Health & Place 35 (2015) 136–146

Contents lists available at ScienceDirect

Health & Place

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

Exploring the forest instead of the trees: An innovative method for defining obesogenic and obesoprotective environments

Claudia Nau[a,*], Hugh Ellis[a,b], Hongtai Huang[a], Brian S. Schwartz[a,c], Annemarie Hirsch[c], Lisa Bailey-Davis[c], Amii M. Kress[a], Jonathan Pollak[a], Thomas A. Glass[a]

[a] Johns Hopkins Bloomberg School of Public Health Global Obesity Prevention Center, 615 N Wolfe Street, Baltimore, MD 21205, USA
[b] Johns Hopkins Whiting School of Engineering, 3400 North Charles Street, Baltimore, MD 21218, USA
[c] Geisinger Center for Health Research, 100 North Academy Avenue, Danville, PA 1728, USA

A R T I C L E I N F O

Article history:
Received 2 March 2015
Received in revised form
14 July 2015
Accepted 1 August 2015

Keywords:
Obesogenic environments
Childhood obesity
Conditional random forest
Physical activity features
Food features
Social features

A B S T R A C T

Past research has assessed the association of single community characteristics with obesity, ignoring the spatial co-occurrence of multiple community-level risk factors. We used conditional random forests (CRF), a non-parametric machine learning approach to identify the combination of community features that are most important for the prediction of obesegenic and obesoprotective environments for children. After examining 44 community characteristics, we identified 13 features of the social, food, and physical activity environment that in combination correctly classified 67% of communities as obesoprotective or obesogenic using mean BMI-*z* as a surrogate. Social environment characteristics emerged as most important classifiers and might provide leverage for intervention. CRF allows consideration of the neighborhood as a system of risk factors.

1. Introduction

The concept of the "obesogenic environment" was first proposed in the late 1990's (Hill and Peters, 1998; Poston and Foreyt, 1999; Swinburn et al., 1999) as a framework for understanding the joint impact of multiple dimensions of place on obesity risk. Through their physical, institutional, or social features, obesogenic environments impede healthy energy balance-related behaviors by promoting inactivity and excess caloric intake. Since the concept was proposed, a rich body of research has linked numerous environmental characteristics with obesity in a variety of populations.

Multilevel studies have shown associations of several features of the built environment such as land use mix and population density (Frank et al., 2007 ; Franzini et al., 2009; Rundle et al., 2009; Schwartz et al., 2011b), as well as food establishments (Casey et al., 2008; Cummins and Macintyre, 2006; Drewnowski, 2004; Fleischhacker et al., 2011; Franco et al., 2008; Fraser et al., 2012; Giskes et al., 2011; Inagami et al., 2006; Lake and Townshend, 2006; Mehta and Chang, 2008; Michimi and Wimberly, 2010; Morland et al., 2006) and physical activity features (Gordon-Larsen et al., 2006; Kipke et al., 2007) with body composition and obesity at the individual level. Some studies control for social and economic characteristics as potential confounders (Meyer et al., 2015). We followed the socio-ecological literature and conceptualized and modeled social and economic community characteristics as key features of the risk landscape for physical inactivity and caloric over-consumption as well as risk-regulators that influence the likelihood of exposure to other obesogenic features of environments (Block et al., 2004; Greves Grow et al., 2010; Janssen et al., 2006; Larson et al., 2009; Nau et al., 2015).

Despite the recognition that obesogenic environments represent a diverse cluster of spatially co-occurring features, many studies presume that there are separate environments for food, social factors and physical activity-related features; this assumption has not been tested. Furthermore, most studies have used regression analysis to assess the independent effect of each "exposure" in isolation. Individuals, however, experience their community environment as a unified ensemble of features that may act jointly to affect health. A variable-by-variable approach risks what the sociologist, Gordon, called the partialling fallacy (Gordon, 1968). That is, the effect of the obesogenic context cannot be fully determined because it involves multiple variables that measure different dimensions of the same construct. Many studies have found small effects across a wide range of community risk factors when examined in isolation. None captures the totality of the impact of the obesogenic environment because that impact

* Corresponding author.
E-mail address: cnau1@jhu.edu (C. Nau).

http://dx.doi.org/10.1016/j.healthplace.2015.08.002
1353-8292/

represents what Marini and Singer call a conjunctive plurality of causes that cannot be represented by the independent effects detected in linear additive regression models (Marini and Burton, 1988). Further, many studies adjust for related features of environments that are facets of the spatially co-occurring structures that shape obesogenic environments. This exacerbates the partialling fallacy and biases observed associations toward the null. While theorists posit obesogenic environments as a complex multidimensional construct, standard regression analysis does not permit us to identify the combination of interacting risk factors that render an environment obesogenic. Researchers have begun to measure multiple environmental risk factors using factor analysis and latent class analysis (Adams et al., 2011; Meyer et al., 2015; Wall et al., 2012).

We expand this new body of work by demonstrating an innovative approach that allows us to identify from a large set of theoretically plausible risk factors those community features that are most important for rendering a community obesogenic. Our method allows us to reorient the focus from understanding *if* a particular risk factor matters to identifying the set of risk factors that *matter most*. We implement a method called conditional random forests (CRF) (Strobl et al., 2008) to analyze and identify community characteristics that together can predict observed rates of obesity at an ecological level. CRF is a supervised machine learning algorithm that has been used in biomedical research to identify, for example, the set of proteins associated with the presence or absence of a particular cancer (Izmirlian, 2004) or the combination of genetic and dietary factors that jointly increase the risk for the development of Metabolic Syndrome (de Edelenyi et al., 2008). Random Forests (RF) and CRF have also been applied in engineering (Kaur and Malhotra, 2008), geography (Pal, 2005) and ecology. To our knowledge, Basu and Siddiqi (2014) are the only authors to date who have used RF to identify risk environments by identifying features of geographic regions with high mortality.

We define obesogenic and obesoprotective environments for children as communities that fall into the highest or lowest quartile of the community level mean BMI *z*-score distribution. We use community level BMI-z because obesogenic and obesoprotective environments are ecological constructs that cannot be reduced to individual characteristics. Our approach is ecological, with the strengths and limits that this approach implies. We adopt an ecological approach because our goal is primarily about identification of constellations of community features. Obesogenicity and obesoprotectiveness are much like other community features, such as “walkability” or deprivation - properties of larger aggregate ecologies - not a function of the individuals who reside in them. In this study, we limit the scope of our inferences to relations between community constructs and average BMI among children in a community. We avoid the ecological fallacy by restricting our inferences to the group level (Schwartz, 1994). Further, this stage of the analysis is for proof of concept in a new method. We plan on a subsequent analysis using a multilevel analytic framework to address how CRF can inform variation in individual risk for obesity. CRF considers the coordinated effect of a large set of community features in order to classify communities. The CRF algorithm ranks the classifying variables in terms of their contribution to classification success. It identifies the combination of risk factors that matter most for differentiating obesogenic from obesoprotective communities.

The present study uses data from a large electronic health record of measured height and weight on children geo-coded to a diverse set of 1288 communities in 37 counties in Pennsylvania. We assembled a large dataset of community features from secondary data sources that have been linked to obesity in prior research. This set consists of 44 characteristics that cover multiple domains of community risk for obesity including social factors, food availability, and physical activity-related features including land use characteristics and physical activity establishments.

Using CRF we are able to examine the joint, spatially co-occurring pattern of features that may constitute obesogenic and obesoprotective environments. Results of these analyses allow us to: (1) identify the combination of features that are most important in rendering an environment obesegenic; (2) determine the relative importance of particular environmental features; (3) identify factors that do not improve classification accuracy; (4) characterize environmental features that may be targets for policy intervention.

2. Methods

2.1. Data and measures

For the outcome of interest, we used electronic health records from the Geisinger Health System on measured height and weight of children ages 10–18 during 2010 (N=22,497). Data was drawn from a dataset that provides information from the Geisinger Health System from 2001 to 2012 for children ages 0-18. Geisinger is the largest healthcare provider in Pennsylvania, serving patients in 37 counties in central and northeastern Pennsylvania. The study population has been found to be representative of the general population in the region (Schwartz et al., 2014). All children ages 2–18 with valid height and weight whose home address could be geo-coded via ArcGIS with a longitude and latitude were included in the sample. Our study population includes the subset of children who saw their healthcare provider in 2010 (Schwartz et al., 2014). Standardized BMI-*z* scores were computed using CDC growth charts to allow comparability across age and sex strata after removing implausible values (CDC, 2014a, 2014b). Prior research has shown that effects of the community environment are stronger for teenagers than younger children (Schwartz et al., 2011a). Therefore, this analysis was limited to children 10–18 years of age.

Children’s geo-coded street addresses were attributed to one of 1288 communities. Communities were operationalized as minor civil divisions (townships and boroughs) in rural areas or census tracts in urban areas (Schwartz et al., 2014). This mixed definition of community has been applied successfully in several other studies (Liu et al., 2013; Nau et al., 2015; Schwartz et al., 2014, 2011b). Townships and boroughs provide meaningful political community definitions in rural areas (Schwartz et al., 2014). Census tracts are the most frequently used community proxy in urban areas (Riva et al., 2007). Communities are assumed to be the dominant socio-geographic context within which an adolescent’s life unfolds. Applying a mixed definition of place provides greatest face validity for our community measure. We calculated average community level BMIs and classified communities as high- and low-obesity based on the top and bottom quartile in the distribution of average BMI-*z* in each community. To have stable estimates of mean BMI-*z*, we only considered communities with at least 50 children with measured BMI in 2010 (number of communities with 50 or more children N=197, number of communities in highest and lowest quartile used in the analysis N=99).

To capture features of the environment across multiple dimensions, we assembled multiple secondary data sources including information on food, social, physical activity and land use features of the environment. Data on physical activity establishments (e.g. exercise facilities, gyms, parks, outdoor recreational facilities) as well as both food service (e.g., full service restaurants, fast food restaurants) and food retail (e.g., convenience stores, grocery stores, specialty food stores) for 2010 were obtained from