013).
26. Demetriou, C., Raaschou-Nielsen, O. *et al.* Biomarkers of ambient air pollution and lung cancer: a systematic review. *Occupational and Environmental Medicine* **69**(9), 619–627 (2012).
27. Lim, W. Y. & Seow, A. Biomass fuels and lung cancer. *Respirology* **17**, 20–31 (2012).
28. Chen, W., Zheng, R. *et al.* Report of incidence and mortality in china cancer registries. *Chinese Journal of Cancer Research* **25**(1), 10–21 (2013).
29. Tran, G. D., Sun, X. D. *et al.* Prospective study of risk factors for esophageal and gastric cancers in the linxian general population trial cohort in china. *International Journal of Cancer* **113**, 456–463 (2005).
30. Diggs, D. L., Huderson, A. C. *et al.* Polycyclic aromatic hydrocarbons and digestive tract cancers: a perspective. *Journal of environmental science and health. Part C* **29**, 324–357 (2011).
31. Tchounwou, P. B., Yedjou, C. G. *et al.* Heavy metals toxicity and the environment. *EXS* **101**, 133–164 (2012).
32. Elgh-Dalgren, K., Arwidsson, Z. *et al.* Bioremediation of a soil industrially contaminated by wood preservatives-degradation of polycyclic aromatic hydrocarbons and monitoring of coupled arsenic translocation. *Water Air and Soil Pollution* **214**(1), 275–285 (2011).



33. I., S. O., V., K. V. & M., B. A. Rhizosphere bacteria *Pseudomonas aureofaciens* and *Pseudomonas chlororaphis* oxidizing naphthalene in the presence of arsenic. *Applied Biochemistry and Microbiology* **46**(1), 38–43 (2011).
34. Kozlova, E. V., Puntus, I. F. *et al.* Naphthalene degradation by *Pseudomonas putida* strains in soil model systems with arsenite. *Process Biochemistry* **39**(10), 1305–1308 (2004).
35. Ali, N., Dashti, N. *et al.* Indigenous soil bacteria with the combined potential for hydrocarbon consumption and heavy metal resistance. *Environmental Science and Pollution Research* **19**(3), 812–820 (2012).
36. Alon, U., Barkai, N. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.* **96**, 6745–6750 (1999).
37. Cavallaro, U. & Christofori, G. Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. *Nature Reviews Cancer* **4**, 118–132 (2004).
38. Paschos, K. A., Canovas, D. & Bird, N. C. The role of cell adhesion molecules in the progression of colorectal cancer and the development of liver metastasis. *Cellular Signalling* **21**, 665–674 (2009).

## Acknowledgements

This research was supported by grants from the China Ministry of Science and Technology (MOST) Project 973 (Grant No. 2012CB955501); the Fundamental Research Funds for the Central Universities.

## Author Contributions

J.A. conceived the experiments and analysis methods; J.J. and J.A. analyzed the data; S.Z. and Q.X. conducted the experiments; J.J. and S.Z. wrote the paper; Q.X. and S.Z. assisted with the interpretation of data. All authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing Interests:** The authors declare no competing financial interests.

**How to cite this article:** Jin, J. *et al.* Identification of risk factors in epidemiologic study based on ROC curve and network. *Sci. Rep.* **7**, 46655; doi: 10.1038/srep46655 (2017).

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



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017