Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2022;8(1): 37-44.

Original Article

Prediction of the breast cancer mortality rate and its effective factors using genetic algorithm and logistic regression

Mahdieh Mirzaie[1, 2], Yunes jahani[1, 2], Abbas Bahrampour[1, 2*]

[1]Modeling in Health Research Center, Institute for Futures Studies in Health,Kerman University of Medical Sciences, Kerman,Iran.
[2]Department of Biostatistics and Epidemiology, School of Public Health, Kerman University of Medical Sciences, Kerman, Iran.

ARTICLE INFO

Received 27.06.2021
Revised 29.08.2021
Accepted 02.12.2021
Published 15.03.2022

Key words:
Breast cancer;
Cross over;
Mutation;
Genetic algorithm;
Logistic regression.

ABSTRACT

Introduction: Logistic regression is one of the most common models used to predict and classify binary and multiple state responses in medicine. Genetic algorithms search techniques inspired by biology have recently been used successfully as a predictive model. The aim of present study was to use the genetic algorithm and logistic regression models in diagnosing and predicting factors affecting breast cancer mortality.
Methods: Data of 2836 people with breast cancer during the years 2014-2018 were examined. Information was registered in the cancer registration system of Kerman University of Medical Sciences. Death status was considered as the dependent variable, while age, morphology, tumor differentiation (grad), residence status, and place of residence were considered as independent variables. Sensitivity, specificity, accuracy, and area under the receiver operating characteristic (ROC) curve were used to compare the models.
Results: The logistic regression model determined factors affecting the breast cancer mortality rate, (with sensitivity (0.60), specificity (0.80), area under the ROC curve (0.70), and accuracy (0.77)), and also genetic algorithm model (with sensitivity (0.21), specificity (0.96), area under the ROC curve (0.58) and accuracy (0.87)) did so.
Conclusion: The sensitivity and area under the ROC curve of the logistic regression model were higher than those of the genetic algorithm, but the specificity and accuracy of the genetic algorithm were higher than those of the logistic regression. According to the purpose of the study, two models can be used simultaneously.

Introduction

The conventional method for prediction in medicine is logistic regression, but since we are dealing with nonlinear problems in the field of medicine, it is not appropriate to use logistic regression. On the other hand, it sometimes not possible to use logistic model alone for a conceptual model including all independent variables. In addition, the conventional statistical techniques are not sufficient for analysis of the big data.[1] In recent

*.Corresponding Author: abahrampour@yahoo.com

years, the tendency to use genetic algorithms has increased in data mining for classification.[2] Genetic algorithms (GA) were first proposed by John Holland as an optimization and a search technique.[3] GA Inspired by the principles of natural selection and natural science according to which the most suitable people in a population are those who have more survival due to easy adaptation to their environmental changes, and there is a possibility for them to reproduce more[4]). The advantage of genetic algorithms over other statistical methods is that they are able to solve problems for which there is no human expertise.[5] Since the algorithm considers several points in the search space in each iteration, the chance of converging to a local maximum is reduced.[6]

Breast cancer is caused by cellular changes which can be based on two theories of stem cells and gene mutations. In the first theory, there are cancer cells in the body from birth. In the second theory, natural cells are affected by environmental cancer factors. Breast cancer occurs when some cells in the breast grow out of control and spread near the tissues.[7] One of the most important mental concerns of the patient is to be aware of the future state of the disease, so determining valid prognostic factors can help physicians make appropriate treatment decisions.[8]

Numerous studies have compared logistic regression in predicting mortality and other consequences of various diseases with other prediction models.

According to the field of study, the diagnostic performance of genetic algorithm and logistic regression models was different. Genetic algorithms and logistic regression have also been used in various studies to diagnose breast cancer, but so far, no study has been conducted to predict death from this cancer using logistic models and genetic algorithms. So, the aim of this study was to compare the performance of genetic algorithm and logistic regression in predicting mortality rate of breast cancer and determining the factors affecting it.

Method

The data of this study was obtained from 2836 patients (whose information was registered in the cancer registration system of Kerman university of Medical Sciences) with breast cancer during the years 2014-2018. Death status was considered a dependent variable, its value was determined according to the presence or absence of death certificate in patients' files when extracting information from the system, while age, morphology, tumor differentiation, place of residence and residence status were considered as independent variables.

Randomly 70% of the data was regarded as the training set and the remaining 30% as the test set; logistic regression models (using R software) and genetic algorithm (using MATLAB software) for detection of variables affecting breast cancer mortality were fitted to the data. sensitivity, specificity, area under the ROC curve and accuracy were used to compare the predictive power of the models.

Logistic regression (LR) analyzes the relationship between a two or multi-level dependent discrete variable with several independent variables. so that, independent variables can be of any kind. Logistic regression is based on the relationship between the probability of membership in a group and one or more predictor variables. Since the predicted probability should be between zero and one, a simple linear regression technique

is not sufficient because the results are inconsistent and out of this range. if p was the probability of an event occurring and x1,…, xn were independent variables, The logistic regression model is as follows

$$\log\left(\frac{\rho}{1-\rho}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

Where $\frac{\rho}{1-\rho}$ is named the odds ratio. α, β1, and βn are model parameters. In the above equation, the logit conversion is used to relate the probability of group membership to a linear function of independent variables. parameters in the logistic regression model are estimated by the maximum likelihood method.

Genetic algorithms

The genetic algorithm encodes the decision variables of a research problem into strings of the finite length of the alphabet. The strings which are the selected answers to the problem are called chromosomes; and each letter in the string is called a gene. To follow the law of natural selection and in fact to select the most appropriate strings, we need a fitness function which determines the relative fitness of an answer.[9]

After encoding the answers and determining a fitness function, the genetic algorithm which is based on an iterative process works to evolve the answers in such a way that first an initial population of chromosomes is randomly generated from the search space. The fitness of all of them is assessed by calculating the fitness function. Following the survival mechanism of the best, the chromosomes with the highest fitness values are selected as parents. Parent chromosomes use genetic operators such as selection, cross over and mutation to generate new-generation chromosomes. Repetition of the generation continues until the algorithm converges to the best chromosome which is the optimal answer.

Selection operator

In a genetic algorithm, the selection is a set of rules that help determine the parents of the next generation. In order to be able to reproduce and produce the children of the next generation, the criterion for choosing parents is their fitness, but it is done randomly.[10]

Cross over operator

A combination operator combines one or more sections of a parent chromosome to produce offspring chromosomes. The produced chromosomes are not similar to any of the parents but have a combination of their characteristics and produce higher quality answers.[9]

Mutation operator

If two parents have the same genes, the cross over operator does not produce new offspring with better chromosomes, and the mutation operator randomly improves responses by modifying genes.[10]

In this study, one gene was considered for each independent variable, and a chromosome with a length of 5 was considered for each individual. A level of "insignificance" was added to the possible values of all genes to indicate the ineffectiveness of a variable in the state of cancer death. We are looking for a chromosome (law) that predicts the death of people with cancer. In fact, how much should be considered for each study variable

that a person's cancer leads to death. For this purpose, the fitness function was considered the average sensitivity and specificity of the chromosome that determines the state of death correctly. Therefore, an initial population of chromosomes are selected from the answer set as the input of the problem. In order to maximize the fit function, genetic operators are applied to them.

Results

The frequency levels of the study variables are summarized in Table 1. According to the clinician, the risk of breast cancer is 50 years. We have divided people into two age groups: over 50 years and under 50 years.

Table 1. Frequency of study variables

	Frequency	Percent
Death status		
Died	238	11.9
No-died	2498	88.1
Residence place		
Kerman	1426	50.3
Other cities	1410	49.7
Degree of tumor differentiation		
Good and moderate	2217	78.2
Poor and bad	619	21.8
Residence status		
Urban	2495	88.0
Rural	341	12.0
Morphology		
Neoplasm	671	23.7
Infiltrating	1887	66.5
Other	278	9.8
Age		
Over 50 years	1437	50.7
Under 50 years	1399	49.3

Logistic regression results

The logistic regression model was fitted to the training data set using R software. Cut-off point was determined based on reality,[11] considering that 12% of the subjects would die. Twelve percent was considered the probability cut-off point for calculating the logistic regression model. All variables were classified by determining the last base level (software default). Model coefficients and p-values for each variable are summarized in Table 2. According to the table 2, the variables of tumor differentiation, location, morphology, and age affect death. For the variable “the degree of tumor differentiation (β = -0.388)”, the odds of death for people with good and moderate differentiation is 32.2% less than those with poor and bad differentiation. For the variable “residence (β = 0.460)”, the chance of death for those living in Kerman is 1.58 times lower than those living in other cities. For the morphology variable, the chance of death for people with neoplasm morphology (β = 2.417) is 11.209 times higher than those with other morphology, and the chance of death for people with infiltrating morphology (β = 0.298) is 34% higher than those who have other morphologies. For the age variable, the chance of death for people under 50 years (β = -0.358) is 33% lower than that for people over 50 years.

Based on the classification table 3 for the logistic regression model, for test dataset the sensitivity was 0.60, the specificity was 0.80, the area under the ROC curve was 0.70 and the accuracy was 0.77.

Also, to increase the accuracy, the training and test sets (100 times) were randomly selected from the data. After fitting the model to the training data set and calculating the indices for the test data set, the average sensitivity was 0.56, the average specificity was 0.82 and the area under the ROC curve was 0.71.

Table 2. Parameter estimates of logistic regression model

variable	β	OR	P-value	Lower CI (OR)	Upper CI (OR)
Constant	-2.908	0.055	0.000	-	-
Rural living (reference level=urban living)	0.278	1.320	0.199	0.864	2.018
good and moderate differentiation (reference level=week and bad)	-0.388	0.678	0.043	0.466	0.988
Resident of kerman (reference level=other cities)	0.460	1.584	0.002	1.176	2.134
Neoplasm morphology (reference level=other)	2.417	11.209	0.000	5.546	22.656
Infiltrating morphology (reference level=other)	0.298	1.347	0.409	0.664	2.734
Age under 50 years (reference level= over 50 years)	-0.358	0.699	0.018	0.520	0.939

Table 3. Classification table for logistic regression model

Predicted	Observed	
	Not death	Death
Not death	592	39
death	145	60

Place of residence	Tumor differentiation	Morphology	Age	Residence status
Kerman	Good and moderate	neoplasm	over 50 years	urban

Figure 1. The law proposed by genetic algorithm

Genetic algorithm result

The genetic algorithm model was fitted to the data using MATLAB software. To implement the genetic algorithm, an initial population size of 50 was considered. The roulette wheel selection method, the single-point cross over method with a probability of 0.6, and the uniform mutation method with a probability of 0.2 were used. According to the fit of the best chromosome, the law proposed by the genetic algorithm is shown in figure1. Which means, algorithm predicts death for a person living in Kerman with good and moderate tumor differentiation, neoplasm morphology, age over 50 years and urban living status. Moreover, the above law has a sensitivity of 0.21, a characteristic of 0.96, a surface below the rock curve of 0.58 and an accuracy of 0.87 for test dataset. In addition to the value of the parameters mentioned above, the algorithm using the tournament selection method, the multi-point, two-point and uniform cross over method was also implemented. The model indices were higher using the parameters mentioned above, but the result didn't change for different values of cross over and mutation probability of 0.8-0.2, 0.6-0.5, 0.9-0.1 and 0.5-0.5.

Discussion

The present study was conducted to compare the logistic regression model and genetic algorithm in predicting the breast cancer mortality rate in Kerman Province in south of Iran. For the logistic regression model and genetic algorithm, sensitivity (0.61 and 0.19), specificity (0.81 and 0.97), area under ROC curve (0.58 and 0.74) and accuracy (0.74, 0.87) were obtained respectively. Therefore, the sensitivity and the area under the ROC indices

of the logistic regression model are higher in predicting death than the genetic algorithm, but the specificity and accuracy indices of the genetic algorithm model are higher than the logistic regression.

Meysam Jahani and Mehdi Mahdavi (2016) used genetic algorithm, neural network and logistic regression in a study to predict diabetes. In this comparison, the accuracy of neural network prediction is higher than other methods and the accuracy of logistic regression model is the least reported[12] which is consistent with the result of present study. In a study to diagnose ovarian cancer, Hey-Jiong Song et al. (2016) pursued two main objectives of marker selection for ovarian cancer diagnosis and classification of individuals in terms of cancer. To select markers, they used the genetic algorithm model, random forest, t-test and logistic regression to achieve the second target using the discriminant analysis method, k-nearest neighbor and logistic regression.[13] The results revealed that logistic regression acted better than the genetic algorithm which is somewhat consistent with the results of the present study. In a study by Der-Ming Liuo and Wei-Pen Chang (2015) they compared genetic algorithm, neural network, logistic regression and tree regression to diagnose breast cancer. In their study, the accuracy of the genetic algorithm was higher than the average accuracy of other methods;[10] it is in line with the results of the present study. Lei Ming Sun (2010) used logistic regression and genetic algorithm to screen people with obstructive sleep apnea and showed that the sensitivity and specificity of genetic algorithm was more than those of logistic regression.[14] Chun-Lang Chang and Ming-YuanhSu (2009) proposed three screening models for pancreatic cancer: genetic algorithm, logistic regression and neural network. They also showed that the area under the ROC curve was not different, but the sensitivity and specificity of the genetic algorithm were higher than those of the other two models. In the present study, however, the sensitivity of logistic regression is less than that of the genetic algorithm.[15] Synthia Arsalanian et al. (2006) decoded the prediction rules which nurses used to classify people suspected of having acute coronary syndrome using logistic regression and genetic algorithm; they concluded that the accuracy of the two models was the same which contradicts the results of the present study.[16]

Conclusion

The logistic regression model revealed that the odds of death for people with good and moderate tumor differentiation is less than those for people with bad and weak one; it is less for people with infiltrating morphology than for others; it is more for people living in Kerman than for those who live in other cities; and it is more for people over 50 than for those who are younger than 50. Predicted death is the law proposed by the genetic algorithm for urban people who have neoplasm morphology, good and moderate degree of differentiation, who live in Kerman and who are over 50 years of age. The sensitivity and area under the ROC curve of the logistic regression model were calculated more than those of the genetic algorithm, but the specificity and accuracy of the genetic algorithm were calculated more than those of the logistic regression. According to the purpose of the study, it is recommended

to use these two models together.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethical statements

This study was approved by the ethics committee of kerman university of medical sciences with id KMU.IR 1400.154.

Acknowledgements

The authors thank the cancer registration system of Kerman University of Medical Sciences for providing the information.

References

1. Chen T-C, Hsu T-C. A GAs based approach for mining breast cancer pattern. Expert Systems with Applications. 2006;30(4):674-81.

2. Al-Maqaleh B, Shahbazkia H. A Genetic Algorithm for Discovering Classification Rules in Data Mining. International Journal of Computer Applications. 2012;41:40-4.

3. Eken C, Bilge U, Kartal M, Eray O. Artificial neural network, genetic algorithm, and logistic regression applications for predicting renal colic in emergency settings. International journal of emergency medicine. 2009;2(2):99-105.

4. Holland JH. Adaptation in natural and artificial systems: MIT Press; 1992.

5. Johnson P, Vandewater L, Wilson W, Maruff P, Savage G, Graham P, et al. Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. BMC Bioinformatics. 2014;15 Suppl 16:S11.

6. Sivanandam SN, Deepa SN. Introduction to Genetic Algorithms: Springer Publishing Company, Incorporated; 2007.

7. Christophides D, Appelt AL, Gusnanto A, Lilley J, Sebag-Montefiore D. Method for Automatic Selection of Parameters in Normal Tissue Complication Probability Modeling. International journal of radiation oncology, biology, physics. 2018;101(3):704-12.

8. Celentano DD, Szklo M, Gordis L. Gordis epidemiology2019.

9. Sastry K, Goldberg D, Kendall G. Genetic Algorithms. In: Burke EK, Kendall G, editors. Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques. Boston, MA: Springer US; 2005. p. 97-125.

10. Liou DM, Chang WP. Applying data mining for the analysis of breast cancer data. Methods in molecular biology (Clifton, NJ). 2015;1246:175-89.

11. Kleinbaum DG, Klein M. Introduction to Logistic Regression. In: Kleinbaum DG, Klein M, editors. Logistic Regression: A Self-Learning Text. New York, NY: Springer New York; 2010. p. 1-39.

12. Jahani M, Mahdavi M. Comparison

of Predictive Models for the Early Diagnosis of Diabetes. Healthcare informatics research. 2016;22(2):95-100.

13. Song HJ, Yang ES, Kim JD, Park CY, Kyung MS, Kim YS. Best serum biomarker combination for ovarian cancer classification. Biomedical engineering online. 2018;17(Suppl 2):152.

14. Sun LM, Chiu HW, Chuang CY, Liu L. A prediction model based on an artificial intelligence system for moderate to severe obstructive sleep apnea. Sleep & breathing = Schlaf & Atmung. 2011;15(3):317-23.

15. Chang C-L, Hsu M-Y. The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer. Expert Systems with Applications. 2009;36(7):10663-72.

16. Arslanian-Engoren C, Engoren M. Using a Genetic Algorithm to Predict Evaluation of Acute Coronary Syndromes. Nursing research. 2007;56:82-8.