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

Claudia Nau, Ph.D.
Post-doctoral fellow, Johns Hopkins Global Obesity Prevention Center
Presented at Annual meetings of the Society for Epidemiologic Research, June 25th, 2014, Seattle, WA (SPOTLIGHT SESSION)

Claudia Nau¹, H. Ellis³, H. Huang³, A. Hirsch², L. Bailey-Davis², B. Schwartz¹,² J. Pollak¹ A. Liu¹, T. Glass¹

¹ Johns Hopkins Bloomberg School of Public Health
² Geisinger Center for Health Research
³ Johns Hopkins Whiting School of Engineering
Obesogenic Environments

• Consider the **joint** impact of

• **Multiple dimensions** of the
  – Social
  – Physical activity and
  – Food environment

• On energy related behaviors and BMI
Prior research has identified numerous community-level risk factors associated with obesity.
Problem: Most prior research studied isolated risk factors, not the entire risk environment

• Limits of regression analysis
  – Assumes each feature as independent effects
  – Ignores geospatial clustering of risk factors
  – Control for other environmental factors potentially on the causal pathway (e.g., community socioeconomic status)
  – “Partialling” fallacy (Gordon, 1968)

• Very few studies measure the OG/OP environments comprehensively
  – Data reduction
  – Are the food, physical and social environments separate?

• “Partialling fallacy” (Gordon 1968)
  — Separate features of the obesogenic environment are not sufficiently distinct
Goal of this analysis:

1. To Identify **the combination of spatially co-occurring food, physical activity, and social features** that best classify environments of children as **obesogenic vs. obesoprotective**
Outcome data:

• The Geisinger Health System
  – Large primary care network in 37 county area of PA
  – Electronic health records, 2010
  – Measured height and weight, BMI z-scores
  – 22,497 children ages 10-18 years of age
  – Geocoded to residence in XXX communities defined as townships, boroughs or census tracts (in urban areas)
  – All communities with 50 or more children

• Obesogenic & Obesoprotective communities:
  – Highest (obesogenic) and lowest (obesoprotective) quartile of average BMI-z among eligible communities
## Independent variables: Community characteristics

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Source</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social features</strong></td>
<td>American Community Survey 5-year estimates (2005-09)</td>
<td>% pop unemployed % pop with less than high school</td>
</tr>
<tr>
<td>8 variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Food features</strong></td>
<td>Dun &amp; Bradstreet and InfoUSA, geocoded (2010)</td>
<td>Count grocery stores Count fast food outlets</td>
</tr>
<tr>
<td>19 variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Physical Activity features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Physical Activity Establishments</strong></td>
<td>Dun &amp; Bradstreet and InfoUSA, geocoded (2010)</td>
<td>Count gyms Count outdoor recreational facilities/clubs</td>
</tr>
<tr>
<td>11 variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Land use Characteristics</strong></td>
<td>American Community Survey 5-year estimates (2005-09) Penndot, Teleatlas</td>
<td>Vehicle miles travelled Population density</td>
</tr>
<tr>
<td>6 Variables</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methods: Why use conditional random forest analysis?

1. A machine learning method
2. Handles a large number of variables
3. No parametric assumptions
4. Iterative method of searching for a combination of variables that classify (predict) an outcome
5. Does not require strong assumptions of other latent variable measurement methods
Methods: Conditional Random Forest (CRF), The general idea

CRF grows classification trees:

• Uses entire set of community characteristics

• Splits communities recursively into groups based on outcome (BMI-z average) to form a tree

• Each successive set of groups is more homogeneous in terms of the outcome than the previous group.

• Identifies variables that contribute most to classification success (variable importance score)
Conditional Random Forests
General idea cont.

• CRF grows many trees (in our case 5000)
• Each tree is grown on
  - A random subsample of communities (training dataset)
  - At each split only a randomly drawn subset of variables is evaluated
  - Each tree used to predict the classes of the unused data (out of bag sample, OOB)
  - Results of predictive success are averaged across all trees
Conditional Random Forest
Principal outcome measures

• Overall and class-specific error rates (OOB Errors)
  – Measures classification success of entire set of independent variables

• Conditional variable importance list
  – Relative ranking of variables in terms of their importance for the classification
  – Conditional on all other variables in the forest
Results: OOB Classification error for analysis of 44 community characteristics

**OOB Errors - Full Set of Variables**

- Overall: 33%
- Obesoprotective: 32%
- Obesogenic: 33%
Conditional Variable Importance ranking of 44 predictors
Conditional Variable Importance ranking of 44 predictors

- 13 variables contribute consistently to the classification
- Top 13 variables come from all three domains
- Social characteristics dominate (7 out of 13)
- Followed by land use characteristics
Conditional Variable Importance ranking of 44 predictors

- After consideration of all variables, some frequently studied variables do not improve prediction accuracy.
Food vs. Physical activity features: Does one provide greater leverage for intervention?

- Food features classify Obesogenic environments well
- Reverse is true for physical activity environment characteristics
Limitations

1. CRF does not provide direction or absolute size of effects
2. Sampling underrepresents low population communities
3. Mean BMI-z may be insensitive measures of obesogenic/obesoprotective environments
4. Measurement error in indicators
5. Community selection as confounder (reverse confounding)
6. Atheoretical?
Summary

1. We identified a combination of 13 variables from multiple domains that classify obesogenic and obesoprotective environments based on mean BMI-z

2. Social characteristics of communities are powerful classifiers

3. Not separate environments

4. Obesogenic environments classified by food features; obesoprotective environments classified by physical activity characteristics

5. CRF is a promising approach for characterizing spatially co-occurring features of the risk environment
Acknowledgement:

**Johns Hopkins Project Team:**
- Thomas A. Glass (Epi), Project leader
- Brian S. Schwartz (EHS), Project co-leader
- Karen Bandeen-Roche (Bio), Biostatistician
- Joseph Bressler (EHS), epigenetics
- Tak Igusa (Engineering), Systems scientist
- Claudia Nau (IH), Post-doctoral fellow, trainee
- Mehdi Jalalpour (Eng), Pre-doctoral fellow, trainee
- Amii Kress (Epi), Post-doctoral fellow, trainee
- Jonathan Pollak (EHS), Data analyst

**Geisinger Health System Team:**
- Annemarie Hirsch, subcontract PI
- Lisa Bailey-Davis, co-investigator
- Dione Mercer, Project coordinator
- Sy Landau, Research assistant
- Joseph DeWalle, GIS analyst

Funding was provided from the Education and Training Core of the Johns Hopkins Global Obesity Prevention Center and by Grant Number U54 HD070725 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD). The project is co-funded by Office of Behavioral and Social Sciences Research (OBSSR).
END OF TALK