Independent Project 3

Bathsheba Aklilu

2025-04-06

Introduction

Breast cancer, like many other health issues, is characterized by inequitable survival rates, particularly among marginalized groups (Giaquinto et al., 2022). Disparities in health outcomes, such as age at diagnosis, quality of life, and overall survival, have been shown to vary significantly by race and ethnicity. These disparities may be exacerbated by the stressors and social determinants of health (SDOH) that individuals from marginalized communities face, in addition to their breast cancer diagnosis. Social determinants of health, such as income and access to healthcare, can contribute to increased inflammation and stress, which are often reflected in biomarkers like C-reactive protein (CRP) and cortisol levels (Antoni et al., 2020). Black and Latina/Hispanic women are disproportionately affected by these health disparities in breast cancer (Yedjou et al., 2019), yet they remain underrepresented in much of oncological research (Duma et al., 2018). C-reactive protein (CRP) is an important biomarker for inflammation, frequently utilized in clinical and epidemiological research to assess risk factors for cardiovascular disease, chronic inflammation, and overall health status. CRP levels can be influenced by various factors, including social determinants of health (SDOH) such as socioeconomic status, education, and access to healthcare. In particular, low socioeconomic status (SES) and limited access to healthcare are often associated with chronic stress, which can contribute to elevated CRP levels. While much of the research has focused on the role of SDOH in these disparities, few studies have explored how stress factors are related to these outcomes, particularly in a diverse cohort such as the NHANES dataset.

This analysis aims to investigate the relationship between CRP levels and social determinants of health among participants with history of breast cancer using the **NHANES 2021-2023 dataset**. We will focus on CRP as the dependent variable, exploring how social determinants of health such as income, education, and marital status impact inflammation. The National Health and Nutrition Examination Survey (NHANES) 2021-2023, includes detailed information on various health indicators, including CRP levels, income, and cancer history. We will examine CRP levels in relation to social determinants of health, including family income and poverty level, as well as demographic factors such as age and race/ethnicity. The primary statistical approaches utilized in this analysis include descriptive statistics, ANOVA, and regression modeling.

Research Question and Hypothesis

- **Research Question**: How do social determinants of health (SDOH) such as income, education, and marital status influence CRP levels among women with hisptory of breast cancer in the NHANES 2021-2023 cohort?
- **Null Hypothesis**: There is no association between C-reactive protein levels and social determinants of health across breast cancer suriviors by racial and ethnic groups.
- **Alternative Hypothesis**: There is a significant association between C-reactive protein levels and social determinants of health across breast cancer suriviors by racial and ethnic groups.
- **Hypothesis (alternative)**: We hypothesize that individuals with lower income, lower education levels, and divorced marital status will have higher CRP levels, reflecting higher inflammation due to chronic stress and poor health.

Aim 1: Determine the association between social determinants of health (income, education, and marital status) and C-reactive protein by race/ethnicity among women with history of breast cancer.

Data Installation

```
# Install and load necessary packages
#install.packages("tidyverse")
library(tidyverse)
library(haven)

# loading data
crp_data <- read_xpt("HSCRP_L.xpt")
demo_data <- read_xpt("DEMO_L.xpt")
income_data <- read_xpt("INQ_L.xpt")
medical_data <- read_xpt("MCQ_L.xpt")
```

Data Cleaning and Preparation

```
# Clean and prepare CRP data (select necessary columns only)
crp_data_clean <- crp_data %>%
  select(SEQN, LBXHSCRP)

# Clean and prepare demographics data (select necessary columns only)
demo_data_clean <- demo_data %>%
  select(SEQN, RIAGENDR, RIDAGEYR, RIDRETH3, RIDRETH1, INDFMPIR, DMDEDUC2, DMDMARTZ)

# Clean and prepare income data (select necessary columns only)
income_data_clean <- income_data %>%
  select(SEQN, INDFMMPI)

# Clean and prepare medical conditions data (select necessary columns only, specifically, participants
breast_cancer_data <- medical_data %>%
  filter(MCQ230A == 14) %>%
  select(SEQN, MCQ230A)

# Merge the datasets by SEQN (participant id)
merged_data <- demo_data_clean %>%
  left_join(crp_data_clean, by = "SEQN") %>%
  left_join(income_data_clean, by = "SEQN") %>%
  inner_join(breast_cancer_data, by = "SEQN")

# Change coding of variables to characters
cleaned_data_labeled <- merged_data %>%
  mutate(
    RIDRETH3 = recode(RIDRETH3,
      `1` = "Mexican American",
      `2` = "Other Hispanic",
      `3` = "Non-Hispanic White",
      `4` = "Non-Hispanic Black",
      `6` = "Non-Hispanic Asian",
```

```
      `7` = "Other Race - Including Multi-Racial",
      .default = NA_character_
    ),
    RIDRETH1 = recode(RIDRETH1,
      `1` = "Mexican American",
      `2` = "Other Hispanic",
      `3` = "Non-Hispanic White",
      `4` = "Non-Hispanic Black",
      `5` = "Other Race - Including Multi-Racial",
      .default = NA_character_
    ),
    DMDMARTZ = recode(DMDMARTZ,
      `1` = "Married/Living with partner",
      `2` = "Widowed/Divorced/Separated",
      `3` = "Never married",
      `77` = "Refused",
      `99` = "Don't know",
      .default = NA_character_
    ),
    DMDEDUC2 = recode(DMDEDUC2,
      `1` = "Less than 9th grade",
      `2` = "9-11th grade (Includes 12th grade with no diploma)",
      `3` = "High school graduate/GED or equivalent",
      `4` = "Some college or AA degree",
      `5` = "College graduate or above",
      `7` = "Refused",
      `9` = "Don't know",
      .default = NA_character_
    )) %>%
  filter(
    DMDMARTZ != "Refused",
    DMDMARTZ != "Don't know",
    !is.na(DMDMARTZ),
    !is.na(RIDRETH1),
    !is.na(RIDRETH3),
    DMDEDUC2 != "Refused",
    DMDEDUC2 != "Don't know",
    !is.na(DMDEDUC2))


# Drop last nas if any

cleaned_compiled_all <- cleaned_data_labeled %>%
  drop_na()

cleaned_compiled_all
```

```
## # A tibble: 108 x 11
##      SEQN RIAGENDR RIDAGEYR RIDRETH3         RIDRETH1 INDFMPIR DMDEDUC2 DMDMARTZ
##     <dbl>    <dbl>    <dbl> <chr>            <chr>       <dbl> <chr>    <chr>
##  1 130392        2       74 Non-Hispanic Wh~ Non-His~     3.04 College~ Married~
##  2 130407        2       73 Non-Hispanic Wh~ Non-His~     4.37 College~ Widowed~
##  3 130523        2       61 Non-Hispanic Wh~ Non-His~     5    College~ Married~
```

```
##  4 130826        2       79 Non-Hispanic Wh~ Non-His~     3.3  Some co~ Married~
##  5 131137        2       67 Other Hispanic   Other H~     4.05 High sc~ Widowed~
##  6 131169        2       78 Non-Hispanic Wh~ Non-His~     5    Some co~ Widowed~
##  7 131342        2       61 Non-Hispanic Wh~ Non-His~     4.67 College~ Married~
##  8 131450        2       80 Non-Hispanic Wh~ Non-His~     3.68 College~ Widowed~
##  9 131509        2       45 Non-Hispanic Wh~ Non-His~     5    College~ Widowed~
## 10 131554        2       69 Non-Hispanic Wh~ Non-His~     5    College~ Widowed~
## # i 98 more rows
## # i 3 more variables: LBXHSCRP <dbl>, INDFMMPI <dbl>, MCQ230A <dbl>
```

##Data Visualization

```r
# Loading Okabe-Ito palette (colorblind-friendly color palette)
okabe_ito_colors <- palette.colors(palette = "Okabe-Ito")

ggplot(cleaned_compiled_all, aes(x = RIDRETH3, y = LBXHSCRP, fill = RIDRETH3)) +
  geom_boxplot() +
  geom_hline(yintercept = 3.0, linetype = "dashed", color = "red") +
  labs(
    x = "Race/Ethnicity",
    y = "High-Sensitivity CRP (mg/L)",
    title = "C-Reactive Protein Levels of Breast Cancer Survivors by Race/Ethnicity"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  ) +
  scale_fill_manual(values = okabe_ito_colors)
```

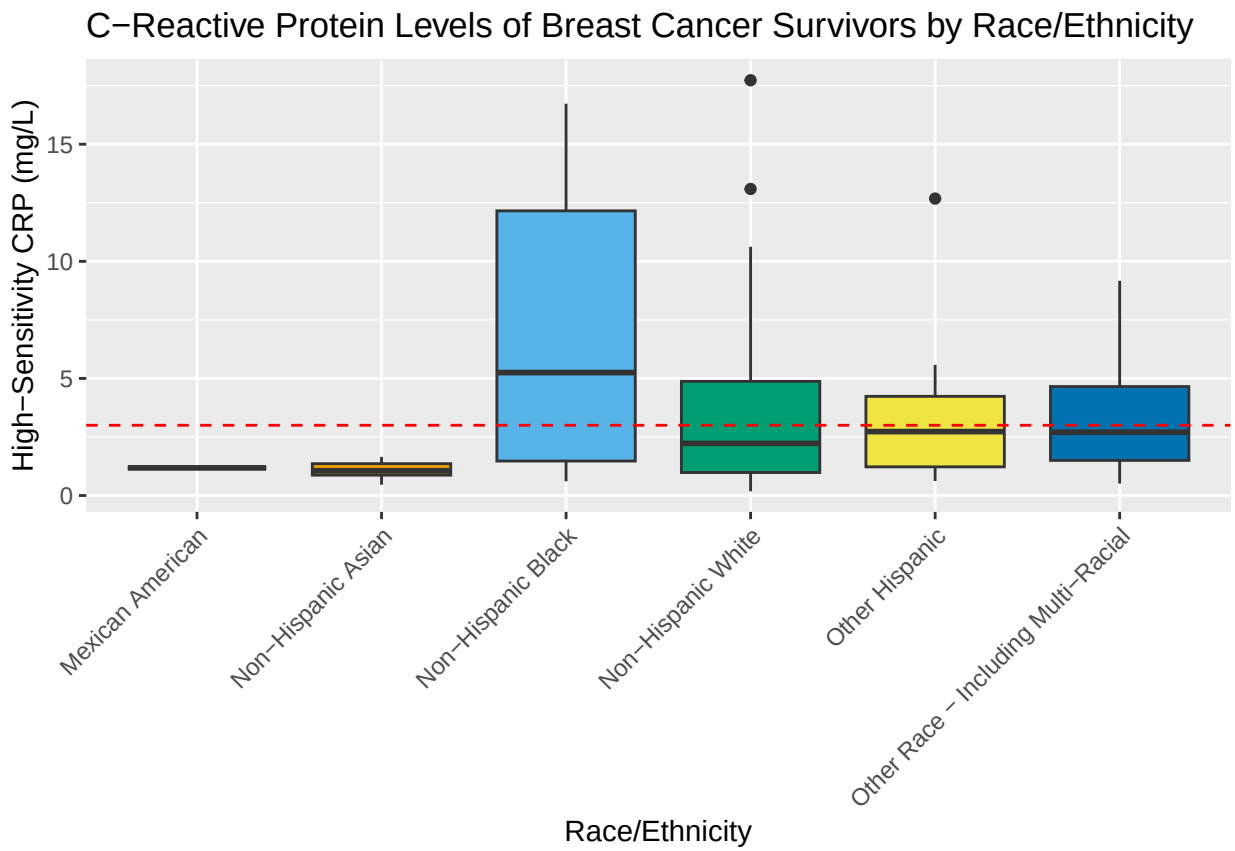

Figure 1: Boxplot depicting high-sensitivity c-reactive protein (CRP) levels (mg/L) in serum or plasma among participants with history of breast cancer stratified by race/ethnicity. C-reactive protein was collected by serum/plasma, not time-specific. Data collected through the CDC NHANES Research Program. Values above 3.0 mg/L suggest elevated cardiovascular or chronic inflammation risk and a dashed red line on the y-axis.

Statistical Analysis

- **Shapiro-Wilk Test**: Identify the skew/normality of dataset

```
# Run ANOVA
crp_aov <- aov(LBXHSCRP ~ RIDRETH3, data = cleaned_compiled_all)

# Get residuals
crp_res <- residuals(crp_aov)

# Shapiro-Wilk test
shapiro.test(crp_res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  crp_res
## W = 0.87274, p-value = 3.658e-08
```

```r
#Because there are less than 2 observations in a few groups, I will filter them out in order to run bar
filtered_data <- cleaned_compiled_all %>%
  group_by(RIDRETH3) %>%
  filter(n() >= 2) %>%
  ungroup()

# Bartlett's test for homogeneity of variances
bartlett.test(LBXHSCRP ~ RIDRETH3, data = filtered_data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  LBXHSCRP by RIDRETH3
## Bartlett's K-squared = 23.156, df = 4, p-value = 0.0001178
```

Shapiro test an batrlett both indicate not-normal distribution.

- **Kruskal-Wallis Test**: We cannot run ANOVA because many groups violate normality.

```r
# Kruskal-Wallis for Education
kruskal.test(LBXHSCRP ~ DMDEDUC2, data = cleaned_compiled_all)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  LBXHSCRP by DMDEDUC2
## Kruskal-Wallis chi-squared = 6.2027, df = 4, p-value = 0.1845
```

```r
# Kruskal-Wallis for Marital status
kruskal.test(LBXHSCRP ~ DMDMARTZ, data = cleaned_compiled_all)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  LBXHSCRP by DMDMARTZ
## Kruskal-Wallis chi-squared = 0.062187, df = 2, p-value = 0.9694
```

```r
# Kruskal-Wallis for Race/Ethnicity
kruskal.test(LBXHSCRP ~ RIDRETH3, data = cleaned_compiled_all)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  LBXHSCRP by RIDRETH3
## Kruskal-Wallis chi-squared = 5.6454, df = 5, p-value = 0.3423
```

- **Spearman's**: Due to lack of normal distribution, we are no longer using pearsons, and instead using spearmans.

```r
# Correlation between CRP and family poverty index
cor.test(cleaned_compiled_all$LBXHSCRP, cleaned_compiled_all$INDFMMPI, method = "spearman")
```

```
## Warning in cor.test.default(cleaned_compiled_all$LBXHSCRP,
## cleaned_compiled_all$INDFMMPI, : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  cleaned_compiled_all$LBXHSCRP and cleaned_compiled_all$INDFMMPI
## S = 260180, p-value = 0.0126
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2393428
```

```r
# Correlation between CRP and income-to-poverty ratio
cor.test(cleaned_compiled_all$LBXHSCRP, cleaned_compiled_all$INDFMPIR, method = "spearman")
```

```
## Warning in cor.test.default(cleaned_compiled_all$LBXHSCRP,
## cleaned_compiled_all$INDFMPIR, : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  cleaned_compiled_all$LBXHSCRP and cleaned_compiled_all$INDFMPIR
## S = 265448, p-value = 0.005683
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2644349
```

References:

Anandapadmanathan, & Kresten. (2019, July 24). How to add dashed horizontal line with label in ggplot [Online forum post]. Stack Overflow. https://stackoverflow.com/questions/57177608/how-to-add-dashed-horizontal-line-with-label-in-ggplot

Antoni, M. H., Lechner, S. C., Kilbourn, K. M., & Phillips, K. A. (2020). Behavioral, physical, and psychological predictors of cortisol and C-reactive protein in breast cancer survivors: A longitudinal study. Psycho-Oncology, 29(8), 1237-1245. https://doi.org/10.1002/pon.5397

Bewick, V., Cheek, L., & Ball, J. (2003). Statistics review 7: Correlation and regression. Critical care (London,England), 7(6), 451–459. https://doi.org/10.1186/cc2401

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Cohen, S., Doyle, W. J., & Baum, A. (2006). Socioeconomic status is associated with stress hormones. Psychosomatic medicine, 68(3), 414–420. Doi: 10.1097/01.psy.0000221236.37158.b9. PMID: 16738073.

Coughlin S. S. (2019). Social determinants of breast cancer risk, stage, and survival. Breast cancer research and treatment, 177(3), 537–548. https://doi.org/10.1007/s10549-019-05340-7

DeSantis, C. E., Ma, J., Sauer, A. G., Newman, L. A., & Jemal, A. (2017). Breast cancer statistics, 2017, racial disparity in mortality by state. CA: A Cancer Journal for Clinicians, 67(6), 439-448. https://doi.org/10.3322/caac.21412

Ding, Z., Mangino, M., Aviv, A., Spector, T., Durbin, R., & UK10K Consortium (2014). Estimating telomere length from whole genome sequence data. Nucleic acids research, 42(9), e75. https://doi.org/10.1093/nar/gku181

Duma, N., Vera Aguilera, J., Paludo, J., Haddox, C. L., Gonzalez Velez, M., Wang, Y., Leventakos, K., Hubbard, J. M., Mansfield, A. S., Go, R. S., & Adjei, A. A. (2018). Representation of Minorities and Women in Oncology Clinical Trials: Review of the Past 14 Years. Journal of oncology practice, 14(1), e1–e10. https://doi.org/10.1200/JOP.2017.025288

Current hematologic malignancy reports, 18(6), 284–291. https://doi.org/10.1007/s11899-023-00717-4

Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, et al.. Breast cancer statistics, 2022. CA: Cancer J Clin (2022) 0:1–18. doi: 10.3322/caac.21754

Guo, L., Liu, S., Zhang, S., Chen, Q., Zhang, M., Quan, P., Lu, J., & Sun, X. (2015). C-reactive protein and risk of breast cancer: A systematic review and meta-analysis. Scientific reports, 5, 10508. https://doi.org/10.1038/srep10508

He, X.-Y., Gao, Y., Ng, D., Michalopoulou, E., George, S., Adrover, J. M., . . . & Egeblad, M. (2023). Chronic stress increases metastasis via neutrophil-mediated changes to the microenvironment. Nature, 616 (7956), 563–572. https://doi.org/10.1038/s41586-023-06020-3

Hopper, J.L., Dite, G.S., MacInnis, R.J. et al. Age-specific breast cancer risk by body mass index and familial risk: prospective family study cohort (ProF-SC). Breast Cancer Res 20, 132 (2018). https://doi.org/10.1186/s13058-018-1056-1

Islami, F., Ward, E. M., Sung, H., Cronin, K. A., Tangka, F. K. L., Sherman, R. L., Zhao, J., Anderson, R. N., Henley, S. J., Yabroff, K. R., Jemal, A., & Benard, V. B. (2021). Annual Report to the Nation on the Status of Cancer, Part 1: National Cancer Statistics. JNCI: Journal of the National Cancer Institute, 113(12), 1648–1669. https://doi.org/10.1093/jnci/djab131

PMC7048405. Mikkelsen, M. K., Lindblom, N. A. F., Dyhl-Polk, A., Juhl, C. B., Johansen, J. S., & Nielsen, D. (2022). Systematic review and meta-analysis of C-reactive protein as a biomarker in breast cancer. Critical 3 Reviews in Clinical Laboratory Sciences, 59(7), 480–500. https://doi.org/10.1080/10408363.2022.2050886

Nazmi, A., & Victora, C. G. (2007). Socioeconomic and racial/ethnic differentials of C-reactive protein levels: A systematic review of population-based studies - BMC Public Health. BioMed Central. https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-7-212

Phelan, J. C., Link, B. G., & Tehranifar, P. (2010). Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. Journal of health and social behavior, 51 Suppl, S28–S40. doi: 10.1177/0022146510383498.PMID: 20943581.

Solorio, S., Murillo-Ortíz, B., Hernández-González, M., Guillén-Contreras, J., Arenas-Aranda, D., Solorzano-Zepeda, F. J., Ruiz-Avila, R., Mora-Villalpando, C., de la Roca-Chiapas, J. M., & MalacaraHernández, J. M. (2011). Association between telomere length and C-reactive protein and the development of coronary collateral circulation in patients with coronary artery disease. Angiology, 62(6), 467–472. doi: 10.1177/0003319710398007. PMID: 21441231.

Wang, F., Giskeødegård, G. F., Skarra, S., Engstrøm, M. J., Hagen, L., Geisler, J., Mikkola, T. S., Tikkanen, M. J., Debik, J., Reidunsdatter, R. J., & Bathen, T. F. (2023). Association of serum cortisol and cortisone levels and risk of recurrence after endocrine treatment in breast cancer. Clinical and experimental medicine, 23(7), 3883–3893. https://doi.org/10.1007/s10238-023-01109-x

Williams, D. R., Priest, N., & Anderson, N. B. (2016). Understanding associations among race, socioeconomic status, and health: Patterns and prospects. Health psychology : official journal of the Division of Health Psychology, American Psychological Association, 35(4), 407–411. doi: 10.1037/hea0000242. PMID: 27018733; PMCID: PMC4817358.

Wong, J. Y., De Vivo, I., Lin, X., Fang, S. C., & Christiani, D. C. (2014). The relationship between inflammatory biomarkers and telomere length in an occupational prospective cohort study. PloS one, 9(1), e87348. doi: 10.1371/journal.pone.0087348. PMID: 24475279; PMCID: PMC3903646.

Yedjou, C. G., Sims, J. N., Miele, L., Noubissi, F., Lowe, L., Fonseca, D. D., Alo, R. A., Payton, M., & Tchounwou, P. B. (2019). Health and Racial Disparity in Breast Cancer. Advances in Experimental Medicine and Biology, 1152, 31. https://doi.org/10.1007/978-3-030-20301-6_3