Exploratory Analysis of Factors Associated with Cancer Mortality in the National Health and Nutrition Examination Survey Dataset

Martin Skarzynski
Prof. Elizabeth Platz

2018-04-14

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 200 Cox proportional hazards models with and without ridge regularization to randomly selected subsets of 30 variables and divided the models into 4 groups of roughly equal size according to their Akaike Information Criterion (AIC) and concordance values. Highly significant variables ($p < 1e\text{-}10$) were most common in the group with the lowest AIC and highest concordance and least common in the group with highest AIC and lowest concordance. Interestingly, some of these highly significant variables showed up only in models with above-median corcordance values or only in ridge-penalized models with below-median AIC values. The The highly significant variables that showed up most frequently in the models we trained were mostly related to advanced age and and smoking status.

As expected, the 100 Cox models with ridge penalties all had lower AIC values than the 100 non-penalized models and ridge-penalized variables tended to have lower hazard ratios.

Conclusions:

Introduction

Methods

Results

H_2O H^2O

Table 1: Description of Highly Significant Variables

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

[1]

Discussion

Cancer susceptibility is influenced by modifiable and non-modifiable factors. Major modifiable cancer risk factors include body mass index (BMI) and cigarette use, whereas 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, the focus of this proposal, is of particular interest, because it is responsible for the highest number of

Figure 1: Concordance (y-axis) and AIC (x-axis) values of 200 Cox models with (triangle) and without (circle) ridge penalties split into quadrants based on

Figure 2: Scatter plot showing the concordance (y-axis) and AIC (x-axis) values of 200 Cox models

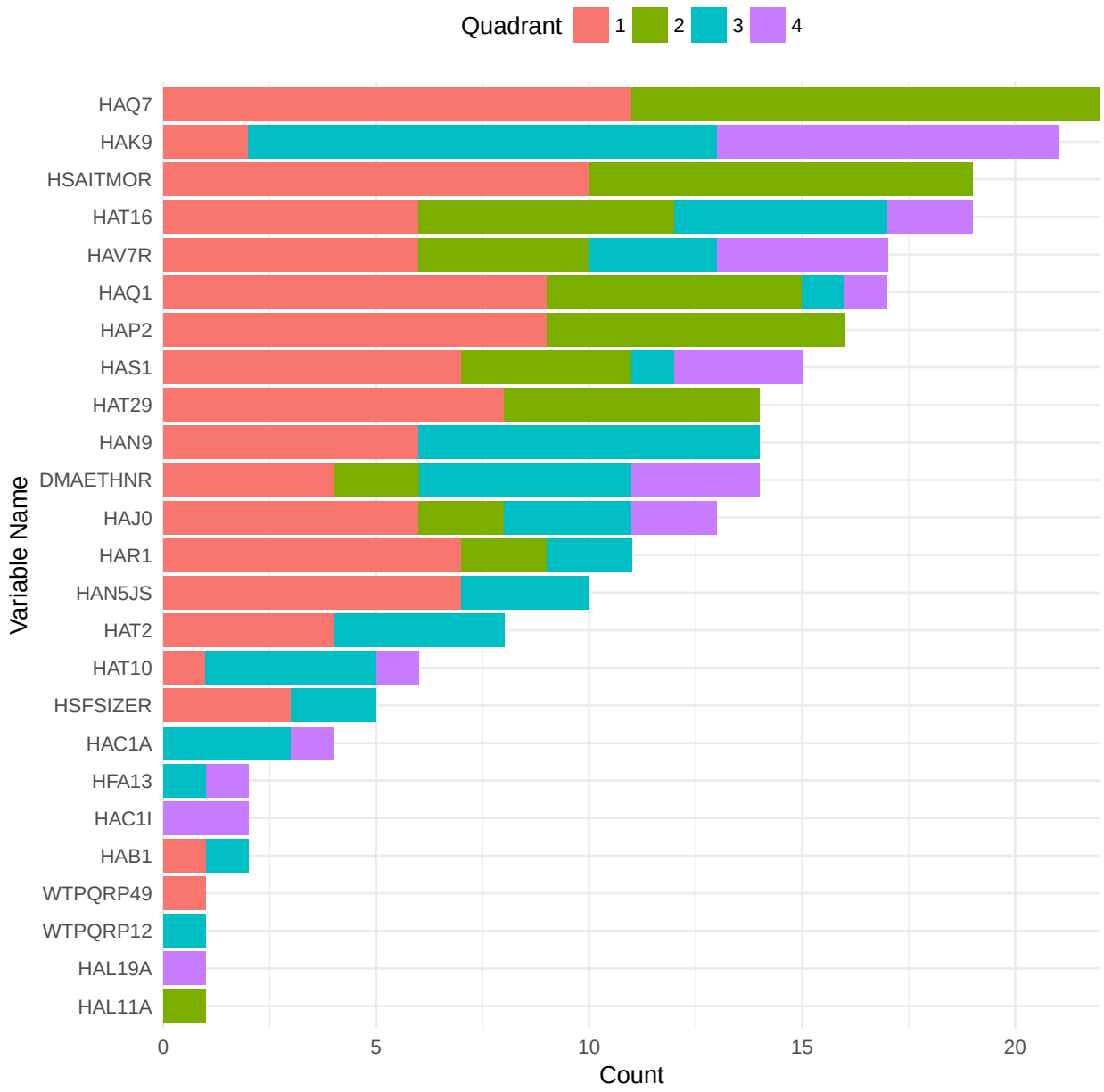

Figure 3: Scatter plot showing the concordance (y-axis) and AIC (x-axis) values of 200 Cox models

deaths in the United States [2] and worldwide [4]. Cigarette smoking is a known cause of lung cancer risk, and 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 [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]. The ARIC study provides a rich, multidimensional dataset and a unique opportunity for cancer etiology and prevention research, including genetic risk prediction [6].

With respect to non-smoking risk factors

1) For the limitation section: no info on indoor radon ... most lung cancer death caused by radon are in ever smokers (because of an interaction): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3673501/

2) Need to check to see if we have info on second hand smoke.

Miranda Jones may know more about the type of info available on air pollution based on residence.

We have occupation. I doubt that many participants will have jobs with known occupational exposures such as to asbestos or arsenic.

Non occupational arsenic exposure... inhaled. Miranda may know if ARIC has attempted to geocode to point to participant residents in higher arsenic areas.

Don't think we have info on radiation to the lung.

Need to consider how to include family history of lung cancer (available at V3)... perhaps a stratified analysis?

We may need to restrict the analysis or perform subanalyses to make lung cancer histology-specific inferences. Recommend restricting to non-small cell and restrict to ever smokers (lung cancer is never smokers will be uncommon). Lung cancer histology: https://www.cancer.org/cancer/lung-cancer.html

5. Main Hypothesis/Study Questions: Are SNPs associated with the residual risk of lung cancer after accounting for known and suspected risk factors in ARIC?

We expect that SNPs associated with residual risk may point to yet unidentified pathways that contribute to lung cancer risk and protection.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Study design We will perform a prospective cohort study of men and women beginning follow-up at Visit 1 through 2012. We will first identify SNPs that are present on the Affymetrix Genomewide SNP Array 6.0, which was used in ARIC. Using these SNPs and those imputed to 1000 Genome (1000G) reference panel [7],

we will fit statistical models that include known and putative risk factors such as cigarette smoking status and pack years of smoking.

Inclusion/exclusion All ARIC participants with adequate genotyping data available will be included in this study. We will exclude all participants with prevalent cancer at baseline and who did not provide informed consent to DNA usage for future studies and/or to studies of other chronic diseases like cancer.

Exposure: SNPs (already genotyped using the Affymetrix Genomewide SNP Array 6.0 and 1000G reference panel [7] imputed.

Covariates - Lung cancer risk factors: age, smoking status (current, former – maybe exclude never smokers?), packyears smoked, [what else will you adjust for?]

Outcome: Lung cancer incidence using the 2012 case file. [irrespective of histology?]

Statistical analysis [describe the actual analysis you will do... All statistical analyses will be performed using the R programming language [8]. The primary analysis method will be Cox proportional hazards regression [no need to say this... the model is routinely used..., which represents a well-established and effective statistical technique for modeling time-to-event data. The coefficients for predictor variables in a Cox model are easily interpretable as the expected log of the hazard ratio relative to a one unit change in the associated predictor variable, holding all other predictors constant.] Additionally, there are other survival analyses that can be done in R, including survival tree analysis using the rpart package [9] and survival random forest models using the randomForestSRC package [10]. The tree-based models like random forests are easy to interpret and allow for the quantification of the proportion of variance explained by variables included in the model. Another statistical method called XGBoost [11] can compute an F-score representing the importance of each variable. These methods can allow for the assessment of the effect of a predictor variable, for example the presence or absence of SNP, on the outcome variable, which in this case is survival.]

Run primary analysis stratified by race (W or B).

Restrict to non-small cell lung cancer?

Restrict to ever smokers?

Methodologic challenges: [add this section... include limitations] Add a discussion on power

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.