Environmental Research 155 (2017) 92–107

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Socioeconomic and air pollution correlates of adult asthma, heart attack, and stroke risks in the United States, 2010–2013

Louis Anthony (Tony) Cox, Jr.

Cox Associates and University of Colorado, 503 N. Franklin Street, Denver, CO 80218, USA

ARTICLE INFO

Keywords:
Asthma
Air pollution
Heart disease
Stroke
Environmental exposure
Causality
Bayesian networks

ABSTRACT

Asthma in the United States has become an important public health issue, with many physicians, regulators, and scientists elsewhere expressing concern that criterion air pollutants have contributed to a rising tide of asthma cases and symptoms. This paper studies recent associations (from 2008 to 2012) between self-reported asthma experiences and potential predictors, including age, sex, income, education, smoking, and county-level average annual ambient concentrations of ozone (O3) and fine particulate matter (PM2.5) levels recorded by the U.S. Environmental Protection Agency, for adults 50 years old or older for whom survey data are available from the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS). We also examine associations between these variables and self-reported heart attack and stroke experience; all three health outcomes are positively associated with each other. Young divorced women with low incomes are at greatest risk of asthma, especially if they are ever-smokers. Income is an important confounder of other relations. For example, in logistic regression modeling, PM2.5 is positively associated ($p < 0.06$) with both stroke risk and heart attack risk when these are regressed only against PM2.5, sex, age, and ever-smoking status, but not when they are regressed against these variables and income. In this data set, PM2.5 is significantly negatively associated with asthma risk in regression models, with a 10 μg/m^3 decrease in PM2.5 corresponding to about a 6% increase in the probability of asthma, possibly because of confounding by smoking, which is negatively associated with PM2.5 and positively associated with asthma risk. A variety of non-parametric methods are used to quantify these associations and to explore potential causal interpretations.

1. Introduction

Asthma in the United States has become an important public health issue. Many physicians, regulators, and scientists have expressed concern that exposures to criterion air pollutants have contributed to a rising tide of asthma cases and symptoms. This paper has the following two major aims: (1) Examine recent associations (from 2008 to 2012) between self-reported asthma experiences and potential predictors, including age, sex, income, education, smoking, and ambient concentrations of ozone (O3) and fine particulate matter (PM2.5) levels, for adults 50 years old or older for whom survey data are available from the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance (BRFSS) System. (2) Apply Bayesian network learning algorithms and other non-parametric machine-learning algorithms to this data set to clarify possible causal interpretations of the observed associations among these variables. We also examine associations between these variables and self-reported heart attack and stroke experience to show whether well-established relations between smoking and heart attack or stroke risks are seen in this data set (Shah and Cole, 2010; Oliveira et al., 2007).

2. Data

To investigate the association between air pollutants (O3 and PM2.5) and self-reported adult asthma, stroke, and heart attack risks, we merged the following data sources: (a) The most recent 5 years of available survey response data from a survey of over 228,000 individuals from 15 states, retrieved from the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance (BRFSS) System (www.cdc.gov/brfss/questionnaires/state2013.htm); and (b) Environmental Protection Agency (EPA) data on O3 and PM2.5 concentrations for the counties in which these individuals lived at the time of the survey, retrieved from the US EPA web site (www.epa.gov/airtrends/pm.html). Counties were used as the common key for merging annual average air pollution levels with individual response data. Table 1 summarizes the number of individual responses from each state for each of several questions. These responses are coded so that a response of “Yes” has a value of 1 and a value of “No” has a value

E-mail address: tcoxdenver@aol.com.

http://dx.doi.org/10.1016/j.envres.2017.01.003
Received 12 September 2016; Received in revised form 4 December 2016; Accepted 3 January 2017

Table 1
Means and frequency counts (*N*) for individual responses to different questions on the BRFSS survey, for 2008–2012. Responses are broken down by states (rows).

State	Sex=Male	Sex=Male N	Asthma ever	Asthma ever N	Flu shot	FluShot N	Health plan	Health plan N	Heart attack ever	Heart attack ever N	Hispanic	Hispanic N
AZ	0.38	8618	0.14	8592	0.55	8379	0.94	8600	0.09	8562	0.10	8559
CA	0.40	25528	0.13	25505	0.52	23146	0.93	25515	0.07	25499	0.16	25463
FL	0.37	9915	0.12	9895	0.52	9557	0.91	9887	0.10	9844	0.10	9814
GA	0.34	1925	0.12	1919	0.53	1850	0.92	1922	0.07	1914	0.02	1915
IL	0.36	4638	0.12	4631	0.50	4532	0.93	4634	0.07	4619	0.05	4620
MA	0.37	49621	0.14	49451	0.57	46565	0.97	49461	0.08	49329	0.06	49319
MI	0.35	10334	0.13	10310	0.50	10098	0.94	10312	0.09	10260	0.02	10276
NJ	0.38	27550	0.12	27466	0.51	26113	0.93	27478	0.08	27420	0.08	27423
NY	0.37	6939	0.12	6912	0.58	6706	0.94	6912	0.07	6888	0.07	6866
NC	0.37	8935	0.12	8916	0.59	8745	0.93	8922	0.08	8894	0.02	8911
OH	0.36	17820	0.12	17761	0.54	17326	0.93	17781	0.09	17690	0.01	17729
PA	0.36	9770	0.12	9735	0.56	9472	0.94	9747	0.09	9705	0.02	9708
TX	0.37	13110	0.13	13074	0.56	12727	0.90	13076	0.08	13020	0.21	12977
VA	0.43	388	0.11	385	0.60	377	0.95	388	0.06	387	0.02	387
WA	0.40	33278	0.15	33172	0.57	32814	0.94	33234	0.07	33033	0.02	33131
All Grps	0.38	228369	0.13	227724	0.55	218407	0.94	227869	0.08	227064	0.07	227098

of zero. Other responses, or non-responses, are coded as missing data. Thus, for example, 38% of the 8618 respondents from Arizona were male (giving a mean value of 0.38 to the variable "Sex=Male" (henceforth abbreviated as "Sex") with values of 1 for men and 0 for women). As suggested by this example, the respondents in the BRFSS do not constitute a simple random sample of the population. The BRFSS survey supplies county weights for reweighting responses to better reflect the entire population. However, this paper does not seek to extrapolate relations outside the surveyed population, but focuses on quantifying conditional relations within this sample, e.g., studying how probability of asthma varies by age and sex and other variables, without considering how to adjust for differences between the joint frequency distribution of these variables in the survey population and in the more general population.

Similarly, not every respondent answered all questions, and there is no guarantee that responses can be extrapolated from those who did to those who did not. Hence, we only consider questions that were answered by almost all of the 228,369 respondents. For the variables in Table 1, for example, over 95% of surveyed individual answered each question.

The BRFSS data consist primarily of either dichotomous (yes-no) variables such as those in Table 1, all of which are coded as binary (0–1) variables with 0 = no, 1 = yes; or categorical variables, including age (50–99 years), income, education, and marital status. To these we added the two continuous pollution variables obtained from EPA: average daily O3 concentration in ppm and average daily PM2.5 concentration in micrograms of fine particulate matter per cubic meter of air. Table 2 lists the complete set of variables analyzed (other than survey year, month, and location) and their means and minimum and maximum values, as well as the number of individuals responding to each question. Table 3 shows the layout of the data (the first 21 of 228,369 records) for individual respondents. Ozone measurements were not available for the county (Apache County, AZ), year, and month of the survey (January 2010) for these 21 individuals. The entire data set is available from the author upon request.

Table 2
Variables, number of records with complete data for each question, and mean, minimum, and maximum values.

Variable	Valid N	Mean	Minimum	Maximum
Age	228369	65.64	50.00	99.00
Sex=Male	228369	0.38	0.00	1.00
Income Code	193321	5.66	1.00	8.00
Education	227945	4.92	1.00	9.00
Marital Status	228087	2.12	1.00	9.00
Smoking	172563	0.51	0.00	1.00
PM2.5	222349	9.39	1.45	31.54
O3	177148	0.04	0.01	0.08
Asthma Ever	227724	0.13	0.00	1.00
Heart Attack Ever	227064	0.08	0.00	1.00
Stroke Ever	227606	0.05	0.00	1.00

In Table 3, the three categorical variables *Income*, *Education*, and *Marital Status* have integer values for responses of 1–8, 1–6, and 1–6, respectively, with higher numerical values representing higher levels for *Income* and *Education*. *Smoking* is a binary variable that indicates whether a respondent reports having smoked at least 100 cigarettes (5 packs) during his or her life to date. The dependent variables *Asthma Ever*, *Heart Attack Ever*, and *Stroke Ever* are answers to the question of whether a doctor, nurse, or other health professional had ever told the respondent that s/he had the corresponding condition, with answers coded as 1 for yes, 0 for no, and blank (missing) for all other values.

3. Methods and analytic strategy

Since most of the variables in this data set other than age, PM2.5, and O3 are dichotomous or categorical, it is useful to examine associations and interactions among them using interaction plots that show how the mean value of one variable varies with the levels of one or more others. The following sections plot the main dependent variables of interest (prevalence of self-reported asthma, stroke, or heart attack) against explanatory variables such as age, income, sex, and average concentrations of O3 and PM2.5 in the counties where respondents lived at the time of the survey. Traditional 95% confidence intervals (mean plus or minus 1.96 sample standard deviations) are indicated visually as vertical bars around the mean values shown in the interaction plots. Such exploratory data analysis can reveal nonlinear patterns of association and does not require any parametric modeling assumptions. However, interaction plots are most useful for examining the relations among only a few explanatory variables and the dependent variables. We also used multiple logistic regression models to quantify associations between multiple explanatory variables and health effects, and used a non-parametric Bayesian network (BN) learning program (the *bnlearn* package in R) to discover and visualize statistical dependence relations (represented by arrows between variables) and conditional independence relations (represented by a lack of arrows between variables) among all variables simultaneously.

As discussed by Cox (2016), potential causal relations in observational data can be clarified using modern nonparametric methods. Many top-performing methods in recent competitions that evaluate the empirical performance of causal discovery and inference algorithms on suites of test problems (e.g., Hill, 2016; NIPS, 2013) use the following ideas: