Developing a Data-driven Approach to inform Planning in County Health and Human Services Departments in the Context of a Case Study on Obesity

Stephanie Shipp  
University of Virginia

Joshua Goldstein  
University of Virginia

Vicki Lancaster  
University of Virginia

Ian Crandell  
Virginia Tech

Follow this and additional works at: https://repository.upenn.edu/admindata_conferences_presentations_2018

Shipp, Stephanie; Goldstein, Joshua; Lancaster, Vicki; and Crandell, Ian, "Developing a Data-driven Approach to inform Planning in County Health and Human Services Departments in the Context of a Case Study on Obesity" (2018). 2018 ADRF Network Research Conference Presentations. 44.  
https://repository.upenn.edu/admindata_conferences_presentations_2018/44

DOI https://doi.org/10.23889/ijpds.x345.1077

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/admindata_conferences_presentations_2018/44

For more information, please contact repository@pobox.upenn.edu.
Developing a Data-driven Approach to inform Planning in County Health and Human Services Departments in the Context of a Case Study on Obesity

Abstract
Since the 1970s, the obesity rate has steadily increased due to growing availability of food and declining physical activity. The existing environments within a community, including active recreation opportunities, access to healthy food options, the built environment, and transportation options, can moderate obesity. In Virginia, Fairfax County Health and Human Services (HHS) system is interested in developing the capacity for data-driven approaches to gain insights on current and future issues, such as obesity, to characterize factors at the county and sub-county level, and to use these insights to inform policy options. In exploring these questions, we developed statistical methods to combined data from a multitude of different sources including local administrative data (e.g., tax assessments, land use, student surveys), place-based data, and federal collections. Using synthetic data methods based on imputation, we recomputed American Community Survey statistics for non-Census tract geographic regions for political districts and high school attendance areas. We combined this with environmental factors, such as land dedicated to parks and recreation facilities, as well as measures of the density of healthy and unhealthy food locations to create a map of potentially obesogenic factors. Finally, we combined these data sources with Fairfax County’s youth survey and trained a random forest model to predict the effects of the environment on healthy food consumption and exercise. Our analysis highlights the need for (administrative) data at a fine scale and recommends policy changes concerning the recording and sharing of local data to better inform the policy and program development.

Comments
DOI https://doi.org/10.23889/ijpds.v3i5.1077

This presentation is available at ScholarlyCommons: https://repository.upenn.edu/admindata_conferences_presentations_2018/44
Developing a Data-driven Approach to Inform Planning in County Health and Human Services Departments: Case Study on Obesity

ADRF November 2018

Sessions - Uses of Administrative Data for Supporting Public Programs and Public Health Policy

Michelle Gregory
Sophia Dutton
*Fairfax County*

Sallie Keller (PI)
Stephanie Shipp
Vicki Lancaster
Joshua Goldstein,
*University of Virginia*

Ian Crandell
*Virginia Tech*
Engagement, Issues, & Questions

**Overarching Goal:** Develop data-driven insights on current issues and build forecasts to inform future issues

- Expand Fairfax County's capacity to access and integrate county, state, and federal data in useful ways to address critical problems

**Project Focus:** Identify the trends in obesity and activities related to obesity across geographies of interest for local policy and program development

- Focus on determinants identified in the literature related to obesity - the built environment, nutrition, physical activity, family support, demographic and economic characteristics
Data Science Framework

Problem Identification: Relevant Theories & Working Hypotheses

Data Discovery

Data Ingestion & Governance

DATA INGESTION & GOVERNANCE

DATA PROFILING
Data Structure, Data Quality, Provenance & Meta Data

DATA PREPARATION
Cleaning, Transformation, Restructuring

DATA LINKAGE
Ontology, Selection & Alignment, Entity Resolution

DATA EXPLORATION
Characterization, Summarization, Visualization

DATA PROFILING

DATA INGESTION & GOVERNANCE

DATA INGESTION & GOVERNANCE

DATA INGESTION & GOVERNANCE

DATA INGESTION & GOVERNANCE

FITNESS-FOR-USE-ASSESSMENT

STATISTICAL MODELING & DATA ANALYSES
Data science innovations to develop *sub-county* data-driven insights

- Data Discovery - identify all types of data
- Synthetic information technology
- Geocoding housing units and locations of interest
- Vulnerability Composite Indicators
- Exploring the data using visualization tools
Overarching Goal: Expand Fairfax County's capacity to access and integrate county, state, and federal data to address critical problems and build forecasts to inform future issues

Project Focus: Identify trends in youth obesity, access to healthy food, and physical activities across geographies of interest for local policy and program development
Data Discovery - Local community Data Map
Healthy eating and physical activity of teens

- Access to healthy food - grocery stores, community gardens, farmers markets, restaurants (fast food, other)
- Living Conditions
- Personal Safety
- Engagement
- Support Networks

- Behavioral Health
- Physical Health
- Social Wellness
- Support Networks

- Education
- English Literacy
- Health Literacy
- Engagement
- Support Networks

- Family Stability
- Income Stability
- Living Conditions
- Health Literacy
- Support Networks
### Data Discovery, Inventory & Acquisition

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Geography</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Community Survey data (Census), 2011-2015 (updating now to 2012-2016)</td>
<td>Census Tracts and Block Groups</td>
</tr>
<tr>
<td>American Time Use Survey (BLS), 2017</td>
<td>National</td>
</tr>
<tr>
<td>Youth Risk Behavior Surveillance System, 2015</td>
<td>State</td>
</tr>
<tr>
<td>County Health Rankings, 2017</td>
<td>County</td>
</tr>
<tr>
<td>Built Environment, e.g., Grocery stores, SNAP retailers, recreation centers, community gardens</td>
<td>Address Level</td>
</tr>
<tr>
<td>Fairfax real estate tax assessment data</td>
<td>Address Level</td>
</tr>
<tr>
<td>Fairfax Open data: Zoning, Environment, water, Parks, Roads</td>
<td>Shapefiles</td>
</tr>
<tr>
<td>Fairfax County Youth Survey, 2016</td>
<td>High School Attendance Area</td>
</tr>
<tr>
<td>8th, 10th, 12th graders</td>
<td></td>
</tr>
<tr>
<td>Virginia Department of Education, 2017</td>
<td>High School</td>
</tr>
<tr>
<td>National Center for Education Statistics, 2014-2015</td>
<td>High School</td>
</tr>
<tr>
<td>Center for Disease Control, 2014-2015</td>
<td>High School</td>
</tr>
</tbody>
</table>

Initial data sources used with geographic specificity

- All are updated as new data are available

![Data Discovery Workflow](image)
Re-Distribution of Data & Estimates Across Geographies

**Problem** - Data do not align with geographies of interest, e.g., Supervisor Districts and School Boundaries

**Solution** - Use data **direct aggregation**, if possible, alternatively develop **synthetic populations** based on data and redistribute

**Synthetic re-distribution** based on variables of interest
- Iterative Proportional Fitting (IPF)
- Multivariate Imputation by Chained Equations (MICE)
Direct aggregation based on location of housing units

Geocoding owner-occupied local housing stock

In general, adding rental units can be a challenge and may require imputation.

Examples of place data:
- All restaurants
- Fast Food restaurants
- Farmer’s Markets
- Community Gardens
- Recreation Centers
- SNAP Retailers
- Parks
Re-distribution of data based on Synthetic Information

- Use American Community Survey (ACS) summaries and PUMS microdata to impute synthetic person data for all people in area of interest
- Re-weight synthetic data according to ACS tables to simultaneously match the relevant distributions, to Census Tracts or Block Groups
  - Age, income, race, and poverty in this case
- Aggregate synthetic person data to compute summaries, and margins of error, over the new geographic boundaries
Fairfax County Vulnerability Indicators

Statistical combination of percent of Households with:

- housing burdens > 50% of Household income
- no vehicle
- receiving Supplemental Nutrition Assistance Program (SNAP)
- in poverty

Source: American Community Survey 2011-2015 aligned to geographies using SDAL Synthetic Technology
Fairfax Profiles by Supervisor Districts
Dashed lines = Average; Supervisor Districts arranged by Vulnerability Index from high to low

Source: American Community Survey 2011-2015 aligned to Supervisor Districts using SDAD Synthetic Technology
Some students in all school areas experience food insecurity every month

- 22% of all students in Fairfax County High Schools experience food insecurity (dashed line)
High School Vulnerability Indexes

Combination of:
- Percentage of student in LEP classes
- Percentage of students that eligible for one of the following:
  - Free/Reduced Meals
  - Medicaid
  - Temporary Assistance for Needy Families
  - Migrant or experiencing Homelessness

Fairfax County - next steps

- **Descriptive analytics** - Prepares and analyzes *historical* data and identifies patterns from samples for reporting of trends

- **Predictive analytics** - Predicts *future* probabilities and trends and finds relationships in data that may not be readily apparent with descriptive analysis

- **Prescriptive analytics** - Prepares models to conduct *scenario analysis* on future policy and program options
Operational realities - Could you briefly share when this project started, when you got data access, and when it was completed? And any other reflections on the operational opportunities or challenges related to using administrative data?

Timeline - overall about a year; actual research about 5-6 months
- Discussions in spring 2017
- Agreement in place by June 2017
- Kickoff in August 2017
- Data work and analysis from August – January 2018
- Briefings during the spring 2018

Data access
Used open data from FX and the web + ACS, federal surveys, place-based data scraped from web + anonymized FX Co. Youth data, which we used for some preliminary modeling.
Research provided increasing awareness about making other data available to researchers through DSAs and processes to ensure the security of the data. Our research group has these processes in place. The next step is to work with the county.
What do you think are the biggest challenges and opportunities for your profile in relation to data access (whether Agency data owners, Federal research who was not the data owner, an outside academic, or another type of profile)? How much did you have to modify your target question based on data access you had?

- Challenges have led to our innovations.
- Rarely have perfect data – in this case, missing key variable, BMI
- Data discovery is important step for us to identify data we ideally want. Provides us with a guide and helps identify gaps.
- Also has led us to develop synthetic technology processed to statistically integrate and aggregate data as well as to present data at different levels of geography than just CTs and CBs, such as political districts and school boundaries.
- **Data Ingestion and Governance part of our DSF** - the profiling, cleaning, linking, and exploring the data - can take 80% of the time of the project. We are developing methods to automate and hence reduce this time – which will lead to reductions in the amount of time to do these activities and increase over time – and hence allow more time for analysis. This is a really important part of our work.