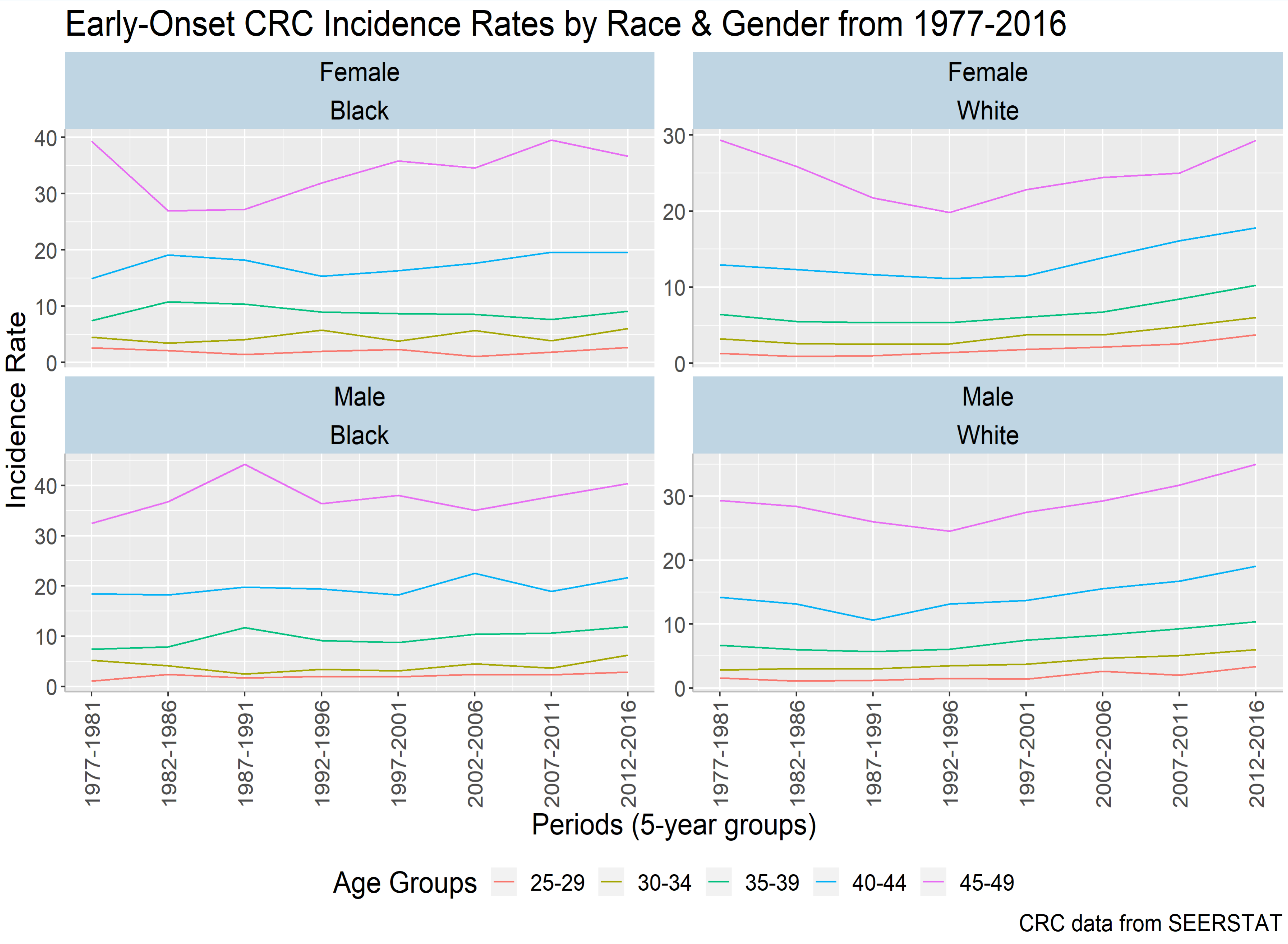
Using ‘Big Data’ To Explore and Identify Potential Risk Factors for Early-Onset Colorectal Cancer

Lulu Zhang, Jianjiu Chen, Piero D. Dalerba, Mary Beth Terry, Wan Yang
Columbia University, Department of Epidemiology
COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Background

- Incidence rates for colorectal cancer (CRC) have been increasing dramatically in younger adults in the past decade
- Established risk factors (RF) come from studies in older adults (50+), reasons for increase in younger cases remains unknown
- Traditional epidemiologic studies face challenges in identifying RF for early-onset cancer due to low absolute risk
- Accurately identifying modifiable RF is important for informing effective prevention in younger adults



Early-Onset CRC Incidence Rates by Race & Gender from 1977-2016

CRC data from SEERSTAT

Methodology

Data Sources

- Large scale survey data representative of the US population:
  - National Health and Nutrition Examination Surveys (NHANES) interviews about 5,000 individuals each year[1]
  - National Health Interview Surveys (NHIS) interviews almost 100,000 individuals each year[2]
  - Cancer incidence data from NCI’s Surveillance, Epidemiology, and End Results (SEER) Program (~34% of US population) [3]

Risk Factoring Coding

- Obtained RF data for White and Black males and females aged 25-49, from 1977 to 2016
- Grouped data into eight 5-year periods and five 5-year age groups
- Combined data from NHIS, NHANES II, NHANES III, and continuous NHANES 1999 to 2016 and revised the weights to increase sample size
- Categorized mean exposure values using quintiles from overall population for each gender[4]

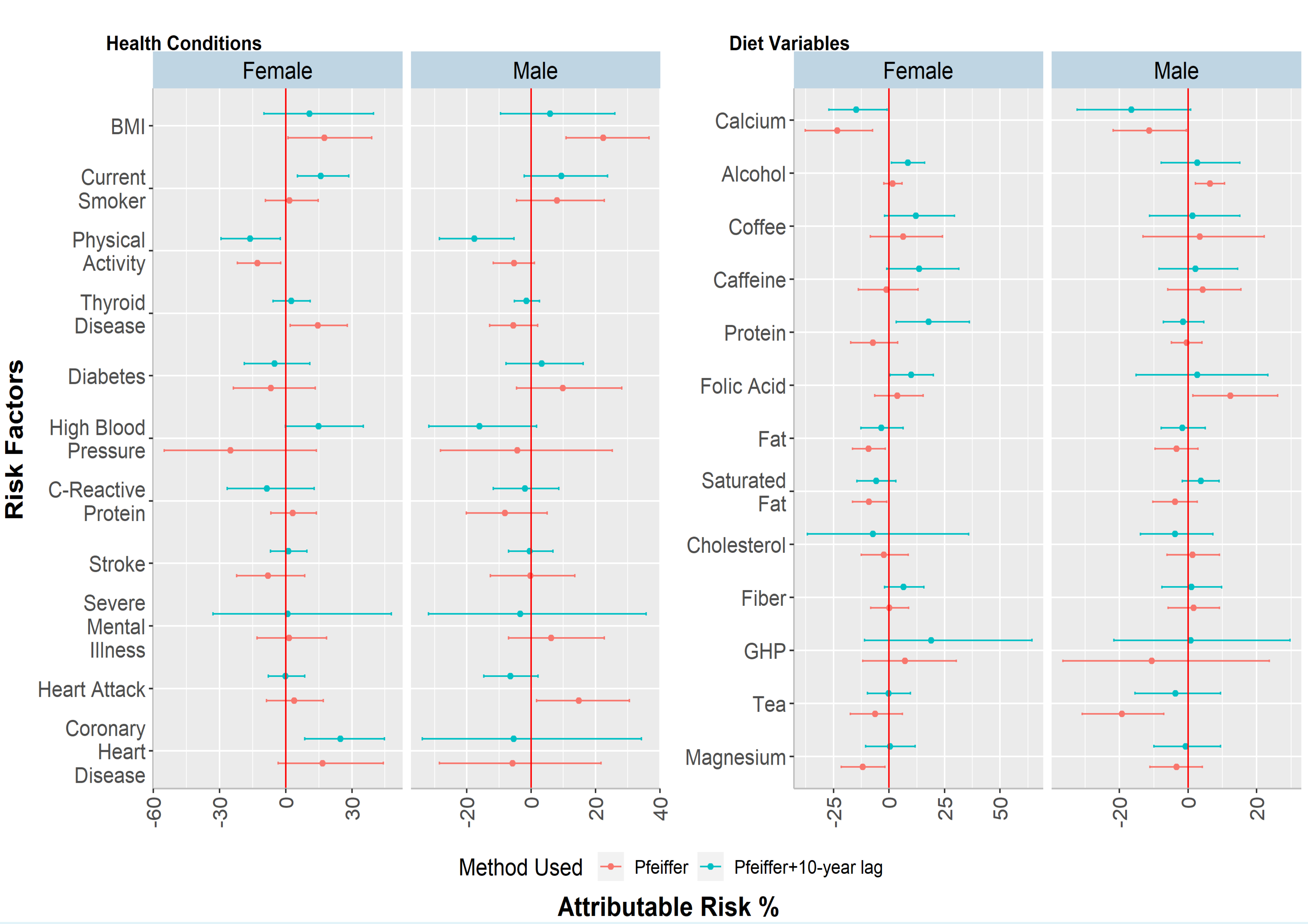
Statistical Analyses

- Ran quasi-Poisson regression for current intake, 10-year lagged intake
- RF lagged by 10 years to match current CRC incidence

Results

- Examined several diet variables and health conditions, both established and unestablished as CRC risk factors (e.g., BMI, cholesterol)
- Final methods selected were current intake and 10-year lagged intake (both quintiles)
- Other methods explored: age-period-race-mean centered current, 10-year lag, 10-year cumulative intake, and non-centered 10-year lag, non-centered current intake (RF modeled as continuous)
- Mixed results: some consistent with literature, others differed
  - E.g., calcium consistent, fat inconsistent
- Attributable risk, 95% CIs, standard errors were obtained for all RF for final methods
- Problems with limited data points (~80 observations per gender) despite use of ‘big data’
- Harmonizing risk factors from different surveys and across years due to changes and differences in design proved a challenge

Attributable Risk and 95% Confidence Intervals for EOCRC Risk Factors, ages 25-49



Future Directions

First stage was hypothesis generating; next steps will be to investigate the underlying mechanism of risk factors that stood out and how it ties in to early-onset CRC to identify risk factors via mechanistic models.

Acknowledgements

This study is supported by the Data Science Institute and Irving Institute for Cancer Dynamics Seed Funds Program at Columbia University.

References

1. NHANES - About the National Health and Nutrition Examination Survey. Published January 8, 2020. Accessed March 18, 2021. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm
2. NHIS - About the National Health Interview Survey. Published September 16, 2020. Accessed March 18, 2021. https://www.cdc.gov/nchs/nhis/about_nhis.htm
3. National Cancer Institute. Overview of the SEER program 2019 [cited 2019 2/13]. Available from: https://seer.cancer.gov/about/overview.html.
4. Pfeiffer RM, Webb-Vargas Y, Wheeler W, Gail MH. Proportion of U.S. Trends in Breast Cancer Incidence Attributable to Long-term Changes in Risk Factor Distributions. *Cancer Epidemiol Prev Biomark*. 2018;27(10):1214-1222. doi:10.1158/1055-9965.EPI-18-0098