

The Impact of Data Quality on Spatial Analysis of Cancer Registry Data: The Example of Missing Stage at Diagnosis and Late-Stage Colorectal Cancer

Recinda Sherman, MPH CTR
University of Miami, Miller School of
Medicine, Public Health Sciences
1120 NW 14th St, Rm 912
Miami, FL 33136
1-305-243-4602
rsherman@med.miami.edu

Kevin Henry, PHD
Rutgers University, School of Public
Health, Dept of Epidemiology
683 Hoes Lane W, Rm 128
Piscataway, NJ 08854
1-732-235-4037
henryk1@sph.rutgers.edu

David Lee, PHD
University of Miami, Miller School of
Medicine, Public Health Sciences
1120 NE 14th St, Rm 1530
Miami, FL 33136
1-305-243-6980
dlee@med.miami.edu

ABSTRACT

Most disease surveillance systems currently geocode case data. This, coupled with advances in geographic analysis tools, has led to a rise in epidemiologic studies on distribution of disease that rely on analysis of secondary data, e.g. from cancer registries. However, while the data and tools are available for performing geospatial analyses, there are challenges with which methodologies to apply, how to interpret and translate results, and how results are impacted by data quality. The issue of data quality is the subject of this paper.

Mapping cancer rates highlights spatial patterns that can help elucidate environmental, clinical, or social causality pathways that drive differences in disease burden by geographic locations. Locating areas with high rates of cancer incidence or variations by stage at diagnoses can help prioritize cancer control efforts. Once the geographic patterns of cancer are mapped, the ideal action is to follow with effective public health interventions for the high risk communities. However, before using results of spatial research to inform public health response, it is important to consider whether the results are spurious due to methodological issues, such as data quality. Missing or incorrect data can distort research conclusions and result in ineffective public health policy.

Using colorectal cancer (CRC) as an example, the impact of missing stage at diagnosis on late-stage at diagnosis cluster detection is evaluated. The impact on cluster detection, area-based modeling, and distance from services analysis is described.

Categories and Subject Descriptors

J.3 [Life and Medical Sciences]: Health

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, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Keywords

Colorectal cancer, cluster detection, stage at diagnosis, data quality, area-based measures, screening disparities.

1. INTRODUCTION

For 2,000 years medicine was concerned with geographic variations of social, cultural, and physical environments, but, in the last century, germ theory, “the doctrine of specific etiology,” moved research questions about disease into a predominantly biomedical, individual-based perspective [1]. But more recently there is renewed interest, driven in large part due to advances in computing, in ecological research seeking to understand the importance of place, both physical and social, and resultant health [2].

The principle behind analyzing the geographic distribution of cancer is that both disease burden and disease risk are not randomly disbursed in physical space [3]. The suitability of spatial applications in sanitation and infectious disease epidemiology is intuitive, but the relevance of a spatial approach to chronic disease epidemiology is less clear. Cancers, for instance, are characterized by long latency between exposures and disease, protracted clinical course, and unclear etiologies which make the interpretation of geographic patterns problematic. However, cancer burden and outcomes also vary by geography, and these deviations have important implications for the development and implementation of prevention strategies, as well as expanding our understanding of cancer etiology [4].

Geocoding cancer cases down to census tract-level is now required for central cancer registries in the United States [5]. This enables all state registries to address cancer prevention and control efforts from a sub-county spatial perspective. While the analysis portion can be incorporated into registry operations, the difficulty is applying methodology that allows straightforward interpretation of results for translation into meaningful public health action. One area that is hampering interpretation is the impact of missing data. There is a growing field of literature on the influence of missing geocoded data, in 2012 a special issue of *Spatial and Spatio-temporal Epidemiology and Transactions in GIS* was dedicated to geocoding quality and influence on health research, but the impact of missing clinical variables is not often addressed from a spatial perspective.

Missing data in health surveillance registries reflects two problems: absence of clinical assessment or the failure of the surveillance system to capture the information [6]. While the North American cancer surveillance system is robust and follows national and international standards, systemic influences on missing data have been reported and can potentially bias results. But beyond percent missing, there is no standard method for presenting quality of cancer registry data (aside from the broad-spectrum registry level North American Association of Central Cancer Registries (NAACCR) Certification Standard) and little documentation of the impact of missing or miscoded data on research results [6].

There are two types of data missingness: random and systematic. Both types can negatively impact researcher's results. Random error occurs when there is no pattern driving the missing data—the probability that variables are missing is not dependent upon the data itself but on chance. Systematic error is when the probability that variables are missing or wrong is dependent upon the data itself; there is a pattern driving the poor data quality.

Random missingness is difficult to address from the data collection standpoint, but fairly straightforward to handle during analysis, i.e. interpolation or random assignment. However, if random error is not addressed appropriately through analytic methods, it can lead to reduced power of the analysis. Reduced power in the analysis can lead to inconclusive or missed associations. Systematic error is difficult to adjust for statistically, but, once the root source is identified, the data collection piece can be resolved. Systematic error not only reduces the power of analysis, but can also lead to erroneous conclusions. The solution to systemic error is to improve data collection, which is a task not easily leveraged by public health researchers who rely on surveillance datasets generally collected via state or federal mandate and not by the researchers themselves.

Systematic error is hypothetically problematic in geospatial epidemiology applications. For instance, Oliver and colleagues found that missing area-based socioeconomic data linked to cancer data was missing based on location as well as disease rates. They coined this issue as potential “cartographic confounding” [7], and it is also commonly referred to as geographic selection bias. One approach to improve the assignment of non-geocoded cases to geographic boundaries is to impute the geocoded data [8]. But geocoded location is not the only important missing variable. Stage of disease at diagnosis, which is a proxy for both screening uptake in a population and prognosis, and histologic grade, which is a proxy for how aggressive a cancer is, often are missing in population-based registries which, not only make the geographic picture incomplete, but may introduce bias [6]. Treatment information as well as information on stage and grade (which often depends upon medical surgery) is related to socioeconomic status [9], and missing stage and grade is higher among blacks than whites in central cancer registry data [10]. Further, research shows that cases reported to the central registry based on information from the death certificate alone, and are not confirmed by a treating or diagnosing facility, are not only patients that cross state lines to receive health care but also are patients more likely to live in areas of lower income, potentially representing patients not connected to the health care system [11]. In general, data quality

reported from affluent areas is higher than in more impoverished areas [10]. The impact of missing clinical variables on geospatial results is not currently understood.

Using CRC cancer in Florida for illustration, this paper investigates the impact of missing stage at diagnosis. CRC cancer is an ideal candidate for this demonstration. CRC is among the most common cancers in the United States and mortality is mitigated, in part, through screening. Because prognosis and quality of life are critically dependent upon the stage of cancer at diagnosis, routine screening can reduce mortality through early detection. And effective screening with colonoscopy can remove precancerous lesions, which makes most CRC potentially eradicable through secondary prevention. Regardless of proven effectiveness, screening rates for all recommended modalities are low. In 2010, 70% of white Floridians aged 50+ reported receiving colonoscopy or sigmoidoscopy in the past five years, however, among blacks and Hispanics rates were 64% and 62%, respectively. And 22% of white Floridians aged 50+ reported receiving a blood stool test in the last two years, 24% and 16% for blacks and Hispanics, respectively [12]. Although screening rates have shown recent improvement, historically blacks and Hispanics have had lower CRC screening rates compared to non-Hispanic whites [12]. Furthermore, CRC mortality rates among non-Hispanic whites have been decreasing over the last few decades, but a similar reduction has not yet been seen among blacks or Hispanics [13, 14].

Under the assumption we can reduce racial/ethnic disparities in early detection and resultant mortality of CRC through targeted screening of at-risk populations, we evaluated the impact of missing stage on the results of spatial applications for three cancer control questions. Where should we target a screening intervention? How should we tailor the intervention to reflect the demographics of the high risk communities? And, are the disparities being driven by unequal proximity to clinical care?

2. METHODS

2.1 Study Population and Data

This study obtained approval under expedited review from the Florida Department of Health Institutional Review Board and the Florida Cancer Registry under two protocols, H12005 and H12010.

This is a population based study on late-stage CRC cancer in Florida. According to the 2010 US Census, Florida comprised 6% of the US population. Florida's population is older (18% 65+ versus 13%), more Hispanic (23% versus 17%), and more foreign born (19% versus 13%) than the general US population. Only 16% of the population is non-Hispanic white, 19% is black, and 65% is Hispanic, the majority of which are Cuban. Over 70% of all US Cubans reside in Florida [15].

Population-based CRC incidence data were obtained from Florida's statewide cancer registry. All CRC cases (ICD-O C18.0-C20.9, C26.0) diagnosed from 2006-2010 in Florida with known age 50+ (to match currently screening guidelines), sex and geocoded address at diagnosis were included. Analysis was conducted on the first CRC primary reported and included only adenocarcinomas, the histologic type most responsive to screening (and 90% of all CRC). Both race and Hispanic origin,

two mutually exclusive variables in cancer surveillance, had to be known for the Hispanic white and non-Hispanic white analyses, but only race was required for the black analysis.

All geocoding was done at the central cancer registry level using a geocoder developed for cancer registry use [16]. Cases unmatched from automated geocoding are not resolved interactively. The FCDS current geocoding procedures do not include any attempts at manual geocoding or address correction. Only cases that were assigned a 2010 census tract were used.

Cases diagnosed *in situ* or localized stage using the Surveillance Epidemiology and End Results (SEER) Summary Staging system were classified as “early” and cases diagnosed at regional or distant stage were classified as “late.” Missing stage at diagnosis was handled three ways for comparison. Because unknown stage has poor prognosis (35% five-year survival rate compared to 90% for local, 70% for regional, and 13% for distant stage [17]), unstaged/unknown cases are often classified as late-stage. This method assumes the majority of unknown cases are due to absence of clinical assessment, cases diagnosed at end of life or among patients with contraindicative comorbidities. Indeed, the percentage unknown in patients over 85 years of age is three times the percentage in the youngest age group (9% and 3%, respectively). This method, unknown=late, was compared to excluding all cases with unknown stage, unknown=exclude, and allocation of the unknowns to early or late based on the distribution by race/ethnicity, sex, and age, unknown=allocate.

2.2 Cluster Detection

SaTScan version 9.1.2 beta, a cluster detection software developed in part by 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 at diagnosis for blacks, Hispanic whites, and non-Hispanic whites. Because we modeled risk of late-stage clustering using American Community Survey (ACS) data, the lowest unit of analysis used was the census-tract, the same level the area-based sociodemographic data used is released by ACS. Census tracts, with an average size of about 4,000 people, represent relatively homogenous communities with respect to socioeconomic characteristics and living conditions.

A Bernoulli model was used to identify communities with high rates of late stage (compared to early stage). A Poisson model was used to identify communities at high risk of CRC rates overall. The Poisson models were adjusted for age and sex, using the US Census 2010 population. The analysis adjusts for multiple testing, and we used 999 Monte Carlo iterations to calculate p-values for each cluster. Analysis was conducted using circular and elliptical shapes concurrently, to allow adjustment for multiple comparisons and to detect compact and linear clusters. Secondary clusters were also evaluated, and we adjusted for most likely clusters ($p=.05$, maximum iterations 15) to find geographically distinct, homogenous clusters.

Preliminary analysis was performed to identify statistically significant areas at high risk of late-stage CRC cancer. Because choice of scale for cluster detection analysis influences results [18], the preliminary analysis was conducted on a range of scales. These results were mapped using ESRI ArcMap 10 and visually compared. A single scale, for each stratified race/ethnic

group for each model (Poisson or Bernoulli) was manually selected based on statistical significance ($p<.05$) and consistency of physical overlap as the areas for further analysis. The initial analysis was conducted classifying all unknown stage as late stage. We replicated this analysis (at the scale selected for each method and race/ethnic stratification) using unknown=exclude and unknown=allocate methods for comparison.

2.3 Area-based measures and analysis

Tract-level area-based socioeconomic measures from the 2006-2010 ACS and county-level screening variables from the 2010 Florida adult Behavior Risk Factor Surveillance System (BRFSS) were modeled. ACS is an ongoing statistical survey administered by US Census used to describe sociodemographic characteristics of a community. BRFSS is the world’s largest ongoing, telephone based health survey.

Hierarchical, logistic regression was conducted to identify sociodemographic risk factors associated with increased risk of a case being diagnosed in a community at high risk of late-stage CRC at diagnosis. Predictor variables chosen were based on previous analysis, again which classified unknown stage as late. The *a priori* analysis evaluated a combination of individual-level (from cancer registry) and area-based measures (from ACS and BRFSS). Only area-based measures were predictive of whether a case lived in a high risk area, so regression analysis was conducted on tract-level variables from the 2006-2010 ACS (% non-white, % Hispanic, % minority, % foreign born, % not high school graduate, % no English spoken, % below poverty), and county-level variables from the 2010 Florida BRFSS (% ever received sigmoidoscopy or colonoscopy, % received fecal occult blood test (FOBT) in past 2 years) to predict case-level location, i.e., does the case live in or out of an area high risk of late-stage at diagnosis CRC. Analysis was conducted in SAS (proc glimmix). The proxy variables for racial/ethnic segregation are two mutually exclusive variables, % non-white and % Hispanic, and % minority, which is the aggregate of the two. Initial analysis included the two separate variables. If the direction for both variables was the same, e.g. predicting an increase of risk, and $p<.10$, the variable %minority was used in the model instead. Final models were selected by removing the non-significant predictors individually, based on size of P , and evaluating the reported fit statistics.

2.4 Distance analysis

Using the NAACCR Shortest Path Finder Tool, the travel distance between a patient’s residence and the reporting facility associated with that residence (generally the diagnosing or diagnosing and treating facility) was calculated. The Shortest Path Tool is a web-based application that calculates travel time and distance between two geographic locations (based on longitude/latitude pairs) using road networks [19]. Median and mean travel time by stage was calculated and compared by case-level race and by quartile of tract-based poverty from the ACS.

3. RESULTS

There were 3,779 cases analyzed for blacks, 4,989 for Hispanic whites, and 28,796 for non-Hispanic whites (Table 1). The difference in percent late stage between the unknown=exclude and the unknown=allocate methods was negligible. The percentage of late stage, by design, was higher for all categories for the unknown=late method.

	Black (n=3,779)			Hispanic White (n=4,989)			Non-Hispanic White (n=28,796)		
	Unknown			Unknown			Unknown		
	=Late	=Exclude	=Allocate	=Late	=Exclude	=Allocate	=Late	=Exclude	=Allocate
Total late cases	2,242	2,058	2,165	3,003	2,778	2,907	16,160	14,782	15,512
All combined	59%	57%	57%	60%	58%	58%	56%	54%	54%
Sex									
<i>female</i>	58%	56%	56%	60%	58%	58%	56%	54%	54%
<i>male</i>	61%	59%	59%	60%	58%	58%	56%	54%	54%
Age									
50-54	61%	60%	60%	62%	60%	60%	58%	57%	57%
55-59	64%	63%	63%	67%	65%	65%	60%	59%	59%
60-64	60%	58%	58%	62%	60%	60%	59%	58%	58%
65-69	57%	55%	55%	58%	57%	57%	56%	54%	54%
70-74	56%	54%	54%	61%	59%	60%	54%	52%	52%
75-79	59%	58%	58%	60%	58%	58%	54%	52%	51%
80-84	57%	53%	53%	57%	55%	55%	54%	51%	51%
85+	60%	55%	54%	58%	54%	54%	57%	53%	53%

Table 1. Distribution of late stage by method of handling unknowns, by race/ethnicity, sex, and age

	# clusters			# cases in cluster(s)			range of RR [^]			range of p		
	=Late	=Exclude	=Allocate	=Late	=Exclude	=Allocate	=Late	=Exclude	=Allocate	=Late	=Exclude	=Allocate
Black (B 30%)	1	1	1	581	58	423	1.2	1.5	1.2	0.05	0.1	0.4
Black (P 40%)	1	1	2	86	284	196	1.5	1.4	1.5- 4.2	0.03	0.1	<.01 - 0.05
HW (B na)	na	na	na	na	na	na	na	na	na	na	na	na
HW (P50%)	2	3	2	1,860	1,555	1,455	1.4 - 1.7	1.4 - 1.7	1.4 - 1.6	<.001 - <.01	<.001 - 0.05	<.001 - 0.01
nonHW (B 15%)	3	2	4	4,692	3,704	3,380	1.1 - 1.2	1.1 - 1.1	1.1 - 1.2	<.001 - 0.03	<.001	<.001 - 0.02
nonHW (P 25%)	12	11	12	9,823	8,048	8,321	1.2 - 5.2	1.2 - 8.8	1.2-5.1	<.001 - 0.03	<.001 - 0.03	<.001 - 0.05

All = allocate B = Bernoulli, P = Poisson, * maximum cluster size (scale), ^RR = Relative Risk, HW = Hispanic White

note: none of the Bernoulli analysis was statistically significant for comparison.

Table 2. Late-stage at diagnosis colorectal cancer clusters, Florida 2006-2010

3.1 Cluster Detection Results

The cluster results for the Bernoulli Method for blacks detected the same number of clusters for all methods, but the allocation method cluster was not statistically significant (Table 2). Comparing to unknown=late method, the cluster location was similar for the unknown=allocate method, but the unknown=exclude method, while statistically significant, was a much smaller cluster (Figure 1). For the Poisson Method, again the unknown=late and the unknown=allocate had similar locations, but the unknown=exclude identified two clusters—one with a slight overlay with the other two methods and one quite a bit farther north. For Hispanic whites, only the Poisson Method resulted in statistically significant clusters in the *a priori* analysis using unknown=late, so no analysis was conducted for the Bernoulli Method. The results were similar among the methods, clustering in the Miami-Dade metro region and in the Tampa Bay metro region. But the unknown=exclude split the Miami-Dade cluster into two separate, statistically significant clusters, and the unknown=allocated method also split the areas in to two cluster but only one of which was statistically significant (this cluster shown in Figure 1 but not Table 2).

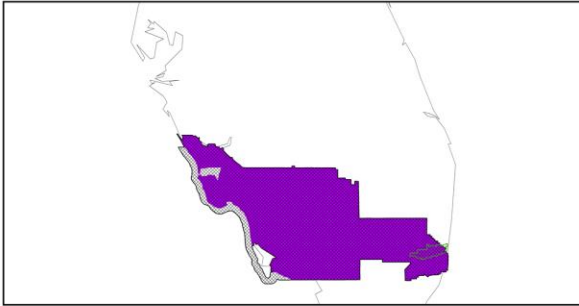
For the Bernoulli Method for non-Hispanic whites, the unknown=late and unknown=exclude were similar, but the unknown=exclude method identified one less statistically significant cluster. The unknown=allocate identified the same north/south cluster in the north of the state, but identified a number of smaller, non-overlapping clusters in the south. There was reasonable overlay among methods for the Poisson Method for non-Hispanic whites, but the methods all identified large areas of the state as high risk. Figure 1 shows the unknown=late clusters by magnitude of relative risk (1.0-1.5 and > 1.5). Using this method would be appropriate to prioritize cancer control efforts. The highest risk clusters in the south of the state can reasonable overlap among methods, but a high risk cluster to the east of the Tampa-Bay metro area was not overlaid by other methods.

3.2 Hierarchical Analysis Results

The results for blacks of the area-based risk modeling were similar among methods for the Bernoulli Method analysis. However, the increasing percent non-high school in a tract was predictive for a case being diagnosed in a high risk cluster only in the unknown=late method and increasing county-level

Potential Geographic Targets for Colorectal Cancer Screening Florida 1996-2010

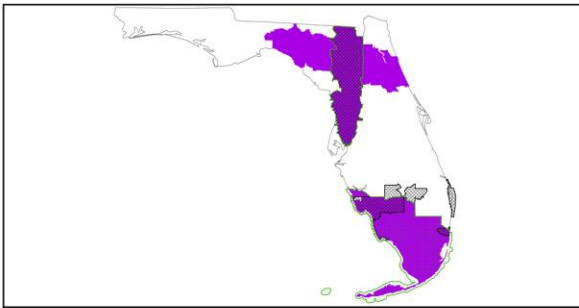
blacks, Bernoulli



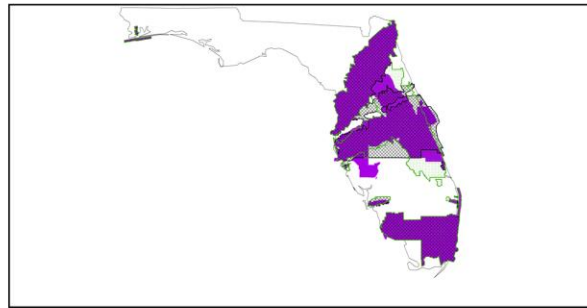
blacks, Poisson



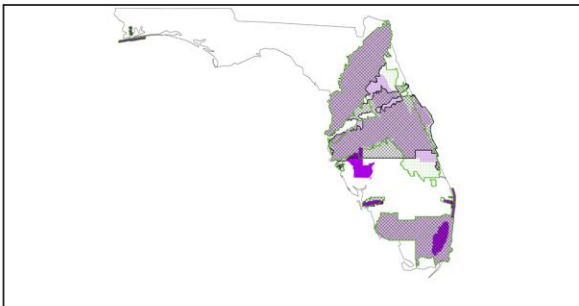
Non-Hispanic whites, Bernoulli



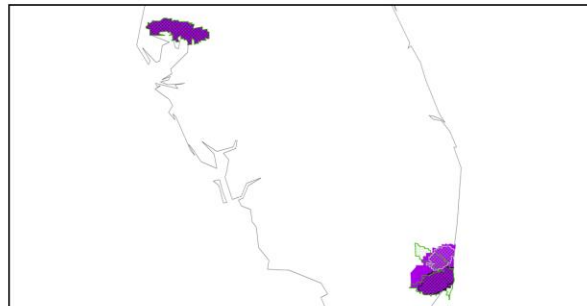
Non-Hispanic whites, Poisson






Non-Hispanic whites, Poisson with relative risk



Hispanic whites, Poisson



Legend:

-  Unknown stage excluded
-  Unknown stage allocated
-  Unknown stage classed as late

 Unknown stage allocated: $p = .20$



-  Unknown stage classed as late: $RR = 1.0 - 1.2$
-  Unknown stage classed as late: $RR > 1.3$

Figure 1. Comparison of tract-based cluster locations by method of handling unknowns, by race/ethnicity and Model type

Black: Bernoulli Method				Unknown=Late			Unknown=Exclude			Unknown=Allocate		
Tract Level				OR	p-value	CI	OR	p-value	CI	OR	p-value	CI
% non-white				1.0	0.00	1.0, 1.0	1.0	0.05	1.0, 1.0	1.0	<.001	1.0, 1.0
% hispanic				<i>not included</i>			<i>not included</i>			<i>not included</i>		
% minority				<i>not included</i>			<i>not included</i>			<i>not included</i>		
% foreign born				1.1	<.001	1.1, 1.1	1.1	<.001	1.1, 1.1	1.1	<.001	1.1, 1.1
% not hs grad				1.1	0.00	1.0, 1.1	<i>not included</i>			<i>not included</i>		
% no English spoken				0.8	<.001	0.8, 0.9	0.8	0.04	0.7, 1.0	0.9	0.01	0.9, 1.0
% below poverty				0.9	<.001	0.9, 1.0	0.9	0.01	0.9, 1.0	0.9	<.001	0.9, 1.0
County Level												
% ever received sigmoidoscopy/colonoscopy				<i>not included</i>			1.1	0.02	1.0, 1.3	<i>not included</i>		
% received fobt last 2 years				0.8	<.001	0.7, 0.9	0.8	0.01	0.7, 1.0	0.8	<.001	0.8, 0.9
Black: Poisson Method				Unknown=Late			Unknown=Exclude			Unknown=Allocate		
Tract Level				OR	p-value	CI	OR	p-value	CI	OR	p-value	CI
% non-white				<i>not included</i>			<i>not included</i>			<i>not included</i>		
% hispanic				<i>not included</i>			<i>not included</i>			<i>not included</i>		
% minority				1.1	<.001	1.0, 1.1	<i>not included</i>			1.0	<.001	1.0, 1.0
% foreign born				<i>not included</i>			0.9	<.001	0.8, 0.9	<i>not included</i>		
% not hs grad				<i>not included</i>			1.0	0.02	1.0, 1.0	<i>not included</i>		
% no English spoken				<i>not included</i>			<i>not included</i>			<i>not included</i>		
% below poverty				0.9	<.001	0.9, 1.0	<i>not included</i>			0.9	<.001	0.9, 1.0
County Level												
% ever received sigmoidoscopy/colonoscopy				0.9	<.001	0.8, 0.9	1.3	<.001	1.2, 1.3	0.9	<.001	0.9, 1.0
% received fobt last 2 years				0.8	0.03	0.6, 1.0	0.9	<.001	0.8, 0.9	0.8	0.03	0.7, 1.0

Table 3a. Heirarchical, area-based risk models for blacks, by method of handling unknowns and Model Type

Hispanic White: Poisson				Unknown=Late			Unknown=Exclude			Unknown=Allocate		
Tract Level				OR	p-value	CI	OR	p-value	CI	OR	p-value	CI
% non-white				<i>not included</i>			1.0	0.00	1.0, 1.0	<i>not included</i>		
% hispanic				1.0	<.001	0.9, 1.0	1.0	0.00	1.0, 1.0	1.0	0.00	1.0, 1.0
% minority				<i>not included</i>			<i>not included</i>			<i>not included</i>		
% foreign born				1.1	<.001	1.1, 1.1	1.1	<.001	1.1, 1.1	1.1	<.001	1.0, 1.1
% not hs grad				<i>not included</i>			<i>not included</i>			<i>not included</i>		
% no English spoken				1.1	0.00	1.0, 1.2	1.1	<.001	1.1, 1.2	1.2	<.001	1.1, 1.3
% below poverty				<i>not included</i>			<i>not included</i>			1.0	0.01	1.0, 1.1
County Level												
% ever received sigmoidoscopy/colonoscopy				0.7	<.001	0.7, 0.8	0.7	<.001	0.7, 0.8	0.6	<.001	0.5, 0.6
% received fobt last 2 years				1.4	<.001	1.3, 1.5	1.5	<.001	1.3, 1.6	1.9	<.001	1.7, 2.1

Table 3b. Heirarchical, area-based risk models for Hispanic whites by method of handling unknowns

Non-Hispanic White:Bernoulli Method				Unknown=Late			Unknown=Exclude			Unknown=Allocate		
Tract Level				OR	p-value	CI	OR	p-value	CI	OR	p-value	CI
% non-white				1.0	<.001	1.0, 1.0	1.0	<.001	1.0, 1.0	1.0	0.04	1.0, 1.0
% hispanic				1.0	0.03	1.0, 1.0	1.0	0.00	1.0, 1.0	not included		
% minority				not included			not included			not included		
% foreign born				1.1	<.001	1.1, 1.1	1.1	<.001	1.1, 1.1	1.0	<.001	1.0, 1.1
% not hs grad				not included			not included			not included		
% no English spoken				0.9	0.00	0.8, 1.0	0.9	0.00	0.8, 1.0	0.9	0.00	0.9, 1.0
% below poverty				1.0	0.00	1.0, 1.1	not included			1.0	0.05	1.0, 1.0
County Level												
% ever received sigmoidoscopy/colonoscopy				0.8	<.001	0.8, 0.8	0.8	<.001	0.8, 0.8	1.0	0.0	0.9, 1.0
% received fobt last 2 years				not included			not included			1.1	<.001	1.1, 1.1
Non-Hispanic White: Poisson Method				Unknown=Late			Unknown=Exclude			Unknown=Allocate		
Tract Level				OR	p-value	CI	OR	p-value	CI	OR	p-value	CI
% non-white				not included			not included			not included		
% hispanic				not included			not included			not included		
% minority				1.1	<.001	1.1, 1.1	1.1	<.001	1.1, 1.1	1.1	<.001	1.0, 1.1
% foreign born				1.1	<.001	1.0, 1.1	not included			1.0	0.02	1.0, 1.1
% not hs grad				not included			1.0	0.01	1.0, 1.0	not included		
% no English spoken				0.8	<.001	0.8, 0.9	0.9	0.04	0.7, 0.8	0.8	<.001	0.8, 0.9
% below poverty				1.0	0.00	1.0, 1.0	1.0	0.03	1.0, 1.0	1.0	0.0	1.0, 1.0
County Level												
% ever received sigmoidoscopy/colonoscopy				0.8	<.001	0.8, 0.8	0.8	<.001	0.8, 0.9	0.8	<.001	0.8, 0.8
% received fobt last 2 years				not included			1.0	0.03	0.9, 1.0	1.0	0.01	0.9, 0.9

Table 3c. Heirarchical, area-based risk models for non-Hispanic whites, by method of handling unknowns and Model Type

		Unknown=Late				Unknown=Exclude				Unknown=Allocate			
		Early		Late		Early		Late		Early		Late	
number of cases		12,556		16,646		12,556		15,310		13,295		15,907	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean
All combined		12.0	17.1	12.0	17.2	12.0	17.1	12.0	17.3	12.0	17.0	12.0	17.2
Race													
	Black	10.8	14.8	10.8	14.5	10.8	14.8	10.8	14.6	10.8	14.7	10.8	14.5
	Hispanic white	12.6	16.4	12.6	15.9	12.6	16.4	12.6	15.9	12.6	16.4	12.6	15.9
	Non-Hispanic white	12.0	17.5	12.0	17.8	12.0	17.5	12.0	17.9	12.0	17.3	12.0	17.9
Poverty													
	Lowest Poverty	13.2	16.7	12.6	17.0	13.2	16.7	12.6	17.2	13.2	16.5	12.6	17.1
	Medium Low	12.0	17.4	12.0	17.3	12.0	17.4	12.0	17.3	12.0	17.3	12.0	17.3
	Medium High	12.0	17.9	12.0	17.9	12.0	17.9	12.0	17.9	12.0	17.8	12.0	17.9
	Highest Poverty	10.2	16.1	10.2	16.3	10.2	16.1	10.2	16.5	10.2	16.0	10.2	16.4
Race and Poverty													
Black	Lowest Poverty	14.4	19.0	15.0	19.1	14.4	19.0	14.4	19.1	15.0	19.1	14.4	19.0
	Medium Low	13.2	17.2	12.6	14.6	13.2	17.2	12.0	14.8	13.2	16.9	12.6	14.8
	Medium High	12.6	16.6	12.0	15.6	12.6	16.6	12.0	15.7	12.6	16.5	12.0	15.6
	Highest Poverty	9.0	12.2	8.4	13.0	9.0	12.2	8.4	13.0	9.0	12.3	8.7	13.0
Hispanic white	Lowest Poverty	15.0	18.3	14.4	19.2	15.0	18.3	14.4	19.4	15.0	18.2	14.4	19.3
	Medium Low	15.6	20.5	13.8	17.9	15.6	20.5	14.4	18.1	15.0	20.2	14.4	18.0
	Medium High	13.2	16.1	12.6	15.4	13.2	16.1	12.6	15.0	13.2	16.1	12.6	15.3
	Highest Poverty	9.6	12.0	9.6	13.1	9.6	12.0	9.6	13.1	10.2	12.3	9.6	13.0
non-Hispanic white	Lowest Poverty	12.6	16.5	12.0	16.6	12.6	16.5	12.0	16.8	12.6	16.3	12.0	16.8
	Medium Low	12.0	17.0	11.4	17.4	12.0	17.0	12.0	17.3	12.0	17.0	11.4	17.4
	Medium High	12.0	18.4	12.0	18.7	12.0	18.4	12.0	18.9	12.0	18.3	12.0	18.8
	Highest Poverty	10.8	18.6	11.1	19.0	10.8	18.6	11.4	19.3	10.8	18.4	11.4	19.2

Figure 4. Travel distance (in minutes) from patient residence to facility

sigmoidoscopy/colonoscopy was a predictive risk only for the unknown=exclude method. The results for the Poisson Method were the same between the unknown=late and the unknown=allocate method, increasing minority tracts were predictive for a CRC case being diagnosed in a high risk area, and increasing poverty and county-level screening being predictive for a CRC case not being diagnosed in a high risk area. The impact of increasing poverty was not statistically significant for the unknown=exclude method, but increasing percent of foreign born and percent of non-high school graduates were. While there were some differences on the statistical significance of some of the variables, all methods resulted in associations in the same direction with similar magnitude, 1/2-1% increase in risk for every 1% increase in area-based measure with one exception: sigmoidoscopy/colonoscopy was predictive of a case living in a high risk cluster for unknown=exclude only in the Poisson Model.

Only the Poisson Model was used for the Hispanic white analysis, because the original analysis of the Bernoulli Models did not identify any statistically significant clusters. Increasing numbers of non-whites in a community was predictive for a case being diagnosed in a high risk cluster only for unknown=exclude and increasing numbers of individuals living below the poverty line was predictive for a case being diagnosed in a high risk cluster only for unknown=allocate. None of the methods resulted in contradictory directions in relative risk.

As with the other race/ethnic stratified analysis, there were some minor differences in predictors due to statistical significance for non-Hispanic whites among the methods. However, the unknown=allocate method resulted in a model with increasing tract-level numbers of non-whites as predictive of a case being diagnosed in a high risk cluster, the opposite of the other two methods

3.3 Distance Analysis Results

A truncated dataset, 29,202 cases, was used for the distance analysis. Cases were deleted from analysis due lack of facility geocoding due to incomplete facility information (n=8,366), ungeocodable facilities, or cases were homeless (n=28), or invalid time calculated by shortest path tool (n=12).

Median and mean travel times by stage are listed in Table 4. Overall, the mean travel time from a patient's residence to reporting facility is under 20 minutes and the median is under 15 regardless of stage, regardless of method used and the differences among methods were nominal. Regardless of method, the travel time was shortest for blacks for both stages. Non-Hispanic whites had the longest mean travel for both stages and Hispanic whites had the longest median travel time. The median travel times decreased with increasing poverty quartile for all stages, race/ethnicities, and unknown stage handling method. Only the non-Hispanic whites highest poverty quartile had longer travel times for late-stage cases, this was consistent for all methods of handling unknown stage. Only blacks lowest poverty quartile had longer travel times for late-stage cases, and this was only for the unknown=late method.

4. CONCLUSIONS

There is currently no standard for how to handle cases with unknown stage at diagnosis in population-based cancer surveillance studies. Presently, most researchers conducting geographic analysis are excluding unknown stage cases from their study. In this study, we examined how results for geospatial analysis may vary based on three differed approaches for handling unknown stage data: unknown=late, unknown=allocate (based on age, sex, race, ethnic distribution), and unknown=exclude. Unknown=late and unknown=allocate produced similar results.

Cases with unknown stage have poor prognosis, so assigning all unknowns to late or allocating based on demographics are both empirical approaches to handling missing stage data. Both approaches can also help limit power loss due to sample size, but they are both subject to an unknown level of misclassification bias. Unknown=exclude produced less consistent results compared to the two other methods. Removing all case of unknown stage not only reduced power due to reduction in cases but can also cause selection bias.

Random misclassification bias in population based epidemiology pushes results towards the null, and results in underestimates of true associations. Non-random misclassification can result in either an over or under-association. Unknown=allocate most likely results in random misclassification for the unknown stage cases. Unknown=late is likely correctly classifying over 50% of the cases, because over 50% of cases overall are late stage, but the remainder may be systematically misclassified. But the selection bias, from unknown=exclude, is a more grievous bias and leads to an overestimation of effects.

Compared to unknown=late, the unknown=exclude method resulted in additional variables being associated with a case being diagnosed in a high risk area for all of the hierarchical modeling except for the non-Hispanic whites Bernoulli analysis, higher relative risks for clusters for Bernoulli blacks and Poisson non-Hispanic whites.

But the unknown=allocate method failed to identify a significant cluster for blacks using the Bernoulli Model and resulted in a contradictory increase of relative risk in for endoscopy screening in blacks. In this example, potentially important clusters of risk of late-stage may have been missed using the unknown-exclude method for both Hispanic whites and non-Hispanic whites. And this method also resulted in a contradictory direction of relative risk for density of non-whites in the non-Hispanic Bernoulli analysis.

Overall, there was little difference in travel time among methods. But where there were deviations, the unknown=late and the unknown=allocate methods had more similar travel times.

Re-abstraction studies or comparisons with clinical datasets to identify the true distribution of unknown stage would be helpful to inform a more precise allocation method. In the absence of a more exact tool, it may be prudent to run late-stage analysis

using both the unknown=late and unknown=allocate method. Future work should also consider using multiple imputation approaches for assigning missing stage data. Results that are consistent among multiple methods can be interpreted with more confidence.

5. ACKNOWLEDGMENTS

We acknowledge the support of Florida Bankhead-Coley Cancer Research Program grant 2BT02 for funding this research. We also acknowledge the Florida Cancer Data System (FCDS). The Florida cancer incidence data used in this report were collected by the FCDS under contract with the Florida Department of Health (FDOH). The views expressed herein are solely those of the authors and do not necessarily reflect those of the FCDS or FDOH.

6. REFERENCES

- [1] Meade MS and Earickson RJ. Medical Geography, 2nd Edition, New York, NY 2000.
- [2] MacIntyre S, Ellaway A. Ecological approaches: rediscovering the role of the physical and social environment in Berkman LF, and Kawachi I. *Social Epidemiology*, New York, NY, 2000.
- [3] Graves ,BA. Integrative literature review: A review of literature related to geographical information systems, healthcare access, and health outcomes. *Perspectives in Health Information Management*. 2008;5(11): published online <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2500173/>.
- [4] Seidman C. An introduction to prostate cancer and geographic information systems. *American Journal of Preventive Medicine* 2006;30(2S):S1.
- [5] NAACCR Data Standards and Data Dictionary (Volume II), March 2013: published online <http://www.naaccr.org/StandardsandRegistryOperations/VolumeII.aspx#>.
- [6] Klassen, AC, Curriero F, Kulldorff M, Alberg, AJ, Platz, EA, Neloms ST. Missing Stage and Grade in Maryland Prostate Cancer Surveillance Data, 1992-1997 *American Journal Preventive Medicine* 2006;30(2S):S77-S87.
- [7] Oliver MN, Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics*. 2005;4(29) published on-line <http://www.ij-healthgeographics.com/content/4/1/29> .
- [8] Henry KA and Boscoe FP. Estimating the accuracy of geographic imputation. *International Journal of Health Geography*. 2008;7(3) published online <http://www.ij-healthgeographics.com/content/7/1/3>.
- [9] Adams J, White M, Forma D. Are There Socioeconomic Gradients in the Quality of Data Held by UK Cancer Registries? *Journal Epidemiology Community Health* 2004;58:1052-3.
- [10] Boscoe FP, Sherman C. On Socioeconomic Gradients in Cancer Registry Data Quality (letter) *Journal Epidemiology Community Health* 2006;60:551.
- [11] Sherman, RL. 2006. Relationship of Community Level Socioeconomic Status on Stage at Diagnosis of Colorectal Cancer (a master's thesis), OHSU, Portland, Oregon 2006.
- [12] Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2013.
- [13] Naishadham D, Lansdorp-Vogelaar I, Seigel R, Cokkinides V, Jemal A. State disparities in colorectal cancer mortality patterns in the United States. *Cancer Epidemiology & Biomarkers*. 2012;20(7):1296-1302.
- [14] Robbins AS, Siegel RL, Jemal A. Racial disparities in stage-specific colorectal cancer mortality rates from 1985-2008. *Journal of Clinical Oncology*. 2011;30:401-405.
- [15] Brown A, Patten E. Hispanics of Cuban Origin in the United States, 2011 Pew Research Center, Washington, DC 2013.
- [16] Goldberg, D. W., Kohler, B., Kosary, C. (2013). The Texas A&M, NAACCR, NCI Geocoding Service. Available online at <http://geo.naaccr.org>.
- [17] Howlader N, Noone AM, Krapcho M, et al., eds. *SEER Cancer Statistics Review, 1975-2008*. Bethesda, MD: National Cancer Institute; 2011 online available http://seer.cancer.gov/csr/1975_2008.
- [18] Chen J, Roth RE, Naito AT, Lengerich EJ, MacEachren AM. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of US cervical cancer mortality. *International Journal of Health Geographics*. 2008;29(4): published online <http://www.ij-healthgeographics.com/content/4/1/29>.
- [19] Goldberg, D. W. (2013). The NAACCR/Komen/USC Shortest Path. Available online at <http://geo.naaccr.org>.