Advancing Precision Medicine - Investigating The Relationship Between Race, Comorbidities, and Cancer Through Network Analysis

Isaac Elijah Kim Jr.

Spring 2018

Brown University

Department of Computer Science

Submission for APMA-CS Concentration

Adviser: Prof. Neil Sarkar

Reader 1: Prof. Sorin Istrail

Reader 1: Prof. Bjorn Sandstede

Contents

Abstract

The development of personalized treatments for diseases is increasingly plausible due to the increased availability of genomic data. For such efforts to be impactful, it is essential that they reflect the population of individuals that may be affected by a given disease. Amidst claims that there may be racial disparities in research populations, there have been no direct studies to explore this disparity in current research projects that involve genomic sequencing. The precise relationship between underrepresentation of certain races in genomic sequencing studies and health outcomes relative to these races is thus unknown. Here I examine the disparities in racial representation of national datasets pertaining to clinical data, mortality rates, and of a major initiative involving genomic sequence analysis (The Cancer Genome Atlas [TCGA]). The results suggest that Black Americans (BA) are severely underrepresented for most cancers in TCGA compared to clinical and mortality datasets, whereas Asian Americans (AA) were significantly overrepresented. Additionally, male Black Americans tend to be especially underrepresented in such genomic sequencing studies compared to their female counterparts. These findings accentuate the importance of targeted efforts to actively recruit representative patient populations into studies involving genomic sequencing.

In conjunction with these efforts to increase representation of racial minorities in genomic sequencing, I use the Health Care Utilization Program (HCUP) clinical dataset to examine the most frequently occurring cancers associated with patients of different races and comorbidities. I find that given the same combination of comorbidities, patients are at higher risk for different cancers based on their race, implying that race plays an important role in the onset of cancer and its progression.

Isaac Elijah Kim Jr.

Chapter 1

Racial Disparity of Genomic Sequencing

1.1 Introduction and Background

1.1.1 Genomic Sequencing

Genomic sequencing initiatives such as TCGA aim to catalogue cancer-associated genetic mutations towards the overall goal to diagnose, treat, and prevent cancer through better genetic understanding. While different genetic variants may result in similar symptoms, they could lead to diseases that require distinct, "personalized" treatments (Aronson, 2015). Genomic sequencing studies have led to major breakthroughs in the understanding of various types of cancer. In particular, research using TCGA has revealed that the genetic mutations responsible for breast cancer can be categorized into four major subtypes.

1.1.2 All of Us Research Program

As precision medicine initiatives are embarked upon, such as the All of Us Research Program in the United States, it will be essential to be aware of the potential gaps in genetic knowledge. The All of Us Research Program is designed to treat patients based on individual differences in lifestyle, environment, and biology including factors such as race. Studies have shown that genetic makeup across races can impact treatment regimens as well as outcomes. For example, some patients with localized prostate cancer are prescribed active surveillance as opposed to immediate treatment. Other studies have shown, however, that Black American candidates for

active surveillance had worse clinicopathological features on final surgical pathology than their white counterparts, suggesting that the criteria for active surveillance should be more rigorous for Black Americans(Ha, 2013).

1.1.3 Precision Medicine

The advancement of precision medicine requires genetic information on patients of all races. Adequate racial representation in these studies will lead to more effective targeted therapies and at least address the issue of racial disparities in national health measures. In this study, I highlight the current level of underrepresentation of racial minorities such as Black Americans in genomic sequencing studies, as well as identifying that Asian Americans are overrepresented, in the hopes of compelling researchers leading these studies to actively ensure appropriate balance of racial minorities. Specifically, a comparison of TCGA racial distribution to three national incidence databases (Health Care Utilization Program [HCUP] from the Agency for Healthcare Research Quality; mortality data mortality data from the Centers for Disease Control and Prevention [CDC]; and the National Cancer Institute's Surveillance, Epidemiology, and End Results Program [SEER]) reveals a notable discordance.

1.1.4 Personal Contributions

I accessed anonymous clinical information from the HCUP database through the BCBI's secure Stronghold network. Using SQL, I extracted the relevant patient records and computed basic counts and ratios for each demographic. Having separated and placed the patient records of each demographic into their own CSV files, I used Julia to parse through the information and determine the sinuosity indices for each cancer. Finally, I used Excel to plot the curves and an online statistics tool to calculate analysis of variance (ANOVA) statistics.

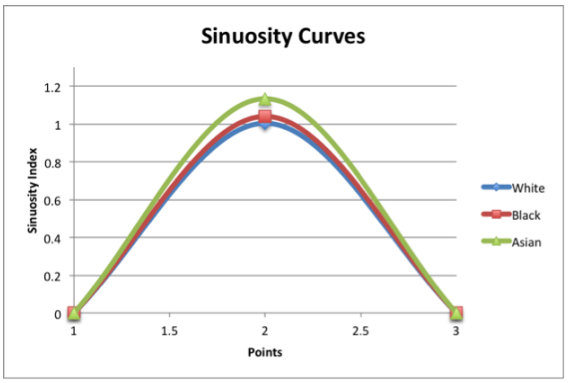


Figure 1.1: Sinuosity Curves

1.2 Results

1.2.1 Representation

For the cancers included in TCGA, black Americans comprised less of the total sample population in TCGA as compared to HCUP, CDC, and SEER for 15 out of 19 cancers, while White Americans and Asian Americans were underrepresented in 11 and 4 out of 19, respectively. While prostate cancer showed the most underrepresentation in TCGA across all examined races, certain female-dominated cancers such as breast and uterus showed the highest representation of Black Americans in TCGA compared to HCUP, CDC, and SEER.

1.2.2 Sinuosity Index

I used sinuosity index, which is a measure of steepness of a curve, as a measure for racial discordance across the studied databases. The relative difference was quantified as the slope between the relative lowest and highest occurrence of a given population group. For the 16 cancers in which data in TCGA and at least two other datasets were listed, the mean sinuosity indices for White Americans, Black Americans, and Asian Americans respectively were $1.00642 \pm 00879[1.00023-1.03608], 1.04298 \pm 02936[1.00128-1.09715], and 1.15744 \pm 09114[1.03703-1.33156]$; the mean slopes were respectively $31.0961 \pm 0.1401[30.9656-31.5034], 32.4931 \pm 1.22428[31.0915-34.7281], 33.5291 \pm 1.0569[31.7173-35.3975]$. Figures 1.1 and 1.2 depicts the sinuosity curves pivoted according to their respective slope angles. As indicated by Figures 1.3-1.6, the p values for the variation between racial groups were 0.0000.

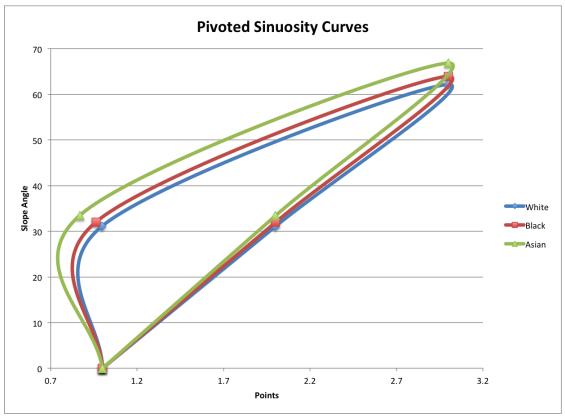


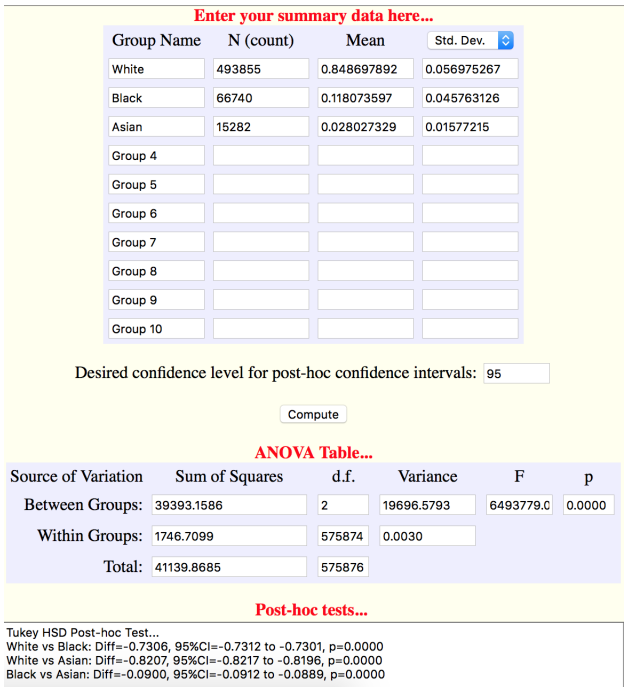
Figure 1.2: Pivoted Sinuosity Curves

1.2.3 ANOVA Statistics

1.3 Discussion

1.3.1 Current Literature

The underrepresentation of Black Americans in genomic sequencing potentially impacts racial disparities in U.S. health care, especially for complex genetic conditions. As a step towards overcoming these challenges toward developing genetically informatics healthcare regimens, it is of the utmost importance that the research community actively recruits more Black Americans into national genomic sequencing efforts, such as the All of Us Research Program. Efforts such as TCGA require patient consent in the procurement of tissue samples, which implies that Black Americans tend to withhold their samples from researchers. Previous studies have shown that Black Americans are significantly less likely than their white counterparts to participate in research that used their DNA, share their DNA with a private company, and permit their DNA to be used to generate cell lines for future research(Dye T., 2016). Moreover, studies have shown that Black Americans are much less likely to want the results of their genetic testing and indicate that the use of genetic testing should be promoted and available for those who seek them. This trend may derive from a general distrust in the medical system by the Black American community due to a history of research abuse and social injustice such as the case of Henrietta Lacks



Enter your summary data here...

Group Name	N (count)	Mean	Std. Dev.
White	493855	0.848697892	0.056975267
Black	66740	0.118073597	0.045763126
Asian	15282	0.028027329	0.01577215
Group 4			
Group 5			
Group 6			
Group 7			
Group 8			
Group 9			
Group 10			

Desired confidence level for post-hoc confidence intervals: 95

Compute

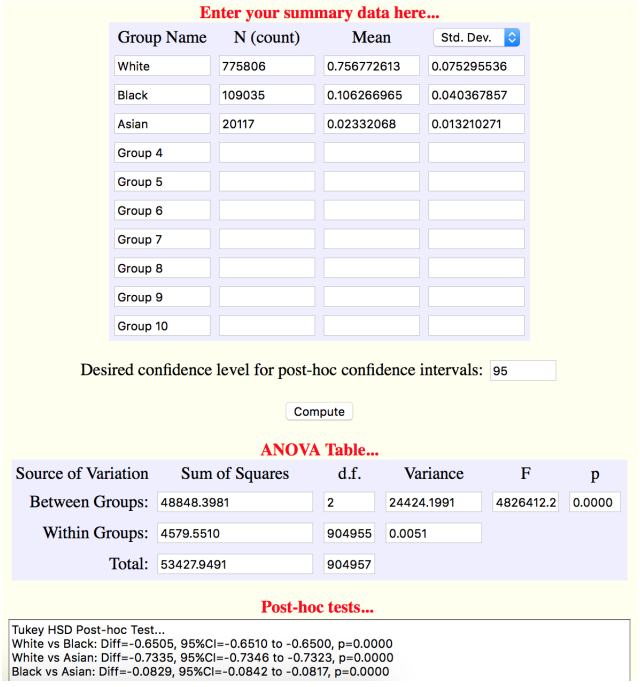
ANOVA Table...

Source of Variation	Sum of Squares	d.f.	Variance	F	p
Between Groups:	39393.1586	2	19696.5793	6493779.0	0.0000
Within Groups:	1746.7099	575874	0.0030		
Total:	41139.8685	575876			

Post-hoc tests...

Tukey HSD Post-hoc Test...
White vs Black: Diff=-0.7306, 95%CI=-0.7312 to -0.7301, p=0.0000
White vs Asian: Diff=-0.8207, 95%CI=-0.8217 to -0.8196, p=0.0000
Black vs Asian: Diff=-0.0900, 95%CI=-0.0912 to -0.0889, p=0.0000

Figure 1.3: ANOVA Statistics for CDC Dataset



Enter your summary data here...

Group Name	N (count)	Mean	Std. Dev.
White	775806	0.756772613	0.075295536
Black	109035	0.106266965	0.040367857
Asian	20117	0.02332068	0.013210271
Group 4			
Group 5			
Group 6			
Group 7			
Group 8			
Group 9			
Group 10			

Desired confidence level for post-hoc confidence intervals: 95

Compute

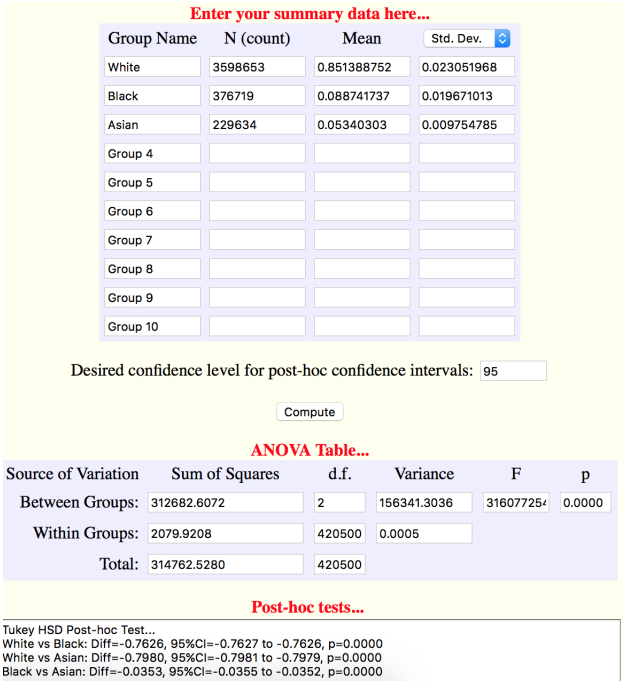
ANOVA Table...

Source of Variation	Sum of Squares	d.f.	Variance	F	p
Between Groups:	48848.3981	2	24424.1991	4826412.2	0.0000
Within Groups:	4579.5510	904955	0.0051		
Total:	53427.9491	904957			

Post-hoc tests...

Tukey HSD Post-hoc Test...
White vs Black: Diff=-0.6505, 95%CI=-0.6510 to -0.6500, p=0.0000
White vs Asian: Diff=-0.7335, 95%CI=-0.7346 to -0.7323, p=0.0000
Black vs Asian: Diff=-0.0829, 95%CI=-0.0842 to -0.0817, p=0.0000

Figure 1.4: ANOVA Statistics for HCUP Dataset



Enter your summary data here...

Group Name	N (count)	Mean	Std. Dev.
White	3598653	0.851388752	0.023051968
Black	376719	0.088741737	0.019671013
Asian	229634	0.05340303	0.009754785
Group 4			
Group 5			
Group 6			
Group 7			
Group 8			
Group 9			
Group 10			

Desired confidence level for post-hoc confidence intervals: 95

Compute

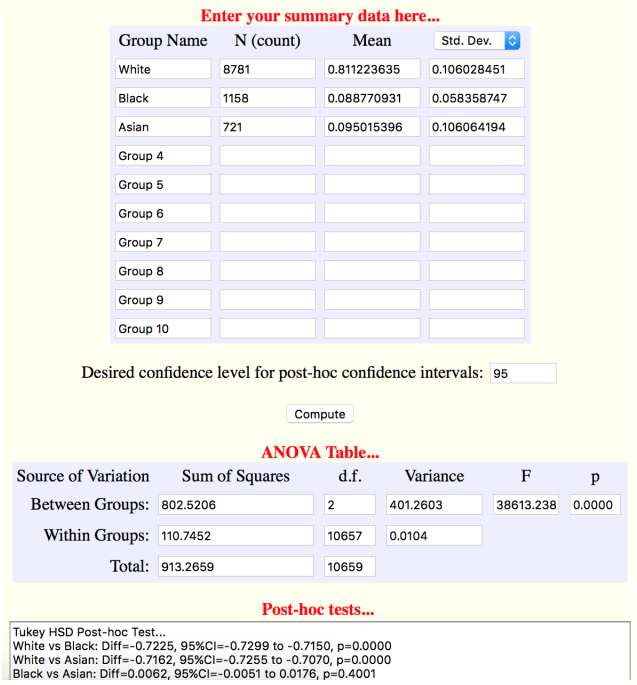
ANOVA Table...

Source of Variation	Sum of Squares	d.f.	Variance	F	p
Between Groups:	312682.6072	2	156341.3036	31607725	0.0000
Within Groups:	2079.9208	420500	0.0005		
Total:	314762.5280	420500			

Post-hoc tests...

Tukey HSD Post-hoc Test...
White vs Black: Diff=-0.7626, 95%CI=-0.7627 to -0.7626, p=0.0000
White vs Asian: Diff=-0.7980, 95%CI=-0.7981 to -0.7979, p=0.0000
Black vs Asian: Diff=-0.0353, 95%CI=-0.0355 to -0.0352, p=0.0000

Figure 1.5: ANOVA Statistics for SEERS Dataset



Enter your summary data here...

Group Name	N (count)	Mean	Std. Dev.
White	8781	0.811223635	0.106028451
Black	1158	0.088770931	0.058358747
Asian	721	0.095015396	0.106064194
Group 4			
Group 5			
Group 6			
Group 7			
Group 8			
Group 9			
Group 10			

Desired confidence level for post-hoc confidence intervals: 95

Compute

ANOVA Table...

Source of Variation	Sum of Squares	d.f.	Variance	F	p
Between Groups:	802.5206	2	401.2603	38613.238	0.0000
Within Groups:	110.7452	10657	0.0104		
Total:	913.2659	10659			

Post-hoc tests...

Tukey HSD Post-hoc Test...
White vs Black: Diff=-0.7225, 95%CI=-0.7299 to -0.7150, p=0.0000
White vs Asian: Diff=-0.7162, 95%CI=-0.7255 to -0.7070, p=0.0000
Black vs Asian: Diff=0.0062, 95%CI=-0.0051 to 0.0176, p=0.4001

Figure 1.6: ANOVA Statistics for TCGA Dataset

and the Tuskegee syphilis experiment(Dye T., 2016).

1.3.2 Recommendations

Racial disparities in genomic sequencing may either be related to or are the direct results of the perception of how healthcare systems and research communities. For results of initiatives like the All of Us Research Program to have practical meaning to the general population, the scientific community must intently ensure that large-scale genomic sequencing efforts are representative of the range of racial backgrounds. For example, the adequate representation of female Black Americans relative to their male counterparts in TCGA indicates that these recruitment efforts into genomic sequencing studies should especially target male Black Americans. For prostate cancer, the sinuosity indices for White Americans, Black Americans, and Asian Americans were 1.01355, 1.08430, and 1.23926, indicating that Black Americans and Asian Americans were more underrepresented in TCGA compared to White Americans. Interestingly, prostate cancer mortality for Black Americans is more than twice the rate observed in White Americans(I.J., 2007). In contrast, Asian Americans showed overrepresentation in genomic sequencing studies, leading to disproportionately high sinuosity indices across examined cancers. My findings suggest that recruiting efforts to target Asian Americans in genomic sequencing studies may not be nearly as vital as those seeking Black Americans.

It should be noted that the ANOVA statistics indicate a statistically significant difference, but it is difficult to fully characterize the magnitude of the difference given the p-values of 0. Thus, I recommend the development and use of alternative statistical measures such as the sinuosity index and slope to quantify this magnitude.

It is worth noting that the difference in the network complexities between races may be due to data availability, not necessary race per se. This issue may be related to the complexity of considering race and analyzing the phenotypes. In particular, are the observed correlations due to genetics, healthcare access, or environmental factors including socioeconomic status? This question raises the argument that race is not biological but is rather a social construct.

1.3.3 Related Studies

The challenge in ensuring diversity in large scale initiatives has been acknowledged. For instance, the 1000 Genomes Project analyzed the genomes of 1,092 individuals from 14 populations ranging from people with African ancestry in Southwest United States to Han Chinese in Beijing, China to British from England and Scotland. Given the importance of racial makeup in patient treatment and health outcomes, all genomic sequencing studies should follow suit to initiatives such as the 1000 Genomes Project and contain racial diversity(Consortium, 2012). However, it is important to note that as initiatives like All of Us launch into recruitment efforts, diversity alone is not sufficient. Diversity must be complemented with ensuring that it is with comparable frequencies relative to actual population. Otherwise, the research and clinical community risk the challenge of arriving at putative treatments that are of little utility to significant portions of the population who would benefit the most from personalized medicine approaches.

1.4 Methods

1.4.1 Datasets

I analyzed and compared four datasets: (1) the 2012 National Inpatient Sample (NIS) from the Healthcare Cost and Utilization Project (HCUP), (2) 2012 mortality data from the Centers for Disease Control and Prevention's (CDC) National Vital Statistics Report, (3) genomic sequencing data from TCGA, and (4) 1973-2014 incidence data from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program (SEER). Based on the disease codes outlined in the Clinical Classifications Software for the International Statistical Classification of Diseases and Related Health Problems, I extracted patient race information according to cancer groups from HCUP and SEER.

HCUP included the races White, Black, Hispanic, Asian or Pacific Islander, Native American, Other, and Invalid, while SEER catalogued a patient's race as White, Black, American Indian, as well as a range of ethnicities belonging to the Asian or Pacific Islander demographic, Other, or

Unknown. I directly analyzed the CDC's mortality data from its corresponding website, while TCGA's genomic sequencing data was found in the National Cancer Institute's Genomic Data Commons Data Portal. CDC divided race into White, Black, American Indian or Alaska Native, and Asian or Pacific Islander, whereas TCGA listed the races White, Black or African American, Asian, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Other, and Not Reported. I graphed the tallies according to overall cancer (see Supplementary Figures 1-27) and cancer, gender, and race for white, black, and Asian Americans (see Supplementary Figures 28-109) using the Plotly.jl Julia package. For certain cancers, data only existed in a single dataset, leading to the omission of the following codes in the plotted graphs: 23 (other non-epithelial cancer of skin), 28 (cancer of other female genital organs), 31 (cancer of other male genital organs), and 34 (cancer of other urinary organs).

For each given cancer and race in which at least three datasets were represented, I sorted all percentage values in ascending order into an array. I then assigned the highest value an adjusted value of 1, while all other values were divided by the highest value.

1.4.2 Sinuosity Index

For the adjusted sinuosity index of each examined cancer and race, in traversing a sorted array of three values, I gave the lowest value an x-value of 1 and set it to the front index of a new 2D array, the second an x-value of 3 and set it to the last, and the third an x-value of 2 and set it to the middle index. In traversing a sorted array of four values, I gave the lowest value an x-value of 1 and pushed it to the front index of a new 2D array, the second an x-value of 4 and set it to the last index, the third an x-value of 2 and to the second index, and the fourth an x-value of 3 and set it to the third index. I calculated the euclidean distance between the first and last points in the sorted 2D array and designated it as the B value. I calculated the sum of the euclidean distances between neighboring points (i.e., first and second, second and third, third and fourth, etc.) in the sorted 2D array and designated it as the A value. I calculated the sinuosity index as A/B.

For the adjusted slope angle of each cancer and race, I gave the lowest value an x-value of 1,

the second lowest an x-value of 2, and so forth. Then, I calculated the euclidean distance between the points with the lowest and highest y-values and designated it as the D value. Next, I calculated the difference between the points with the lowest and highest x-values and designated it as the C value. I calculated the slope angle using cos(C/D).

For each race, I plotted the sinuosity curve based on the median sinuosity index and then pivoted it according to its respective median slope angle. First, I graphed the sinuosity curve by assigning two points with coordinates (1,0) and (3,0) and a third point with an x-value of 2 and y-value of the median sinuosity index. Then, I pivoted the sinuosity curve so that the line connecting the points (1,0) and (3,0) now corresponded to the median slope angle.

1.4.3 ANOVA Statistics

I used an online statistics tool to calculate analysis of variance (ANOVA) statistics and compute p-values (Pezzullo).

1.5 Supplementary Tables

Table 1.1: Sinuosity Index and Slope Angle

Cancer (ICD-9-CM Code)	**Race**	**%-Adjusted Sinuosity Index**	**%-Adjusted Slope Angle (Degrees)**
Head and Neck (11)	White	1.00438	31.0737
Head and Neck (11)	Black	1.01063	31.3418
Head and Neck (11)	Asian	1.12017	33.1793
Esophagus (12)	White	1.01167	31.2422
Esophagus (12)	Black	1.06297	33.6537
Esophagus (12)	Asian	1.33156	35.3975
Stomach (13)	White	1.00507	31.1776
Stomach (13)	Black	1.07589	34.3908
Stomach (13)	Asian	1.14658	33.5137
Colon (14)	White	1.00142	30.9756
Colon (14)	Black	1.03610	31.5895
Colon (14)	Asian	1.08045	31.7173
Liver and Intraheptic Bile Duct (16)	White	1.03608	31.5034
Liver and Intraheptic Bile Duct (16)	Black	1.04369	33.2769
Liver and Intraheptic Bile Duct (16)	Asian	1.24043	34.5932
Pancreas (17)	White	1.00770	31.1781
Pancreas (17)	Black	1.07147	33.6909
Pancreas (17)	Asian	1.07419	33.1019
Bronchus; Lung (19)	White	1.00149	31.0110
Bronchus; Lung (19)	Black	1.00697	31.0915
Bronchus; Lung (19)	Asian	1.10552	32.3497
Skin Melanomas (22)	White	1.00023	30.9656
Skin Melanomas (22)	Black	1.09715	34.7281
Skin Melanomas (22)	Asian	1.09512	34.0309

Table 1.2: Sinuosity Index and Slope Angle

Cancer (ICD-9-CM Code)	**Race**	**%-Adjusted Sinuosity Index**	**%-Adjusted Slope Angle (Degrees)**
Breast (24)	White	1.00198	30.9822
Breast (24)	Black	1.04249	31.6611
Breast (24)	Asian	1.03703	31.9721
Uterus (25)	White	1.00066	30.9778
Uterus (25)	Black	1.03717	32.2554
Uterus (25)	Asian	1.06960	32.7279
Cervix (26)	White	1.00128	30.9941
Cervix (26)	Black	1.06337	31.9875
Cervix (26)	Asian	1.06368	32.9994
Ovary (27)	White	1.00344	31.0783
Ovary (27)	Black	1.01553	31.7026
Ovary (27)	Asian	1.15474	33.2923
Prostate (29)	White	1.01355	31.1755
Prostate (29)	Black	1.08430	33.8490
Prostate (29)	Asian	1.23926	34.0319
Bladder (32)	White	1.00247	31.0048
Bladder (32)	Black	1.01656	31.1584
Bladder (32)	Asian	1.24861	34.8000
Kidney and Renal Pelvis (33)	White	1.00571	31.0647
Kidney and Renal Pelvis (33)	Black	1.05650	31.9157
Kidney and Renal Pelvis (33)	Asian	1.27815	34.6854
Brain and Nervous System (35)	White	1.00565	31.1334
Brain and Nervous System (35)	Black	1.02892	31.5980
Brain and Nervous System (35)	Asian	1.23394	34.0738

Chapter 2

Race-Comorbidity-Cancer Network

2.1 Introduction and Background

2.1.1 Related Studies

Numerous recent studies have demonstrated the connection between race and higher risk for certain diseases. Foy *et al.* discovered that African American women have a 42% higher breast cancer death rate relative to white women. Even after adjusting for confounding factors including age, grade, and surgery, Foy *et al.* determined that overall survival probability of breast cancer was significantly associated with race (Foy K.C., 2018). Relatedly, Tal $etal$. found that pubertal timing, infertility, outcomes after assisted reproductive technology treatment, and reproductive aging were all impacted by racial differences (Tal R., 2013). Beyond studies involving women, Hirsch $etal$. found that blacks and Mexican Americans generally had more excess heart age in relation to white where heart age acted as an estimate for the age of a person's cardiovascular system (Hirsch, 2018).

2.1.2 Controversy over Relationship Between Race and Disease

Such studies connecting race and variable risk for certain diseases, however, have generated considerable controversy, as the correlation between these two factors has long been disputed among members of the scientific community and the general public. In a 2016 Science article entitled *Taking race out of human genetics*, Yudell *et al.* boldly asserted, "We believe the use

of biological concepts of race in human genetic research - so disputed and so mired in confusion - is problematic at best and harmful at worst. It is time for biologists to find a better way" (Yudell, 2016). This widely distributed article made headlines, as the authors beseeched the U.S. National Academies of Sciences, Engineering and Medicine to direct researchers away from the preconceived notion of race in genetics research. To further support these claims, other scientists and writers cited the works of past researchers such as W.E.B. Du Bois who argued that "the human species so shade and mingle with each other that... it is impossible to draw a color line between black and other races" (Yudell, 2014).

In contrast, Harvard Professor of Genetics David Reich points out in his New York Times op-ed *How Genetics Is Changing Our Understanding of 'Race'* that it is time for the scientific community to legitimize and incorporate these claims into genetic studies while working to provide the same freedoms and opportunities to individuals irrespective of their race. Reich cites recent genetic studies, which have shown distinct differences across populations in traits such as bodily dimensions and susceptibility to diseases. For instance, clinical studies have demonstrated that multiple sclerosis is more prevalent in European-Americans, while end-stage kidney disease is more common in African-Americans. Reich reiterates his own findings that prostate cancer occurs 1.7 times more often in African-Americans than European-Americans. More importantly, Reich and his colleagues located a portion of the genome that contains 2.8 percent more African ancestry than the average and also entails seven independent risk factors for prostate cancer. Reich concludes his article with the counterclaim that "it will be impossible - indeed, anti-scientific, foolish, and absurd - to deny the differences [between populations]" (Reich, 2018).

2.1.3 Personal Contributions

Much of the data extraction for this Chapter was derived from the work in Chapter 1, namely the use of anonymous clinical information from the HCUP database through the BCBI's secure Stronghold network with the programming language SQL. To determine statistically significant combinations of co-morbidities from the data within a reasonable runtime, I used the Association Rules Package written by Dr. Paul Stey from the Brown Center for Biomedical Informatics. I then used the Gephi software to illustrate the networks corresponding to each demographic

and cancer.

2.2 Results

2.2.1 Sample Network

A sample network containing three kinds of cancer is shown below. The nodes at the center of the clusters represent cancers (each color represents a different cancer). Connected to the cancer nodes by directed edges are nodes that represent combinations of comorbidities found frequently associated with their respective cancers.

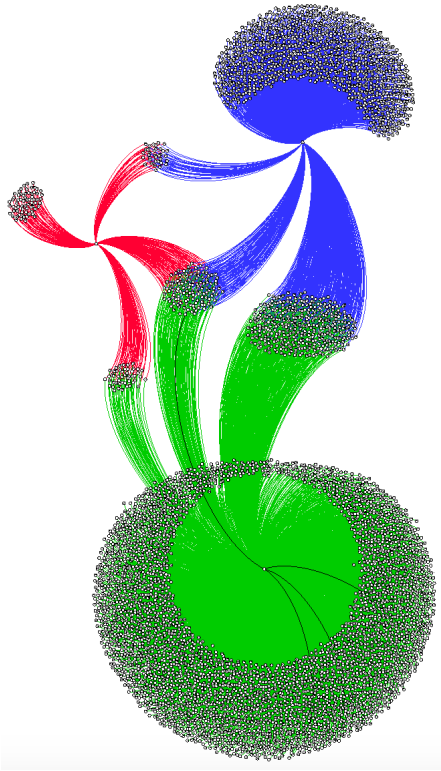


Figure 2.1: White Network for Brain, Bone, and Cervix Cancer

2.2.2 Tables of Notable Comorbidity Combinations

The following tables depict notable comorbidity combinations associated with different cancers for the white, black, Hispanic, and Asian demographics. From the tables, some key differences across the races should be further noted:

1. I found that white patients with nervous disorders and hypertension frequently had cancers of the head and neck, esophagus, stomach, rectum, liver, pancreas, gastrointestinal tract, bone, skin, uterus, cervix, ovary, female genital, testis, brain, thyroid, Hodgkin's lymphoma, myeloma, and primary. In contrast, black patients had thyroid cancer, while Hispanic patients had testis and brain cancer.

2. It seemed that white patients with mental health disorders and hypertension frequently had cancers of the head and neck, esophagus, and uterus. Black patients had cancers of the head and neck, stomach, rectum, liver, and thyroid, while Hispanic patients had rectal cancer.

3. Given secondary malignancies and hypertension, black patients had cancers of the rectum, liver, uterus, and thyroid, while Hispanic patients had cancers of the pancreas and ovary.

Table 2.1: Notable Comorbidity Combinations Associated With Cancers for White Demographic

Comorb Codes	Comorbidities	Cancer Codes	Cancer
{97, 99}	{cardiomyopathy, hypertension}	{11, 15, 40}	{head, rectum, myeloma}
{95, 98}	{nervous disorder, hypertension}	{11, 12, 13, 15, 16, 17, 18, 21, 22, 25, 26, 27, 28, 30, 35, 36, 37, 40, 41}	{head, esophagus, stomach, rectum, liver, pancreas, GI, bone, skin, uterus, cervix, ovary, female genital, testis, brain, thyroid, Hodgkin's myeloma, primary}
{58, 651, 663, 98}	{metabolic, anxiety disorders, mental health, hypertension}	{11, 12, 25}	{head, esophagus uterus}
{58, 651, 657}	{metabolic, anxiety, mood disorders}	{11, 12, 13, 18, 25, 26, 27, 36, 40}	{head, esophagus, stomach, GI, uterus cervix, ovary, thyroid, myeloma}
{158, 244, 53}	{kidney, ext injuries, lipid}	{11, 16, 40, 41}	{head, liver, myeloma, primary}
{106, 108, 131}	{cardiac dysrhythmias, heart failure, respiratory failure}	{11, 12, 13, 15, 16, 27, 40, 41}	{head, esophagus, stomach, rectum, liver, ovary, myeloma, primary}

Table 2.2: Notable Comorbidity Combinations Associated With Cancers for Black Demographic

Comorb Codes	Comorbidities	Cancer Codes	Cancer
{95, 98}	{nervous disorder, hypertension}	{36}	{thyroid}
{663, 98}	{mental health, hypertension}	{11, 13, 15, 16, 36}	{head, stomach, rectum, liver, thyroid}
{158, 99}	{kidney disease, hypertension}	{11, 13, 16, 40}	{head, stomach, liver, myeloma}
{106, 138}	{cardiac dysrhythmia, esophogeal}	{16, 36}	{liver, thyroid}
{257, 98}	{aftercare, hypertension}	{16, 25, 36}	{ liver, uterus, thyroid}
{42, 98}	{secondary malignancies, hypertension}	{15, 16, 25, 36}	{rectum, liver, uterus, thyroid}

Table 2.3: Notable Comorbidity Combinations Associated With Cancers for Hispanic Demographic

Comorb Codes	Comorbidities	Cancer Codes	Cancer
{95,98}	{nervous disorder, hypertension}	{30,35}	{testis, brain}
{663,98}	{mental health, hypertension}	{15}	{rectum}
{158,98}	{kidney disease, hypertension}	{15,17,35}	{rectum, pancreas, brain}
{257,98}	{aftercare, hypertension}	{18}	{GI}
{42,98}	{secondary malignancies, hypertension}	{17,27}	{pancreas, ovary}
{55,98}	{electrolyte disorder, hypertension}	{15,16,17,27}	{rectum, liver, pancreas, ovary}

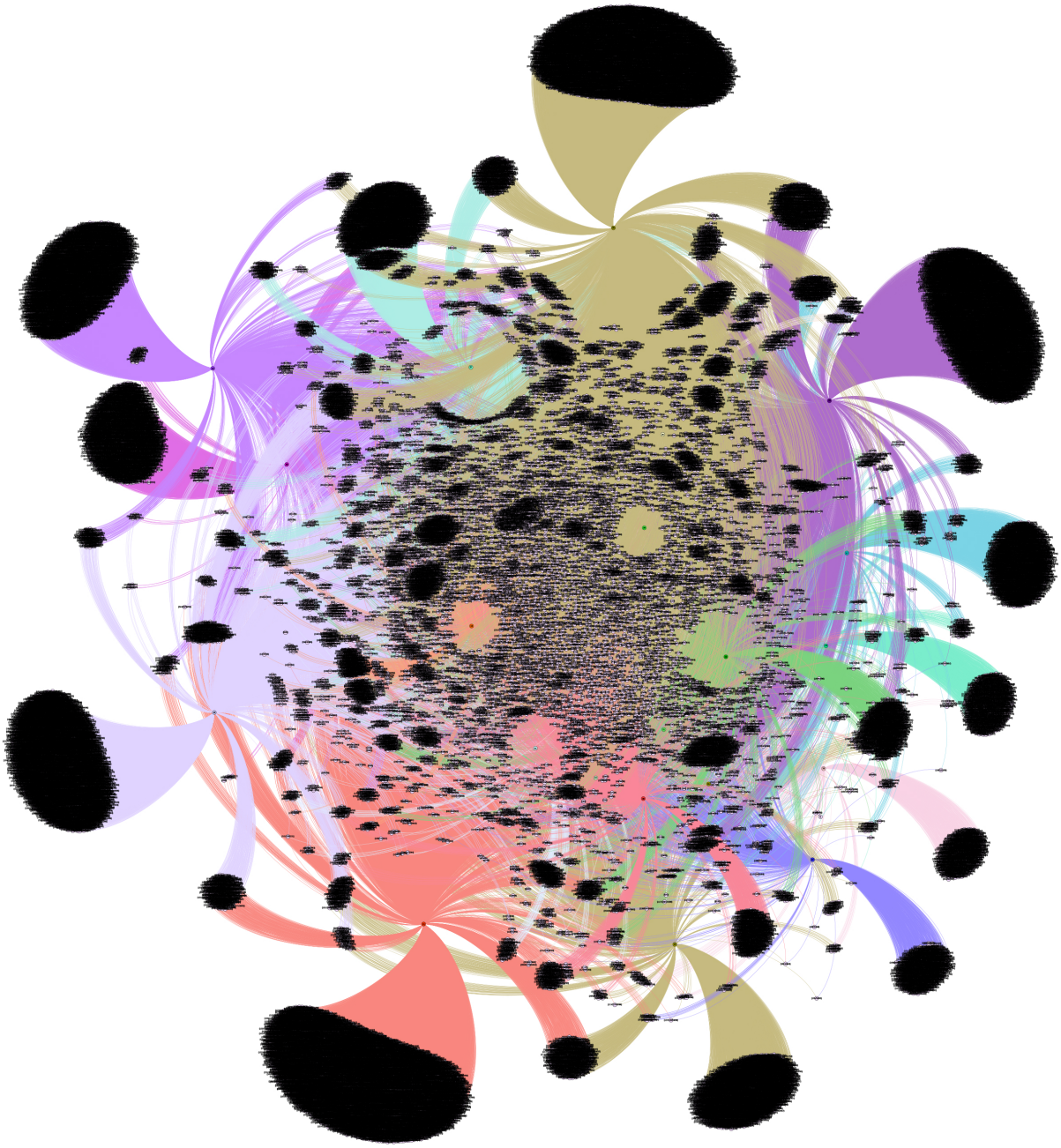
Table 2.4: Notable Comorbidity Combinations Associated With Cancers for Asian Demographic

Comorb Codes	Comorbidities	Cancer Codes	Cancer
{42}	{secondary malignancies}	{16,27}	{liver, ovary}
{151,257}	{liver disorders, aftercare}	{16}	{liver}
{42,55}	{secondary malig, electrolyte dis}	{16}	{liver}
{257,59}	{aftercare, anemia}	{16}	{liver}
{49,55}	{diabetes, electrolyte disorders}	{16}	{liver}
{151,49}	{GI disorders, diabetes}	{16}	{liver}

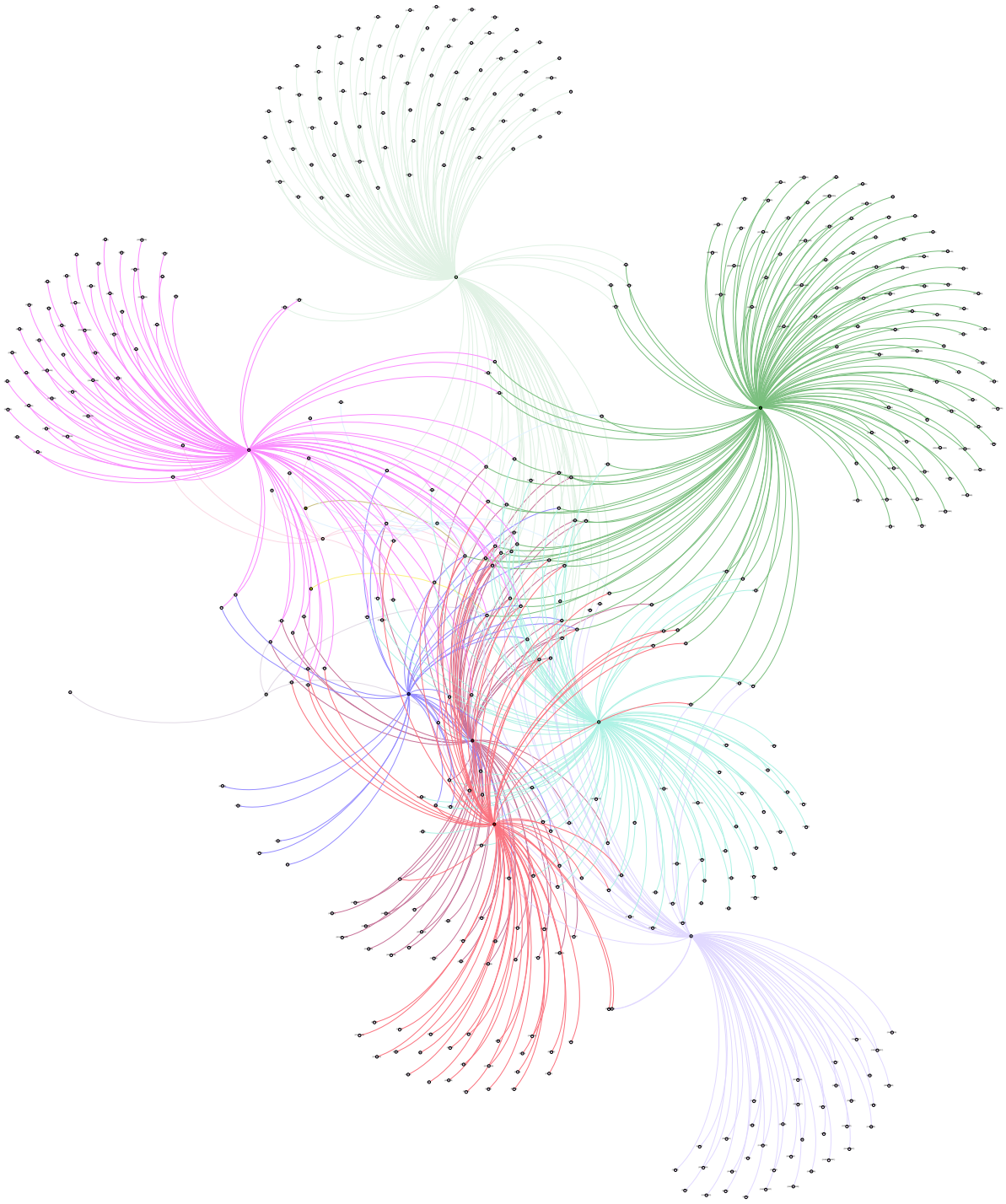
2.2.3 Full Networks

From the national HCUP data set as compared to the Rhode Island data set, I mapped out the following networks, which contain the most frequently found comorbidity combinations associated with different cancers across the white, black, Hispanic, and Asian demographics:

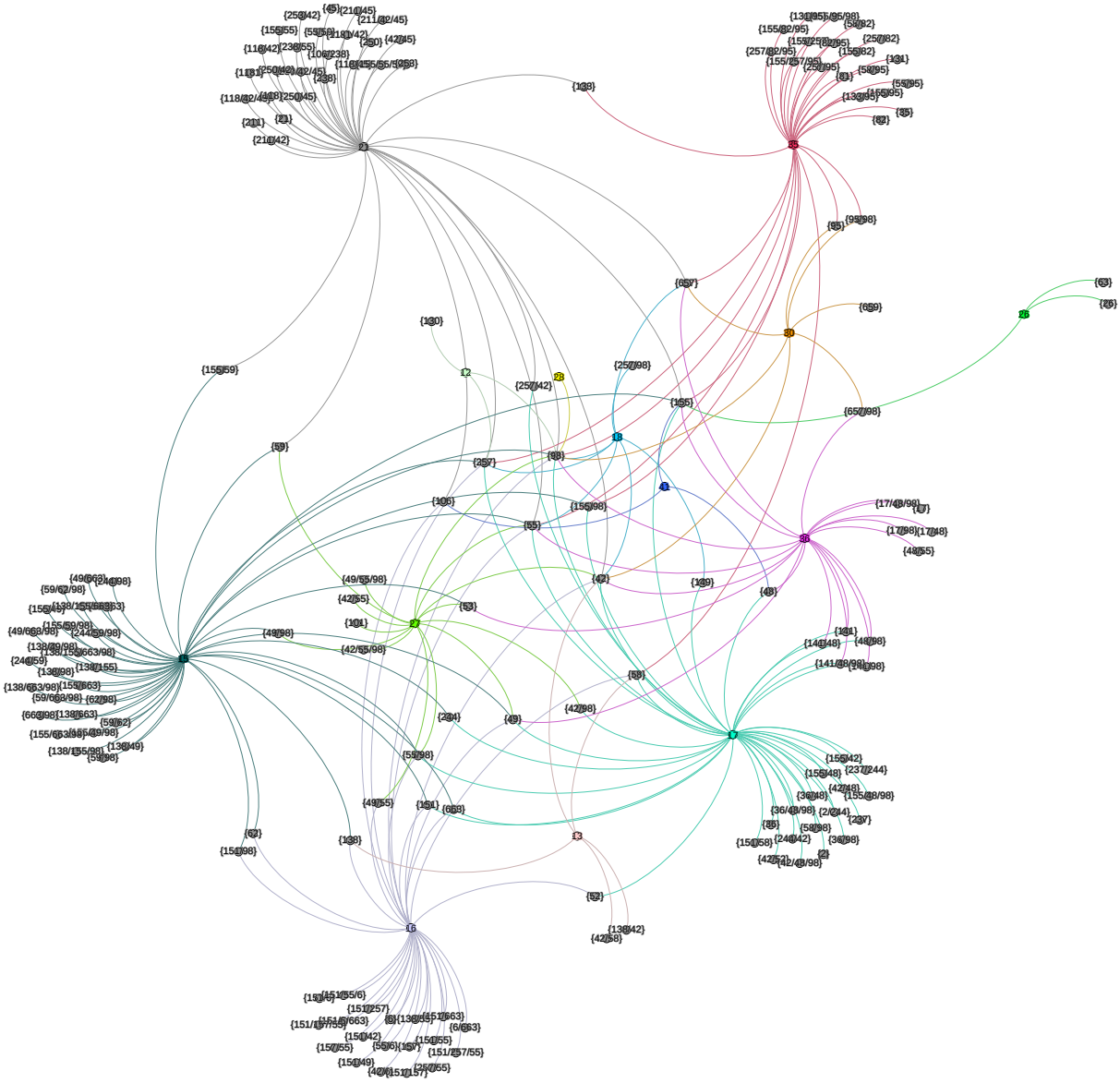
White



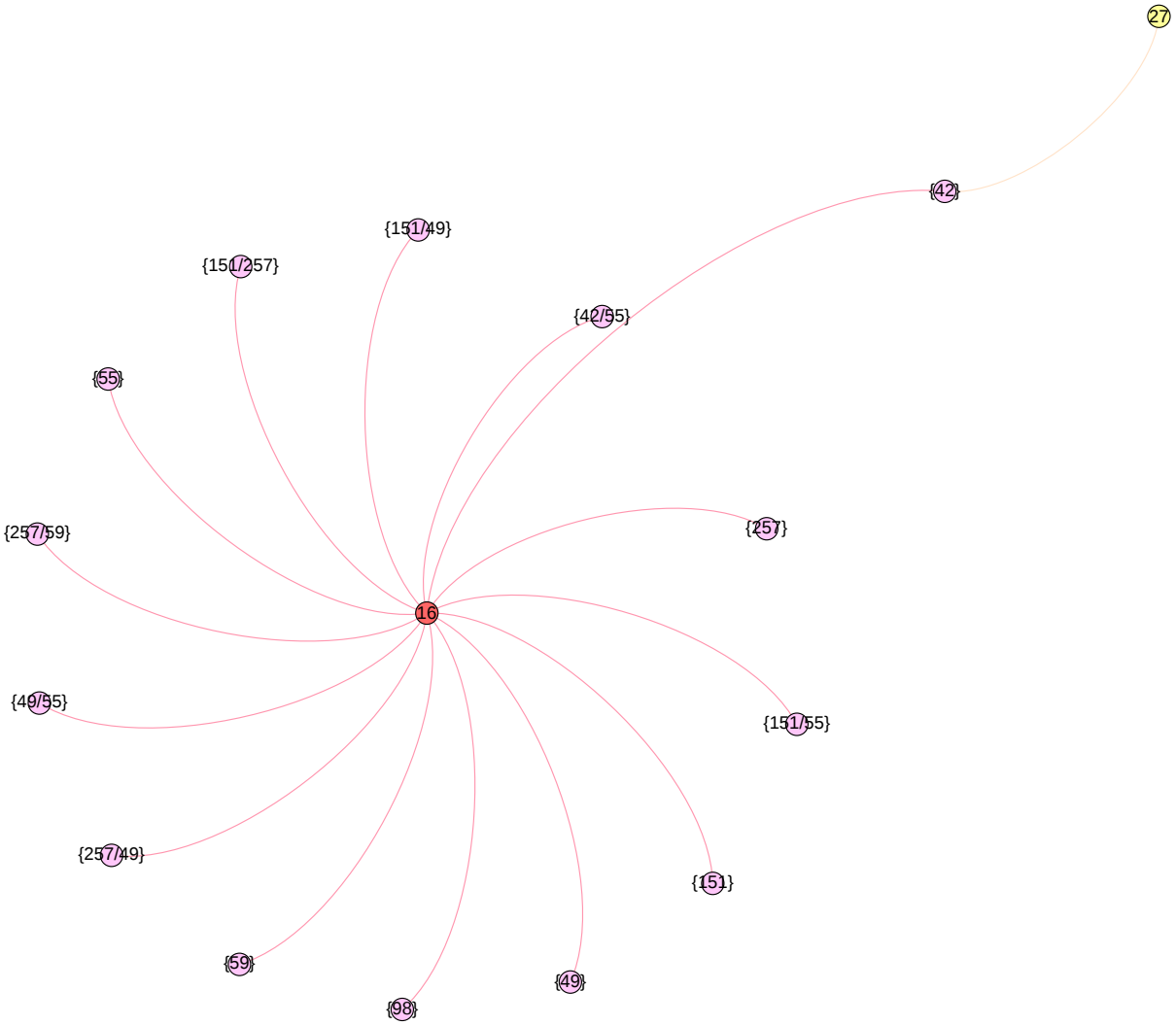
Black



Hispanic



Asian


27
{42}
{151/49}
{151/257}
{42/55}
{55}
{257/59}
{257}
16
{49/55}
{151/55}
{257/49}
{151}
{59}
{49}
{98}

2.3 Discussion

2.3.1 Current Studies and Literature

The term "precision medicine" has garnered much support, as the National Institutes of Health has heavily invested in recent studies related to precision medicine. In particular, the National Institute on Minority Health and Health Disparities (NIMHD) has placed particular emphasis on the following research areas:

1. Developing analytic methods to integrate patient data and other factors to influence health outcomes
2. Developing tools to identify biomarkers for disease progression and drug responses
3. Understanding disease mechanisms that lead to differential health outcomes in minorities

In conjunction with the NIH, institutions such as the Vanderbilt University Medical Center and Stanford University have been working towards establishing centers dedicated to developing precision medicine and further research in the aforementioned areas. In particular, Vanderbilt University Medical Center will investigate genetic and phenotype markers for asthma, pre-term birth, cancer, and Body Mass Index in African American and Hispanics/Latino populations (nih, 2016).

2.3.2 Recommendations

In modern medicine, treatment for cancer is usually based on the type, the size, and whether it has metastasized. As my results show, however, patients are at higher risk for different cancers based on their race and comorbidities. In particular, for the same combination of comorbidities (nervous disorder and hypertension), patients were at higher risk for different cancers based on their race. My results suggest that the future of medicine is in precision medicine through which patients will have treatments that are tailored to them based on their unique attributes such as race and pre-existing chronic conditions. Perhaps more importantly, my results imply that patients should take preventative measures for the cancers that they are at higher risk

for given their race and comorbidities.

My results reinforce the growing wave of support for precision medicine by the government and academic institutions, specifically the NIMHD's aim to understand disease mechanisms that lead to differential health outcomes in minorities. Evidently, controlled for the same combination of comorbidities, there is a genetic component associated with race that affects the kinds of cancer that patients are at risk for.

In response to Reich's article, one writer argued that Reich talked about regions, not people. To illustrate this further, individuals often self-select their racial identity in census and survey data, and the link between self-identification and genome data is not firmly established. Furthermore, many Black Americans are from socially disadvantaged backgrounds. It is not clear then whether stratifying my data using socioeconomic status would not exhibit the same differences found in my study, which would indicate that it is indeed race that is solely responsible for the differences identified in my cancer studies. Thus, I would like to add caution to the interpretation of my findings as pinpointing race as the sole determinant of comorbidities and cancer.

2.4 Methods

2.4.1 Association Rule-Based Machine Learning

Please note that the use of and explanation on association rule-based machine learning was primarily derived from Tan et al (Tan, 2018). Additionally, all figures and examples were obtained from Tan et al (Tan, 2018):

Potential Issues

In determining the most frequently occurring comorbidities and cancers associated with any given race, there were two major issues:

1. Determining the associations of co-morbidity combinations from an expansive data set

proved to be computationally expensive and entailed an impractical runtime. Note that each patient had up to 25 possible co-morbidities, meaning that combinations could have up to 25 different co-morbidities.

2. Some associations could simply have arisen from chance. These random associations then needed to somehow be distinguished and removed.

Thus, the algorithm incorporated association analysis, a methodology designed to uncover notable relationships hidden in large data sets. Referred to as association rules, these relationships defined sets of frequent items and in the context of the network, implied patients of a particular race and combination of co-morbidities were at a significantly higher risk for certain kinds of cancer.

Association Rule and Support

An association rule is defined as an expression in the form $X --> Y$ where X and Y are disjoint itemsets. Association rules can be measured based on the support and confidence. Support denotes how frequently a rule is found in the data set, while confidence relates how often items in Y appear in patients' diagnoses that contain X. Rules with low support values can be interpreted as occurring due to random chance. We define support and confidence as follows:

Support $s(X --> Y) = \frac{\sigma(X \cup Y)}{N}$
Confidence $c(X --> Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

From the HCUP data set containing patients' diagnoses, rules were found having support $\geq minsup$ and confidence $\geq minconf$ where $minsup$ and $minconf$ were defined as the threshold values for support and confidence.

Theory

In the first iteration of this algorithm, I used a brute-force approach to determine all possible k combinations of comorbidities where $0 \leq k \leq 25$ and compute the support and confidence for every possible rule. Given a data set of d items, the total number of possible rules is defined as

$R = 3^d - 2^{d+1} + 1$, showing an exponential runtime. This greedy algorithm took over 24 hours to run on a standard data set where $k = 3$

To significantly reduce the runtime and candidate itemsets, I incorporated the Apriori principle was incorporated into my algorithm which states that if an itemset is frequent, then all of its subsets must also be frequent. As a result, if an itemset is infrequent, then all of its supersets must also be infrequent too.

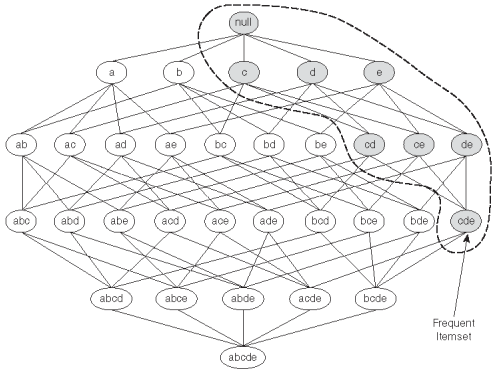


Figure 2.2: Demonstrates that if itemset $\{c, d, e\}$ is frequent, then all subsets of itemset are frequent

By using this strategy known as **support-based pruning**, the Apriori algorithm was the first association rule mining algorithm to use this strategy to significantly reduce the exponential growth of itemsets.

As seen in Figure 2.4, every item is initially considered a 1-itemset. Since $\{Cola\}$ and $\{Eggs\}$ only appear twice and once, respectively, they are removed from the 2-itemsets in the next iteration of the Apriori algorithm. This is because all supersets of infrequent 1-itemsets will also be infrequent. The above steps are repeated for the next iteration when going from the 2-itemsets to 3-itemsets.

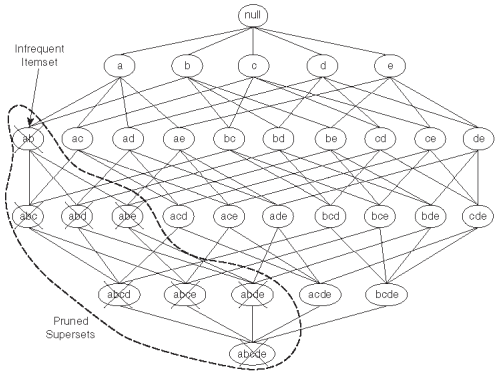


Figure 2.3: Demonstrates that if itemset $\{a, b\}$ is frequent, then all supersets of $\{a, b\}$ are infrequent

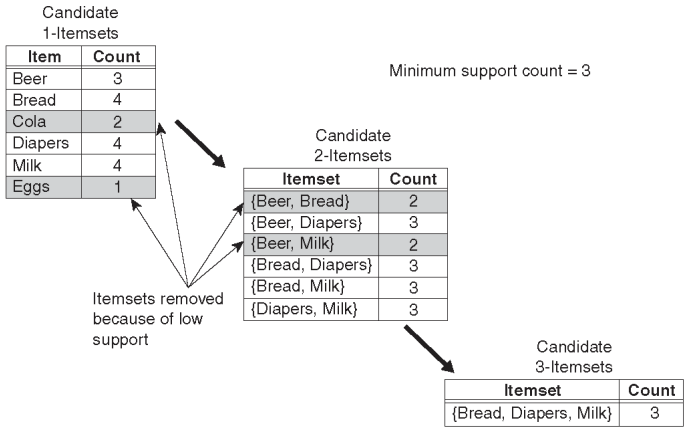


Figure 2.4: Illustrates the Apriori algorithm using an example containing market goods

We should note that without Apriori, we would need to do $\binom{6}{1} + \binom{6}{2} + \binom{6}{3} = 41$ calculations. With Apriori, we get $\binom{6}{1} + \binom{4}{2} + 1 = 13$ calculations.

General Outline of Algorithm

1. The entire dataset must be examined once to calculate the support of each item, after which the set of all frequent 1-itemsets will be known.

2. Iteratively, the algorithm calculates new $k-$itemsets using the frequent $(k-1)-$itemsets determined from the previous iteration (this is the Apriori principle)

3. Each time the algorithm needs to determine the support of the itemsets, the algorithm must go through the entire dataset once.

4. For any given itemset, if its support count is less than the user's inputted $minsup$, that itemset is removed (and by extension, its supersets in later iterations).

Application and Observation

For this research project, I used an Association Rules Package written by Dr. Paul Stey from the Brown Center for Biomedical Informatics (Stey). With this package, the algorithm took less than 20 minutes to run on that same standard data set where $k = 3$.

2.4.2 Data Sets

To compile the race-comorbidity-cancer network, I primarily used the HCUP database, as in Section 1: Racial Disparity of Genomic Sequencing. From HCUP, I extracted national patient records from patients of the races White, Black, Hispanic, and Asian and placed them into CSV files using SQL. Using the programming language Julia and Dr. Stey's Association Rules package, I computed the most frequently occurring k combinations of comorbidities associated with each race and cancer where $k \leq 6$. It should be noted that I only examined the following cancers to limit computational runtime: bone, brain, cervix, esophagus, female genital, gastrointestinal tract, head and neck, Hodgkin's lymphoma, liver, male genital, multiple myeloma, ovary, pancreas, primary, rectum and anus, respiratory, skin, stomach, testis, thyroid, urinary, and uterus. Then, I compared these combinations from national records to those calculated from the Rhode Island patient records. Frequent combinations that were found in both data sets were considered significant.

2.4.3 Data Visualization Using Gephi

For each race, I re-formed these significant combinations to be visualized as a network using Gephi, an open-source network analysis and visualization software package.

Chapter 3

Summary and Recommendations for Further Work

3.1 Summary and Conclusions

My studies show that race plays an important component in determining disease progression and even what kinds of diseases patients are at higher risk for. Thus, while genomic sequencing studies have led to novel discoveries of disease progression, they have inevitably led to better treatments for those races most represented in these studies. Evidently, there has been a shortage of minorities in genomic sequencing studies, thereby contributing to racial disparities in health outcomes. From these findings, I conclude that researchers must actively recruit Black Americans in genomic sequencing efforts.

3.2 Recommendations for Further Work

Further work should be conducted on the genetic basis for why race seems to play an important factor in the onset of different types of cancer. Just as Professor Reich discovered portions of the genome containing risk factors for prostate cancer and associated with race, next steps for these studies should investigate the human genome. For instance, given nervous disorders and hypertension, what portions of the genome are responsible for inducing thyroid cancer in Black Americans and testis and brain cancer in Hispanic Americans?

Acknowledgment

First and foremost, I would like to thank Dr. Neil Sarkar for his extraordinary guidance throughout these past couple years. It is to him that I owe my interest in bioinformatics and long-term desire to conduct research in the field as an academic. He has been incredibly generous and inspirational, and I am truly grateful for having had the opportunity to work with such an accomplished and kind professor like him.

Next, I'd like to thank my academic advisor, Dean Julianne Ip for her mentorship throughout my entire undergraduate career. She has always been there to answer any questions and assuage any concerns that I had regarding my academic interests. She was actually the one to first introduce me to Dr. Sarkar, and I am so thankful.

I'd also like to thank my concentration advisor and reader Professor Bjorn Sandstede as well as my long-time professor and co-reader Professor Sorin Istrail. They have been incredibly kind and constructive in helping me write this thesis.

Along with these professors, I am very grateful to Drs. Paul Stey and Elizabeth Chen of the Brown Center for Biomedical Informatics for providing technical support in various aspects of my project.

Lastly, I'd like to thank my family and friends for their continued support in my research endeavors, particularly my father for first instilling an interest in research in me. Special thanks to Jonathan Chang and Justin Lee who also struggled through the process of writing a thesis with me. Glad we made it together, guys.

I.E.K.

Bibliography

(2016). Nih funds precision medicine research with a focus on health disparities. *National Institutes of Health*.

Aronson, S.J. Rehm, H. (2015). Building the foundation for genomics in precision medicine. *Nature*, 526:336–342.

Consortium, T. . G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65.

Dye T., Li D., D. M. G. S. . F. D. (2016). Sociocultural variation in attitudes toward use of genetic information and participation in genetic research by race in the united states: implications for precision medicine. *Journal of American Medical Informatics Association*, 23(4):782–786.

Foy K.C., Fisher J.L., L. M. G. D. D. C. P. E. (2018). Disparities in breast cancer tumor characteristics, treatment, time to treatment, and survival probability among african american and white women. *NPJ Breast Cancer*, 4:7.

Ha, Y.S., S. A. K. M. S. E. K. J. H. M. P. A. K. W. L. D. . K. I. (2013). Increased incidence of pathologically nonorgan confined prostate cancer in african-american men eligible for active surveillance. *Journal of Urology*, 81(4):831–835.

Hirsch, J.R., W. G. L. Y. S. E. (2018). Racial differences in heart age and impact on mortality. *Journal of the National Medical Association*, 110:2:169–175.

I.J., P. (2007). Epidemiology and pathophysiology of prostate cancer in african-american men. *Journal of Urology*, 177(2):444–449.

Pezzullo, J. C. Analysis of variance from summary data. *http://statpages.info/anova1sm.html*.

Reich, D. (2018). How genetics is changing our understanding of 'race'. *The New York Times*.

Stey, P. Association rules julia package. *https://github.com/bcbi/association_rules_jl*.

Tal R., Seifer, D. (2013). Potential mechanisms for racial and ethnic differences in antimüllerian hormone and ovarian reserve. *International Journal of Endocrinology*.

Tan, P.N., S. M. K. A. K. V. (2018). *Introduction to Data Mining*. Pearson.

Yudell, M. (2014). A century after w.e.b. du bois, science still gets race wrong. *The Inquirer*.

Yudell, M., R. D. D. R. T. S. (2016). Taking race out of human genetics. *Science*, 351:6273:564–565.