Abstract

Context: Large epidemiologic cohort studies, such as the National Health and Nutrition Examination Survey (NHANES), collect copious high-dimensional data that allow for examination of multiple exposures in relation to a given outcome.

Objective: To explore the exposures measured in the Third National Health and Nutrition Examination Survey (NHANES III) dataset in search of factors associated with cancer mortality data obtained from the National Death Index (NDI) and to assess methods for lethal cancer risk prediction model variable selection.

Design, Setting and Participants: NHANES III collected data on 33,994 participants aged 2 months and older from 1988 to 1994 in the United States. The data, which include Interview, Medical Examination, and Laboratory components, were collected and linked with Mortality data from NDI death certificate records by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC). From the initial pool of participants, we selected 16404 adult participants that were cancer-free at baseline and that had no missing values for follow-up time since interview, NDI mortality, primary sampling units (PSU), stratification, and sampling weight variables.

Exposures: The initial publicly available dataset contained 3544 exposures from the Interview, Medical Examination, Laboratory, and Mortality components. After removing variables that were non-numeric, missing any values, only had one unique value, or had correlation to another variable greater than 0.82, we obtained the final set of 290 exposures. The analysis described herein did not involve multiple imputation nor utilize the NHANES III Multiply Imputated Data Set.

Main Outcome Measure: Among the 16404 patients, there were 964 cancer deaths and 280891 total years of follow-up since the initial Interview data were collected. The cancer deaths and follow-up time were used as the outcome (survival) in Cox proportional hazards regression analysis.

Results: We fit 960 Cox proportional hazards models with and without ridge regularization to randomly selected subsets of up to 48 variables and divided the models into 4 groups based on their Akaike Information Criterion (AIC) and concordance values. Applying domain knowledge to the variable descriptions, we selected a subset (exact number) of the most frequent highly significant variables and trained a final model that performed well compared to the randomly generated models.

Conclusions: The work described here constitutes an exploratory analysis of the NHANES III dataset that employs an iterative strategy to generation of cancer risk prediction models. In this approach, a large number of models are generated randomly to inform variable selection and guide training of models

in future iterations. In addition to providing insight into cancer risk factors measured in the NHANES III dataset we hope to develop a general methodology that can be applied to large, high-dimensional cohort study data.

Introduction

Cancer susceptibility is influenced by modifiable and non-modifiable factors. Modifiable cancer risk factors include Body Mass Index (BMI) and cigarette use, whereas the non-modifiable factors include Single Nucleotide Polymorphisms (SNPs) and family history of disease. According to a 2018 study by Islami and colleagues [1], modifiable risk factors are responsible for 42% of all cancer cases and 45% of all cancer deaths. This finding suggests that cancer prevention strategies that target modifiable risk factors have the potential to almost halve cancer incidence and mortality in the United States. A near two-fold reduction in cancer cases and deaths may seem far-fetched, but cancer incidence and mortality in United Status have been declining by ~1.5% every year from 2009-2014 and 2001-2015, respectively [2]. Taken together, these data indicate that while tremendous progress has been made, there is still great potential for cancer prevention approaches to decrease cancer incidence and mortality. Cancer risk prediction methods are paramount to maximizing the benefit of cancer prevention policies and programs. To achieve the best performance, cancer risk prediction models must include both modifiable and non-modifiable risk factors. In 2016, Maas and colleagues [3] demonstrated that cancer risk prediction models based on known epidemiologic risk factors can be improved when genetic information such as SNPs are included in the models. Importantly, the combined model provided better risk stratification than the models containing only epidemiologic risk factors or only genetic variables. The 2016 Maas study [3] focused on breast cancer, but the methodology can be applied to other cancers. Lung cancer is of particular interest, because it is responsible for the highest number of deaths in the United States [2] and worldwide [4]. Lung cancer risk is tightly linked with cigarette smoking, which was the strongest modifiable risk factor in the 2018 study by Islami and colleagues [1].[Is this statement for all cancer or lung cancer?] In fact, Islami and colleagues determined that 19% of cancers cases and roughly 29% of deaths can be attributed to cigarette smoking [1].In never smokers, causes include... https://www.cancer.org/cancer/non-small-cell-lung-cancer/causes-risks-prevention/what-causes.html The scale of cancer burden in the United States is staggering. Siegel and colleagues estimate that in 2018 there will be 1.7 million newly diagnosed cancer cases and roughly 600 thousand cancer deaths [2]. Cancer risk prediction models can help policymakers and cancer prevention practitioners develop more effectice interventions and to channel limited resources towards people at the greatest risk. The challenge of cancer risk prediction is complex and will require cancer-type specific strategies that integrate multiple types of data and explore various modeling methods. As part of the effort to tackle this challenge, we propose to analyze data from a genome-wide association study (GWAS) that was performed as part of the Atherosclerosis Risk

in Communities (ARIC) study [5] to fit prediction models that first incorporate known and putative epidemiologic risk factors and then correlate the residual risk with genetic data, such as SNPs and DNA methylation patterns. [add a focus on lung cancer in this paragraph... The scale of cancer burden in the United States is staggering. Siegel and colleagues estimate that in 2018 there will be 1.7 million newly diagnosed cancer cases and roughly 600 thousand cancer deaths [2]. Cancer risk prediction models can help policymakers and cancer prevention practitioners develop more effective interventions and to channel limited resources towards people at the greatest risk. The challenge of cancer risk prediction is complex and will require cancer-type specific strategies that integrate multiple types of data and explore various modeling methods.] As part of the effort to tackle this challenge, we propose to analyze data from a genome-wide association study (GWAS) that was performed as part of the Atherosclerosis Risk in Communities (ARIC) study [5 the GWAS is not in this ref published in 1989] to fit lung cancer risk prediction models that first incorporate known and putative epidemiologic risk factors, including cigarette smoking, and then associate SNPs with the residual (i.e., not explained by the known/putative risk factors) risk of lung cancer. This approach may provide insight into the contribution of genetic factors to lung cancer risk and could lead to the discovery of novel SNPs and pathways that play a contributing or protective role in lung carcinogenesis and may explain these observations: About 80-90% of lung cancer cases are due to cigarette smoking [http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html], yet only 10% of smokers develop lung cancer [need to find refs]. This approach may provide insight into the contribution of genetic factors to cancer risk and could lead to the discovery of novel SNPs that play a role in carcinogenesis. The ARIC study provides a rich, multidimensional dataset and a unique opportunity for cancer prevention research [6]. The ARIC study provides a rich, multidimensional dataset and a unique opportunity for cancer etiology and prevention research, including genetic risk prediction [6].

Methods

Results

Example variable table: Table needs to be updated.

Description of Highly Significant Variables that Appeared Most Frequently

Name

Median HR

n

Description

HAQ7

0.280

22

50 years or older (binary)

HAK9

1.195

21

times per night you get up to urinate

HAT16

1.753

19

In the past month, did you lift weights

HSAITMOR

1.000

19

Age in months at interview (screener)

HAQ1

1.070

17

Describe natural teeth: excellent. . . poor

HAV7R

1.000

17

Number of weeks lived at this address

HAP2

0.795

16

Do you use glasses, contacts, or both

HAS1

1.665

15

Past 2 wks, did you work at job/business

DMAETHNR

1.157

14

Ethnicity

HAN9

1.801

14

20 years or younger (binary)

HAT29

2.228

14

30 years or younger (binary)

HAJ0

2.050

13

17-74 years old versus 75 and older (binary)

HAR1

0.635

11

Have you smoked 100+ cigarettes in life?

HAN5JS

0.998

10

Flour tortillas - times/month

HAT2

1.733

8

In the past month, did you jog or run

HAT10

1.558

6

Past month, did you do other dancing

HSFSIZER

0.923

5

Family size (persons in family)

HAC1A

0.604

4

Doctor ever told you had: arthritis

Discussion

1. Islami F, Sauer AG, Miller KD, Siegel RL, Fedewa SA, Jacobs EJ, et al. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the united states. CA: A Cancer Journal for Clinicians. 2018;68:31–54. doi:10.3322/caac.21440.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA: A Cancer Journal for Clinicians. 2018;68:7–30. doi:10.3322/caac.21442.

3. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the united states. JAMA Oncology. 2016;2:1295. doi:10.1001/jamaoncol.2016.1025.

4. Center M, Siegel R, Jemal A. Global cancer facts & figures 3rd edition. Atlanta: American Cancer Society. 2015;1–61.

5. THE ATHEROSCLEROSIS RISK IN COMMUNITIES (ARIC) STUDY: DESIGN AND OBJECTIVES. American Journal of Epidemiology. 1989;129:687–702. doi:10.1093/oxfordjournals.aje.a115184.

6. Joshu CE, Barber JR, Coresh J, Couper DJ, Mosley TH, Vitolins MZ, et al. Enhancing the infrastructure of the atherosclerosis risk in communities (ARIC) study for cancer epidemiology research: ARIC cancer. Cancer Epidemiology Biomarkers & Prevention. 2017;27:295–305. doi:10.1158/1055-9965.epi-17-0696.