Environmental Research 134 (2014) 435–452

Contents lists available at ScienceDirect

Environmental Research

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

Developing small-area predictions for smoking and obesity prevalence in the United States for use in Environmental Public Health Tracking

Alberto M. Ortega Hinojosa [a,*], Molly M. Davies [a], Sarah Jarjour [a], Richard T. Burnett [b], Jennifer K. Mann [a], Edward Hughes [c], John R. Balmes [a,d], Michelle C. Turner [e], Michael Jerrett [a]

[a] University of California, Berkeley, School of Public Health, Division of Environmental Health Sciences, 50 University Hall MC7360 Mail Drop, Berkeley, CA 94720-7360, United States
[b] Health Canada, Population Studies Division, Environmental Health Sciences and Research Bureau, Environmental and Radiation Health Sciences Directorate, Healthy Environment and Consumer Safety Branch, Room 134, Environmental Health Center, 50 Columbine Driveway, Ottawa, Ontario, Canada K1A 0K9
[c] Edward Hughes Consulting, 350 Queen Elizabeth Dr., #102, Ottawa, Ontario, Canada K1S 3N1
[d] University of California, San Francisco, School of Medicine, 1001 Potrero Ave, SFGH 30, San Francisco, CA 94143, United States
[e] McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, 1 Stewart Street, Ottawa, Ontario, Canada K1N 6N5

ARTICLE INFO

Article history:
Received 18 September 2013
Received in revised form
29 July 2014
Accepted 31 July 2014

Keywords:
Smoking
Obesity
Small-area
GIS
Multilevel model

ABSTRACT

Background: Globally and in the United States, smoking and obesity are leading causes of death and disability. Reliable estimates of prevalence for these risk factors are often missing variables in public health surveillance programs. This may limit the capacity of public health surveillance to target interventions or to assess associations between other environmental risk factors (e.g., air pollution) and health because smoking and obesity are often important confounders.
Objectives: To generate prevalence estimates of smoking and obesity rates over small areas for the United States (i.e., at the ZIP code and census tract levels).
Methods: We predicted smoking and obesity prevalence using a combined approach first using a lasso-based variable selection procedure followed by a two-level random effects regression with a Poisson link clustered on state and county. We used data from the Behavioral Risk Factor Surveillance System (BRFSS) from 1991 to 2010 to estimate the model. We used 10-fold cross-validated mean squared errors and the variance of the residuals to test our model. To downscale the estimates we combined the prediction equations with 1990 and 2000 U.S. Census data for each of the four five-year time periods in this time range at the ZIP code and census tract levels. Several sensitivity analyses were conducted using models that included only basic terms, that accounted for spatial autocorrelation, and used Generalized Linear Models that did not include random effects.
Results: The two-level random effects model produced improved estimates compared to the fixed effects-only models. Estimates were particularly improved for the two-thirds of the conterminous U.S. where BRFSS data were available to estimate the county level random effects. We downscaled the smoking and obesity rate predictions to derive ZIP code and census tract estimates.
Conclusions: To our knowledge these smoking and obesity predictions are the first to be developed for the entire conterminous U.S. for census tracts and ZIP codes. Our estimates could have significant utility for public health surveillance.

1. Introduction

The principal aim of this study is to develop a national small-area predictive model for two important health risk factors and potential confounders of the relationships between environmental exposures of interest and health outcomes: smoking and obesity prevalence. Both smoking and obesity have established effects on numerous health outcomes including mortality, various cancers, cardiovascular disease, metabolic disorders, muscular-skeletal conditions,

Abbreviations: BMI, body mass index; BRFSS, Behavioral Risk Factor Surveillance System; CDC, Centers for Disease Control and Prevention; EPHT, Environmental Public Health Tracking; MSE, Mean Squared Error; NCI, National Cancer Institute; NHANES, National Health and Nutrition Examination Survey; RDD, random digit dialing

* Corresponding author. Fax: +1 510 642 5815.
E-mail address: alberto.ortega@berkeley.edu (A.M. Ortega Hinojosa).

http://dx.doi.org/10.1016/j.envres.2014.07.029
0013-9351/© 2014 Published by Elsevier Inc.

and mental health (CDC, 2009c; National Cancer Institute, 2013b). Local-area estimates are of public health importance for multiple reasons including surveillance, measuring inequalities and examining contextual influences. In particular, policies and interventions to abate smoking and obesity are often designed and implemented at the local level, and not controlling for these confounders in studies can result in specious associations between exposures of interest and the health outcomes. For example, air pollution and health associations may be confounded by smoking, and lack of control for smoking may wrongly attribute risks to air pollution that are partly from smoking.

The Harvard Public Health Disparities Geocoding Project Monograph has identified the census tract as the most apt geographic level for monitoring health-related socioeconomic inequalities and recommends the analysis of health surveillance data at this level (Krieger et al., 2004). Nevertheless, currently, only limited national prevalence estimates exist for obesity rates at the county level (discussed below), and estimates at a finer spatial resolution (ZIP code and census tract) are not available nationally. Other national estimates such as the National Health and Nutrition Examination Survey (NHANES) have much smaller sampling (approximately 5000 per year) compared to the Behavioral Risk Factor Surveillance System (BRFSS) and offer only regional representation (CDC, 2013e).

1.1. The Environmental Public Health Tracking network and the geography of risk

Starting fiscal year 2002, the U.S. Congress has provided funding to establish the Environmental Public Health Tracking (EPHT) system and a corresponding EPHT network spanning from local and state governments to a centralized national repository. The EPHT was established to narrow the environmental health gap. Central to the EPHT goals is understanding how location affects health and well-being through differences in the environment—built or natural, physical or social; communicating current knowledge of the environment and health relationship to all stakeholders; informing policymaking; and tracking progress in attaining improvements in the population's health (Environmental Health Tracking Project Team, 2000).

Fundamental to achieving the EPHT goals is understanding the geography of risk, which can be characterized through three intersecting elements: (1) geography of susceptibility; (2) geography of exposure; and (3), geography of adaptation (Jerrett et al., 2010). That is, where the populations are most likely to be susceptible to a given exposure; how is the exposure of interest distributed in space; and how resilience and response are distributed by location? Small-area estimates of major health determinants such as smoking and obesity are essential to understanding the geography of risk.

1.2. Smoking, obesity and the geography of susceptibility

Of particular interest to the EPHT program is identifying populations at risk from the effects of smoking and higher body mass index (BMI) status. Understanding of the relationship between environment and health is hindered by a general lack of population health risk data on these two important confounders. Recent assessments of the global burden of disease identify smoking and obesity as leading risk factors for death and disability; smoking ranked first in North America and second globally, while high BMI respectively ranked second and fifth (Lim et al., 2012).

Because high exposure and high susceptibility often occur together, it is critical to control for confounding variables that share spatiotemporal overlap with potentially harmful environmental exposures such as air pollution. This overlap of high susceptibility and high exposure also makes identifying populations at risk of smoking and higher BMI status—and controlling that risk—an environmental justice issue (O'Neill et al., 2003).

Obesity and smoking are thought to have geographic patterns arising from the confluence of individuals with similar traits and exposures to similar social and physical environments such as race, socioeconomic status, environmental exposures and social norms (CDC, 2009d; Corsi et al., 2012; Cureton, 2011; Leal and Chaix, 2011; Liao et al., 2009; Pearce et al., 2012; Poland et al., 2006). The predictive model we develop here is designed specifically to exploit spatial and temporal patterns in these individual and lifestyle factors, which augment the predictions from fixed effects variables such as age, gender, education level and others that may influence smoking and obesity.

2. Methods

2.1. Overview

We used publicly available data from the U.S. Census and the BRFSS telephone survey to construct prevalence estimates of smoking and obesity for ZIP codes and census tracts. In addition to generating ecologic level prevalence predictions of smoking and obesity, we evaluated whether limited socioeconomic, gender and race data were sufficient to develop improved estimates relative to those that are currently available. We compare a basic terms variable set (i.e., no interactions or transformations) Generalized Linear Model (GLM) approach with no random effects to more sophisticated models including a lasso-based variable selection procedure, *GLMNet* and a two-level random effects model (*GAMEPHIT*). We also expand the variable space to include two-way interactions and log transformations.

The BRFSS data were aggregated to create county-level variables used in our models. These data were used to derive the independent (e.g., age, gender) and dependent variables in our models (obesity and smoking prevalence). To extend our estimates for small-area prediction, we compiled census data that matched the predictors in our final models and used the coefficients from the county-level model to create predictions at the census tract and ZIP code areas scales. Extensive analyses were conducted to understand the spatial patterns in the estimates and to compare these to other available estimates from similar studies or from the primary data. Fig. 1 shows the process followed in our study.

2.2. Data sources

BRFSS is a national random digit dialing (RDD) survey established by the Centers for Disease Control and Prevention (CDC). All 50 U.S. States and major territories are sampled, and data are publicly available at the individual level through the CDC (CDC, 2013a). Geographic scope (part of or the entire state) and resolution (finest geospatial scale assigned to an individual) vary by state. Estimates from BRFSS are thought to be representative of the population at the state-level, and county is normally the smallest area that is publicly available. Most states, however, redact the identity of rural counties to avoid individual identifiability of respondents. The CDC provides standard interview forms, but each state agency is responsible for conducting their own telephone interviews. From 1990 to 2011, participation rates (using the Council of American Survey Research Organizations weighted response formula) have fluctuated, ranging from a low of 48.9% in 2000, and a high of 71.4% in 1995 (CDC, 1991, 1996, 2001, 2006, 2011). We conducted extensive cleaning and compilation of county-level data from BRFSS because several geographic identifiers were missing from the publicly available data set (available from CDC (1991–2010)).

Census variables were obtained from the 2007 Business Analyst package (ESRI, Redlands, CA) for the 2000 Census for tract, ZIP code and county levels. ZIP codes were assigned county codes using ArcMap's (ESRI, Redlands, CA) feature to point and spatial join tools. The ZIP code polygon shapefile was converted to a point shapefile using the polygon's centroid (specifying that the centroid be contained within the originating polygon area). We downloaded 1990 census variables from the National Historical Geographic Information System (Minnesota Population Center, 2011). County codes were assigned ZIP codes by matching fields from the ZIP code file prepared for the 2000 Census data as described above.

2.3. Health outcomes

The BMI is commonly used as the measure of body composition to determine overweight and obesity (Gallagher et al., 1996; Hu, 2008). BMI measures body composition defined as an individual's weight in kilograms divided by the square of their height in meters (i.e., kg/m^2). In most individuals, BMI is highly correlated with percent body fat and provides a valid estimate of adiposity (Gallagher et al., 1996). Additionally, BMI is highly correlated with risk factors for obesity-related diseases (Hu, 2008) and with biomarkers of adiposity, such as high levels of leptin (Jurimae et al., 2003). Overweight for adults is defined as a BMI score equal to or greater than 25 and less than 30; obesity among adults is defined as having a BMI greater than or equal to 30 (CDC, 2009a).