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

Martin Skarzynski Prof. Elizabeth Platz

`r Sys.Date()`

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = FALSE,
include = TRUE, warning = FALSE, message = FALSE)
```

```
{r libraries, include=FALSE} library(knitr) library(readr)
```

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

Methods

Results

Example variable table: Table needs to be updated.

Table 1: 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

Name	Median HR	n	Description
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

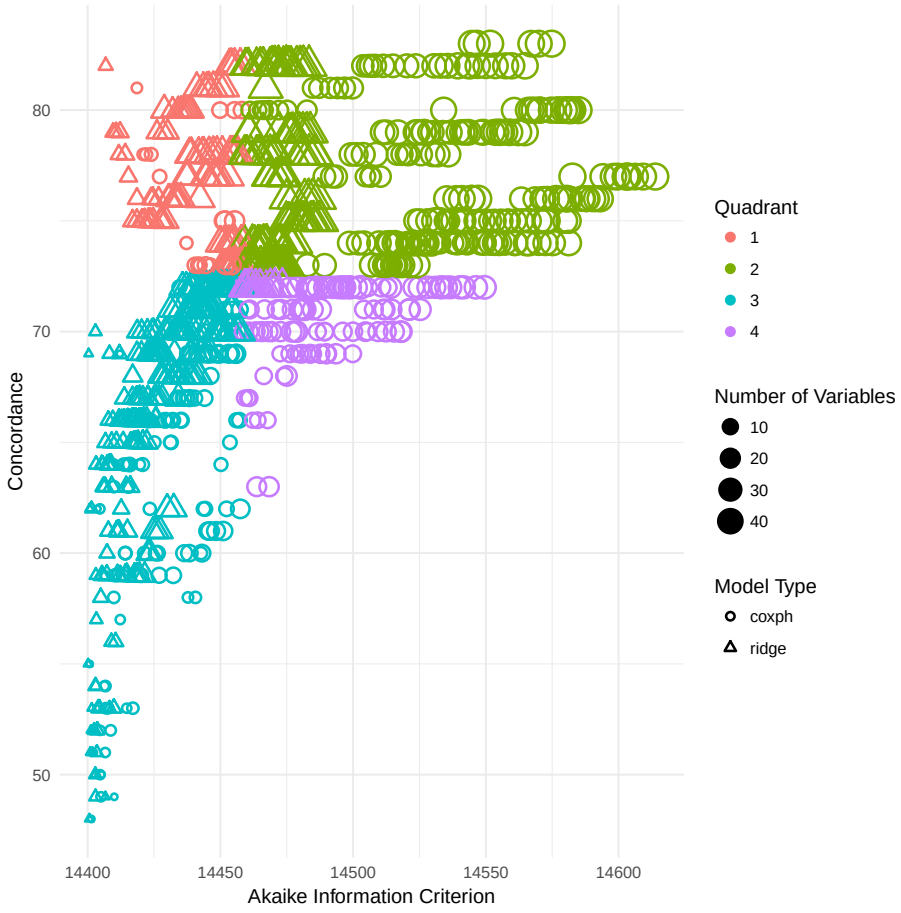

Figure 1: Cox Proportional Hazards models (n=960) with (triangle) and without (circle) ridge penalties on 1 to 48 variables, split into quadrants based on Concordance (y-axis) and AIC (x-axis) values

Figure 2: Significance (y-axis) and Hazard Ratios (x-axis) of all variables, labeled with the quadrants of the 960 Cox models as in Figure 1

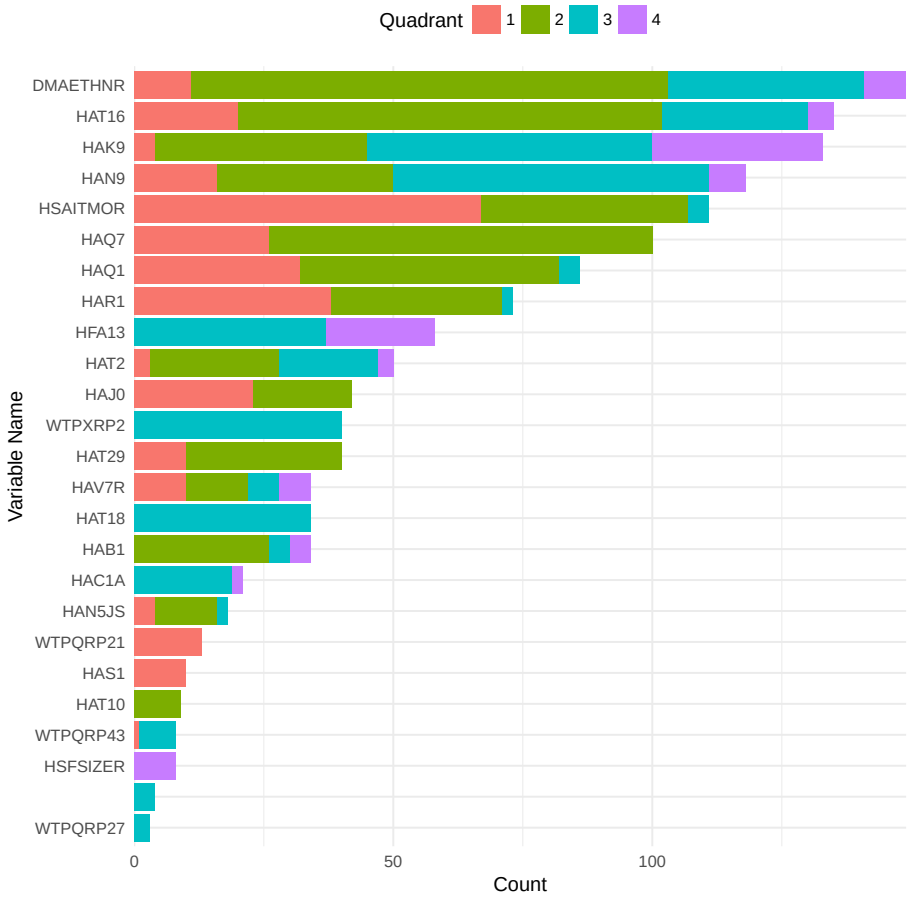

Figure 3: Variables that appeared most frequently, labeled with the quadrants of the 960 Cox models as in Figures 1 and 2