PATHOGENESIS NETWORK FOR BREAST CANCER

Jangho Kim, Yoobin Baik, Wonyoung Jang

Department of Data Science
Seoul National University, Korea
{smkjh1028, banca321, jwy4888}@snu.ac.kr

ABSTRACT

This project aims to understand the brief pathogenesis of breast cancer. First, we selected six factors that are expected to cause breast cancer. SNPs (Single Nucleotide Polymorphism) that influence these factors were then selected through GWAS, and the relationship between these SNPs and breast cancer was identified as a network through additional analysis such as logistic regression. Finally, Mendelian randomization was performed to find out how much each connection exerts its influence on breast cancer. In addition, we conducted the same analysis only for male breast cancer patients. As a result of the analysis, for all patients, we found that several pathways through years of education, white blood cell count, and alcohol intake frequency had direct or indirect effects on breast cancer. For male patients, several pathways through all factors except TDI had direct or indirect effects on breast cancer.

Keywords: Pathogenesis Network; Mendenlian Randomization; GWAS; UK Biobank

1 Introduction

Pathogenesis refers to the biological mechanisms that lead to disease states. Pathogenesis networks can explain the origin and development of diseases and whether the diseases are acute, chronic, or recurring. At the beginning of the project, we considered pathogenesis analysis for other diseases such as asthma, anxiety disorder, and arthritis, but we finally decided to build a network for breast cancer.

Breast cancer is one of the most likely known diseases to be genetic. Breast cancer results from genetic and environmental factors leading to the accumulation of mutations in genes [1]. Also, though the probability is much smaller than of women, many are unaware with the fact that men are also likely to develop breast cancer. In the hopes of raising awareness, we considered breast cancer as an adequate candidate for our project.

In this project, 350,000 unrelated White British were used as samples, which are from UK Biobank. We used called genotype variants (800,000) in autosomal chromosomes. Covariates were age, sex, and 10 principal component scores. GWAS procedure is implemented with plink2 software.

2 GWAS and Logistic Regression for Risk Factors

First, we selected six factors that are expected to affect breast cancer. The selected factors are as follows.

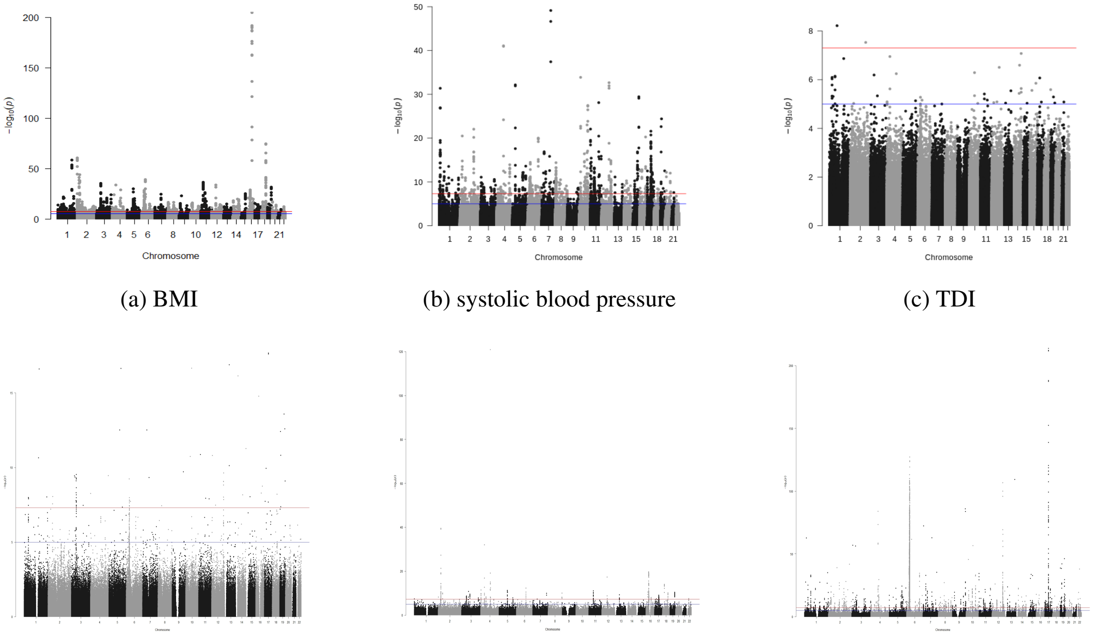
- BMI (Body Mass Index)
- systolic blood pressure
- Alcohol intake frequency
- White blood cell count
- Years of education
- TDI (Townsend Deprivation Index)

BMI was selected because there are many studies showing that increasing body weight increases the incidence of breast cancer. Though BMI is considered as an ideally suited measure for obesity, it has its limits. BMI can lead to an inaccurate assessment, because it is unable to distinguish lean muscle from fat mass. To supplement this weakness, another method was also considered, WHR(Waist-Hip Ration). WHR measures the ratio of one's waist circumference to one's hip circumference. Through this measurement, validation of obesity was expected. However, due to lack of data, we were not able to use this measurement as a risk factor.

Blood pressure was selected as a representative phenotype that causes various complications, and alcohol intake frequency was also selected based on the results of numerous researches on the association between alcohol and breast cancer. The white blood cell count was chosen because a study has shown that if a person with breast cancer has a large number of white blood cells, white blood cells are rather a factor that prevents breast cancer from recovering.

Years of education and TDI(Townsend Deprivation Index) were selected as social factors, each representing educational and economical conditions. Several researches has proposed that adults with higher educational attainment have better health and lifespans compared to their less-educated peers [2]. So the years of education was selected as a factor to represent the level of education each person received. TDI, a measure of material deprivation within a population, was selected to represent each persons economic level, considering the quality of life in a society is one of the most powerful determinants of health [3].

GWAS was performed for each risk factor. Figure 1 shows Manhattan plots for each risk factor.



(a) BMI

(b) systolic blood pressure

(c) TDI

(d) Years of Education

(e) Alcohol Intake Frequency

(f) White Blood Cell Count

Figure 1: Manhattan Plots for Risk Factors

For network construction, we selected SNPs with a p-value of 5×10^{-8} or less for each risk factor. In this project, five SNPs that recorded the smallest p-value in different chromosomes were selected. While only two SNPs met the criteria for TDI, significant amount of SNPs were detected in all other factors. Table 1 shows the list of SNPs selected in this way. For each risk factor, SNPs are arranged in ascending order of p-value.

Then, logistic regression was performed to determine the relationship between each risk factor and breast cancer, performing a total of six logistic regression analyses. As a result of the analyses, the p-values of years of education, alcohol intake frequency, white blood cell count, and systolic blood pressure were notable. However, TDI and BMI did not show statistically significant relationship with breast cancer.

3 Structuring Pathogenesis Network

The pathogenesis network was built based on the SNPs selected through GWAS for risk factors and the results of logistic regression analysis. The network was built using two algorithms - PC (constraint based) [4] and Hill Climbing (score based) [5].

BMI	systolic blood pressure	TDI
rs1421085	rs1173771	rs1526480
rs489693	rs12258967	rs6704753
rs574367	rs1458038	
rs62104180	rs2392929	
rs987237	rs2681492	

Years of Education	Alcohol Intake Frequency	White Blood Cell Count
Affx-52133101	rs1229984	rs8078723
Affx-89014260	rs780094	rs9271588
rs138043244	rs2726032	rs118203416
Affx-92042995	rs4800162	rs3184504
Affx-89014459	rs6884089	Affx-80237936

Table 1: Selected SNPs for Each Risk Factor

PC network was implemented with `pcalg` R package and Hill Climbing network was implemented with `bnlearn` R package. Visualization of the network in both packages is performed with `Rgraphviz` R package. PC algorithm is the first practical application of the Inductive Causation algorithm. The first step of PC is to search for the Markov blanket of each node. Each node is conditionally independent with other nodes given its Markov blanket. Undirected arc is drawn between two nodes if they do not share set of nodes which makes them conditionally independent. The second step is to find local v-structure (e.g A -> C <- B). If the adjacent node C is not in the set of nodes which makes the nodes A and B independent, the node C tends to be the common child of the nodes A and B. Lastly, the algorithm determines the direction of the arcs.

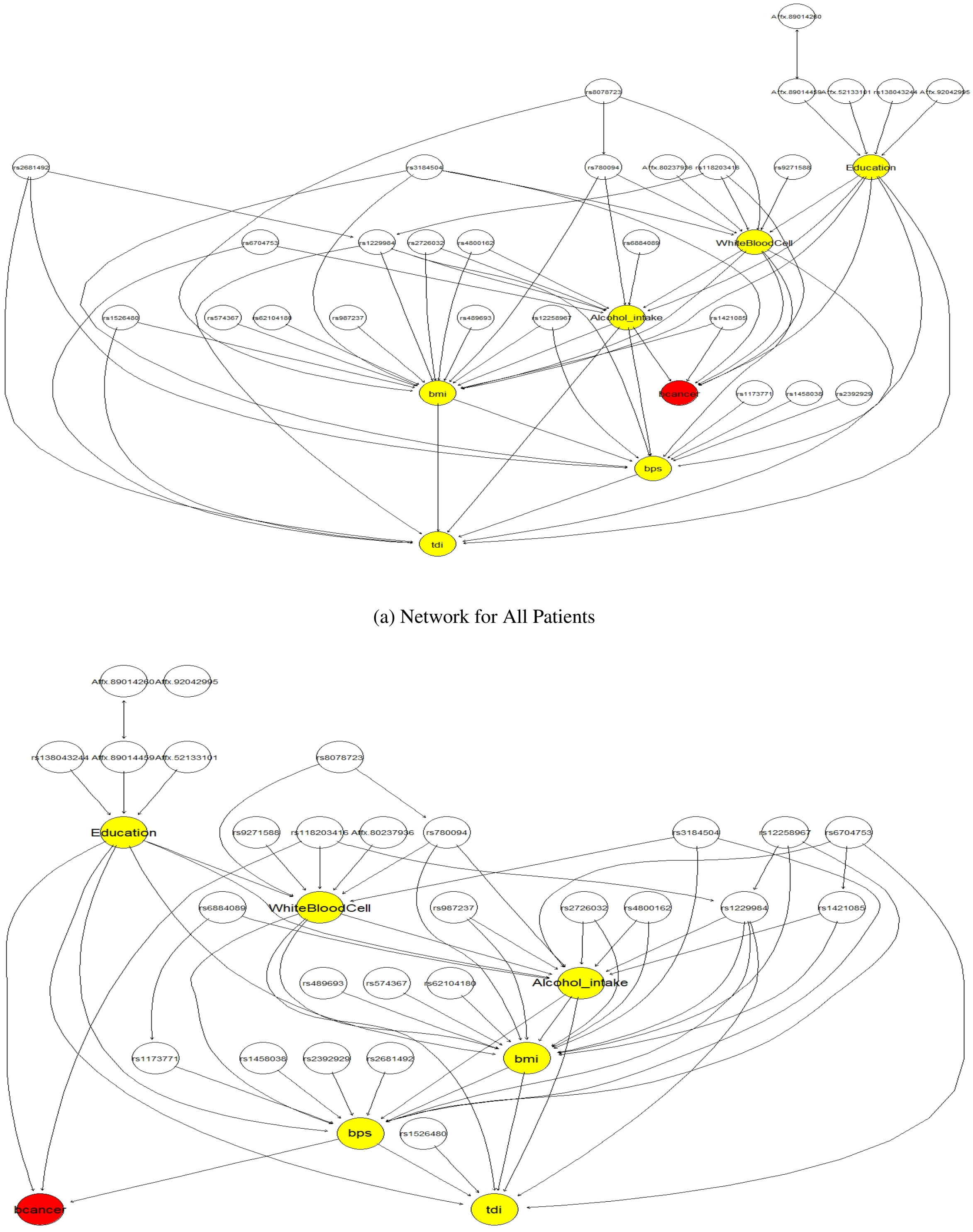
On the other hand, score based algorithms focus on whole structure of the graph and make scores for each equivalent class. Hill climbing is a type f greedy search algorithm. It adds or deletes arc one at a time and see whether the Bayesian Information Criterion (BIC) score can increase further. The critical drawback of score based algorithms is that the score can fall into local optima. Two algorithms (PC and Hill Climbing) was performed on whole sample and male only sample with 34 variables - breast cancer, six risk factors, and 27 SNPs. The missing values are filled with the mean of each variable. The dependence structure of network with Hill Climbing algorithm was unreasonable because the direction of the arcs headed from breast cancer to SNPs and there is no parent node of breast cancer. As a result, the network plotted with PC algorithm is selected for finding principal paths.

Figure 2 shows network structures from the PC algorithm. Figure 2a shows the network for all patients. The alcohol intake frequency directly affected breast cancer, while white blood cell counts and years of education had indirect effects. There were a total of 13 SNPs that indirectly affected through these risk factors. Figure 2b shows the network for male patients. In this case, several pathways through all factors except TDI had direct or indirect effects on breast cancer. Specifically, years of education and systolic blood pressure had direct effects on breast cancer. In addition, we could make an inference that alcohol intake frequency would make BMI higher and it leads to higher blood pressure, which can induce breast cancer.

4 Mendelian Randomization

After calculating the Mendelian Randomization for each selected SNPs, we were able to identify significant SNPs for each significant risk factor: white blood cell count, alcohol intake frequency, years of education. For white blood cell count, with the value 0.5163, SNP rs4800162 contributed to white blood cell count, which ultimately effected the incidence of breast cancer. SNP rs2726032 and rs489693 respectively with the value of -0.4552 and -0.3560 was concluded to affect breast cancer through alcohol intake frequency. Finally, with the value of -0.3697 and -0.3489, SNP rs138043244 and Affx.52133101 influence breast cancer through years of education. Table 2 shows Mendelian randomization results for significant SNPs. In Table 2b, TDI's SNP rs6704753 is included as it affects breast cancer pathway through alcohol intake frequency.

However, when experimenting with specifically only male breast cancer patients, we were able to discover different characteristics. Out of 223,467 male, 3,816 were diagnosed with breast cancer, and with this data, we were able to come across that, white blood cell count, years of education, alcohol intake frequency, and BMI were all significant



(a) Network for All Patients

(b) Network for Male Patients

Figure 2: Pathogenesis Network for Breast Cancer

SNP	White Blood Cell Count	Breast Cancer	Mendelian Randomization
rs4800162	0.0107	0.0055	0.5163
rs780094	-0.0472	-0.0011	0.0239
rs8078723	0.1612	0.0029	0.0181
rs9271588	0.1194	0.0097	0.0815

(a) Mendelian Randomization for White Blood Cell Count

SNP	Alcohol Intake Frequency	Breast Cancer	Mendelian Randomization
rs2726032	-0.0325	0.0148	-0.4552
rs489693	-0.0177	0.0063	-0.3560
rs6884089	-0.0271	-0.0068	0.2519
rs4800162	0.0264	0.0055	0.2082
rs780094	0.0481	-0.0011	-0.0234
rs6704753(TDI)	-0.0221	-0.0042	0.1899

(b) Mendelian Randomization for Alcohol Intake Frequency

SNP	Years of Education	Breast Cancer	Mendelian Randomization
Affx-89014459	8.3513	2.4693	0.2957
Affx-52133101	18.3501	-6.4028	-0.3489
rs138043244	17.3269	-6.4061	-0.3697
Affx-92042995	6.8270	-7.4060	-1.0848

(c) Mendelian Randomization for Years of Education

Table 2: Mendelian Randomization for The Significant Risk Factors

only through hypertension(high systolic blood pressure). Through this, we were able to ascertain that hypertension plays a huge role in breast cancer restricted to only males.

5 Conclusion

In this project, we investigated the brief pathogenesis of breast cancer. We found that several pathways through risk factors had direct or indirect effects on breast cancer not only for female patients but also for male patients. By building a pathogenesis network targeting limited SNPs for a few factors with a constraint-based PC algorithm and score-based Hill-Climbing algorithm, we were able to note that alcohol intake frequency, white blood cell count, and years of education had a significant influence on the incidence of breast cancer. Also, emphasize the fact that the risk of breast cancer for males is not something that one should ignore. Though the likelihood is subtle compare to females, through this experiment, we were able to discover a different characteristic; all other risk factors affect the incidence of breast cancer through hypertension(high systolic blood pressure). With more accumulated data, in the future we hope to build a network with many more various factors and SNPs, resulting in more diversified and dynamic causal inferences on breast cancer.

6 Contribution Statement

Jangho Kim shared research experiences and contributed to the network structuring. Yoobin Baik shared background knowledge for selecting appropriate risk factors. Wonyoung Jang took the lead in writing the report and presentation materials. All participants carried out experiments such as GWAS, and regression analysis. All participants provided critical feedback and helped shape the research, analysis, and report. All members contributed equally to this project.

References

[1] Katherine N Nathanson, Richard Wooster, and Barbara L Weber. Breast cancer genetics: what we know and what we need. *Nature medicine*, 7(5):552–556, 2001.

[2] Viju Raghupathi and Wullianallur Raghupathi. The influence of education on health: an empirical assessment of oecd countries for the period 1995–2015. *Archives of Public Health*, 78:1–18, 2020.

[3] Monica Townson. *Health and wealth: How social and economic factors affect our well being*. James Lorimer & Company, 1999.

[4] Marco Scutari and Jean-Baptiste Denis. *Bayesian networks: with examples in R*. CRC Press, Taylor amp; Francis Group, 2015.

[5] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.