The characteristics of breast cancer research using big data analysis techniques: recurrence gene study or utilization behavior of hospital but sample statistic error.

Han-Jun Cho[1], Eui Seok Jeong[2]

1. Department of Biomedical Institute for Convergence at SKKU, Suwon, 16419, South Korea

2. Department of Ecological Science, Kyungpook National University, Sangju 37224, South Korea

Abstract

Breast cancer is a major cause of female death, and various big data analysis methods have been applied to breast cancer. This study lists cases in which big data analysis was applied to breast cancer research. In addition, statistics and percentages from each specific sample were proposed. However, research on the use of big data has a blind spot that relies on sample characteristics. Therefore, before sampling big data, statistical inference should be discussed more precisely through pre-examination and sample statistical errors should be reduced by professional statistical evaluation of the analysis method.

Introduction

Breast cancer(BRCA) is one of the most common cancers found in women. Also, according to the results reported in the National Cancer Center for 2021, one in four people dead from breast cancer[1]. Recently on the according to the results of the research, there has been great progress in the treatment technology of breast cancer. These methods are breast cancer research using big data. In addition, with the convenience and economy of the national health insurance system, as women's interest in breast cancer and health increases, more patients come to the hospital at an early stage and it is possible to detect it early[2]. In this study, we report the results of analysis of recurrence characteristics using machine learning(ML) and analysis of usage behavior using data provided by the The Cancer Genome Atlas(TCGA) in USA and, Health Insurance Review & Assessment Service in Korea[3].

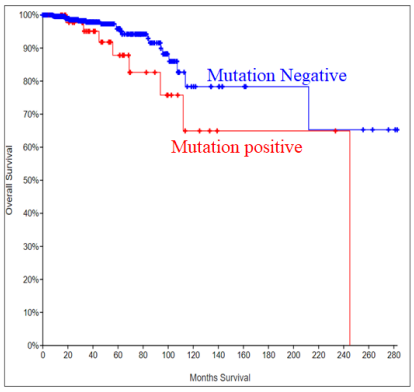
Methods

The Cancer Genome Atlas-BRCA provided data for 652 BRCA patients with somatic non-silent mutations and clinical information. They divided into two Disease Free/Recurred groups. To identify recurrence-related mutations, four feature selection methods (Information Gain, Chi-squared test, MRMR, Correlation) and four classifiers (Naïve Bayes, K-NN, SVM, Decision Tree) were used[4]. We

performed 5 fold-validations to find out the efficient algorithm. The network analysis for medical utilization was conducted using Cytoscape version 3.7.2. Dataset: Health Insurance Review & Assessment Service total patient sample(HIRA-NPS-2016, 2017, HIRA-APS-2016, 2017)[5].

Results

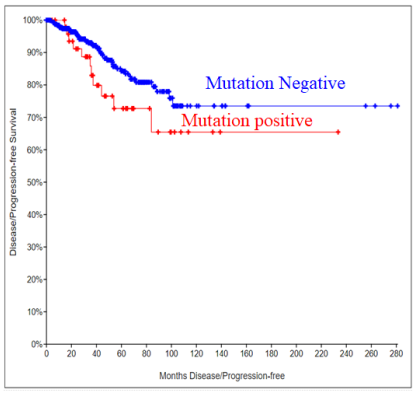
Case 1. A study using machine learning and mutated genes



▷ **Overall survival rate of 40 recurrence-related common genes (Information gain, Chi-squared test, MRMR, Correlation)**

- P-value : 0.0521

	Total	Decease	Median survival (Month)	Decease (%)
Mutation Positive	53	8	244.91	15.1 %
Mutation Negative	599	21	NA	3.5 %
				Δ11.6 %



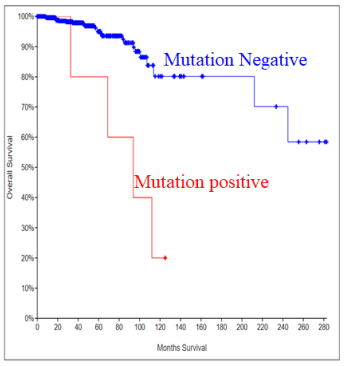
▷ **Disease free survival rate of 40 recurrence-related common genes (Information gain, Chi-squared test, MRMR, Correlation)**

- P-value : 0.107

	Total	Decease	Median survival (Month)	Decease (%)
Mutation Positive	53	11	NA	20.8 %
Mutation Negative	599	52	NA	8.7 %
				Δ12.1 %

Figure 1. Total 7 Recurrence of Kaplan-meier specific from Recurrence-related genes

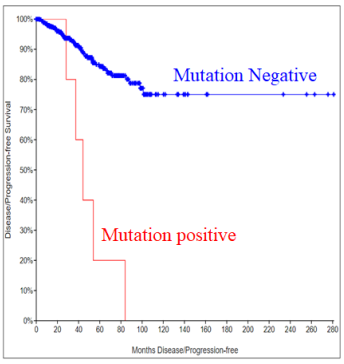
Among the 40 genes extracted by machine learning, 7 genes(ACSF3, ARID3B, KHSRP, LUZP2, RPL18A, TPI1, VWA5B2) highly related to breast cancer patients(supplementary table 2-1, 2-2). According to Kaplan Meyer statistics, the recurrence and survival prediction rates that would be expressed in all 7 specific mutant genes were rather closely related to the survival rate.



▷ **Overall survival rate of 3 recurrence-specific common genes (Information gain, Chi-squared test, MRMR, Correlation)**

- **P-value** : 8.384e-5

	Total	Decease	Median survival (Month)	Decease (%)
Mutation Positive	6	4	93.76	66.7%
Mutation Negative	646	25	NA	3.9%
				Δ62.8%



▷ **Disease free survival rate of 3 recurrence-specific common genes (Information gain, Chi-squared test, MRMR, Correlation)**

- **P-value :** 2.37e-7

	Total	Decease	Median survival (Month)	Decease (%)
Mutation Positive	6	5	44.12	83.3%
Mutation Negative	646	58	NA	9.0%
				Δ74.3%

Figure 2. Kaplan-meier rates(OS and DFS) for 3 survival-specific genes from 4 feature selection methods.

The Kaplan Meyer curve, which measured the expression rate of mutations in breast cancer patients with 3 (KHSRP, LUZP2, VWA5B2) genes, showed a very high predictive rate. This indicates that it can be easier, and only three genes can predict the stage of breast cancer patients.

Case 2. A study on how to apply diagnosis and algorithm using mutation feature selection in machine learning

Table 1. Optimization of 1-500(Option: random selection number) Derivation of Genetic Equivalence for Machine learning gene titration.

K	Accuracy	Precision	Recall	Classification error	Correlation
1	87.61%	43.80%	50.00%	12.39%	0.00%
2	87.95%	93.95%	51.37%	12.05%	15.52%
3	88.12%	94.03%	52.05%	11.88%	19.02%
6	88.29%	94.10%	52.74%	11.71%	21.98%
12	87.78%	93.88%	50.68%	12.22%	10.96%
22	88.46%	94.18%	53.42%	11.54%	24.60%
42	88.79%	94.33%	54.79%	11.21%	29.16%
77	87.78%	74.01%	51.86%	12.22%	13.37%
144	87.44%	60.52%	50.49%	12.56%	4.55%
269	87.61%	43.80%	50.00%	12.39%	0.00%
500	87.78%	93.88%	50.68%	12.22%	10.96%

The optimal algorithm combination was Information gain-Naïve bayes, and when diagnosis using 22-42 mutation-specific genes out of 40 genes, breast cancer can be detected early with an 88.79% probability. This problem arises because of the low proportion of relapsed patients and many non-recurring patients among all breast cancer patients in the data provided by the TCGA.

Case 3. A study on analyzing medical facility usage behavior using big data network technique.

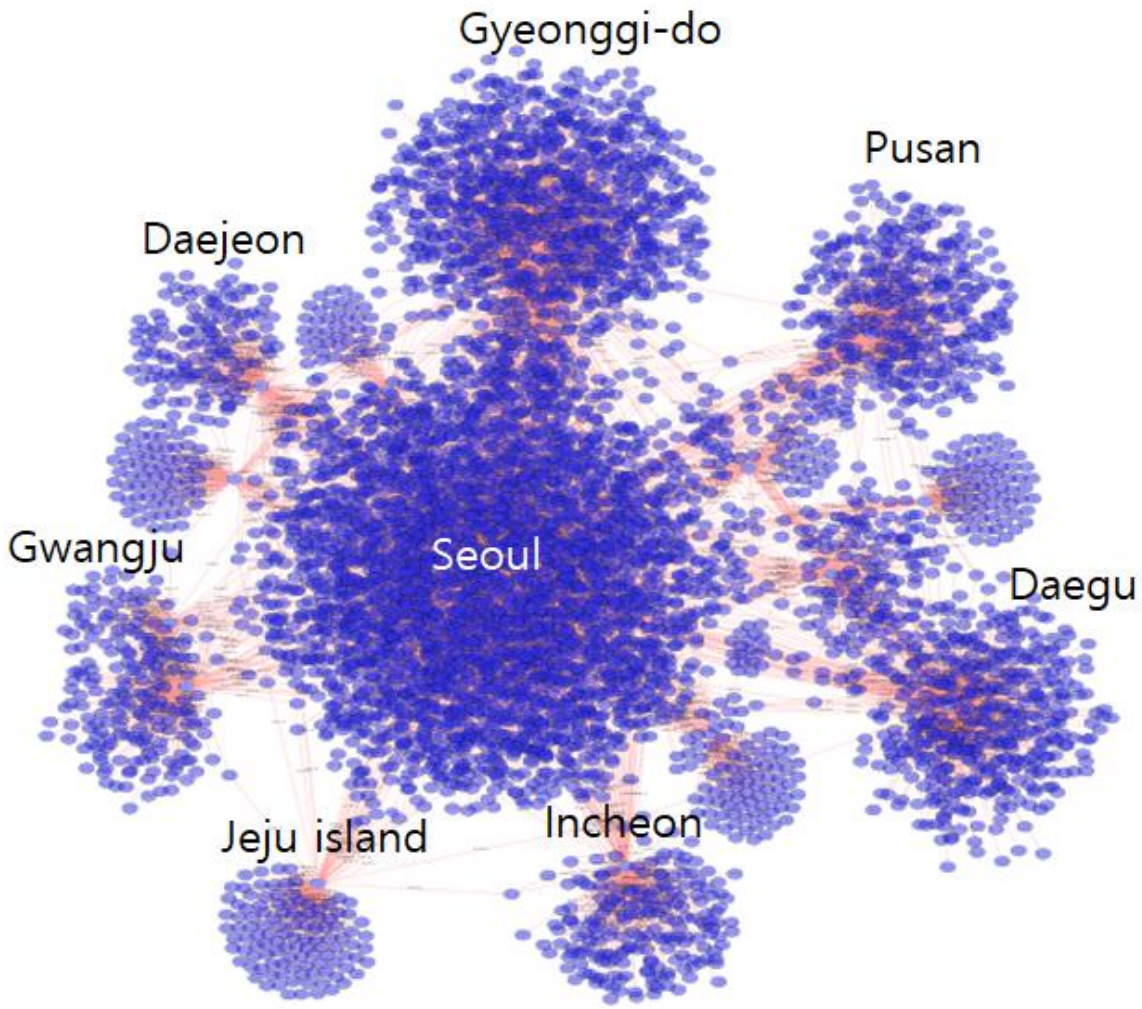


Figure 3. The regional distribution of BRCA patients and hospitals, its network

19% of total hospitals are located in Seoul, whiles 42% of total breast cancer patients visit hospitals in Seoul. This indicates strong seoul-centerism. Busan, Incheon, Daegu provinces also have higher patient visits compared to the percentage of hospitals. It all indicates metropolitan cities' domination in hospital utilization. The network indicates that only few metropolitan cities attract most of the breast cancer patients.

Discussion

The use of big data in cancer research is increasing day by day[6]. However, setting the sample itself is very important for big data. In the case of Figures 1 and 2 mentioned above, mutant genes that can be used as biomarkers show high predictive values of recurrence and survival rates, but in machine learning using real big data, as shown in Table 1, the appropriate number of mutant genes is determined, and The ratio and the expression amount of a particular gene are very important characteristics[7]. In addition, it is difficult to apply to cancer patient treatment because cancer patient's hospital used behavior and network analysis techniques show the regional characteristics of each country. In addition, since it depends on the population density shown in the sample, it is difficult to apply it to improving hospital use and service in countries with low population density[8].

It is good to try to utilize the big data that accumulates every day, but it is necessary to balance the data in order to be used in cancer research. In other words, even when using big data, the more data is used, the lower the accuracy, and the closer to disorder. However, filtering reduces the reliability of the data because the total amount in the sample is reduced[9]. Because there is such a prisoner's dilemma as the Nash equilibrium of big data research, when using big data, it is necessary to reset the sample that is combined with the ratio of the sample rather than the simple population being formed by the organization that provides it[10-11].

In order to overcome these shortcomings of research using big data, first, the use of cancer patient data must be openly open and a clear sample range must be established. Second, deviating from the research methodology, each journal needs an evaluation team to evaluate whether the use of big data is the right analysis method. Third, it was the utilization of big data increases, it can be applied to various fields, so essential big data education of experts in various fields is required.

Conclusion

In this study, genetic defects were important in the study of mutant genes using machine learning in breast cancer. In addition, when looking at the results of patients' hospital use behavior through network analysis, various studies using big data are possible. However, uncertainty in the data remains. An out-of-balance especially in the proportions of the sample warns of the danger. The statistic that supports the result is very important, but it can be a statistic that applies only to a specific sample.

Acknowledge

The authors thank Dong Hyeon Lee, Da Hyun Song, You Jeong Hong and Young Geon Ji for their technical assistance with collection of data.

References

1. Rakovitch, E., Sutradhar, R., Nofech-Mozes, S., Gu, S., Fong, C., Hanna, W., & Paszat, L. (2021). 21-Gene Assay and Breast Cancer Mortality in Ductal Carcinoma in Situ. JNCI: Journal of the National Cancer Institute, 113(5), 572-579.

2. Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Scientific Reports, 11(1), 1-13.

3. Yerrapragada, G., Siadimas, A., Babaeian, A., Sharma, V., & O'Neill, T. J. (2021). Machine learning to predict tamoxifen nonadherence among US commercially insured patients with metastatic breast cancer. JCO Clinical Cancer Informatics, 5, 814-825.

4. Cho, H. J., Lee, S., Ji, Y. G., & Lee, D. H. (2018). Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. PLoS One, 13(11), e0207204.

5. Shim, E. J., Lee, J. W., Cho, J., Jung, H. K., Kim, N. H., Lee, J. E., ... & Kim, Y. S. (2020). Association of depression and anxiety disorder with the risk of mortality in breast cancer: a National Health Insurance Service study in Korea. Breast cancer research and treatment, 179(2), 491-498.

6. Rehman, A., Naz, S., & Razzak, I. (2021). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimedia Systems, 1-33.

7. Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., ... & Schumacher, S. E. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell, 173(2), 338-354.

8.Herwartz, H., & Schley, K. (2018). Improving health care service provision by adapting to regional diversity: an efficiency analysis for the case of Germany. Health Policy, 122(3), 293-300.

9. Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H. P. (2016). Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710.

10. Campbell, H. R., & Lodder, R. A. (2021). Innovative Computational Methods for Pharmaceutical Problem Solving a Review Part I: The Drug Development Process.

11. World Health Organization. (2020). WHO report on cancer: setting priorities, investing wisely and providing care for all.