ABSTRACT
This paper describes an analysis of spatial clustering of colorectal cancer (CRC) in Miami-Dade County, Florida. The objective was to identify geographically based targets for colorectal cancer screening interventions for Blacks and Hispanic Whites, two groups with demonstrated disparities in stage at diagnosis and mortality for CRC. The initial cluster detection analysis identified areas with high risk of late stage CRC, however, none of the results were statistically significant.

The analysis was not based on an academic research question, but instead was an application intended to guide appropriate and targeted strategies for high risk populations. Only about 50% of the general population receives CRC screening, so, while all groups would benefit from increased CRC screening, high risk communities may potentially benefit the most. Because public health resources are limited, geographically targeting high risk populations for enhanced screening efforts is pragmatic public health policy.

Despite the lack of statically significant results, we still needed to develop a helpful answer to the question, where should we market a screening intervention?

The selected geographic areas must have real potential for attenuating excess CRC burden through increased screening efforts. Through evaluating a combination of clusters of late stage and overall CRC risk (two separate models of cluster detection), probable communities with low CRC screening uptake were identified. Although they did not meet statistical significance, they were determined to have public health importance.

Categories and Subject Descriptors
Conference proceedings

General Terms
Measurement

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM SIGSPATIAL HEALTHGIS '12, November 6-9, 2012, Redondo Beach, CA, USA.
Copyright © 2012 ACM ISBN 978-1-4503-1703-0/12/11…$15.00.

Keywords
Colorectal cancer clusters, stage at diagnosis, public health significance, SaTScan, screening disparities

INTRODUCTION
Many epidemiologic studies on the distribution of disease are empirical and rely predominantly on the analysis of secondary data[1,2]. This has been driven by renewed interest in the importance of place in health, expanded availability of geocoded cancer data, and advances in computing capacity [3]. While the analysis can be executed with relative ease, what is less straightforward is which methodology to apply, which level of aggregation to use, understanding the impact of data quality, and how to interpret and apply the results.

The development of cancer registries and national requirements to geocode cancer cases has expanded the scope of cancer surveillance beyond cancer mortality to include the routine use of maps for examining geographic variation in incidence, stage at diagnosis, and survival [4,5,6,7,8,9,10,11,12,13]. Mapping cancer rates is important for visualizing spatial or spatial-temporal patterns that may help identify differences in disease burden for different geographic locations. Locating areas with high rates of cancer can help prioritize cancer control efforts, including implementation of community interventions designed to modify risk behaviors [14]. Mapping geographic variation in stage at diagnosis is also routinely conducted by health officials as a surrogate for cancer screening utilization [5,14,15,16,17,18]. Geographic patterns of disease can be fairly easily mapped, but one challenge is to identify which areas have public health significance. This challenge is the subject of our paper.

CRC is a common cancer in industrialized countries. It is the third most commonly diagnosed invasive cancer, among men and women combined, and the third leading cause of cancer-related death in Florida and the United States [19,20]. CRC mortality rates among non-Hispanic Whites have been decreasing over the last few decades, yet a correspondingly favorable reduction has not been observed among Blacks or Hispanics [21,22]. Reducing the incidence and mortality of CRC is a national health priority area and reflected in the Centers for Disease Control and Prevention (CDC) Healthy People 2020 objectives.
Because prognosis and quality of life is critically dependent upon the stage of cancer at diagnosis, routine screening can reduce mortality due to CRC through early detection [23]. Although evidence-based consensus is lacking on which screening regimen is best, routine CRC screening by fecal occult blood test (FOBT), sigmoidoscopy, or colonoscopy for individuals aged 50 to 75 years of age is a Grade A Recommendation (the highest) by the US Preventive Services Taskforce [24]. And because effective screening by colonoscopy can lead to the identification and removal of precancerous lesions, CRC is potentially eradicable through secondary prevention [25]. Therefore, a diagnosis of CRC, particularly a late stage diagnosis, can be viewed as a preventable, adverse health outcome.

Enhanced screening efforts can lead to a decrease in the burden on of CRC and is, therefore, the focus of a number of national initiatives including Healthy People 2020 and Agency for Healthcare Research and Quality Priority Areas for National Action. The characterization and identification of populations in need of enhanced CRC screening the most is an important component for population-based cancer control.

This paper describes an analysis of spatial clustering of CRC in Miami-Dade County using SaTScan’s spatial scan statistic. The spatial scan is routinely used in epidemiology and public health to identify geographic areas where disease rates (e.g. cancer incidence) or health events (e.g. cancer screening) occur at unusually high or low levels compared to the rate of the background population [26]. The spatial scan statistic creates an infinite number of discreet, circular or elliptical windows across a geographic area. Each circle or ellipse is evaluated as a possible cancer cluster; the null hypothesis being that there is no increased risk inside the scanning window.

Miami-Dade County is the most populous county in the state of Florida. The County is a highly urban, uniquely ethnically and racially diverse with the majority of the population foreign born with over 70% of persons over the age of 5 speaking a non-English language in the home [27]. Only 16% of the population is non-Hispanic White, 19% is Black, and 63% is Hispanic, the majority of which are Cuban [28]. A high percentage, 17%, of Miami-Dade residents are living in poverty [27].

Although their recent screening rates have been improving, non-Hispanic Blacks and Hispanics have significantly lower CRC screening rates compared to non-Hispanic Whites in Miami-Dade County. For persons aged 50 and above, 49% of non-Hispanic Whites reported ever receiving an FOBT and 80% reported ever receiving either a sigmoidoscopy or colonoscopy. Non-Hispanic Blacks reported 29% and 68%, respectively, and Hispanics, 28% and 63%, respectively [30]. Miami-Dade has a statistically higher CRC incidence rate than Florida as a whole, 45.1 cases per 100,000 compared to 41.8, and a statistically similar mortality rate as the rest of Florida of 14.2 per 100,000 [20].

The aim of the analysis was to identify geographic targets for colorectal cancer screening interventions for Blacks and Hispanic Whites, two populations with screening rates well below the state average [29]. Originally, areas of high risk Haitian and Cuban populations were intended to be targeted, but due to lack of small area-based population data and other issues, the majority of the analysis was conducted on the more general populations of Blacks and White Hispanics of Miami-Dade.

Operating under the hypothesis that disparities in risk of late stage diagnosis are amendable through increased access to population-based CRC screening, this study did not investigate any risk factors for late stage diagnosis other than location. This location only approach is appropriate when attempting to identify geographic areas potentially lacking accessible CRC screening resources, such as a low provider to population ratio. And it is also defensible by applying an extension of the Homophily Principle that states similarity breeds connection, that is, people who are similar in sociodemographic, behavior and intrapersonal characteristics tend to associate together. After the areas of high risk have been determined, additional analysis is planned to determine if screening interventions should focus on increasing physical access, e.g. increasing number of providers or clinics, or on behavior modification.

Because the risk of CRC overall varies by race, ethnicity, and socioeconomic factors [5, 20], and because we are interested in a secondary screening intervention versus a primary prevention, we did not use rate maps to inform our geographic selection. Instead we relied on cluster detection, particularly a model that evaluates the ratio of risk for late stage compared to early stage.

METHODS

Population-based CRC incidence data were obtained from Florida’s statewide cancer registry. All CRC cases (ICD-O C18.0-C20.9, C26.0 excluding histologies 9140, 9050-9055, and 9590-9989) diagnosed from 2006-2009 in Miami-Dade County with known age, sex and geocoded address at diagnosis were included. Both race and Hispanic origin, two mutually exclusive variables in cancer surveillance, had to be known for the Hispanic White analysis, but only race was required for the Black analysis. Only White Cubans were included in the Cuban analysis.

All geocoding was done at the central cancer registry level by a proprietary vendor, Nielsen Claritas. Cases that are not matched during the automated portion of the geocoded process are not resolved interactively. The FCDS current geocoding procedures do not include any attempts at manual geocoding or address correction.

For spatial cluster analysis on late stage risk, cases with unknown stage were excluded unless they were reported as unknown by the original treating or diagnosing facility or the case was identified at time of death. This was intended to exclude cases that had an unknown stage due to problems with data collection or reporting but intended to include cases that had progressed to the point that a staging work-up was likely not medically warranted.
analysis showed that areas with low risk of a late stage CRC diagnosis appear if the analysis excludes all cases of unknown stage. It is suspected these are areas with better reporting to the cancer registry and not relevant for determining areas that would benefit from increased CRC screening. About 25% of the cases of unknown stage were excluded from the risk of late stage analysis. SEER Summary Stage 2000 was used to dichotomize cases as early *(in situ and local)* and late (regional, regional extension, metastatic, & unknown— if the case was reported by the diagnosing or treating facility).

SaTScan ver 9.1.1, a cluster detection software developed in part by the grants from the National Cancer Institute (NCI) and the Center for Disease Prevention and Control (CDC), was used to detect clusters of high risk of late stage CRC as well as clusters of high and low risk of a diagnosis of CRC.

A Bernoulli model was used to identify communities with high rates of late stage (compared to early stage) CRC for Blacks, White Hispanics, and Cubans. A discrete, Poisson-based model was used to identify communities of either high or low risk of CRC rates overall for Blacks and White Hispanics. The Poisson models were adjusted for age and sex. For both models, we used a circular scanning window, set the maximum cluster size set at 20% of the population at risk, and ran 9,999 Monte Carlo simulations to calculate p values and to adjust for multiple comparisons. The choice of parameter settings for SaTScan analysis is not consequential and can influence results [32]. The selection of a maximum cluster size of 20% was based on previous analysis conducted by the research group on tobacco-associated cancers. The 20% cut-off represents a replicable standard that was used for a series of cancer cluster analyses that largely resulted in cluster sizes that could be managed by local or regional public health agencies as well as generally homogenous individual location risk, as compared to the overall cluster risk.

The units of analysis were based on the census boundaries for which population data are available, census tract and block group. To address the modified area unit problem (the potential for different results based to aggregation of the data [33]), each analysis was conducted using both the block group and the census tract as the lowest level of analysis. Population data (sex by 18 age groups) from the US Census 2010 was used for the Poisson model. Because the Bernoulli model is a ratio comparing rates of late stage CRC to rates of late stage, only case level data was necessary and, thus, enabled the additional Cuban analysis.

The resulting clusters were mapped in ESRI ArcMap 10. Communities selected for screening intervention were determined first by using the intersection overlay function for areas identified as high risk of late stage CRC, cluster determined using the Bernoulli model. Then a manual review of the overlap of both the Bernoulli and Poisson models was conducted and additional communities selected based on statistical significance, overall risk of CRC, number of cases in the clusters and population size/density. The manual review was a post-hoc decision based on the limited results obtained with the geospatial overlay of areas at high risk for late stage diagnosis. Finally, the areas selected for intervention were evaluated based on quality of geocoded data.

### RESULTS

Table 1. Case Characteristics, Colorectal Cancers

<table>
<thead>
<tr>
<th>Year</th>
<th>Hispanic Male</th>
<th>Hispanic Female</th>
<th>Black Male</th>
<th>Black Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>2.1%</td>
<td>1.4%</td>
<td>1.1%</td>
<td>0.7%</td>
</tr>
<tr>
<td>2006</td>
<td>2.5%</td>
<td>1.9%</td>
<td>1.0%</td>
<td>0.6%</td>
</tr>
<tr>
<td>2007</td>
<td>2.9%</td>
<td>1.7%</td>
<td>1.9%</td>
<td>0.7%</td>
</tr>
<tr>
<td>2008</td>
<td>3.0%</td>
<td>1.5%</td>
<td>1.3%</td>
<td>0.5%</td>
</tr>
<tr>
<td>2009</td>
<td>2.9%</td>
<td>1.3%</td>
<td>0.9%</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

Table 2. Geocoding Characteristics, Colorectal Cancers

<table>
<thead>
<tr>
<th>Year</th>
<th>No code</th>
<th>PO Box</th>
<th>Zipcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>2.1%</td>
<td>0.5%</td>
<td>1.1%</td>
</tr>
<tr>
<td>2006</td>
<td>2.5%</td>
<td>0.3%</td>
<td>1.0%</td>
</tr>
<tr>
<td>2007</td>
<td>2.9%</td>
<td>0.6%</td>
<td>1.9%</td>
</tr>
<tr>
<td>2008</td>
<td>3.0%</td>
<td>0.6%</td>
<td>1.3%</td>
</tr>
<tr>
<td>2009</td>
<td>2.9%</td>
<td>0.1%</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

Table 3. Summary Statistics, Colorectal Cancers

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic, White</td>
<td>2.6%</td>
<td>1.9%</td>
<td>2.1%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Black</td>
<td>2.0%</td>
<td>1.4%</td>
<td>1.8%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Cuban</td>
<td>2.6%</td>
<td>1.9%</td>
<td>2.1%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Haitian</td>
<td>1.0%</td>
<td>0.0%</td>
<td>0.7%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Stage at Diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Late</td>
<td>2.3%</td>
<td>1.2%</td>
<td>2.1%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Unknown</td>
<td>4.9%</td>
<td>2.0%</td>
<td>4.2%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

---

**RESULTS**
### Table 3. Risk of Late Stage Colorectal Cancer, Miami-Dade, FL 2005-2009

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>n</th>
<th>Late</th>
<th>Level</th>
<th>ID</th>
<th>Cluster</th>
<th>(Late)</th>
<th>Persons per Sq Mi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuban</td>
<td>1679</td>
<td>61%</td>
<td>BG</td>
<td><strong>A</strong>* 1.65</td>
<td>0.94</td>
<td>12</td>
<td>na</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>B</strong> 1.17</td>
<td>0.99</td>
<td>228</td>
<td>na</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>CT</strong> C* 1.33</td>
<td>0.13</td>
<td>83</td>
<td>na</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>D</strong> 1.65</td>
<td>0.999</td>
<td>8</td>
<td>na</td>
</tr>
<tr>
<td>White, Hispanic</td>
<td>3887</td>
<td>59%</td>
<td>BG</td>
<td><strong>E</strong>* 1.19</td>
<td>0.29</td>
<td>276</td>
<td>700,980</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>CT</strong> F* 1.17</td>
<td>0.56</td>
<td>249</td>
<td>676,220</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>G</strong> 1.50</td>
<td>0.90</td>
<td>23</td>
<td>118,575</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>H</strong> 1.39</td>
<td>0.98</td>
<td>32</td>
<td>40,600</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>I</strong> 1.52</td>
<td>0.999</td>
<td>17</td>
<td>41,526</td>
</tr>
<tr>
<td>Black</td>
<td>889</td>
<td>52%</td>
<td>BG</td>
<td><strong>J</strong> 1.63</td>
<td>0.93</td>
<td>12</td>
<td>45,857</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>K</strong> 1.43</td>
<td>0.99</td>
<td>27</td>
<td>79,378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>L</strong> 1.44</td>
<td>0.999</td>
<td>22</td>
<td>68,383</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>CT</strong> <strong>M</strong> 1.46</td>
<td>0.999</td>
<td>28</td>
<td>116,358</td>
</tr>
</tbody>
</table>

*Clusters have significant spatial overlap

---

**Figure 1.** Selection based on Intersect, Cubans and Hispanic Whites
Figure 2. Colorectal Cancer Clusters Cubans and Hispanics Whites. Note: crisscrossed hatches indicate high risk at both the tract and block group level.

Figure 3. Colorectal Cancer Clusters, Blacks
There were 4,066 Hispanic White CRC cases total with 3,955 cases meeting selection criteria and 904 Black cases total with 889 cases meeting selection criteria (Table 1). Because there were fewer than 100 cases identified as Haitian, no analysis was conducted on this sub-population. There were a few important differences identified between the total cases and the selected cases. Nearly 41% of the Hispanic White cases were Cuban, but only 31% of the cases met selection criteria. And the percentage of unknown stage for Blacks was higher among the selected cases, 10.8% total and 13.4% selected.

Slightly higher percentages of Cubans and the elderly were excluded from the analysis due to lack of geocoded address at diagnosis (Table 2). And the percentage of cases with unknown stage that were not geocodable was significantly higher than any other category, 4.9%.

For Cubans, there were four (4) clusters of risk of late stage CRC detected, two (2) based on block group level analysis and two (2) based on census tract level analysis (Table 3). For Hispanic Whites, there was one (1) cluster detected using the block group level analysis and four (4) using the census tract level. However, there were no statistically significant clusters of risk of late stage for Cubans or Hispanic Whites. One (1) area had significant overlay with each analysis and the result of the intersection overlay is shown in Figure 1. This area of risk of late stage diagnosis for both Cubans and White Hispanics at both the census tract and block group level is located in an area of statistically significant higher than expected rates of overall CRC among White Hispanics (RR 15.0, p < 0.01, 2,428 cases in cluster, block group level analysis) (Figure 2). The overlap of high risk of CRC in the same community also at high risk for late stage at diagnosis for CRC implies, despite lack of statistical significance for the risk of late stage cluster, an area that likely would benefit from enhanced screening efforts. After manual review, a second area of interest was identified. Cluster G, an area of high risk of late stage CRC cancer among Hispanic Whites, overlays an area of statistically significant low rates of CRC (RR 0.65, p = .09, 87 cases in cluster). This suggests a community with artificially low rates of CRC due to limited population based CRC screening. A review of the quality of the geocoded data match indicated that the clusters did not contain a higher percentage of PO Box or partial geocoded matches than the county as a whole. Please note, in order to preserve group level confidentiality, the cluster maps presented here are not provided with any point of reference that can identify the location of the clusters within Miami-Dade County.

DISCUSSION

Cluster detection of areas at high risk of late versus early stage CRC cancer (Bernoulli model) alone did not provide enough substantiation to empirically target geographic areas for CRC screening interventions for Blacks or Hispanic Whites in Miami-Dade County. However, with the inclusion of the Poisson models of overall CRC risk, useful target areas were determined. It is intuitive that the lack of statistical significance in the Bernoulli models may indicate a lack of power for ratio based cluster detection for these populations. The use of the two models combined identified four (4) target areas for a screening intervention program—two (2) areas appropriate for targeting Hispanics Whites (a combined population of 794,795 Hispanic Whites), and two areas appropriate for targeting Blacks (a combined area with a population of 114,240 Blacks).

Limitations:

There are a number of limitations to this study that are common in the application of spatial techniques to the field of cancer epidemiology. For instance, the lack of detailed population data released at the tract or block group level. We were unable to conduct the analysis we planned with the Haitian population of interest and had to rely on analysis of the more general, total Black population instead. Not only is the underlying denominator data for Haitians unavailable from the US Census at a level that enables the calculation of age-adjusted rates, but the cancer record does not record Haitian origin unless the patient has died and the place of birth is indicated as Haiti on the death certificate. This issue is not unique to the Florida Registry, but it is a particular problem for sub-population analysis in such an ethnically diverse metropolitan areas as Miami-Dade. Also, the impact of geocoding quality could not be assessed. Although the percent of PO Box or Partial matched geocodes were not higher in the areas of interest, indicating the presence of a cluster is not spurious due to address problems lumping cases into the centroid of a zip code, we cannot rule out if incomplete geocoding biased our results. Incomplete geocoding is associated with risk factors for cancers, such as age and education level, and the lower match rates substantially reduce power in addition to potentially biasing results [34,35]. Also, we are unable to assess the impact of missing stage on the results. This analysis included a unique attempt to include unknown stage as late when potentially appropriate in an attempt to map clusters that are driven by the distribution of cancer risk and not possible regional reporting differences. Analysis was run including all unknowns in late stage as well as excluding all late stage cases (data not shown). Analysis including all unknowns highlighted similar areas for risk of late stage CRC, but the results moved even farther away from statistical significance. Analysis excluding all unknowns identified additional statistically significant areas of low risk of late stage—which are potentially areas of higher quality cancer reporting.

Other limitations were our choice of parameter settings for the SaTScan analysis. First, we used a circular scanning window versus an elliptic because our underlying data was not in the required projection. An elliptic scanning window shape provides greater power for clusters that are long and narrow [37] and may be appropriate for a state like Florida with a substantial coastline. Also, we evaluated clusters of up to 20% of the population at risk. This was selected based on the researchers need to use a standard cut-point for a series of analyses as well as an attempt to identify mostly homogenous, in terms of local risk, clusters of a size that
could be managed by local cancer control programs. The 20% cut-off avoided detection of large clusters covering the physical majority of the study area, but the 20% selection may not have been the scale for our data. Chen et al document that both location and statistical significance change with maximum cluster size and suggest a series of analyses to be conducted with 1% iterations of the scanning window size, followed by comparison of the results to determine the most likely clusters [31]. The approach may have resulted in an empirical selection of target areas and may have eliminated the need to overlay the Bernoulli model results with the Poisson model.

Public Health Significance:
CRC is a significant public health problem in the United States. But not only are the three recommended screening methods effective at reducing CRC mortality in all population groups, all methods have also consistently demonstrated cost-effectiveness compared to no screening [38]. While all population groups would likely see a mortality benefit from increased CRC screening, higher risk populations like Blacks and Hispanic Whites would benefit the most. Targeting high risk communities for enhance screening efforts should be incorporated into public health policy. However, due to limited public health resources, identifying local community targets may be more effective than county-wide efforts for large, highly dense, and diverse counties like Miami-Dade. An alternative approach for identifying at risk communities would be to select a geographic target based on clustering of demographic profile risk versus cancer outcomes.

Applied spatial analysis of cancer data is a complex undertaking. Protocol should to be established so that the “art” of selection does not outweigh the evidence, following the example already established by the CDC for responding to community identified cancer cluster concerns [38]. Such a protocol ideally would provide a set of best practices, such as the need to include multiple assessments, with different scales and parameter settings, and compare results. For instance, areas with high late stage risk but low incidence may be more amenable to screening interventions than areas with both high late stage risk and high incidence. But both are appropriate candidates for targeted interventions. However, areas of low cancer incidence but high risk of late stage should also be selected after evaluating the combination of cluster models because that is one profile of a community with low screening uptake. At a minimum, such a protocol could inform cancer registry personnel and academic researchers of the basic issues to address during such analysis.

Finally, in order to translate the results of spatial analysis research into reduction of cancer burden, risk communication messages must be dovetailed with publication and presentation of results. Further, public health partners must follow-up with targeted, appropriate, and successful intervention programs.

ACKNOWLEDGEMENTS
We acknowledge the support of Florida Bankhead-Coley Cancer Research Program grant 2BT02 for funding this research.

REFERENCES


