SOCIO-SPATIAL INEQUALITIES IN LATE-STAGE CANCER DIAGNOSIS IN ILLINOIS: SPATIOTEMPORAL TRENDS AND METHODOLOGICAL CHALLENGES

BY

LAN LUO

DISSERTATION

Submitted in partial fulfillment of requirements
for the degree of Doctor of Philosophy in Geography
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Sara McLafferty, Chair & Director of Research
Associate Professor Shaowen Wang
Associate Professor Marilyn Ruiz
Assistant Professor Diana Grigsby

ABSTRACT

This dissertation examines the effects of social and spatial inequalities on late-stage diagnosis of colorectal and breast cancer, and it addresses several methodological challenges surrounding the use of ZIP codes as a study unit in analyzing late-stage cancer at diagnosis. Given that my dissertation follows the 'three-paper' format, the abstract section is divided into three parts to describe each paper respectively.

The first paper entitled "*Spatial Distribution of Late-Stage Colorectal Cancer in Illinois from 1988 to 2002: Associations with Social-Spatial Covariates*", examines spatial patterns of late-stage colorectal cancer diagnosis over time in Illinois during a period of increasing screening, and it analyzes the varying associations between social, demographic and spatial risk factors and late-stage colorectal cancer diagnosis within the same period. The Bernoulli-based spatial scan statistic was used to detect clusters of late-to-early stage cancer ratios at the ZIP code level in Illinois during two periods: 1988 to 1992, and 1998 to 2002. Then the whole state was divided into three study region: Chicago city, Chicago suburbs, and other areas. For each region in each time period, hierarchical logistic regression models were estimated to assess the associations between demographic, social and spatial factors and late-stage colorectal cancer risk. ZIP code level risk factors include three indicators of socio-economic status and the shortest travel time to the nearest colonoscopy facility and individual-level factors including age, race, and gender. The socio-economic indicators were created using factor analysis.

The results show some changes over time in the spatial distribution of late-stage colorectal cancer and the impacts of risk factors at the ZIP code and individual levels. Specifically, results of the Bernoulli-based spatial scan statistic find statistically significant clusters of late-stage colorectal cancer in the Chicago metropolitan area and rural region in southern Illinois in the period of 1988 to 1992. In the later time period, the cluster outcomes were no longer statistically significant. The change indicates that late-stage risk of colorectal cancer has become more evenly distributed in Illinois over time. In terms of the hierarchical logistic regression results, both individual-level demographic factors and zip-code level covariates present variously important impacts on the risk of the late-stage colorectal cancer diagnosis in different study regions in the two time periods. The risk of late-stage diagnosis is higher among younger colorectal cancer patients. Gender has contradictory impacts on risk in Chicago city and its suburb. The shortest travel time to the nearest cancer screening providers is positively associated with late-stage diagnosis risk outside the Chicago region, suggesting that spatial access to screening services may be an important barrier to early detection in rural areas of the state. One socio-economic status indicator, Minority Disparities, demonstrated a significantly positive relationship with late-stage diagnosis risk outside the Chicago region. Similar to the effects of gender, Factor 3 (Cultural-Language Barriers) also

had contradictory effects in Chicago city and suburbs. Overall, the results showed no clear trends over time in the effects of various factors on late-stage risk, and few strong and statistically significant results. The inconsistent findings suggest the need for more detailed and localized information.

The second paper is titled "*Analyzing Spatial Aggregation Error in Statistical Models of Late-Stage Cancer Risk: A Monte Carlo Simulation Approach*". This paper examines the effect of spatial aggregation error on statistical estimates of the association between spatial access to health care and late-stage cancer. Monte Carlo simulation was used to disaggregate breast cancer cases for two Illinois counties from ZIP codes to census blocks in proportion to the age-race composition of the block population. After the disaggregation, a hierarchical logistic model was estimated examining the relationship between late-stage breast cancer and risk factors including travel distance to mammography, at both the ZIP code and census block levels. Model coefficients were compared between the two levels to assess the impact of spatial aggregation error.

Spatial aggregation error is found to influence the coefficients of regression-type models at the ZIP code level, and this impact is highly dependent on the study area. In one study area (Kane County), block-level coefficients were very similar to those estimated on the basis of ZIP code data; whereas in the other study area (Peoria County), the two sets of coefficients differed substantially raising the possibility of drawing inaccurate inferences about the association between distance to mammography and late-stage cancer risk. The paper reveals that spatial aggregation error can significantly affect the coefficient values in statistical models of the association between cancer outcomes and spatial and non-spatial variables and thus affect inferences drawn from these models. Relying on data at the ZIP code level may lead to inaccurate findings on health risk factors, and the effects are likely to vary from one study area to another.

The third paper, titled "*The Impact of Spatial Aggregation Error on Spatial Scan Analysis: A Case Study of Colorectal Cancer*," aims to examine the effect of spatial aggregation error on results of the spatial scan statistic by geographically and statistically comparing results at the ZIP code level and three reference (census tract, census block group and census block) levels. Data on colorectal cancer cases in Cook County, IL for a 5-year interval (1998-2002) were used. The Monte Carlo simulation approach from the second paper was applied to disaggregate the cancer data from the ZIP code level to each reference level. The Bernoulli-based spatial scan statistic was implemented in SaTScan to detect primary clusters based on cancer data at the four levels. An interactive procedure involving SAS and Java programming, was designed to automatically run SaTScan hundreds of times. Characteristics of clusters at each reference level were compared to those of the ZIP code level cluster to observe differences related to spatial aggregation.

The comparison reveals that the ZIP code level spatial scan statistic can generate reliable clusters at the global level in areas with a large number of cases. Nonetheless, the ZIP code analysis sometimes fails to detect clusters in areas with a lower density of cases. Spatial aggregation error is minimized in areas with sizeable numbers of cases. In the absence of cancer data at a lower level, the ZIP code level data can be used effectively to implement the spatial scan statistic and identify large and dominant clusters. However, smaller clusters located in areas with a relatively low density of cases may be missed. Given that this study focused on a highly urbanized and populated area, future research should assess the influence of spatial aggregation error on spatial scan analysis in suburban and rural regions.

Heading: To my mother and father

ACKNOWLEDGEMENTS

This dissertation work has been supported by the National Cancer Institute (NCI), National Institutes of Health (NIH), under Grant No. 1-R21-CA114501-01, in which Drs. Sara McLafferty and Fahui Wang are the principal investigators. Points of view or opinions in this dissertation are mine and do not necessarily represent the official position or policies of NCI.

I am heartily thankful to my adviser, Dr. Sara McLafferty, whose encouragement, supervision, and support from the preliminary to the concluding level enabled me to grow myself fast and strongly through my PhD graduate study at University of Illinois at Urbana-Champaign.

I would like also to thank my advising committee members, Dr. Shaowen Wang, Dr. Marilyn Ruiz and Dr. Diana Grigsby, for their kind advisement in the completion of this dissertation.

TABLE OF CONTENTS

Page

INTRODUCTION

My dissertation consists of three papers to examine the effects of social and spatial inequalities on late-stage diagnosis of colorectal cancer (CRC) and breast cancer, and to address several methodological challenges surrounding the use of zip codes as a study unit in analyzing late-stage cancer. Late-stage cancer is the term used to describe cancer tumors that, when first diagnosed, are large in size and /or have already spread beyond the initial site to nearby or distant tissues, organs or lymph nodes. Survival rates of cancer patients are highly dependent on cancer stage: patients diagnosed with early-stage cancer have much higher survival rates and much healthier prognostic outcomes, compared to those diagnosed with late-stage cancer. Thus, examining the risk determinants of late-stage cancer and addressing the methodological challenges researchers face in spatially analyzing late-stage cancer are crucial to identifying the main spatial and social barriers that hinder early detection.

The studies in my dissertation are tied to two important themes in recent research on medical and health geography, that is, the examination of geographic inequalities in cancer diagnosis and survival (Haynes et al., 2008, Palmer and Schneider, 2005, Whynes et al., 2003, Woods et al, 2006), and the utilization of spatial statistics and GIS in health geographic research with an emphasis on cancer studies (Jacquez and Greiling, 2003, Meliker et al., 2009, Pollack et al., 2006, Short et al., 2002). In terms of the first theme, it is well established that cancer incidence and mortality rates vary from place to place. These geographical inequalities are an important topic in the field of health/medical geography, given that understanding such inequalities is the 'first step' in establishing 'target areas' where cancer issues have significant spatial disparity. Specifically, these target areas are usually correlated with local issues in the social and spatial domains. Finding these areas can help researchers and policy makers to address these local problems and design corresponding plans for alleviating the spatial disparity. In addition to targeting areas for intervention, researchers have analyzed the impacts of social and spatial variables on geographic inequalities in late-stage cancer risk (Haynes et al., 2008, Wang et. al., 2008). The results of these studies are sometimes contradictory, and the influence of spatial access to cancer screening services on late-stage risk has not been clearly established. It is likely that these influences will vary not only geographically, but also over time with the expansion of cancer screening services and improvements in cancer awareness, education and outreach. Understanding how social and spatial variables affect geographic inequalities in late-stage cancer in diverse study areas and time periods remains an important issue in health geographic research.

For the second theme, analyzing the geographical inequalities in cancer incidence and mortality rates involves using spatial statistics and GIS to visualize and identify the significant spatial clusters and

detect the causes that can include socio-economic factors, healthcare accessibility, demographics, and other risk factors (Rushton, 2004, Thomas and Carlin, 2003). Spatial statistics and GIS provide a systematic and quantitative way to detect the spatial variation in cancer and to model the effects of risk factors. However, these kinds of spatial analysis also pose a series of challenges that stem from the geographic characteristics of the data analyzed. Specifically, most researchers have to use cancer data that is spatially aggregated to areas like ZIP codes or counties so that the privacy and confidentiality of individual health records will be protected. The extent and implications of this spatial aggregation error are topics of increasing interest in health/medical geography research (Shi, 2009).

The three papers in this dissertation build on these themes. The results and subsequent conclusions clearly demonstrate the social-spatial determinants of late-stage CRC diagnosis in Illinois, and provide evidence that can help guide public administrators to generate appropriate cancer prevention strategies. Additionally, these analytical results also reveal that the reliability of the ZIP code as a study unit is case-sensitive, and that spatial aggregation error differs from one geographic context to another. This is an important finding for researchers who are analyzing geographic variation in cancer based on aggregated data, and the conclusions can potentially inform policies that govern release of cancer data at various geographic levels. Each paper contains detailed information about the relevant literature, variable selection and methodologies, and for that reason I do not devote much attention to these topics in the introduction section.

The first paper entitled "*Spatial Distribution of Late-Stage Colorectal Cancer in Illinois from 1988 to 2002: Associations with Social and Spatial Covariates*" examines the spatial patterns of late-stage CRC diagnosis within urban-rural settings of Illinois in two 5-year intervals (from 1988 to 1992 and from 1998 to 2002). The goal is to analyze how spatial inequalities in late-stage risk have changed over time and to evaluate the influence of demographic-social-spatial risk factors. The study period witnessed the rapid expansion of screening services for CRC and increases in education and awareness about the importance of early detection.

To examine the spatial distribution of late-stage CRC, the Bernoulli-based spatial scan statistic is used to detect clusters with significantly high late-to-early CRC ratios in the whole state, the Chicago metropolitan area and non-Chicago metropolitan area. The results reveal the existence of statistically significant clusters in the former period and the disappearance of these clusters in the latter period, indicating that late-stage CRC diagnoses have become more evenly distributed across the state. Afterwards, covariates hypothesized to influence late-stage risk are collected and preprocessed. The covariates represent three main medical geographic dimensions: 1) demographic factors consisting of

race, age and gender, 2) socio-economic indicators (Socio-Economic Disadvantages, Minority Disparities, and Cultural-Language Barriers), and 3) spatial access to health care, measured by the shortest travel time to hospitals with colonoscopy services and accessibility to primary care physicians. Hierarchical logistic regression is applied to investigate the different impacts of these demographic, social and spatial factors on late-stage CRC diagnoses during the two time periods in Illinois.

The analytical results from the hierarchical logistic regression clearly uncover risk factors of late-stage CRC at diagnosis. The negative relationship of age and late-stage CRC at diagnosis reveals the high risk of late-stage diagnosis in young and middle-aged CRC patients, different from the well-established positive association between age and all-staged CRC cases in previous studies (Brawarsky et al., 2003, CDC 1999, CDC 2001, Cokkinides et al., 2003, Cooper et al., 1995, Nelson et al., 1999, Palmer and Schneider, 2005, Schneider, 2009, Wingo et la., 1998). The contrasting impacts of gender in neighboring study areas (Chicago city and Chicago suburbs) suggest no consistent gender discrepancy in screening and early detection. The strongly positive relationship between the SES factor, Minority Disparities, and late-stage CRC in the non-Chicago metropolitan area suggests that the black population in that setting is not receiving timely CRC screenings and early diagnoses as compared to other racial groups. The contradictory influences of another SES factor (Cultural-Language Barriers) in Chicago city and the Chicago suburbs, shows that the immigrant effect is also not consistent across the study area and may be highly dependent on immigrant demographic characteristics, such as marital status, socio-cultural characteristics and length of residence in the U.S. Spatial access to colonoscopy services is only significantly associated with late-stage risk for patients living in areas outside the Chicago metropolitan region. Shortest travel time exhibits a significantly positive association with late-stage CRC, indicating that accessibility to specialized CRC facilities is lower and more variable in low population-density areas than in more densely settled regions.

These associations between demographic-social-spatial variables and late-stage diagnosis of CRC can provide suggestions for both health care policies and cancer research. Specifically, policies can be designed for targeting CRC patients younger than 50 years old to ensure them receiving cancer prevention strategies and diagnosis in time. Cancer screening and prevention policies need to focus on vulnerable groups, such as black people living in impoverished and racially segregated regions and residents in remote rural areas. The contradictory findings for both gender and one SES factor (Cultural-Language Barriers) indicate that more localized risk variables, such as individual health insurance status, personal behaviors, and biological characteristics may be important in affecting late-stage diagnosis. These detailed personal characteristics are rarely obtainable from health datasets, because of the concern to protect privacy and personal identities. Providing researchers access to more localized health data without

violating privacy restrictions is an important policy challenge. Furthermore, the non-significant results from the interactions between ZIP code level variables and individual-level demographic predictors suggests that ZIP codes may be too large for studying rare events like CRC and statistical analysis at this level may be strongly affected by spatial aggregation error. The second and third chapters in this dissertation address the topic of spatial aggregation error.

In the second paper "*Analyzing Spatial Aggregation Error in Statistical Models of Late-Stage Cancer Risk: a Monte Carlo Simulation Approach*", the influence of spatial aggregation error on ZIP code level statistical analysis is estimated. The study is motivated by the fact that health studies using predefined study units are often confronted with the analytical biases caused by the spatial disparity between the research interests and the predefined geographic level of health data. Spatial aggregation error happens when utilizing a large area or a single point to represent spatially distributed individuals. The aggregation of individuals to zones such as ZIP codes or counties creates error in measurement of variables like distance to the nearest hospital facility. A sizeable literature has identified types of spatial aggregation error and their specific impacts on spatial analysis (Bonner et al., 2003, Fortney et al., 2000, Gregorio et al., 2005, Hewko et al., 2002, Hillsman and Rhoda, 1978, Jacquez and Waller, 1999, Krieger et al., 2001, McElroy et al., 2003). However, the impact of spatial aggregation error on coefficient estimates of statistical analysis at the ZIP code level has been rarely studied. Thus the major contribution of this paper is to analyze the impact of spatial aggregation error on parameters estimated in ZIP code level statistical models.

To evaluate the impact of the spatial aggregation error, this study designs a Monte Carlo simulation method to proportionally disaggregate cancer cases from the ZIP code level to the census block (reference) scale, with an age-race demographic link to connect the two levels. The statistical coefficients at the ZIP code level are compared with coefficients from the same statistical analysis based on the same cancer data at the block level. If the coefficients at the two levels display a significant disparity, then spatial aggregation error may have a recognizable influence on the ZIP code level statistical analysis. Two study areas, Kane and Peoria, provide the opportunity to determine if the effect of spatial aggregation error differs between study regions.

The Monte-Carlo simulation approach provides a new method of generating 'valid' datasets at a more localized geographic scale when data about individual case locations are inaccessible. The approach involves generating 1,000 sets of cancer cases at the block level to maximally encompass possible spatial distribution patterns. Hierarchical logistic regression is selected to generate parameter estimates for a statistical model predicting late-stage breast cancer risk as a function of individual-level demographic

predictors and ZIP code level shortest travel distance to the nearest mammography facility. This study also evaluates the computational capacity of the widely-used statistical software package, SAS, because the macro-level syntaxes built in SAS automatically compute coefficients for the hierarchical logistic regression model 1,000 times. All the methods used in the paper provide a good model of interdisciplinary application of GIS visualization, intensive computing techniques, and applied statistics in cancer research.

The results show a dramatic difference in spatial aggregation error between the Kane and Peoria study areas, indicating that the influence of spatial aggregation error is highly case-sensitive. The outcomes suggest characteristics associated with high spatial aggregation error. These characteristics include large-sized study units, the uneven distribution of healthcare providers, highly uneven and segregated residential distributions, spatial autocorrelation of age-and racially-categorized populations, and a bifurcated rural-urban pattern centered on a densely population city. The findings from the paper also emphasize the need to develop methods and procedures for minimizing the impact of spatial aggregation error. The analytical results also infer that the ZIP code may not be a reliable study unit for cancer research in some cases. While recognizing the need to protect individual confidentiality, these results demonstrate the benefits to be gained by releasing cancer datasets at a smaller area level for research purposes.

The third paper, "*The Impact of Spatial Aggregation Error on the Spatial Scan Analysis: A Case Study of Colorectal Cancer*", evaluates the effect of spatial aggregation error on the spatial scan statistic at the ZIP code level. Many studies have utilized spatial scan statistics to detect spatial clusters of various diseases using different areal units (Gregorio et al., 2002, Gregorio et al., 2003, Gregorio et al., 2004, Jemal et al., 2002, Kulldorff et al., 1997, Pollack et al., 2006, Roche et al., 2002, Rushton et al., 2004, Seeff et al., 2003, Thomas and Carlin, 2003). However, few researchers have examined the effect of spatial aggregation error on the spatial scan statistic, and there are no guidelines about how the optimum areal unit for spatial scan statistics might be chosen. Hence, this paper serves as a beginning phase in evaluating the impact of spatial aggregation error on spatial scan statistics. The third paper is also highly related to the first and second papers, using data and methods from each paper.

A non-significant late-stage CRC cluster at the ZIP code level (1998 to 2002) identified in the first paper stimulated the question: does this cluster accurately represent clusters that exist when data are disaggregated to the small area level? This third paper uses data on CRC cases from the first paper, and the Bernoulli-based spatial scan statistic from the first paper is also applied to discover clusters with significantly high late-to-early ratios. Using the Monte Carlo method developed in the second paper,

cancer data are disaggregated to three census geographical units, census tract, census block group and census block (referred to as "reference scales"), for comparison with the ZIP code level spatial scan analysis. The study area is limited to Cook County in order to reduce problems caused by inconsistent boundaries between ZIP codes and census tract/block group in sparsely populated areas.

The Monte Carlo simulation approach from the second paper is applied to generate 100 patterns of CRC cases at each reference scale, linking with the cancer data at the ZIP code level by 12 age-race-gender demographic combinations. Each run of the Bernoulli-based spatial scan statistic in SaTScan requires four input files with particular contents and file extensions, and separate import/output commands. These time-consuming steps challenge repeated manual runs of the spatial scan statistic in SaTScan. Thus, an important contribution of this study is to design a cost-effective and efficient way to auto-run spatial scan statistics in SaTScan. The auto-run procedure is composed of a macro-level SAS program to automatically generate input files for each SaTScan run, and a Javascript to automatically produce each parameter file with the corresponding import commands and non-duplicated output names for each SaTScan run. The procedure designed in this paper is simple and straightforward and can easily be adjusted for other spatial scan statistics by minor changes.

Another contribution from this paper is the finding that the reliability of the ZIP code as a study unit is highly dependent on the local spatial density of cases. Generally, the Bernoulli-based spatial scan statistic based on ZIP code level health data can identify the primary cluster found based on small-area data in areas with a large sample size of cases. In areas with a low density of cases, utilizing ZIP code as the study unit misses clusters that exist on the basis of small-area data. Thus, as in the second paper, spatial aggregation error is found to be highly context-dependent for the spatial scan statistic at ZIP code level. Spatial aggregation error has a minimal effect in study regions with a high density of cases, while in areas with fewer cases spatial aggregation error has an unavoidable impact on zip-code level spatial scan analysis. In each spatial scan study, the specific locations and densities of study cases are the main factor affecting the results of spatial scan statistics. Given that the majority of health data is published at a predefined area level, the release policy requires flexibility to provide data for variable study units to match different case-distribution circumstances.

The main contributions of my dissertation can be divided into two domains. The first one emphasizes the development of innovative spatial analytical strategies. The papers integrate multidisciplinary methods from the fields of the spatial statistics, geovisualization, statistical analysis, epidemiology, and public health to examine the effects of social and spatial inequalities on late-stage diagnosis of colorectal cancer. I developed methods to better evaluate and understand the effects of spatial

aggregation error on health data analysis based on data for predefined study units. These methods include a Monte Carlo simulation method to disaggregate cancer case data from larger areal units to smaller ones, a macro-level SAS program to automate the generation of SaTScan input files, and the java program to produce the SaTScan parameter files automatically. These programs make it possible to auto-run SaTScan for use in Monte Carlo simulation modeling. The other domain focuses on the analytical results and subsequent conclusions from the three papers. Specifically, the risk factors of late-stage CRC were examined in different regions of Illinois. The results indicate that statistically significant risk factors vary over space and time, and that spatial clustering of late-stage CRC diminished during the 1990s. Spatial variation was also noted in the results from the second paper: The effects of spatial aggregation error on ZIP code level statistical analysis were found to differ substantially from one study area to another. In the third paper, ZIP code was proved to be 'not too bad' for implementing spatial scan analysis and detecting spatial clusters of late-stage cancer in areas with large numbers of study cases. However, in areas with fewer cases, using health data at the ZIP code level was unable to detect some small and statistically significant clusters.

References

Bonner M, Han D, Nie J, Rogerson P, Vena J, Freudenheim J: **Positional Accuracy of Geocoded Addresses in Epidemiologic Research.** *Epidemiology*, 2003, **14**: 408-412.

Brawarsky P., Brooks D.R., and Mucci L.A. **Correlates of Colorectal Cancer Testing in Massachusetts Men and Women**. *Preventive Medicine*, 2003, **36**: 659-668.

Centers for Disease Control and Prevention. **Screening for Colorectal Cancer – United States, 1997**. *Morbidity and Mortality Weekly Report*, 1999, **48**: 116-121.

Centers for Disease Control and Prevention. **Trends in Screening for Colorectal Cancer – United States, 1997 and 1999**. *Morbidity and Mortality Weekly Report*, 2001, **50**: 162-166.

Cokkinides V.E., Chao A., Smith R.A., Benon S.W., and Thun M.J. **Correlates of Underutilization of Colorectal Cancer Screening among US Adults, Age 50 Years and Older**. *Preventive Medicine*, 2003, **36**: 85-91.

Cooper G.S., Yuan Z., Landefeld C.S., Johanason J.F., and Rimm A.A. **A National Population-Based Study of Incidence of Colorectal Cancer and Afe. Implications for Screening in Older Americans.** *Cancer*, 1995, **75**: 775-781.

Fortney J, Kathryn R, Warren J: **Comparing Alternative Methods of Measuring Geographic Access to Health Services**. *Health Services & Outcomes Research Methodology*, 2000, **1(2):** 173-184.

Gregorio D.I., Kulldorff M., Barry L., and Samociuk H. **Geographic Differences in Invasive and in Situ Breast Cancer Incidence according to Precise Geographic Coordinates, Connecticut, 1991-1995**. *International Journal of Cancer*, 2002, 100: 194-198.

Gregorio D.I., and Samociuk H. **Breast Cancer Surveillance Using Gridded Population Units, Connecticut, 1992 to 1995**. *Annals of Epidemiology*, 2003, **13**: 42-49.

Gregorio D.I., Kulldorff M., Sheehan T.J., and Samociuk H. **Geographic Distribution of Prostate Cancer Incidence in the Era of PSA Testing, Connecticut, 1984 to 1998**. *Urology*, 2004, 63: 78-82.

Gregorio D, DeChello L, Samociuk H, Kulldorff M: **Lumping or Splitting: Seeking the Preferred Areal Unit for Health Geography Studies.** *International Journal of Health Geographics*, 2005, **4**: 6.

Haynes R., Pearce J., and Barnett R. **Cancer Survival in New Zealand: Ethnic, Social and Geographical Inequalities.** *Social Science & Medicine*, 2008, **67(6):** 928-937.

Hewko J, Smoyer-Tomic KE, Hodgson MJ: **Measuring Neighborhood Spatial Accessibility to Urban Amenities: Does Aggregation Error Matter?** *Environment and Planning A*, 2002, **34**: 1185-1206.

Hillsman E, Rhoda R: **Errors in Measuring Distances from Populations to Services Centers**. *Annals of Regional Science*, 1978, **12**: 74-88.

Hodgson MJ, Shmulevitz F, Kőrkel M: **Aggregation Error Effects on the Discrete-Space *P*-Median Model: The Case of Edmonton, Canada**. *The Canadian Geographer* 1997, **41**: 415-428.

Jacquez G.M., and Greiling D.A. **Local Clustering in Breast, Lung and Colorectal Cancer in Long Island, New York.** *International Journal of Health Geographics*, 2003, **2(3):** 1-12.

Jacquez GM, Waller LA: **The Effect of Uncertain Locations on Disease Cluster Statistics. In Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing**. Mowrer HT, Congalton RG eds., *Chelsea MI: Sleeping Bear Press* 1999, 53-64.

Jemal A., Kulldorff M., Devesa S.S. Hayes R.B., and Fraumeni J.F.Jr. **A Geographic Analysis of Prostate Cancer Mortality in the United States, 1970-89**. *International Journal of Cancer*, 2002, **101**: 168-174.

Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW: **On the Wrong Side of the Tracts? Evaluating the Accuracy of Geocoding in Public Health Research.** *American Journal of Public Health,* 2001, **91**: 1114-1116.

Kulldorff M. **A Spatial Scan Statistic**. *Communications in Statistics-Theory and Methods*, 1997, **26(6):** 1481-1496.

Meliker J.R., Jacquez G.M., Goovaerts P., Gopeland G., and Yassine M. **Spatial Cluster Analysis of Early Stage Breast Cancer: A Method for Public Health Practice Using Cancer Registry Data.** *Cancer Causes Control*, 2009, **20:** 1061-1069.

McElroy JL, Remington P, Trentham-Dietz A, Robert SA, Newcomb PA: **Geocoding Addresses from A Large Population-Based Study: Lessons Learned.** *Epidemiology*, 2003, **14**: 399-407.

Nelson D.E., Bolen J., Marcus S., Wells H.E., and Meissner H. **Cancer Screening Estimates for U.S. Metropolitan Areas**. *American Journal of Preventive Medicine*, 2003, **24**: 301-309.

Palmer R.C., and Schneider E.C. **Social Disparities across the Continuum of Colorectal Cancer: A Systematic Review**. *Cancer Causes and Control*, 2005, **16:** 55-61.

Pollack L.A., Gotway C.A., Bates J.H., Parish-Patel A., Richards T.B.,Seeff L.C., Hodges H., and Kassim S. **Use of the Spatial Scan Statistic to Identify Geographic Variations in Late Stage Colorectal Cancer in California (United States).** *Cancer Causes & Control*, 2006, **17**:449-457.

Roche L.M., Skinner R., and Weinstein R.B. **Use of a Geographic Information Systems to Identify and Characterize Areas with High Proportions of Distant Stage Breast Cancer**. *Journal of Public Health Management and Practice*, 2002, 8: 26-32.

Rushton G., Peleg I., Banerjee A., Smith G., and West M. **Analyzing Geographic Patterns of Disease Incidence: Rates of Late-Stage Colorectal Cancer in Iowa**. *Journal of Medical Systems*, 2004, **28**: 223-236.

Schneider E.C. **Chapter 7 Disparities and Colorectal Cancer.** *Toward the Elimination of Cancer Disparities*, H.K. Koh ed., DOI 10.107/978-0-387-89443-0_7. Springer Science+Business Media, LLC, 2009.

Shi, X: **A Geocomputational Process for Characterizing the Spatial Pattern of Lung Cancer Incidence in New Hampshire**. *Annals of the Association of American Geographers,* 2009, **99(3):** 521-533.

Thomas A., and Carlin B.P. **Late Detection of Breast and Colorectal Cancer in Minnesota Counties: An Application of Spatial Smoothing and Clustering**. *Statistics in Medicine*, 2003, **22**: 113-127.

Short M., Carlin B.P., and Bushhouse S. **Using Hierarchical Spatial Models for Cancer Control Planning in Minnesota (United States).** *Cancer Causes and Control*, 2002, **13(10):** 903-916.

Wang F.H., McLafferty S., Escamilla V., and Luo L. **Late-Stage Breast Cancer Diagnosis and Health Access in Illinois**. *The Professional Geographer*, 2008, **60**: 54-69.

Whynes D.K., Frew E.J., Manghan C.M., Scholefield J.H., and Hardcastel J.D. **Colorectal Cancer, Screening and Survival: The Influence of Socio-Economic Deprivation.** *Public Health*, 2003, **17(6):** 389-395.

Wingo P.A., Ries L.A., Parker S.L., and Heath C.W. Jr. **Long-Term Cancer Patient Survival in the United States**. *Cancer Epidemiological Biomarkers & Prevention*, 1998, **7**: 271-282.

Woods L.M., Rachet B., and Coleman M.P. **Origins of Socio-economic Inequalities in Cancer Survival: A Review.** *Annals of Oncology*, 2006, **17**:5-19.

Chapter I

Spatial Distribution of Late-Stage Colorectal Cancer in Illinois from 1988 to 2002: Associations with Social and Spatial Covariates

1. Introduction

Colorectal cancer (CRC) is a public health priority in the United States because of its high incidence and mortality. CRC is the third leading cause of cancer mortality for both men and women in the United States (Schneider, 2009). In 2007, 142,672 people (52.7 per 100,000) were diagnosed with colorectal cancer and 53,219 people (20.0 per 100,000) died of the disease (CDC, 2008).

Like many other cancers, CRC involves a complicated set of risk factors, including biological, behavioral, social and geographical conditions. Because CRC prognosis is strongly and negatively associated with the tumor stage at time of diagnosis (Rossi et al., 1990), early detection is a critical determinant of CRC patient survival rates (Wang et al., 2010). Regular screening helps to detect tumors at an early and treatable stage. However, mortality rates of CRC demonstrate significant geographic variation (Devesa et al., 1999). Given that the greatest proportion of colorectal cancer deaths (90.2%) occurs among patients diagnosed at an advanced (late) stage, geographic variation in mortality indicates that the rates of late-stage CRC diagnosis vary from place to place. This significant geographic variation in late-stage CRC also suggests that in some regions, CRC screening remains well below 'ideal' screening levels (McMahon and Gazelle, 2002). Geographic barriers such as long travel times and distances to screening facilities may play an important role in late-stage diagnosis especially in remote and impoverished rural areas and among populations with limited access to transportation (Amey et al., 1997, Fazio et al., 2005, Jemal et al., 2005, Lengerich et al., 2005, Rushton et al., 2004).

Recent studies suggest that socio-economic status (SES), demographic factors, and spatial accessibility to healthcare are critical predictors of late-stage CRC diagnosis (Gomez et al., 2007, Henry et al., 2009, Palmer and Schneider 2005). However, with the increase in cancer screening facilities and expansion of health insurance coverage for screening, access to screening has improved over time, so it is likely that the impacts of social and spatial factors on late-stage CRC diagnosis will show a spatial-temporal variation. Thus, the primary objective of this study is to examine the clustering patterns of late-stage CRC at diagnosis over time in Illinois during a period of increasing cancer screening, and the varying associations between social-demographic-spatial risk factors and late-stage CRC diagnosis in the same period. This paper addresses three questions: 1) Are there spatial clusters of late-stage CRC in

Illinois, and did clustering change over time? 2) How do SES and geographic factors influence the risk of late-stage CRC? 3) Did these associations change over time during the period from 1988 to 2002?

2. Background

Late-stage CRC is the term used to describe CRC tumors that, when first diagnosed, are large in size and/or have already spread beyond the colon and rectum to nearby or distant tissues, organs or lymph nodes (NCI, 1999). As discussed in introduction section, early-stage diagnosis of CRC is very critical to reducing CRC mortality. In terms of diagnostic methods, it is generally agreed that colonoscopy screening is the preferred screening test, because of its simultaneous abilities to visualize the entire colon in most (80% ~ 90%) cases and to remove precursor polyps (Godreau 1992, Lieberman and Smith 1991). In addition to specialized cancer screening, primary healthcare is also important for early CRC diagnosis, because it is the beginning step for CRC patients to seek professional help (Mullins 1999). Therefore, timely and accessible colonoscopy screening services and primary healthcare directly improve prognostic outcomes of CRC patients, and high rates of late-stage CRC may be an indication of poor access to and underutilization of CRC screening services and primary health care.

Although CRC mortality has declined since the 1970s (Ries et al., 2000), mortality disparities appear to be increasing among population groups differentiated by race, gender, age, SES, and health insurance status. One can also infer that late-stage diagnosis varies among specific sub-population categories. The risk factors that make people delay colonoscopy screening and result in late-stage diagnosis/death are explicitly described in the following parts of the background section.

2.1. Race and Ethnicity

Research has found that, compared to their white counterparts, blacks are more likely to be diagnosed with an advanced stage of CRC and have a much lower survival rate (Alexander et al., 2007, Amey et al., 1997, Doubeni et al., 2007, Du et al., 2007, Henry et al., 2009, Huang et al., 2007, Krieger et al., 1999, Mayberry et al., 1995, Marcella and Miller 2001; Le et al., 2009, Lengerich et al., 2005, Pagano et al., 2003, Palmer and Schneider, 2005). Both individual and area-based factors contribute to this racial disparity. Lack of health insurance and related knowledge about cancer screening is the major hindrance for blacks to access cancer screening in a timely manner (Bryant and Mah, 1992, Carr et al., 1996, Elnicki et al., 1995, Giovannucci et al., 1995, Thun et al., 1992, Willett et al., 1990). Blacks living in impoverished areas, either rural or urban, are most vulnerable for late-stage CRC diagnosis (Amey et al., 1997, Lengerich et al., 2005, Paquette and Finlayson, 2007). The CRC mortality rate in this country has declined, since Medicare began covering colorectal cancer screening in 1998 (CMS, 2011) and use of

colonoscopy screening has increased. However, the decline in CRC mortality from 1992 to 2002 was much smaller for blacks (0.8%) than for whites (1.9%) (Schneider, 2009). Furthermore, Chien et al. (2005) found that, in addition to blacks, American Indians, Hawaiians, and Mexicans had higher hazard ratios for stage-adjusted CRC mortality, compared to whites. First generation Asian and South Asian immigrants have a relatively lower incidence and mortality of CRC; however, the next acculturated generation experiences a much higher rate (Blesch et al., 1999, Flood et al., 2000, Le Marchand et al., 1997). Given its significant and interactive effects on late-stage CRC diagnosis, treating race and ethnicity as important risk factors in late-stage CRC diagnosis study is a must.

2.2. Gender

In the gender domain, studies have provided contradictory results with regard to whether men or women have a higher incidence of late-stage CRC diagnosis (Wu et al., 2001, Mandelblatt et al., 1996; 9:1). Wu et al., (2001) concluded that men-to-women rate ratios dramatically increase from local-stage to advanced-stage CRC. However, Mandelblatt et al. (1996) suggested that women are more likely than men to be diagnosed with late-stage CRC. Another study found that both sexes have an equal likelihood of CRC mortality (Ries et al., 1999). Callcut et al. (2006) suggested that gender differences in late-stage CRC may be caused by different screening rates within different age groups. Uncertainty still remains about the gender disparity in late-stage CRC at diagnosis and the disparity is expected to vary with screening accessibility.

2.3. Age

A sizeable body of research has shown a positive relationship between age and CRC incidence (CDC, 2011, Nelson et al., 1999, Cokkindes et al., 2003, Brawarsky et al., 2003, Cooper et al., 1995, Wingo et al., 1998). Specifically, the incidence of CRC is more than 10 times higher in people 60-64 years old than in those 40-44 years old (Ries et al., 1999). In the U.S., the median age of people who die from CRC is 75 years (Schneider, 2009). However, although only 8.5% of CRC cases are diagnosed in people under 50 years old, the incidence of CRC is increasing among this sub-population group (Fairley et al., 2006, Pine et al., 2007). Moreover, individuals within this group diagnosed with CRC are likely to be diagnosed with the disease at an advanced stage (Schneider, 2009). Younger CRC patients have lower rates of screening than older CRC patients (Palmer and Schneider, 2005).

2.4. Socio-Economic Status

Conceptually, SES is a combinatorial variable that reflects the availability of resources needed for a healthy and prosperous life. It is composed of factors like educational attainment, occupation,

household income, and employment. These factors influence access to resources, including healthcare access and health insurance status. Although SES is a characteristic of individuals and households, it is often measured at the area level because of the lack of data on individual SES. In cancer research, studies have generally shown that area-based SES is associated with cancer outcomes (Singh et al., 2003). Social epidemiologists have suggested that, while the area-based SES has limits for individual-level cancer research, it is very reasonable and acceptable to study the association between area-based SES and cancer outcomes (Kawachi and Berkman, 2003, Krieger et al., 2006). Furthermore, some literature concludes that area-based SES plays a critical role in influencing prognostic outcomes of CRC, independent of SES at the individual level (Diez-Roux et al., 1998, Diez-Roux et al., 2001, Gomez et al., 2007, Krieger et al., 2002). Studies show that people living in high-poverty areas are less likely to regularly utilize cancer screening services and more likely to present with a late-stage diagnosis than are individuals residing in wealthy areas (Abe et al., 2006; Mackinnon et al., 2007; Pollack et al., 2006; Roche et al., 2002, Sheehan et al., 2004). In these studies, a spatial correlation has also been discovered: clusters of late-stage CRC often coincide with low SES areas. Other research indicates that SES may be an important mediator of the black-white disparity in late-stage CRC at diagnosis (Du et al., 2007, Marcella and Miller 2001, Mayberry et al., 1995, Polite et al., 2006). In the US, individuals residing in low SES areas often lack adequate health insurance, are unable to afford screening services, and/or have difficulty travelling to screening services. This results in a higher incidence of late-stage CRC diagnosis (Walsh and Terdiman, 2003). However, measuring area-based SES lacks a 'gold-standard' for defining appropriate neighborhood areas, and results are highly case-specific. Different studies (Cokkindides et al., 2003, Holmes-Rovner et al., 2002, Marcella and Miller 2001) select different variables to represent SES, which makes it difficult to assess the quantitative influence of area-based SES on late-stage CRC diagnosis. Additionally, area-based SES is often correlated with other demographic risk factors.

2.5. Spatial Disparity

CRC incidence and mortality also vary among geographic contexts (Ries et al., 2000, Cooper et al., 1997). In general, the largest difference is observed between rural and urban settings. Some studies find that individuals living in remote rural areas suffer a higher risk of CRC (CDC, 1999, CDC, 2001, Cooper et al., 1996, Coughlin et al., 2002, Hawlet et al., 2001, Nelson et al., 2003, Ries et al., 2000, Thomas and Carlin, 2003). However, other research shows a higher incidence in urban areas. Hsu and Mas (2006) have shown that within Texas, the greatest excess mortality of CRC occurred in the urban areas of Houston and Dallas. A recent study revealed that the risk of late-stage CRC diagnosis is higher in urban areas, after controlling for spatial and demographic factors (Paquette and Finlayson, 2007). McLafferty and Wang (2009) created a continuous rural-urban gradient on the basis of Rural/Urban

Commuting-Area (RUCA) codes to investigate the geographic variation of late-stage CRC diagnosis in Illinois. They concluded that geographical differences in late-stage CRC are mainly explained by demographic and social-spatial characteristics in different areas. More research is needed to understand the importance of spatial disparity on late-stage CRC diagnosis.

2.6. Healthcare Accessibility

Healthcare accessibility is the ease with which people can receive healthcare services at a given location. Healthcare accessibility has been divided into two major categories: revealed accessibility and potential accessibility (Joseph and Phillips, 1984, Phillips 1990, Thouez et al., 1988). While the former measures utilization of services, the latter estimates the availability of services in an area. Given the difficulty in measuring revealed accessibility for population-based research, many studies focus on potential accessibility (Schneider, 2009, Guagliardo, 2004, Luo and Wang 2003, Luo 2004, Wang et al., 2010). Potential accessibility can be measured in many different ways including distance or travel time to service facilities, or more complex measures that assess the supply of services in relation to demand (Guagliardo 2004, Huff 2000, Joseph and Bantock 1982, Joseph and Phillips 1984, Shen 1998, Wang and Minor 2002, Weibull 1976, Yang et al., 2006). Studies have found that potential accessibility to primary care is an important predictor of cancer diagnostic or prognostic outcomes (Harold and Winder, 2000, Pandya et al., 1985, Winchester et al., 1979). Several studies have used network analysis to examine potential accessibility to colonoscopy screening services at the neighborhood level, and have evaluated the impacts of spatial accessibility on late-stage cancer (Luo et al. 2010, Wang, et al., 2010).

Healthcare accessibility also depends on transportation. Zenk et al. (2006) showed that travel time using public transportation was, on average, at least 10 minutes longer than that using a private automobile. Paskett et al. (2004) found a relationship between healthcare accessibility and lack of a countywide public transportation system in a rural county of North Carolina. Rushton et al. (2004) concluded that high rates of late-stage CRC at diagnosis appeared in places where the average distances to diagnostic facilities were lengthy. Furthermore, the types of transportation available and the amount of travel time are often associated with neighborhood poverty level. Residents in neighborhoods with greater poverty levels are more likely to rely on public transportation, which is associated with higher travel times to healthcare facilities (Paskett et al., 2004).

2.7. Health Insurance Status

In CRC research, evidence has been mounting to indicate that health insurance status has a relationship with late-stage CRC at diagnosis (Ayanian et al., 1993, Chen et al., 2007, Committee on the

Consequences of Uninsurance, 2002, Halpern et al., 2007, Halpern et al., 2008, Roetzheim et al., 1999,). Research shows that individuals without health insurance have worse prognostic outcomes compared to those with private health insurance (Ayanian et al., 1993, Chen et al., 2007, Halpern et al., 2007, Halpern et al., 2008, Roetzheim et al., 1999). Individuals without health insurance are likely to be diagnosed with late-stage CRC, given that they are less likely to participate in regular colorectal cancer screening (Palmer and Schneider, 2005). Individuals covered by Medicaid also tend to present with late-stage CRC at diagnosis, compared to people with private insurance (Cooper et al., 1997, Mayberry et al., 1995, Thomas and Carlin, 2003, VanEenwyk et al., 2002). However, differences in health insurance status fail to explain other socio-demographic disparities in CRC studies, especially racial discrepancies (Klabunde et al., 2006, McMahon et al., 1999, O'Malley et al., 2005, Seeff et al., 2004, Shih et al., 2006). Moreover, privacy restrictions create barriers for researchers to access data on insurance status. Thus it is extremely difficult to incorporate health insurance status in studies of cancer disparities.

In summary, research shows that a variety of demographic, SES and geographic factors influence the risk of late-stage CRC. However, although each aforementioned risk factor has been thoroughly studied, it remains unclear how all these demographic, social and geographical factors work together to affect disparities in late-stage cancer, and how their effects have changed over time. Since 1990, the number of CRC screening facilities has increased in the U.S., awareness of the importance of CRC screening has also increased, and insurance coverage of CRC screening has improved. These trends may reduce social and spatial inequalities in late-stage CRC and change the associations between risk factors and late-stage CRC. The purpose of this paper is to integrate all these risk determinants together to study their influences on late-stage CRC in two time periods, from 1988 to 2002. The following conceptual framework illustrates how late-stage cancer risk is influenced by demographic, SES and spatial factors at the individual and area levels (Figure 1).

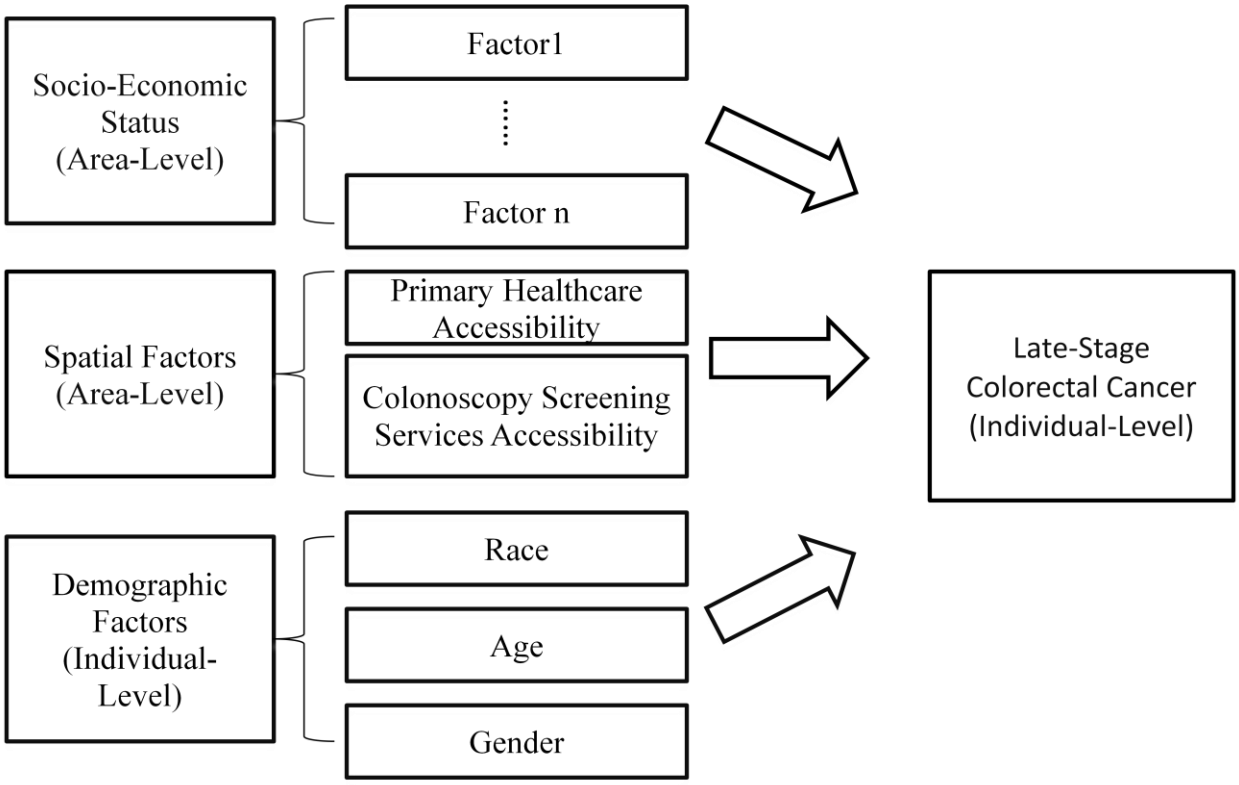

Figure 1. Conceptual Framework Showing the Relationships between Multilevel Predictors and Late-Stage Colorectal Cancer

3. Dataset and Methodology

This study area covers the state of Illinois. Illinois is an ideal study area, because of its vast spatial-demographic variation, from the highly populated Chicago metropolitan area to remote rural regions (McLafferty and Wang, 2009). This research uses four main data sets: 1) CRC cancer cases, 2) SES characteristics of residential areas, 3) locations of primary care and CRC screening services, and 4) the road networks linking populations and services. The following sections introduce the four main datasets, the corresponding methods to create variables for analysis, and the subsequent statistical models to address the research questions in this paper.

3.1. Colorectal Cancer Data

The CRC case data used in this spatial-temporal study were obtained from the Illinois State Cancer Registry (ISCR). ISCR follows the guidelines of the National Cancer Institute (NCI) Surveillance Epidemiology and End Results (SEER) Registries, and it includes data on all cancers diagnosed in Illinois residents. Illinois residents who were diagnosed in neighboring states, such as Missouri and Wisconsin,

are also included in the registries, and the completeness of case ascertainment is estimated as 98% (Lehnerr and Havener, 2002).

The data are for 5-year time intervals. To focus on the 1990s, a period of rapid growth in CRC screening, CRC cases within the time periods, the time intervals 1988 to 1992, and 1998 to 2002 were analyzed. The cancer dataset consists of individual records, which are geocoded to the ZIP code of residence. No cancer data were available for smaller areas like census tracts or blocks because of privacy and confidentiality restrictions. For each case, data on cancer type, age group, sex, race, diagnosis stage and year are included. The ISCR uses a classification scheme consistent with SEER summary stage to measure diagnosed cancer stage (Young et al., 2001). Early-stage CRC cases comprise the localized categories (0 and 1), and late-stage CRC cases include regional and distant categories (stages 2 to 7). Unstaged or unknown-staged cases were excluded.

In addition to staging information, data on gender, race and age are also included. Gender is divided into male and female. The race data identify black and non-black populations; no other racial or ethnic designations are included in the dataset. Age information is stored as 5-year groups from 0 to over 80 years old. To clearly differentiate the age disparity in advanced-stage CRC diagnoses, the age information was re-divided into three groups: age less than 50 years old, age between 50- and 70-years old, and age above 70 years old. Using 50 as the cut-off point makes sense because health insurance plans start to cover regular colorectal cancer screening when the insured person reaches 50 years old. Choosing 70 as another dividing point separates senior patients from young and mid-age patients for examining the influence of age on late-stage CRC at diagnosis. Cases with unknown gender, unknown racial group, and missing age information are all excluded from this study.

3.2. Socio-Economic Status

Based on prior measures of area-based SES, I analyzed SES using a series of census variables to represent local concentrations of vulnerable population groups. The variables include race, immigrant status, language barrier, education attainment, economic level, marriage status, occupation, and physical limitation at zip-code level. Note that the midpoints of the CRC data time intervals match the census years. The original variables were extracted from US Census Bureau decennial Summary Files 3 (SF3) in 1990 and 2000 (US Census Bureau, 1990 and 2000). The detailed information of the 14 SES indicators is explained below:

Table 1. Socio-Economic Variables in 1990 and 2000

Variable Name	Description
RACE	Percentage of black people among the total population
IMMIGRANT	Percentage of people born overseas among the total population
LANGUAGE BARRIER	Percentage of people who speak English as the secondary language
EDUCATIONAL ATTAINMENT	Percentage of people without a high school diploma among the total population ≥ 25 years old
MEDHHINC	Median household income
HOUSEHOLD W/O PLUMBING	Percentage of households without complete plumbing facilities among the total households
HOUSEHOLD W/O VEHICLE	Percentage of households not owning a vehicle in the total households
POPULATION BELOW POVERTY LEVEL	Percentage of population below the poverty level
SINGLE STATUS	Percentage of people ≥ 15 years old who are unmarried (single status) among the total population ≥15 years old
DIVORCE STATUS	Percentage of people ≥ 15 years old who are divorced among the total population ≥15 years old
POPULATION IN WORKING CLASS	Percentage of people employed in working class occupations (clerical and blue collar jobs) among the total employed population
UNEMPLOYED POPULATION	Percentage of people unemployed among the total population in the labor force
PHYSICAL LIMITATION	Percentage of people with limited mobility and/or self-care among the total population
COMMUNTING TIME	Percentage of people ≥ 16 years old who do not work at home, travelling ≥ 30 minutes

It is likely that these variables are correlated with each other which can potentially cause multicollinearity in statistical analysis. Thus factor analysis (FA) was used to examine the intercorrelations among the variables, and to identify latent dimensions that aggregate correlated variables together. *PROC FACTOR* in SAS (SAS, 2011) was implemented to run the factor analysis. Varimax, one of the common orthogonal rotations in factor analysis, was chosen to maximize the differences in variable loadings to clearly describe different dimensions of SES. Following the Kaiser criterion, factors whose eigenvalues were less than 1 were excluded (Griffth and Amrhein, 1997). In each factor, the widely used cutoff value, ±0.5 was chosen to identify variables that load highly on a factor.

For 1990, the SES variables were available at the ZIP code level, while in 2000 these variables were only available at the census-tract level. Areal interpolation (Fisher and Langford, 1996, Tobler, 1979), specifically the areal-weighting transformation, was employed to translate the SES factors from census tracts to ZIP code areas in 2000.

3.3. Healthcare Accessibility

Healthcare accessibility varies across geographical areas, given the uneven distribution of healthcare providers and patients with geographical and financial limitations. This study focuses on potential spatial accessibility because data are not available for the study periods on actual health care utilization. I analyze three components of potential accessibility: spatial access to primary healthcare, access to the nearest colorectal cancer screening services, and health professional shortage areas (HPSAs).

3.3.1. Spatial Access to Primary Healthcare

Access to primary care doctors is very critical to prevent late-stage diagnosis of CRC, because these physicians are the first contacts for patients and the gateways for referral to specialty care (Guagliardo, 2004, Wang et al., 2010). Two key factors, physician supply and population demand, greatly influence primary healthcare accessibility. Physicians in primary healthcare are family physicians, general practitioners, general internists, obstetricians-gynecologists, and physician specialists such as oncologists (Cooper, 1994). The data for Illinois primary care physicians in 1990 and 2000 were obtained from the Physician Master File of the American Medical Association (AMA). Constrained by the geographical accuracy in the AMA data (a significant amount of physicians have P.O.Box addresses which are useless for street address geocoding), physicians were geocoded to the zip code level. The population-weighted centroid of a ZIP code was used to approximate the location of physicians whose general addresses are within the ZIP code area. To represent demand for physicians, because individual street addresses for all Illinois residents in 1990 and 2000 were not obtainable, the population-weighted centroids of census tracts were treated as population residential locations. The travel time along the street network from residential locations to primary care physicians was calculated using ArcGIS Network Analyst (ESRI, 2006). Population data in 1990 and 2000 was obtained from US Census Bureau Summary File 1 (SF1) at the census tract level. The street networks in the two periods came from Tiger/Line 2000 (US Census Bureau, 2000), because 2000 was the starting point to systematically and nationally publish electronic road network data. The speed limit along each street segment was estimated using street type and population density around the segment, according to a previous study (Luo and Wang, 2003). Given that people mainly access primary healthcare during regular business hours, the speed estimation also considered possible traffic congestion. Consequently, the speed limit ranged from 65 miles/hour on interstate highways to 5 miles/hour on unpaved dirt roads (Table 2). Then the Enhanced Two-Step Floating Catchment Area (E2SFCA) method was applied to evaluate primary healthcare accessibility in 1990 and 2000.

The E2SFCA measures healthcare accessibility by calculating physician to population ratios for small areas. E2SFCA is applied in two steps: (1) define the travel time zones around each physician (ZIP code), and search all population locations that are within the threshold travel time; then compute the weighted physician-to-population ratios in each travel time zone and sum the values; (2) for each population location (census tract), search all physician locations that are within the threshold travel time, and sum the weighted physician-to-population ratios. Detailed information about the E2SFCA method is provided in Luo and Qi (2009). Travel times are calculated based on network distances and estimated speeds. In this paper, 30-minutes served as the maximum travel time threshold (Lee, 1991). Subsequently, each area was separated into three travel time zones: 0-10min, 10-20min, and 20-30min. The distance decay weight for each zone was computed according to a Gaussian function, generating the values: 1.0 - 0.42 - 0.03. The end result was a population to physician ratio for each census tract representing the local supply of primary care physicians relative to local population. Spatial interpolation, specifically the areal weighting method, was used to transform the healthcare accessibility values from the census tract to ZIP code level (Flowerdew and Green, 1994).

Table 2. Estimated Speed Limits on Different Road Categories in Illinois

Road Category(CFCC)	Population Density (per km-square)	Area	Speed Limit (MPH)
A1*(Interstate Highway)	>=100	Urban and Suburban	55
	<100	Rural	65
A2* (US and State Highway)	>=1000	Urban	35
	1000>density>=100	Suburban	45
	<100	Rural	55
A3* (State Highway)	>=1000	Urban	35
	1000>density>=100	Suburban	45
	<100	Rural	55
A4*(Local Road)	>=1000	Urban	20
	1000>density>=100	Suburban	25
	<100	Rural	35
A5*(Vehicle Trail: one-lane dirt road)	No Matter where		5
A6*(Road with Special Utilities: intersections, bicycle line etc.)	>=1000	Urban	15
	1000>density>=100	Suburban	20
	<100	Rural	25

*: Any integral number between 0 and 9.

3.3.2. Spatial Access to Colonoscopy Screening

Cancer screening services are also important components of accessibility to health care for CRC patients. To accurately measure accessibility to CRC screening in 1990 and 2000, effort was put into obtaining the street addresses of screening facilities. I focused on colonoscopy because it is the most recommended screening method (McMahon and Gazelle, 2002). This service is typically performed in

specialized hospitals rather than at individual doctors' offices, so finding specific locations of the hospitals providing colonoscopy screening services was crucial. After numerous contacts and consults with different health data providers, the addresses of hospitals in 1990 and 2000 were extracted from American Hospital Association (AHA) yearly guides for 1991 and 2001. The AHA guide is an annually-updated encyclopedia, containing hospital system profiles, hospital listings and healthcare statistics (AHA, 2011). Based on recommendations from managers at the National Cancer Data Base, hospitals coded as 2 (2 designates a cancer program approved by American College of Surgeons) in the cancer category were identified as facilities that offer colonoscopy screening services (Anderson, 2010). Data from AHA was verified by comparing data for 2000 with the same year's data on colonoscopy screening facilities from the U.S. Food and Drug Administration (FDA) (Wang et al., 2010).

Ninety-five hospitals in 1990 and 194 hospitals in 2000 provided colonoscopy screening services in Illinois, reflecting the expansion of screening services across the state. Hospitals in 2000 were more evenly-distributed, while hospitals were mainly concentrated in the Chicago Metropolitan area in 1990 (Figures 2 and 3). Travel distance and travel time along the road network from each ZIP code residential location to the nearest colonoscopy screening service were calculated.

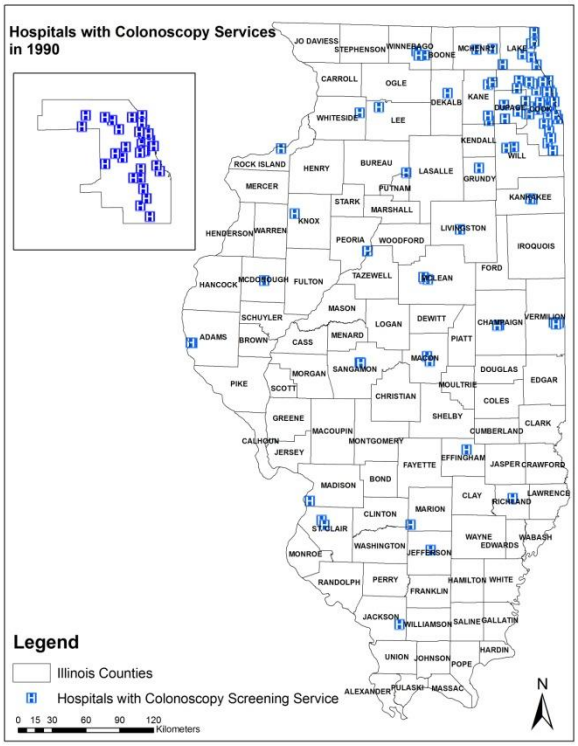

Figure 2. Locations of Hospitals with Colonoscopy Screening Services in 1990

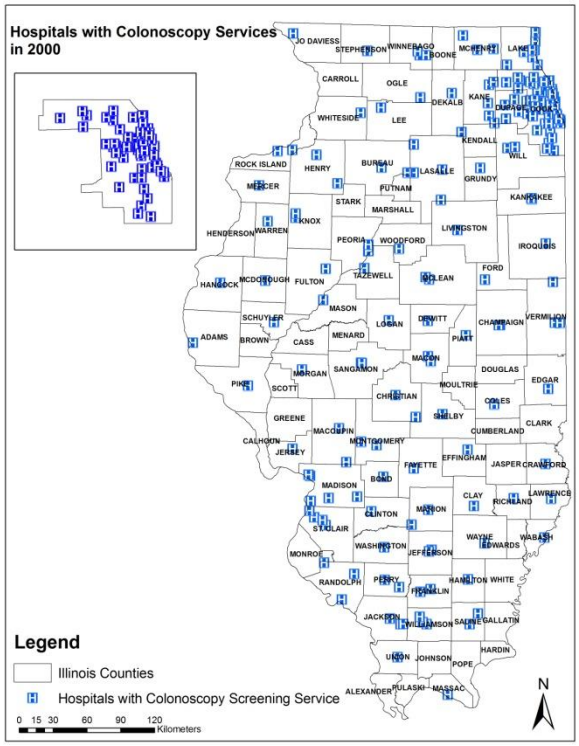

Figure 3. Locations of Hospitals with Colonoscopy Screening Services in 2000

It is also interesting to examine the impact of living in a Health Professional Shortage Area (HPSA) on late-stage colorectal cancer risk. "Health Professional Shortage Areas (HPSAs) are designated by HRSA (Health Resources and Services Administration) as having shortages of primary medical care, dental or mental health providers and may be geographic (a county or service area), demographic (low income population) or institutional (comprehensive health center, federally qualified health center or other public facility)." (U.S. DHHS , 2011). HPSA's are defined based on the ratio of population to full-time-equivalent (FTE) physicians within an area and the availability of providers in contiguous areas (U.S. DHHS, 2011). Primary care HPSAs in Illinois in 1990 and 2000 were selected. Like other historical datasets, HPSA suffers from inconsistent terminology. The areas were named Health Manpower Shortage Area (HMSA) in 1990, as verified by HRSA staff through emails. The geographic components of HPSAs are civil townships and census tracts. The corresponding boundary shapefiles were obtained from Cartographic Boundary Files, County subdivisions section in US Census Bureau website (U.S. Census Bureau, 1990 and 2000). According to these files, 490 townships and census tracts were located in HMSAs in1990, and 287 townships and census tracts were designated as primary care HPSAs in 2000. A dummy variable designating HPSA location (or not) was defined for each year.

3.4. Study Units

To maintain consistency with the geographical characteristics of the CRC dataset from ISCR, data on CRC risk factors were obtained at two geographic levels: individual and ZIP code. Demographic information (race, gender, and age), was available at the individual-level from the cancer dataset. Other social-spatial covariates, such as SES and healthcare accessibility, were measured at the ZIP code level to match information about the ZIP code of residence for cancer cases. ZIP codes, 5-digit numeric codes, were designed by U.S. Postal Service to facilitate mail delivery, and each ZIP code consists of a collection of mail distribution points and routes. The spatial analysis of late-stage CRC in this paper requires ZIP codes as polygons. One particular debate in using ZIP code level data is which ZIP code zones – ZIP code boundaries created by zonal interpolation, or ZIP Code Tabulation Areas (ZCTA) developed by US Census Bureau (U.S. Census Bureau, 2000) – are more accurate in reflecting postal ZIP codes. Each ZCTA was built by aggregating census blocks whose addresses shared a given ZIP code. Grubesic (2008) thoroughly examined the differences between both systems, and concluded that the regular ZIP-code scheme is more appropriate for research. Thus, this study utilized ZIP code zones created by interpolation by TeleAtlas based on mailing addresses and roads (Grubesic and Matisziw, 2006). To avoid variation in boundaries and size caused by periodic updates of ZIP code areas, ZIP codes in 2000 were selected as the basic areal units. This may cause some inaccuracy in the 1990 analysis, if particular ZIP code boundaries changed between 1990 and 2000. The original data were improved by

deleting isolated islands, merging small-sized areas, and dissolving separated parts to build a continuous topology. After this, there were 1,236 ZIP code polygons.

3.5. Spatial Clustering Analysis of Late-Stage Colorectal Cancer Cases

An exploratory statistical method, the spatial scan statistic, was performed to identify spatial clusters of late-stage CRC cases during the two study periods. Many researchers have used spatial scan statistics, implemented in SaTScan software, to examine geographic differences in incidence and mortality of cancer (Gregorio et al., 2002, Jemal et al., 2002, Kulldorff et al., 1997, Seeff et al., 2003, Roche et al., 2002, Thomas and Carlin, 2003). SaTScan's free availability and powerful detection techniques make it a popular choice (Kulldorff et al., 2003). The Bernoulli-based scan statistic was used to detect spatial clustering based on ratios of late-to-early CRC diagnoses at the ZIP code level. The analysis was performed separately for the state as a whole and for two sub-areas: the Chicago metropolitan area and non-Chicago metropolitan area because of the large and dense concentration of population and cases in Chicago. To apply the spatial scan statistic, a circular scanning window whose maximum size contains 33% of CRC cases was selected as the optimal parameter, after multiple trials in each study region. Within each circular window, the maximum likelihood method was utilized to test deviations from the null hypothesis of equality between the late-to-early ratio inside the circle and that outside. Subsequently, a Monte Carlo permutation technique (999) was applied to evaluate the statistical significance of clusters. Finally, clusters with high late-to-early ratios of CRC cases were mapped in ArcGIS.

3.6. Hierarchical Logistic Regression

Hierarchical logistic regression was applied to examine the relationship of late-stage CRC diagnosis with multilevel risk determinants within the two time periods (1988 to 1992 and 1998 to 2002). A two-level logistic model was implemented to maximally use demographic information about individual-level CRC cases (age, gender, race), as well as ZIP code level covariates. The dependent variable in the multilevel logistic model was late-stage CRC diagnosis; the micro-level predictors included patient's age, gender, and race categories. Specifically, age was modeled as the two younger groups (Age < 50, Age 50 -70), and the reference category was the oldest group (≥70). Race was treated as a binary variable: Black and Non-Black (reference category). Gender was divided into male and female (reference category). The macro-level predictors were: (1) factor scores representing the three dimensions of socio-economic status (SES Disadvantage, Minority Disparities, and Cultural-Language Barriers), (2) shortest travel time to the nearest hospital with colonoscopy screening services, and (3) accessibility to

primary healthcare. HPSA was included in initial models, but its coefficients were never statistically significant and therefore it was dropped from the final models.

The specific formulation is a 2-level random intercept model, which specifies the fixed-effect of individual variables across ZIP code areas and the random-effect of intercepts at the ZIP-code level caused by the variation of macro-level predictors across ZIP codes. This study also utilized the strategy of entering 'blocks' variables, separately entering ZIP code level predictors and individual-level covariates, to examine their respective influences on late-stage CRC diagnosis. Subsequently, independent variables at both levels were combined in a final model. The 'proc glimmix' statement in SAS (SAS, 2006) was used to build these multilevel logistic models. The formulations of the hierarchical logistic model are shown below:

The micro specification (level 1) is:

$$Logit(\Pr ob(Y_{ij} = latestageCRC)) = \beta_{0j} + \beta_{1j}(Race)_{ij} + \beta_{2j}(Age)_{ij} + \beta_{3j}(Gender)_{ij} + R_{ij} \quad (1)$$

where the β s denote the intercept and regression coefficients of the predictors at the individual level, i=1,…,n_j denotes individuals within different ZIP code areas, and j=1,…,J denotes ZIP code areas. The R_{ij} are micro errors with independent normal distributions, $R_{ij} \sim N(0, \sigma^2)$.

The macro stage (level 2) model is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(SESDisadvantages)_j + \gamma_{02}(MinorityDisparity)_j + \gamma_{03}(CulturalLanguageBarrier)_j + \gamma_{04}(ShortestTravelTime)_j + \gamma_{05}(\Pr imaryHealthcareAccessibility)_j + U_{oj} \quad (2)$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

where Us represent random effects of the intercept at the ZIP-code level, $U_{0j} \sim N(0, \tau_0^2)$ and they are independent over j and with R_{ij}. Equations (1) and (2) can be combined to form a single-equation model (3):

$$Logit(\Pr ob(Y_{ij} = latestage)) = \gamma_{00} + \gamma_{10}(Race)_{ij} + \gamma_{20}(Age)_{ij} + \gamma_{30}(Gender)_{ij}$$
$$+ \gamma_{01}(SESDisadvantages)_j + \gamma_{02}(MinorityDisparity)_j + \gamma_{03}(CulturalLanguageBarrier) \quad (3)$$
$$+ \gamma_{04}(ShortestTravelTime)_j + \gamma_{05}(\Pr imaryHealthcareAccessibility)_j + U_{oj} + R_{ij}$$

Given the large geographic and social differences across Illinois, I expect that the impact of social and spatial variables on late-stage risk might differ among geographic contexts. To address this, the state was divided into three zones: Chicago city, Chicago suburbs, and other areas. To differentiate Chicago suburbs from other areas, the Rural-Urban Commuting Areas (RUCA) ZIP code approximation scheme, developed by the Office of Rural Health Policy was used. That scheme categorizes areas into 4 general taxonomies, urban core areas, suburban areas, large town areas, and small town and isolated rural areas, based on urbanized population and commuting flow (Hart et al., 2005). All contiguous ZIP codes classified as "urban" or "suburban" and located in areas neighboring Chicago were grouped into the Chicago suburbs category. The RUCA ZIP code approximation codes for Illinois in 1990 and 2000 were derived from the following website (http://depts.washington.edu/uwruca/ruca-data.php). Comparing the classifications in 1990 and 2000, shows little difference in the number of ZIP codes in the three geographical zones between the two years, except for a small decrease in the Other Areas category and a corresponding increase in the Chicago suburbs (Table 3). Multilevel models were estimated for each of the three regions in each year.

Table 3. Number of RUCA ZIP-Codes in the Three Regions of Illinois in 1990 and 2000

Time Frame	Chicago City	Chicago Suburbs	Other Areas
1990	57	224	955
2000	57	242	937

4. Results and Discussion

4.1. Late-Stage Colorectal Cancer Incidence

The number of CRC cancer cases in Illinois increased during the 1990s, from 32,843 cases in 1988 to 1992 to 36, 053 in 1998 to 2002 (Table 4). The number of late-stage CRC cases also increased, but the percent late-stage cases declined from 61.19% to 57.34%. Additionally, cases in the youngest age group showed an increasing trend, while the numbers in the other two age groups declined. CRC cases in the black population increased from 3,381 (10.29%) to 4,000 (11.09%), which is consistent with findings from other studies (Alexander et al., 2007, Amey et al, 1997, Doubeni et al., 2007, Du et al., 2007, Huang et al., 2007, Henry et al., 2009, Krieger et al., 1999, Le et al., 2009, Lengerich et al, 2005, Pagano et al., 2003, Palmer and Schneider, 2005, Mayberry et al., 1995, Marcella and Miller 2001). In the period of

1988 to 1992, the percentages of late-stage CRC cases in Chicago city and Chicago suburbs are all higher than the state-average late-stage percentage, while the percentage in the other area is lower (Table 5). Similar geographic differences exist in the period of 1998 to 2002. Although the number of CRC cases increases from 1988-92 to 1998-2002 in each study region, the percentage of late-stage cases decreases, indicating that the risk of late-stage CRC has improved as a result of enhanced cancer detection technology, wider coverage of CRC screening by health insurance plans, and increased public awareness of the importance of regular screening. However, the decreases are small, revealing that prevention of late-stage CRC and early detection still have a long way to go.

Table 4. Descriptive Statistics for Colorectal Cancer in Illinois, 1988-1992, and 1998-2002

Time Period	Total Cases	Stage (No. of Cases and Percentage)		Age (No. of Cases and Percentage)		Race (No. of Cases and Percentage)		Gender (No. of Cases and Percentage)	
88 to 92	32,843	Early	10711 (32.61%)	<50	1918 (5.84%)	Black	3381 (10.29%)	Male	16299 (48.63%)
		Late	20095 (61.19%)	50 ~ 70	12063 (36.73%)	Non-Black	29352 (89.37%)	Female	16544 (51.37%)
				>70	18862 (57.43%)				
		Unknown	2037 (6.20%)	Unknown	0 (0.00%)	Unknown	110 (0.33%)	Unknown	0 (0.00%)
Time Period	Total Cases	Stage (No. of Cases and Percentage)		Age (No. of Cases and Percentage)		Race (No. of Cases and Percentage)		Gender (No. of Cases and Percentage)	
98 to 02	36,053	Early	12195 (33.83%)	<50	2490 (6.90%)	Black	4000 (11.09%)	Male	16546 (45.89%)
		Late	20672 (57.34%)	50 ~ 70	11787 (32.69%)	Non-Black	28669 (79.52%)	Female	16321 (45.27%)
				>70	18590 (51.56%)				
		Unknown	3186 (8.84%)	Unknown	3186 (8.84%)	Unknown	3384 (9.39%)	Unknown	3186 (8.84%)

Table 5. Percentages of Late-Stage Colorectal Cancer in the Three Study Areas

Time Period	Study Regions	Total CRC Cases	Late-Stage CRC Cases and Percentage (% of Total Cases in Each Region)
88 to 92	Chicago City	4868	3067 (63.00%)
	Chicago Suburbs	12032	7821 (65.00%)
	Other Areas	15943	9207 (57.75%)
Time Period	Study Regions		Late-Stage CRC Cases and Percentage (% of Total Cases)
98 to 02	Chicago City	3256	2019 (62.00%)
	Chicago Suburbs	13721	7958 (58.00%)
	Other Areas	19076	10695 (56.07%)

4.2. Spatial Clustering of Late-Stage Colorectal Cancer

The spatial scan statistic analysis shows two primary clusters with high late-to-early ratios for each time period (Figures 4 and 5). Specifically, in the early time period, the first cluster (p-value = 0.037) is located in the southern part of Illinois, and only two hospitals are located within that cluster (Figure 4). The secondary cluster, with a 0.094 p-value, occurs inside the Chicago metropolitan area, covering the eastern part of DuPage County and a small tip of Cook County (Figure 5). In the later period, the first cluster moves into Cook County, potentially indicating growing urban disadvantage. The secondary cluster is located on the western edge of Illinois. However, these clusters are not statistically significant as indicated by their non-significant p-values. This finding indicates that in 1998-2002, late-stage CRC cases were more evenly distributed than they were in the early time period. The potential reasons are: over time, the disparity of seeking professional help for CRC in different Illinois regions may have gradually decreased, and there may be much more equitable access to health facilities across the state. Thus the spatial inequality of late-to-early ratios across Illinois has diminished over time.

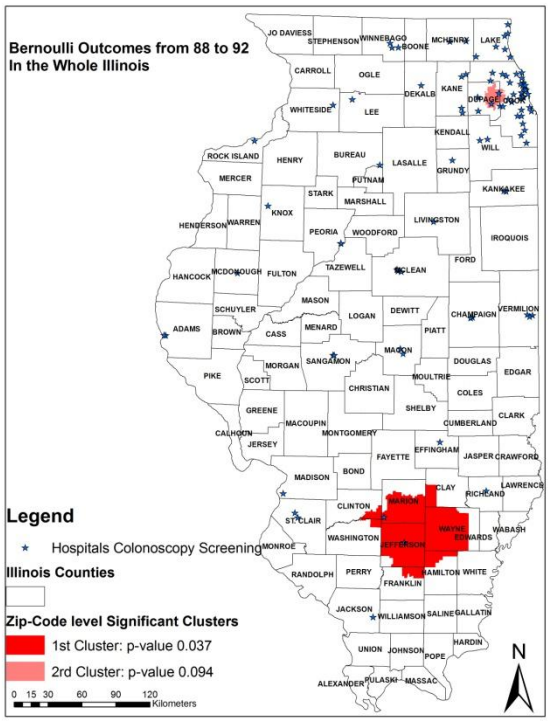

Figure 4. Spatial Clusters of High Ratios of Late- to Early-Stage Colorectal Cancer from 1988 to 1992 in Illinois

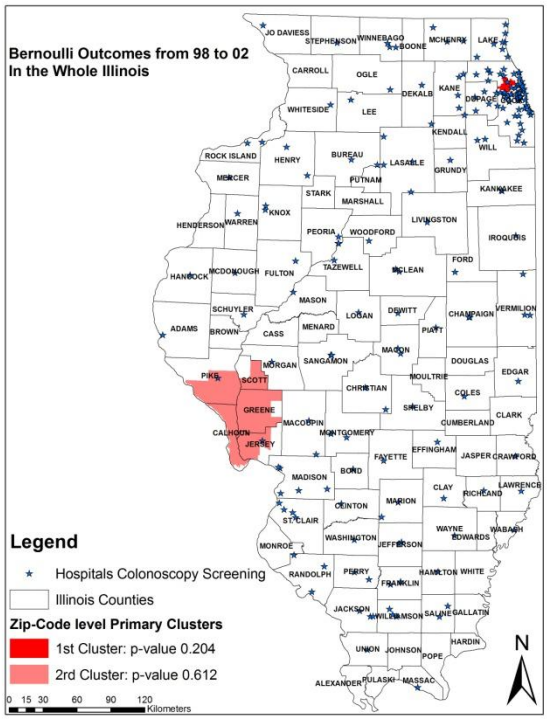

Figure 5. Spatial Clusters of High Ratios of Late- to Early-Stage Colorectal Cancer from 1998 to 2002 in Illinois

SaTScan was separately applied within Chicago metropolitan area, and only the first cluster in each period was illustrated, given that other secondary clusters had very large p-values. The locations of clusters are overlaid with the ones for the whole state. In the earlier period, the primary cluster mainly appears in the eastern section of DuPage County and a small portion of neighboring Cook County. In the later period, the cluster covers the northwestern tip of Chicago city and its surrounding areas (Figures 6 and 7). The size of the cluster diminished over time with the p-value increasing from 0.031 to 0.0916, indicating a more even geographic distribution of late-to-early ratios in the latter period. Hospitals are plentiful within and around these two clusters. Immigrant concentration may account for the spatial clustering. Specifically, the west and north sides of Chicago contain large concentrations of Polish and Hispanic immigrants. Chicago is the largest Polish city outside of Poland, with the Polish community concentrating in the northwest side (The Polish American Association, 2004). Previous studies (Newman and Spengler, 1984, Staszewski and Haenszel, 1965) observed that Polish immigrants have a higher CRC incidence and mortality, compared to their native counterparts. The northwestern cluster may reflect barriers to early diagnosis in the Polish community, a topic that requires more study in the future.

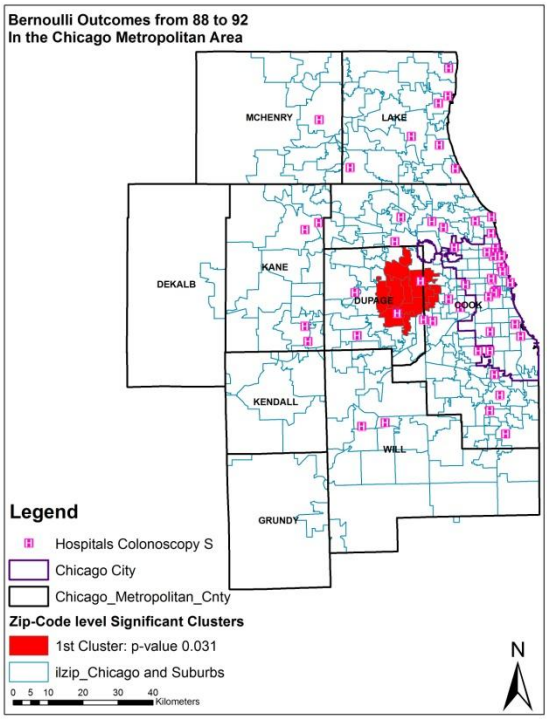

Figure 6. Spatial Clusters of High Ratios of Late- to Early-Stage Colorectal Cancer from 1988 to 1992 in the Chicago Metropolitan Area

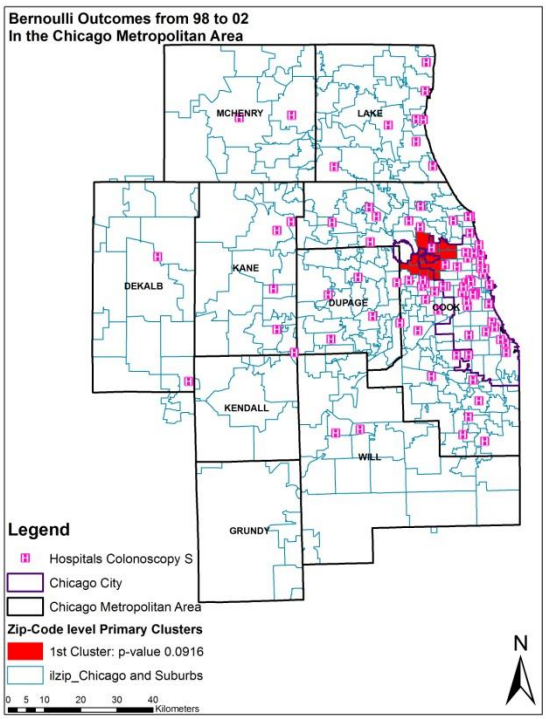

Figure 7. Spatial Clusters of High Ratios of Late- to Early-Stage Colorectal Cancer from 1998 to 2002 in the Chicago Metropolitan Area

Similar to the clusters in Chicago metropolitan area, the primary clusters in the non-Chicago metropolitan area are similar to the ones for the whole state (Figures 8 and 9). The cluster in 1988-1992 is highly statistically significant (p=0.01) whereas that for 1998-2002 is not significant (p=0.365). This indicates that geographical inequality of late-stage CRC decreased substantially during the 1990s. The southern region, where the primary cluster appears between 1988 and 1992, is mainly composed of rural areas, indicating localized rural disadvantage during that period. The disappearance of this cluster in later period suggests that the rural conditions have improved a lot in terms of CRC late-stage prevention, which may be a result of improvements in access to CRC screening.

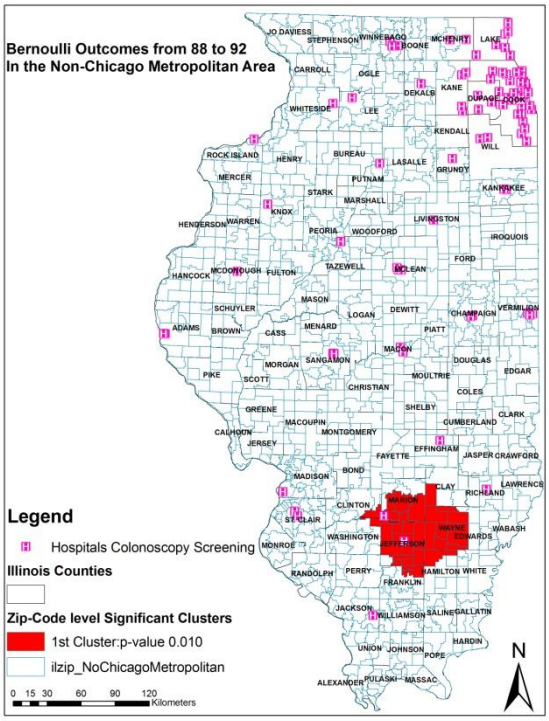

Figure 8. Spatial Clusters of High Ratios of Late- to Early-Stage Colorectal Cancer from 1988 to 1992 in the Non-Chicago Metropolitan Area

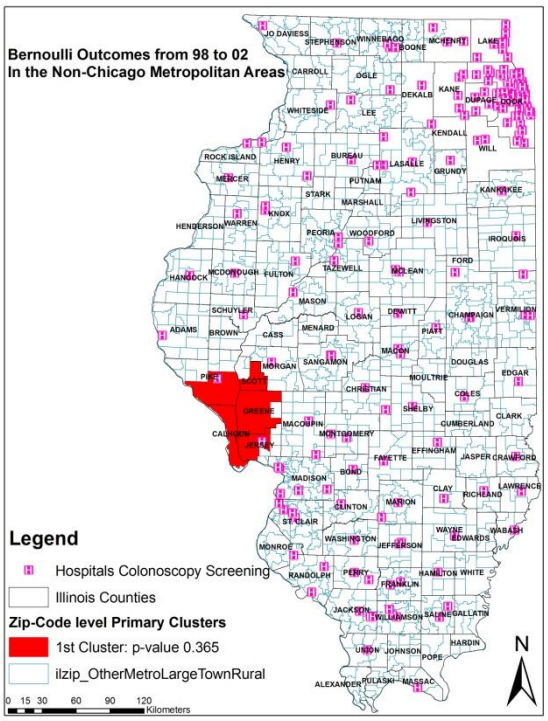

Figure 9. Spatial Clusters of High Ratios of Late- to Early-Stage Colorectal Cancer from 1998 to 2002 in the Non-Chicago Metropolitan Area

4.3. Measures of Socio-Economic Status

To construct variables representing SES, factor analysis was performed on the original SES variables, and 11 and 12 variables were respectively selected out of the original 14 variables in 1990 and 2000 (Table 6). In 1990, the three factors explain 51.86% of total variance, and in 2000 the factors account for 62.20% of the total variance. Based on the characteristics of variables contained in each factor,

the three factors are named: 'Socio-Economic Disadvantage', 'Minority Disparity', and 'Cultural-Language Barriers'. The component-structures of factors in 1990 and 2000 are quite similar: Factor 3 contains the same variables in the two years. Most variables in Factor 1 in each year are the same except for '% unemployed in 1990, and % household without plumbing in 2000. Factor 2 in 2000 has one more variable than in 1990, adding % unemployed. The high similarity makes the factors generally comparable between the two years.

Table 6. ZIP-Code Level Factor Structure of Socio-Economic Status Variables

Variables	1990			Variables	2000		
	Factor 1: SES Disadvantage	Factor 2: Minority Disparity	Factor 3: Cultural-Language Barriers		Factor 1: SES Disadvantage	Factor 2: Minority Disparity	Factor 3: Cultural-Language Barriers
Median household income ($)	-0.512	-0.0920	0.103	Median household income ($)	-0.737	-0.0944	0.224
Population ≥25 without high school diploma (%)	0.819	0.170	0.0793	Population ≥25 without high school diploma (%)	0.795	0.118	0.222
Population below poverty level (%)	0.631	0.540	-0.0508	Population below poverty level (%)	0.710	0.637	-0.0292
People with high care needs (%)	0.603	0.385	0.0495	People with high care needs (%)	0.762	0.235	-0.0241
% working class (%)	0.713	-0.240	-0.165	% working class (%)	0.684	-0.264	-0.120
% workers unemployed (%)	0.535	0.498	-0.0282	Household w/o plumbing (%)	0.496	0.173	0.0366
Black (%)	0.128	0.777	0.0601	Black (%)	0.148	0.739	0.0696
Households w/o a vehicle (%)	0.216	0.765	0.283	Households w/o a vehicle (%)	0.249	0.815	0.308
People ≥15 years old are single (%)	-0.197	0.683	0.326	People ≥15 years old are single (%)	0.00596	0.780	0.376
Population born overseas (%)	-0.207	0.176	0.868	% workers unemployed (%)	0.578	0.641	0.0262
Linguistically isolated households (%)	0.0281	0.109	0.875	Population born overseas (%)	-0.102	0.160	0.945
				Linguistically isolated households (%)	0.117	0.117	0.958

Figures 10 and 11 illustrate the spatial distribution of SES disadvantage in the two years. In 1990, the most disadvantaged areas are dispersed in the urbanized areas across the state, with the highest disadvantage scores occurring in ZIP codes within Cook County and in other metropolitan areas which mainly spot in the middle and southern sections of Illinois. The majority of rural areas have a moderate level of the SES disadvantage. Areas with the lowest SES disadvantage scatter within Chicago metropolitan area and other urbanized regions across the state. In 2000, the Chicago suburbs, display a low level of SES disadvantage. Other urbanized areas, located in central Illinois, also experience a low level of SES disadvantages. Regions located in the western and southeastern sections of Illinois suffer a moderate level of disadvantage. Within the Cook County, areas with the highest SES disadvantage scatter in the middle and southeastern sections in 1990, while the highest scores show up in the northwestern tip and middle parts in 2000.

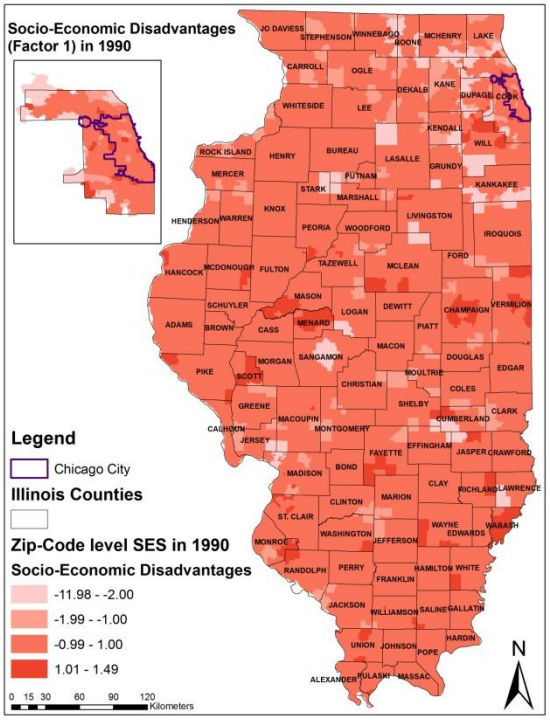

Figure 10. The Distribution of Socio-Economic Disadvantages in Illinois from 1988 to 1992

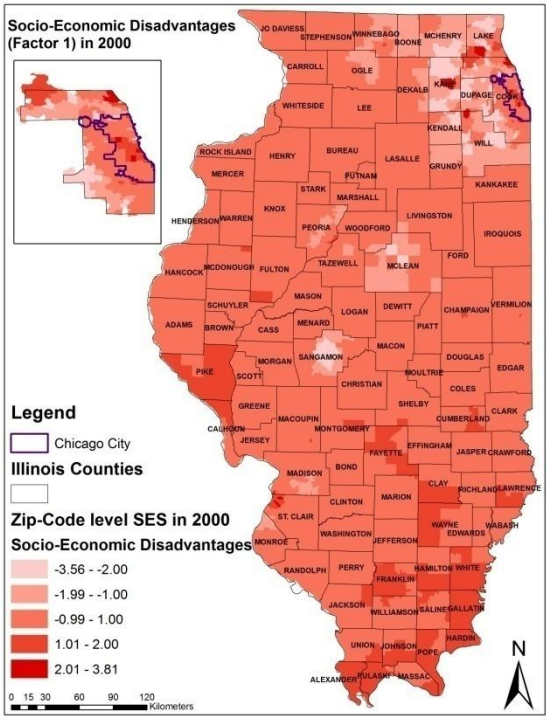

Figure 11. The Distribution of Socio-Economic Disadvantages in Illinois from 1998 to 2002

The minority disparities factor shows concentrations of minority population and associated variables mainly in and around Chicago in both years (Figures 12 and 13). In 1990, areas featured by the highest minority disparities are mainly focused in the southern part of Chicago city, and a small area in northwestern St. Clair County. The urbanized areas near Chicago city and other metropolitan areas also

have relatively high scores on minority disparities. The majority of rural areas present a low level of minority disparities in 1990. In 2000, minority disparities continue to be concentrated in the southern section of Chicago city, but also expand into the northern part of neighboring Will County. Other parts of the state show higher levels of minority disparities in 2000 than in 1990.

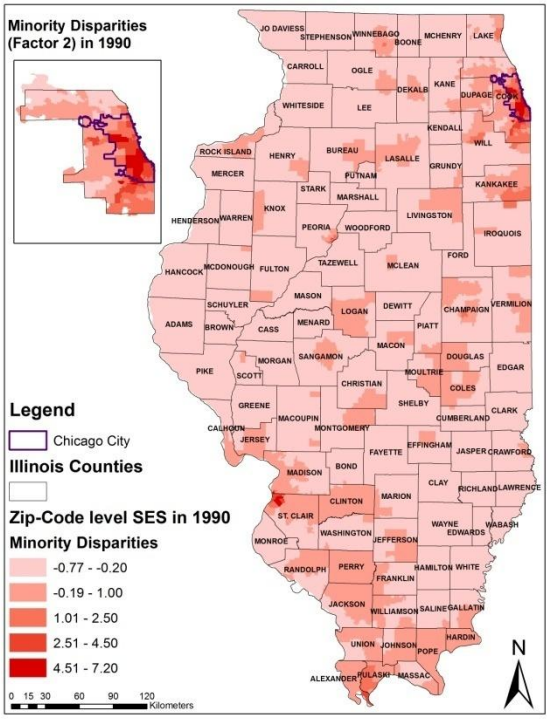

Figure 12. The Distribution of Minority Disparities in Illinois from 1988 to 1992

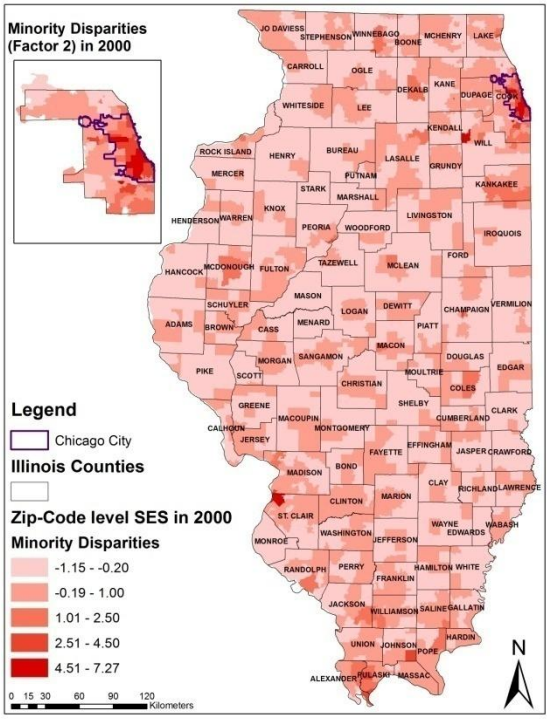

Figure 13. The Distribution of Minority Disparities in Illinois from 1998 to 2002

Factor 3, representing cultural-language barriers, has similar patterns in Illinois during the two time periods (Figures 14 and 15). Particularly, areas in the first and second highest cultural-language barriers categories mainly are located in the Chicago metropolitan area, including Cook, McHenry, Lake, DuPage, some portions of Kane and Will in 1990. A few urbanized areas, scattered in central Illinois, also present a moderate level of cultural-language barriers. In 2000, the areas in the first two highest categories of Factor 3 still focus on the Chicago metropolitan area. The middle category shows up in other urbanized areas outside the Chicago region, mainly in the north and central regions, indicating dispersal of immigrant populations to other parts of the state.

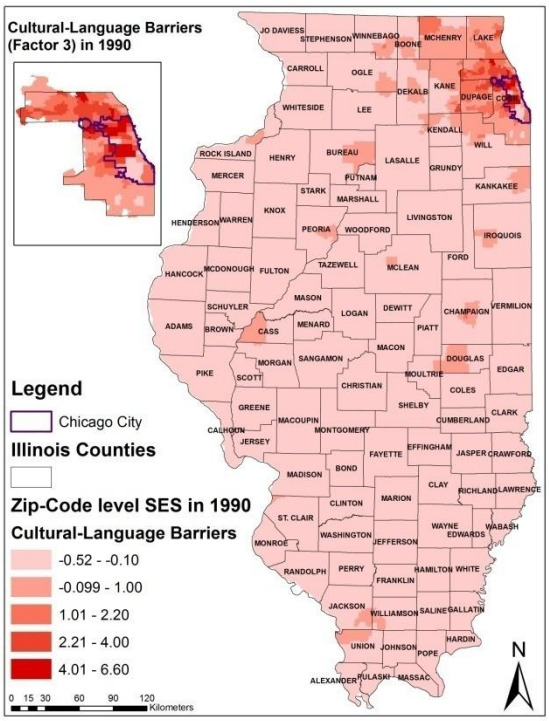

Figure 14. The Distribution of Cultural-Language Barriers in Illinois from 1988 to 1992

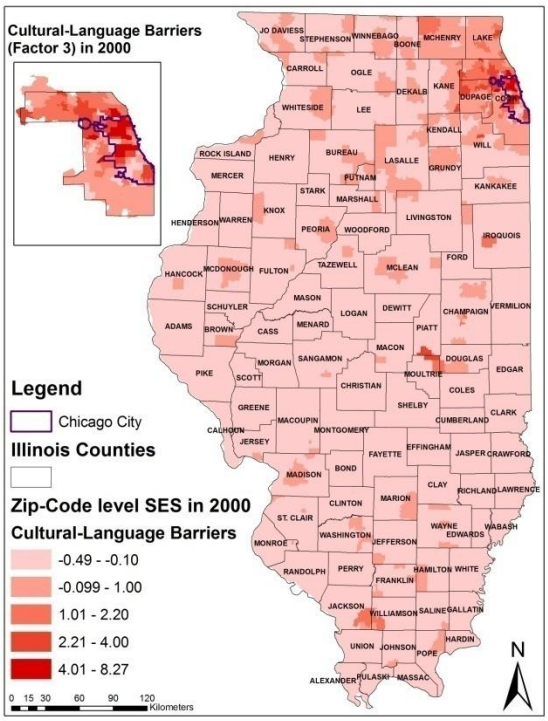

Figure 15. The Distribution of Cultural-Language Barriers in Illinois from 1998 to 2002

4.4. Healthcare Accessibility

Figures 16 and 17 show geographic variation in spatial accessibility to primary health care in 1990 and 2000. Comparing the maps indicates that accessibility generally improved over time. In 1990, most regions with high primary healthcare accessibility are in the Chicago metropolitan area, especially in Cook and DuPage Counties. Specifically, most of Cook County has high accessibility, except the south and western areas of Chicago city and neighboring areas in Cook County. Other ZIP codes falling into the highest category are located in metropolitan areas across the state. In 2000, more sections in the Chicago metropolitan area fall into the highest category of primary healthcare accessibility, including the eastern part of McHenry, eastern and southern parts of Lake, and neighboring parts of Kane, DuPage and Will counties. Improvements in spatial access also occur in other metropolitan areas. Primary healthcare accessibility increased because the number of primary care physicians more than doubled between 1990 and 2000. However, the improvement seems to be concentrated in metropolitan areas, while rural areas in Illinois still suffer from low primary healthcare accessibility. The disparity between urban and rural

reveals spatial barriers for people living in rural areas that can prevent them from receiving timely diagnosis and treatment for CRC.

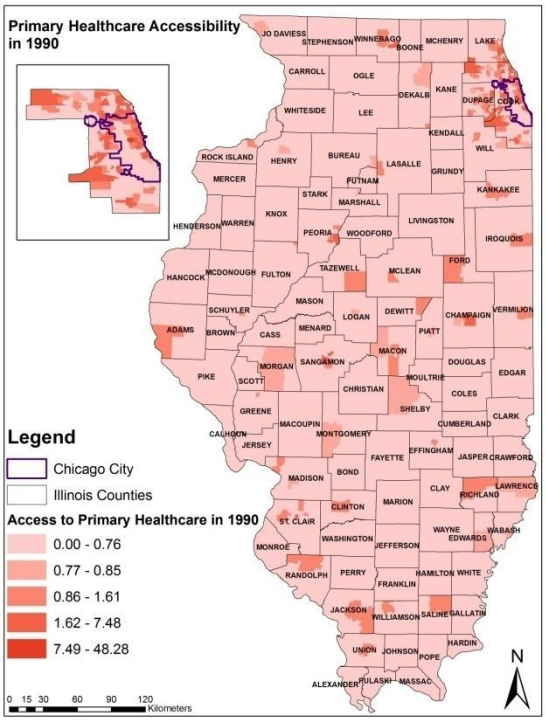

Figure 16. The Distribution of Primary Healthcare Accessibility in 1990

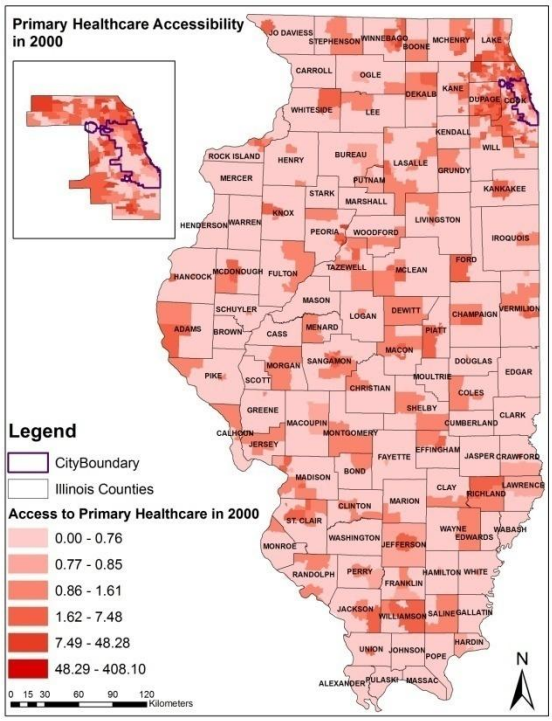

Figure 17. The Distribution of Primary Healthcare Accessibility in 2000

Table 7 contains the descriptive statistics for network distance and travel time to the nearest hospitals with colonoscopy screening services in 1990 and 2000. Given the high correlation between travel distance and travel time, only the shortest travel time, which is more relevant for travel decisions, is selected for the multilevel logistic analysis to avoid multicollinearity. The difference between the minimum and maximum values of travel time decreased substantially from 1990 to 2000, falling from 94.89 minutes in 1990 to 41.01 in 2000. The average travel time also improved over time: in 1990, ZIP codes in Illinois were 25 minutes on average from the nearest hospital with colonoscopy services, while that travel time decreased to 14 minutes in 2000.

Table 7. Descriptive Statistics of Shortest Travel Distance/Time to Colonoscopy Screening in Illinois

Variable	Minimum	Maximum	Mean	Std. Error
Shortest Travel Distance in 1990 (Meters)	164.13	128063.49	3.08×10^4	624.66
Shortest Travel Time in 1990 (Minute)	0.20	95.09	25.48	0.46

Table 7. (cont.)

Shortest Travel Distance in 2000 (Meters)	98.06	55219.43	1.64×10^4	300.59
Shortest Travel Time in 2000 (Minute)	0.15	41.16	14.18	0.23

Figures 18 and 19 illustrate HPSA locations in 1990 and 2000. Compared with HPSAs in 1990, HPSAs cover more areas in 2000. Residents living in HPSAs generally suffer low primary healthcare accessibility. Because of the discontinuous boundaries between ZIP code areas and HPSAs units, an arbitrary decision was made to determine whether a ZIP code area belongs to HPSA category or not. Specifically, if a HPSA covers 50% or more area within a ZIP code, that ZIP code is classified as HPSA, and vice versa. In 1990, 50 ZIP code areas are totally classified to HPSAs, and 65 ZIP code areas are categorized as HPSAs in 2000. The HPSA variable was excluded from the final hierarchical logistic regression models, because of its non-significant effect in the many statistical models.

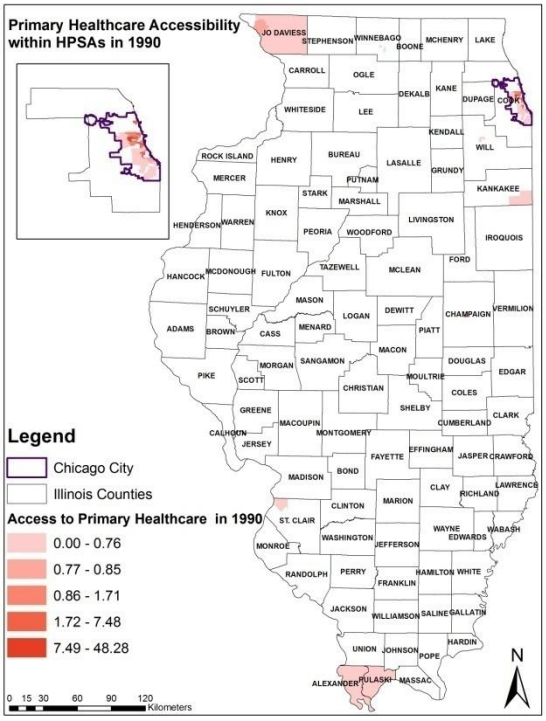

Figure 18. The Distribution of Primary Healthcare Accessibility in HPSAs in 1990

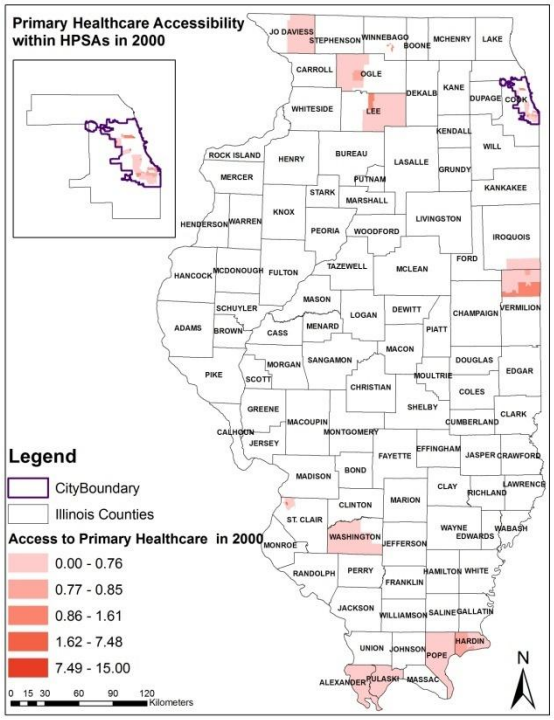

Figure 19. The Distribution of Primary Healthcare Accessibility in HPSAs in 2000

4.5. Hierarchical Logistic Regression Outcomes in Illinois and the Three Sub-Regions

Results from separate 'block-input' models and the complete, fully-specified multilevel models are consistent in each study region. Therefore, only results from the complete multilevel logistic models are discussed in these sections. Also, a high correlation (coefficient >0.5, p-value <0.0001) was detected between shortest travel time and primary healthcare accessibility, so to avoid multicollinearity, only the former variable was retained in the analysis, given its statistically significant influence in the model for the non-Chicago metropolitan region (the "other areas"). Using the whole state as the study region, the predictors in the hierarchical logistic models do not show any statistically significant influence in the risk of late-stage CRC diagnosis within the two study periods. Thus the corresponding analytical outcomes are not discussed in this paper. The subsequent sections mainly focus on describing and interpreting the analytical results in the three regions during the two periods. In all models, the ratio of the generalized chi-square statistic and its degrees of freedom is close to 1, indicating that the predicators are properly modeled and overdispersion only mildly exists (Schabenberger, 2007). The predictors that have a statistically significant influence on late-stage CRC diagnosis are highlighted in gray in each table below.

4.5.1. Hierarchical Logistic Model Results for Places outside the Chicago Region (Non-Chicago Metropolitan Area)

In the model for areas outside the Chicago Metropolitan region in 1988 to 1992, factor 2 (Minority Disparities) has a positive and statistically significant influence on late-stage CRC at diagnosis (Table 8). As mentioned earlier, the highest level of this factor is seen in urbanized areas (Figure 12). Thus the positive influence reveals the vulnerability of people living in metropolitan areas with high percentages of black population and households without vehicles, consistent with previous studies (Krieger et al., 1999, VanEenwyk et al., 2002). The finding can be treated as a specific form of 'urban disadvantage' for people living in racially segregated and low SES areas in cities outside of Chicago metropolitan area. Beyond this, none of the other predictors, whether at the individual or ZIP code level, has a statistically significant effect.

The impacts of independent variables on late-stage CRC diagnosis differ for the 1998- 2002 time period (Table 9). Specifically, people in young- and mid-age groups are more likely to be diagnosed with late-stage CRC, compared to people in the oldest group, consistent with previous studies (Palmer and Schneider, 2004, Schneider, 2009). People in the two younger age groups may not receive regular cancer screening services. In addition, shortest travel time shows a positive and statistically significant relationship with the risk of CRC late-stage diagnosis. Rushton et al. (2004) suggested that areas with high rates of late-stage CRC are often characterized by lack of local cancer screening providers. Given

that this study region covers areas outside of the Chicago metropolitan region, the significant impact of the travel-time reveals that hospitals with colonoscopy screening services in these areas are unevenly distributed; and long distances to care are associated with higher late-stage risk. More specialized services for CRC screening need to be established outside of Chicago metropolitan area. Taking cost and profit into consideration, mobile colonoscopy screening services can be implemented in the remote rural areas in Illinois.

Table 8. Coefficient Estimates for ZIP-Code level and Individual Predictors in Hierarchical Logistic Regression: Non-Chicago Metropolitan Area, 1988 to 1992

Variables	Estimates	Std. Error	t-value	p-value	Alpha	90% CI Lower	90% CI Upper
Intercept	0.36	0.054	6.79	2.63E-11	0.1	0.27	0.45
Race (Black)	-0.031	0.22	-0.14	0.89	0.1	-0.40	0.33
Age (<50)	0.22	0.17	1.30	0.19	0.1	-0.058	0.50
Age (50 ~ 70)	0.075	0.071	1.07	0.29	0.1	-0.041	0.19
Gender (Male)	0.058	0.066	0.87	0.38	0.1	-0.052	0.17
Factor 1 (SES Disadvantages)	-0.043	0.048	-0.91	0.36	0.1	-0.12	0.035
Factor 2 (Minority Disparities)	0.083	0.048	1.74	0.082	0.1	0.0044	0.16
Factor 3 (Cultural-Language Barriers)	0.049	0.047	1.05	0.29	0.1	-0.028	0.13
Shortest Travel Time	-0.011	0.040	-0.26	0.80	0.1	-0.078	0.056

Table 9. Coefficient Estimates for ZIP-Code level and Individual Predictors in Hierarchical Logistic Regression: Non-Chicago Metropolitan Area, 1998 to 2002

Variables	Estimates	Std. Error	t-value	p-value	Alpha	90% CI Lower	90% CI Upper
Intercept	0.30	0.060	4.93	1.09E-06	0.1	0.20	0.39
Race (Black)	-0.021	0.11	-0.19	0.85	0.1	-0.20	0.16
Age (<50)	0.59	0.16	3.73	0.00020	0.1	0.33	0.85
Age (50 ~ 70)	0.25	0.078	3.17	0.0015	0.1	0.12	0.38
Gender (Male)	-0.075	0.073	-1.02	0.31	0.1	-0.20	0.046
Factor 1 (SES Disadvantages)	-0.066	0.043	-1.53	0.13	0.1	-0.14	0.0050
Factor 2 (Minority Disparities)	-0.016	0.044	-0.37	0.71	0.1	-0.090	0.057
Factor 3 (Cultural-Language Barriers)	0.028	0.042	0.65	0.52	0.1	-0.042	0.097
Shortest Travel Time	0.10	0.040	2.47	0.013	0.1	0.033	0.17

4.5.2. Hierarchical Logistic Model Results for Chicago Suburbs

Within the Chicago suburbs, in the earlier period, the youngest age group, has a positive and significant association with late-stage CRC risk at diagnosis (Table 10). This is the only statistically significant variable in the model. This finding that younger patients have a higher late-stage risk is

consistent with previous research (Fairley et al., 2006, Pine et al., 2007), and it suggests low use of screening in this age group. The possible reasons include lack of coverage of screening services by health insurance plans and lack of awareness of colonoscopy screening services among young people. Most health insurance plans, including Medicaid, start covering the colonoscopy screening services when people reach 50 years old. Young people may be unable to afford costly colonoscopy screening services. Beyond age, no other variables have statistically significant associations with late-stage diagnosis in the early time period.

In the period 1998 to 2002, male patients have a higher risk of late-stage CRC (Table 11), and this finding is consistent with some research (Wu et al., 2001) and contradictive with others (Mandelblatt et al., 1996). The gender disparity in late-stage CRC at diagnosis may relate to differences in screening rates between men and women. Additionally, Factor 3 is also positively and significantly associated with the late-stage risk of CRC diagnosis, indicating that people living in Chicago suburban areas with higher cultural-linguistic disparity have a higher possibility to be diagnosed with late-stage CRC.

Table 10. Coefficient Estimates for ZIP-Code level and Individual Predictors in Hierarchical Logistic Regression: Chicago Suburbs, 1988 to 1992

Variables	Estimates	Std. Error	t-value	p-value	Alpha	90% CI Lower	90% CI Upper
Intercept	0.48	0.066	7.27	9.33E-12	0.1	0.37	0.59
Race (Black)	0.14	0.22	0.64	0.52	0.1	-0.22	0.49
Age (<50)	0.47	0.18	2.62	0.0087	0.1	0.18	0.76
Age (50 ~ 70)	0.11	0.081	1.33	0.19	0.1	-0.026	0.24
Gender (Male)	-0.050	0.077	-0.64	0.52	0.1	-0.18	0.077
Factor 1 (SES Disadvantages)	-0.061	0.066	-0.92	0.36	0.1	-0.17	0.048
Factor 2 (Minority Disparities)	-0.014	0.069	-0.200	0.84	0.1	-0.13	0.10
Factor 3 (Cultural-Language Barriers)	-0.066	0.047	-1.42	0.16	0.1	-0.14	0.010
Shortest Travel Time	0.061	0.048	1.27	0.013	0.1	-0.018	0.14

Table 11. Coefficient Estimates for ZIP-Code level and Individual Predictors in Hierarchical Logistic Regression: Chicago Suburbs, 1998 to 2002

Variables	Estimates	Std. Error	t-value	p-value	Alpha	90% CI Lower	90% CI Upper
Intercept	0.31	0.059	5.26	3.31E-07	0.1	0.21	0.41
Race (Black)	0.015	0.11	0.13	0.89	0.1	-0.17	0.20
Age (<50)	0.22	0.15	1.48	0.14	0.1	-0.025	0.47
Age (50 ~ 70)	0.042	0.075	0.56	0.57	0.1	-0.081	0.17
Gender (Male)	0.13	0.070	1.83	0.067	0.1	0.013	0.24
Factor 1 (SES Disadvantages)	-0.00079	0.044	-0.018	0.99	0.1	-0.072	0.071

Table 11. (cont.)

Factor 2 (Minority Disparities)	0.042	0.044	0.95	0.35	0.1	-0.031	0.11
Factor 3 (Cultural-Language Barriers)	0.072	0.043	1.68	0.093	0.1	0.0015	0.14
Shortest Travel Time	0.0060	0.044	0.14	0.89	0.1	-0.067	0.079

4.5.3. Hierarchical Logistic Model Results for Chicago City

In the context of Chicago city, both young- and middle-age groups demonstrate a significantly higher risk of being diagnosed with late-stage CRC in the time span of 1988 to 1992 (Table 12). Since late-stage CRC at diagnosis is a direct consequence of underutilization or inadequate cancer screening, people in these two age groups apparently received or accessed colonoscopy screening services less frequently than those in the old age group (>70 years old) in Chicago city. Another demographic factor, gender shows a clear disparity between men and women: men have a lower risk of late-stage CRC diagnosis than women in this inner-city setting. This finding is totally contradictory with the one in the Chicago suburb region from 1998 to 2002. None of the area-level SES or spatial factors is significantly associated with late-stage risk, suggesting that contextual factors had little influence in Chicago during the early time period.

The effects of individual-level variables in the 1998 to 2002 model are similar. Age and gender are statistically significant, with younger patients having a higher risk of late-stage CRC and men a lower risk (Table 13). The conflicting findings for gender between two neighboring areas (Chicago city and suburbs) demonstrate that results are very case-sensitive. Factor 3 shows a significantly negative association with late-stage CRC at diagnosis, representing that people residing in Chicago ZIP codes characterized by a high level of cultural-language barriers are less likely than others to be diagnosed with late-stage CRC. This is an unexpected finding that contradicts results from the Chicago suburbs. Chicago city has a very large immigrant population. First-generation immigrants sometimes display the 'healthy immigrant' effect (Schneider, 2009). Also, Chicago city contains a large concentration of economically disadvantaged population, and compared to that population, immigrants may be healthier. Several factors can possibly explain why living in immigrant areas has the opposite outcomes between Chicago city and suburbs. Immigrants living in inner-city areas are mainly first generation with strong healthy immigrant effect, while immigrants living in Chicago suburban areas may consist of second generation, acculturated populations whose health may have deteriorated (Blesh et al., 1999, Flood et al., 2000, Le Marchand et al., 1997). Also, the majority population in the suburbs consists of people of moderate to high SES who are likely to have better access to cancer screening than immigrants in the suburbs. Furthermore, in the suburbs, the areas with high cultural-language barriers could be the same areas associated with

impoverishment and racial segregation, which needs more detailed research. It is interesting, however, that living in economically disadvantaged area did not emerge as a statistically significant predictor of late-stage risk in either the Chicago suburbs or city in either period.

Table 12. Coefficient Estimates for ZIP-Code level and Individual Predictors in Hierarchical Logistic Regression: Chicago City, 1988 to 1992

Variables	Estimates	Std. Error	t-value	p-value	Alpha	90% CI Lower	90% CI Upper
Intercept	0.54	0.083	6.50	5.17E-08	0.1	0.40	0.68
Race (Black)	-0.076	0.15	-0.52	0.60	0.1	-0.32	0.16
Age (<50)	0.39	0.19	2.01	0.045	0.1	0.070	0.71
Age (50 ~ 70)	0.20	0.094	2.14	0.032	0.1	0.047	0.36
Gender (Male)	-0.28	0.087	-3.22	0.0010	0.1	-0.42	-0.14
Factor 1 (SES Disadvantages)	0.063	0.080	0.79	0.43	0.1	-0.068	0.20
Factor 2 (Minority Disparities)	-0.0037	0.11	-0.030	0.97	0.1	-0.18	0.18
Factor 3 (Cultural-Language Barriers)	0.034	0.062	0.55	0.58	0.1	-0.068	0.14
Shortest Travel Time	0.021	0.054	0.39	0.70	0.1	-0.068	0.11

Table 13. Coefficient Estimates for ZIP-Code level and Individual Predictors in Hierarchical Logistic Regression: Chicago City, 1998 to 2002

Variables	Estimates	Std. Error	t-value	p-value	Alpha	90% CI Lower	90% CI Upper
Intercept	0.49	0.081	6.01	2.40E-07	0.1	0.35	0.63
Race (Black)	0.12	0.15	0.85	0.40	0.1	-0.12	0.37
Age (<50)	0.34	0.190	1.80	0.072	0.1	0.029	0.65
Age (50 ~ 70)	0.056	0.11	0.53	0.60	0.1	-0.12	0.23
Gender (Male)	-0.19	0.098	-1.91	0.056	0.1	-0.35	-0.026
Factor 1 (SES Disadvantages)	0.052	0.10	0.51	0.61	0.1	-0.11	0.22
Factor 2 (Minority Disparities)	-0.21	0.14	-1.43	0.15	0.1	-0.44	0.031
Factor 3 (Cultural-Language Barriers)	-0.24	0.11	-2.27	0.023	0.1	-0.41	-0.066
Shortest Travel Time	-0.042	0.056	-0.74	0.46	0.1	-0.13	0.051

5. Conclusion

In this study, the spatial patterns of late-stage CRC at diagnosis in Illinois were examined during the two periods, 1988 to 1992, and 1998 to 2002. The presence of significant spatial clusters of late-to-early ratios in the early period represents geographical inequality of late-stage CRC diagnosis, suggesting inequalities in access to cancer screening services in the Chicago metropolitan area and southern Illinois. In the later time period, the absence of statistically significant clusters shows that spatial inequality in late-stage CRC diagnosis has dramatically declined across the state. This may reflect improvements in the

availability of CRC screening services, expansion of insurance coverage for screening, and improvements in education about the need for screening.

This paper also applied hierarchical logistic regression to detect the varying impacts of social-demographic-spatial risk determinants on the risk of late-stage CRC diagnosis in three study regions (Chicago city, Chicago suburb, and non-Chicago metropolitan area) during the two time spans. Among the individual risk factors, age was confirmed as a critical indicator of late-stage CRC diagnosis. Young- and middle-aged groups showed a consistent trend from the analytical results in the two time periods: people who were in the two age categories were more likely to present with late-stage CRC diagnosis in the three regions during both periods, consistent with previous studies (Brawarsky et al., 2003, CDC, 1999, CDC, 2001, Cokkinides et al., Cooper et al., 1995, 2003, Nelson et al., 1999, Wingo et al., 1998). The striking divergence in late-stage CRC diagnosis with age has also been observed by Mandelblatt et al., (1996) and they concluded that the trend is more pronounced in low SES areas. Their conclusion can explain the consistent disparity in risk among the young-aged group for the two periods, within Chicago city which has a high concentration of low SES neighborhoods. However, interaction variables of SES indicators at the ZIP code level and age did not show a statistically significant influence on late-stage diagnosis. This failure may indicate that ZIP code areas are not fine enough to show neighborhood-scale variation in SES. Smaller neighborhood areas that more closely represent the spaces people experience on a daily basis need to be applied to more accurately measure area-based SES.

The effects of gender have been found to be highly case-sensitive in previous studies (Chen et al., 2007, Mandelblatt et al., 1996, Ries et al., 1999). This research presents similar results with the contrasting impacts on late-stage CRC diagnosis in two neighboring areas, the Chicago suburbs and Chicago city. Some research suggests that different screening rates by age may cause the gender disparity (Callcut et al., 2006). However, incorporating interactions of age groups and gender in the multilevel logistic regressions did not provide any significant results, indicating that other unknown variables affect gender variation. These variables could include area-based SES at a finer scale, individual health insurance status, or more complicated combined effects of patients' individual and social behaviors. More detailed and localized research is a must to find the reasons for observed gender disparities.

Although many studies have found that black race has a significant and positive relationship with late-stage CRC (Alexander et al., 2007, Amey et al., 1997, Doubeni et al., 2007, Du et al., 2007, Henry et al., 2009, Huang et al., 2007, Krieger et al., 1999, Mayberry et al., 1995, Marcella and Miller 2001; Le et al., 2009, Lengerich et al., 2005, Pagano et al., 2003, Palmer and Schneider, 2005), the black variable did not show any statistically significant relationship with late-stage CRC in this study. There are many

potential reasons. One traces back into the colorectal cancer dataset obtained from ISCR: Error may exist in the coding of racial information, and this kind of error can vary among the different regions given that it is highly dependent on the experiences and perceptions of doctors and the varying qualities of data administration in different health facilities. Since Factor 2 (Minority Disparities) which includes the percent of black population at the ZIP code level shows a statistically significant relationship with late-stage CRC outside the Chicago region, the black variable may be influential at an aggregated level rather than individual level. Other possible reasons need to be studied in future research.

In terms of the ZIP code level spatial risk factors, shortest travel time to the nearest hospital with colonoscopy screening services, exhibited a significant and positive association with late-stage CRC diagnosis risk in areas outside the Chicago metropolitan region for 1998 to 2002. Even though potential accessibility to screening has improved over time, its significant impact in the later period demonstrates the increasing disparity of spatial accessibility to specialized providers between the Chicago-metropolitan area and other parts of the state. These other areas report some of the longest travel times to colonoscopy services, suggesting that long travel times may present a barrier to early CRC diagnosis in rural areas. Many factors may confound the influence of potential accessibility, such as individual health insurance plans, access to transportation, and specific hospital policies. However, limited by data availability, this paper can only emphasize that spatial accessibility to screening services is important.

ZIP code level SES factors have been shown to have varying impacts on the risk of late-stage CRC diagnosis in the three geographic contexts. Surprisingly, Factor 1(SES Disadvantage) was not associated with the dependent variable in any context or time period. Factor 2 (Minority Disparities) was mainly related to late diagnosis outside the Chicago metropolitan area, and Factor 3 (Cultural-Language Barriers) had varying impacts between Chicago city and suburbs. The lack of significance for SES may relate to the coarse, ZIP code scale of analysis, as discussed above, or to other inadequacies in measuring SES variables. Furthermore, bias was introduced by interpolating from the census-tract level to ZIP-code scale for SES variables in 2000, and this bias may have distorted the influence of SES on the risk of late-stage CRC diagnosis. Nonetheless, the availability of reliable SES variables for large population-based studies is very limited, and it is quite difficult to collect and measure SES indicators at a finer scale than census tract.

In addition to demographic, SES, and healthcare accessibility factors, the risk of developing CRC has also been found to be highly correlated with personal behaviors, such as dietary habits (high fat intake and high levels of red meat consumption), physical inactivity, and the consumption of alcohol and tobacco (Almendingen et al., 2000, Chen et al., 1997, Giovannucci et al., 1995, Inoue et al., 2001, Terry

et al., 2001, Thun et al., 1992). Personal vital characteristics, such as family history of CRC, obesity, and diabetes also increase the risk of CRC (Ekbom et al., 1990, Fuchs et al., 1994, Gatof and Ahnen 2002, Winawer et al., 1996). These personal risk factors may also be correlated with the risk of being diagnosed with late-stage CRC (Greenwald et al., 1996, Kern et al., 1989). Without detailed, individual-level data, it is unclear which of these well-established variables are correlated with the risk of late-stage CRC diagnosis and whether they can explain the disparities related to demographic factors, SES, and healthcare accessibility observed in this study. Additionally, personal health insurance status directly influences the accessibility of primary healthcare and CRC cancer screening services (Ayanian et al., 1993, Chen et al., 2007, Halpern et al., 2007, Halpern et al., 2008, Roetzheim et al., 1999, Palmer and Schneider, 2005). Inadequate transportation to primary care doctors' offices or colonoscopy screening facilities can also delay diagnosis (O'Malley et al., 2004, Paskett et al., 2004, Rushton et al., 2004, Zenk et al., 2006). Other barriers such as the confusing characteristics of many modern oncology centers and poor communication between oncologist and patients may inhibit patients from accessing needed cancer screening services in time (Christie et al., 2008, Dohan and Schrag, 2005, Fowler et al., 2006). Lack of data for these and other relevant variables limited the statistical analysis. These omitted variables may partly explain the inconsistent impacts of SES and gender in neighboring areas. Collecting data for these individual- and areal-level variables and incorporating them in statistical models of late-stage CRC is an important objective for future research. Additionally, Illinois has limitations as a study area, because the Chicago metropolitan area includes a large number of CRC cases compared to other areas of the state. Potential risk factors cannot be fully investigated in the non-Chicago metropolitan area because of the small sample size of cancer cases. It would be useful to study other places, such as a state with a more evenly-distributed pattern of cancer cases, to accurately investigate the risk factors of late-stage CRC in areas with low population density.

6. References

Abe T., Martin I.B., and Roche L.M. **Clusters of Census Tracts with High Proportions of Men with Distant-Stage Prostate Cancer Incidence in New Jersey, 1995 to 1999**. *American Journal of Preventive Medicine*, 2006, **30**: S60-S66.

AHA. AHA Guide 2011 Edition. American Hospital Association, Chicago, IL, 2011. Available at: http://www.ahadata.com/ahadata/html/AHAGuide.html (Accessed: January 28th, 2011)

Alexander D.D., Waterbor J., Hughes J., Funkhouser E., Grizzle W., Manne U. **African-American and Caucasian Disparities in Colorectal Cancer Mortality and Survival by Data Source: An Epidemiologic Review.** *Cancer Biomark,* 2007, **3(6)**: 301-313.

Almendingen K., Hofstad B., Teygg K. Hoff G., Hussain A., and Vatn M.H. **Smoking and Colorectal Adenomas: a Case-Control Study**. *European Journal of Cancer Prevention,* 2000, **9**: 193-203.

Amey C.H., Miller M.K., and Albrecht S.L. **The Role of Race and Residence in Determining Stage at Diagnosis of Breast Cancer**. *The Journal of Rural Health*, 1997, **13**: 99-108.

Anderson S. Senior Manager, Nation Cancer Data Base. Personal Communication by phone on July 8th, 2010.

Ayanian J.Z., Kohler B.A., Abe T., and Epstein A.M. **The Relationship between Health Insurance Coverage and Clinical Outcomes among Women with Breast Cancer**. *New England Journal of Medicine*, 1993, **329**: 326-331.

Blesch K.S., Davis F., and Kamath S.K. **A Comparison of Breast and Colon Cancer Incidence Rates among Native Asian Indians, US Immigrant Asian Indians, and Whites**. *Journal of the American Dietetic Association*,1999, **99**: 1275-1277.

Brawarsky P., Brooks D.R., and Mucci L.A. **Correlates of Colorectal Cancer Testing in Massachusetts Men and Women**. *Preventive Medicine*, 2003, **36**: 659-668.

Bryant H., and Mah Z. **Breast Cancer Screening Attitudes and Behaviors of Rural and Urban Women**. *Preventive Medicine*, 1992, **21**: 405-418.

Carr W.P., Maldonado G., Leonard P.R. Halberg J. U., Church T.R., Mandel J.H., Dowd B., Mandel J.S. **Mammogram Utilization among Farm Women**. *Journal of Rural Health*, 1996: **12(4 suppl)**: 278-290.

Callcut R.A., Kaufman S., Stone-Newsom R. Remington P., and Mahvi D. **Gender disparities in Colorectal Cancer Screening: True or False?** *Journal of Gastrointestinal Surgery*, 2006, **10**: 1409-1417.

Centers for Disease Control and Prevention. **Screening for Colorectal Cancer – United States, 1997**. *Morbidity and Mortality Weekly Report*, 1999, **48**: 116-121.

Centers for Disease Control and Prevention. **Trends in Screening for Colorectal Cancer – United States, 1997 and 1999**. *Morbidity and Mortality Weekly Report*, 2001, **50**: 162-166.

Centers for Disease Control and Prevention. **Fast Facts about Colorectal Cancer**. CDC, Atlanta, Georgia, 2008. Available at: http://www.cdc.gov/cancer/colorectal/basic_info/facts.htm. (Accessed: January 13th, 2011).

Centers for Medicare & Medicaid Services. **Colorectal Cancer Screening Overview**. CDC, Atlanta, Georgia, 2011. Available at: https://www.cms.gov/ColorectalCancerScreening/. (Accessed: January 15th, 2011).

Chen V.W., Fenoglio-Preiser C.M., Wu X.C. Coates R.J., Reynolds P., Wickerham D.L., Andrews P., Hunter C., Stemmermann G., Jackson J.S., and Edwards B.K. **Aggressiveness of Colon Carcinoma in Blacks and Whites. National Cancer Institute Black/White Cancer Survival Study Group**. *Cancer Epidemiology, Biomarkers & Prevention*, 1997: **6**: 1087-1093.

Chen A.Y., Schrag N.M., Halpern M.T., and Ward E.M. **The Impact of Health Insurance Status on Stage at Diagnosis of Oropharyngeal Cancer**. *Cancer*, 2007, **110**: 395-402.

Chien C., Morimoto L.M., Tom J., Li C.I. **Differences in Colorectal Carcinoma Stage and Survival by Race and Ethnicity.** *Cancer*, 2005, **104(3)**: 629-639.

Christie J., Hooper C., Redd W.H. Winkel G., DuHamel K., Itzkowitz S., and Jandorf L.**Predictors of Endoscopy in Minority Women**. *Journal of National Medical Association*, 2005, **97**: 1361-1368.

Cokkinides V.E., Chao A., Smith R.A., Benon S.W., and Thun M.J. **Correlates of Underutilization of Colorectal Cancer Screening among US Adults, Age 50 Years and Older**. *Preventive Medicine*, 2003, **36**: 85-91.

Committee on the Consequences of Uninsurance. **Institute of Medicine. Care Without Coverage: Too Little, Too Late**. National Academy Press, Washington DC, 2002.

Cooper R.A. **Seeking a Balanced Physician Work-Force for the 21st Century**. *Journal of the American Medical Association*, 1994, **272**: 680-687.

Cooper G.S., Yuan Z., Landefeld C.S., Johanason J.F., and Rimm A.A. **A National Population-Based Study of Incidence of Colorectal Cancer and Afe. Implications for Screening in Older Americans.** *Cancer*, 1995, **75**: 775-781.

Cooper G.S., Yuan Z., Rimm A.A. **Racial Disparity in the Incidence and Case-Fatality of Colorectal Cancer: Analysis of 329 United States Counties**. *Cancer Epidemiological Biomarkers & Prevention*, 1997, **6**: 283-285.

Coughlin S.S., Thompson T.D., Seeff L., Richards T., and Stallings F. **Breast, Cervical, and Colorectal Carcinoma Screening in a Demographically Defined Region of the Southern U.S**. *Cancer*, 2002, **95**: 2211-2222.

Devesa S.S., Grauman D.J., Blot W.J., Pennello G. A., Hoover R.N., and Fraumeni J.F. **Atlas of Cancer Mortality in the United States: 1950-94**. National Institute of Health, National Cancer Institute (NIH Publication No.99-4564), 1999.

Diez-Roux A.V. **Bringing Context Back into Epidemiology: Variables and Fallacies in Multilevel Analysis**. *American Journal of Public Health*, 1998, **88(2)**: 216-222.

Diez-Roux A.V., Keife C.I., Jacobs D.R.Jr, Haan M., Jackson S.A., Neto F.J., Paton C.C., and Schulz R.**Area Characteristics and Individual-level Socioeconomic Position Indicators in Three Population-Based Epidemiologic Studies**. *Annals of Epidemiology*, 2001, **11(6)**: 395-405.

Dohan D., and Schrag D. **Using Navigators to Improve Care of Underserved Patients: Current Practices and Approaches**. *Cancer*, 2005: **104**: 848-855.

Doubeni C.A., Field T.S., Buist D.S. Korner E.J., Bigelow C., Lamerato L., Herrington L., Quinn V.P., Hart G., Hornbrook M.C., Gurwitz J. H., and Wager E.W. **Racial Differences in Tumor Stage and Survival for Colorectal Cancer in an Insured Population**. *Cancer*, 2007, **109**: 612—620.

Du X.L., Meyer T.E., and Franzini L. **Meta-analysis of Racial Disparities in Survival in Association with Socioeconomic Status among Men and Women with Colon Cancer**. *Cancer*, 2007, **109(11)**: 2161-2170.

Ekbom A., Helmick C., Zack M. Adami H-O. **Ulcerative Colitis and Colorectal Cancer: A Population-based Study**. *The New England Journal of Medicine*, 1990, **323**: 1228-1233.

Elnicki D.M., Morris D.K., Shockcor W.T. **Patient-perceived Barriers to Preventive Health Care among Indigent, Rural Appalachian Patients**. *Archives of Internal Medicine*, 1995, **155**: 421-424.

ESRI. **ESRI Data & Maps 2000. An ESRI White Paper.** ESRI, Redland, California, 2001. Available at: http://www.geobotany.org/library/pubs/ESRI2000_digchtworld.pdf (Accessed: January 13th, 2010).

ESRI. (2006). **ArcGIS 9: ArcGIS Network Analyst Tutorial**. ESRI, Redland, California.

Fazio L., Cotterchio M., Manno M., McLaughlin J., and Gallinger S. **Association between Colonic Screening, Subject Characteristics, and Stage of Colorectal Cancer.** *The American Journal of Gastroenterology*, 2005, **100**: 2531-2539.

Fairley T.L., Cardinex C.J., Martin J., Alley L., Friedman C., Edward B., and Jamison P. **Colorectal Cancer in U.S. Adults Younger than 50 Years of Age, 1998-2001**. *Cancer*, 2006, **107**: 1153-1161.

Fisher P.F., and Langford M. **Modeling Sensitivity to Accuracy in Classified Imagery: A Study of Areal Interpolation by Dasymetric Mapping**. *Professional Geographer*, 1996, **48**: 299-309.

Flood D.M., Weiss N.S., Cook L.S., Emerson S.M., Schwartz, and Potter J.D. **Colorectal Cancer Incidence in Asian Migrants to the United States and Their Descendants**. *Caner Causes Control*, 2000, **11**: 403-411.

Flowerdew R., and Green M. **Areal Interpolation and Types of Data**. In: Fotheringham A.S., and Rogerson P. Editors, Spatial Analysis and GIS, Taylor and Francis, London, 1994: 121-145.

Fowler T., Steakley C., Garcia A.R. Kwok J., and Bennett M. **Reducing Disparities in the Burden of Cancer: the Role of Patient Navigators**. *PLoS Medicine*, 2006, **3**: e193.

Fuchs C.S., Giovannucci E.L., Colditz G.A. Huner D.J., Speizer F.E., and Willett W.C. **A Prospective Study of Family History and Risk of Colorectal Cancer**. *The New England Journal of Medicine*, 1994, **331**: 1669-1674.

Gatof D., and Ahnen D. **Primary Prevention of Colorectal Cancer: Diet and Drugs**. *Gastroenterology Clinics of North America*, 2002, **31**: 587-623, xi.

Godreau C.J. **Office-Based Colonoscopy in a Family Practice**. *The Family Practice Research Journal*, 1992, **12**: 313-320.

Giovannucci E., Rimm E., Ascherio A., Stampfer M. J., Colditz G.A., and Willett W.C. **Alcohol, Low-Methionine-Low-Folate Diets, and Risk of Colon Cancer in Men**. *The Journal of National Cancer Institute*, 1995, **87**: 265-273.

Gomez S.L., O'Malley C.D., Stroup A., Shema S.J., and Satariano W.A. **Longitudinal, Population-based Study of Racial/Ethnic Differences in Colorectal Cancer Survival: Impact of Neighborhood Socioeconomic Status, Treatment and Comorbidity**. *BMC Cancer*, 2007, **7(193)**: 1-19.

Greenwald H.P., Borgatta E.F., McCorkel R., and Polissar N. **Explaining Reduced Cancer Survival among the Disadvantaged**. *The Milbank Quarterly,* 1996, **74(2):** 215-238.

Gregorio D.I., Kulldorff M., Barry L., and Samociuk H. **Geographic Differences in Invasive and in Situ Breast Cancer Incidence according to Precise Geographic Coordinates, Connecticut, 1991-95**. *International Journal of Cancer*, 2002, **100**: 194-198.

Griffith D., and Amrhein C. **Multivariate Statistical Analysis for Geographers**. New Jersey: Prentice-Hall, 1997.

Grubesic T.H., and Matisziw T.C. **On the Use of Zip Codes and Zip Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data**. *International Journal of Health Geographics*, 2006, **5**:58.

Grubesic T.H. **Zip Codes and Spatial Analysis: Problems and Prospects**. *Socio-Economic Planning Sciences*, 2008, **42**: 129-149.

Guagliardo M.F. **Spatial Accessibility of Primary Care: Concept, Methods and Challenges**. *International Journal of Health Geographics*, 2004, **3**:3.

Huang L., Pickle L.W., Stinchcomb D., and Feuer E.J. (2007). **Detection of Spatial Clusters: Application to Cancer Survival as a Continuous Outcome**. *Epidemiology*, 2007, **18(1)**: 73-87.

Halpern M.T., Bain J., Ward E.M., Schrag N.M., and Chen A.Y. **Insurance status ad Stage of Cancer at Diagnosis among Women with Breast Cancer**. *Cancer*, 2007, **110**: 403-411.

Haplern M.T., Ward E.M., Pavluck A.L., Schrag N.M., Bain J., and Chen A.Y. **Association of Insurance Status and Ethnicity with Cancer Stage at Diagnosis for 12 Cancer Sites: a Retrospective Analysis**. *Lancent Oncology*,2008, **9**: 222-231.

Harold J. B. and Winder E.P. **Primary Care for Survivors of Breast Cancer**. *The New England Journal of Medicine*, 2000, **10**: 1086-1094.

Hart L.G., Larsen E.H., and Lishner D.M. **Rural Definitions for Health Policy and Research**. *American Journal of Public Health*, 2005, **95**: 1149-1155.

Hawley S.T., Foxhall L., Vernon S.W., Levin B., and Young L.E. **Colorectal Cancer Screening by Primary Care Physicians in Texas: A Rural-Urban Comparison**. *Journal of Cancer Education*, 2001, **16**: 199-204.

Henry K.A., Sherman R., and Roche L.M. **Colorectal Cancer Stage at Diagnosis and Area Socioeconomic Characteristics in New Jersey**. *Health & Place*, 2009, **15**: 505-513.

Holmers-Rovner M., Williams G.A., Hoppough S., Quillan L., Bulter R., and Given C.W. **Colorectal Cancer Screening Barriers in Persons with Low Income.** *Cancer Practice*, 2002, **10**: 240-247.

Huang L., Pickle L.W., Stinchcomb D., and Feuer E.J. **Detection of Spatial Clusters: Application to Cancer Survival as a Continuous Outcome.** *Epidemiology*, 2007, **18(1)**: 73-87.

Huff D.L. **Don't Misuse the Huff Model in GIS**. *Business Geographies*, 2000, **8(8)**, 12.

Hsu C.E., and Mas F.S. **Surveillance of the Colorectal Cancer Disparities among Demographic Subgroups: A Spatial Analysis**. *Southern Medical Journal*, 2006, **99**: 949-956.

Inoue H., Kiyohara C., Shinomiya S. **Glutathione S-transferase Polymorphisms and Risk of Colorectal Adenomas**. *Cancer Letters*, 2001, **91**: 585-587.

Jemal A., Kulldorff M., Devesa S.S., Hayes R.B., and Fraumeni J.F. Jr. **A Geographic Analysis of Prostate Cancer Mortality in the United States, 1970-89**. *International Journal of Cancer*, 2002, **101**: 168-174.

Jemal A., Ward E., Wu X., Martin H.J., McLaughlin C.C., and Thun M.J. **Geographic Patterns of Prostate Mortality and Variations in Access to Medical Care in the United States**. *Cancer Epidemiology Biomarkers & Prevention*, 2005, **14**: 590-595.

Joseph A.E. and Bantock P.R. **Measuring Potential Physical Accessibility to General Practitioners in Rural Areas: A Method and Case Study**. *Social Science and Medicine*, 1982, **16**: 85-90.

Joseph A.E., and Phillips D.R. **Accessibility and Utilization-Geographical Perspectives on Health Care Delivery**. Happer & Row Publishers, New York, 1984, 214pp.

Kern S.E., Fearon E.R., Tersmette K.W.F., Enterline J.P., Leppert M., Nakamura Y., White R., Vogelstein B., Hamilton S.R. **Allelic Loss in Colorectal Carcinoma**. *The Journal of the American Medical Association*, 1989, **261(21):** 3099-3103.

Kawachi I., and Berkman L.F. **Neighborhoods and Health**. New York, Oxford University Press, 2003.

Klabunde C.N., Schenck A.P., and Davis W.W. **Determinants of Black/White Differences in Colon Cancer Survival**. *American Journal of Preventive Medicine*, 2006, **30**: 313-319.

Krieger N., Quesenberry C. Peng T., Pamela H_R., Stewart S., Brown S., Swallen K., Guillermo T., Suh D., and Alvarez-Martinez L et al. **Social Class, Race/Ethnicity, and Incidence of Breast, Cervix, Colon, Lung, and Prostate Cancer among Asian, Black, Hispanic, and White Residents of the San Francisco Bay Area, 1988-92 (United States).** *Cancer Causes Control*, 1999, **10**: 525-537.

Krieger N., Chen J.T., Waterman P.D., Rehkopf D.H., Yin R., and Coull B.A. **Race/Ethnicity and Changing US Secioeconomic Gradients in Breast Cancer Incidence: California and Massachusetts, 1978-2002 (United States)**. *Cancer Causes Control*, 2006, **17(2)**: 217-226.

Kulldorff M., Feuer E.J., Miller B.A., and Freedman L.S. **Breast Cancer Clusters in the Northeast United States: a Geographic Analysis**. *American Journal of Epidemiology*, 1997, **146**: 161-170.

Kulldorff M., Tango T., and Park P. **Power Comparisons for Disease Clustering Tests**. *Computational Statistics & Data Analysis*, 2003, **42**: 665-684.

Lee R.C. **Current Approaches to Shortage Area Designation**. *Journal of Rural Health*, 1991, **7**: 437-450.

Le H., Ziogas A., Taylor T., Lipkin S., and Zell J. **Survival of Distinct Asian Groups Among Colorectal Cancer Cases in Californi**a. *Cancer*, 2009, **115**: 259-270.

Lehnerr M., and Havener L. **Assessment of Interstate Exchange of Cancer Data: Illinois, 1986-1998**. Springfield, Ill: Illinois State Department of Public Health, 2002.

Le Marchand L., Wilkens L.R., Kolonel L.N., Hankin J.H., and Lun L.-C. (**Associations of Sedentary Lifestyle, Obesity, Smoking, Alcohol Use, and Diabetes with the Risk of Colorectal Cancer**. *Cancer Research*, 1997, **57**: 4787-4794.

Lengerich E.J., Rubio A., Brown P.K., Knight E.A., and Wyatt S.W. **Results of Coordinated Investigations of a National Colorectal Cancer Education Campaign in Appalachia** [serial online]. *Preventing Chronic Disease*, 2005, **3**: A32.

Luo W., and Wang F.H. **Measure of Spatial Accessibility to Health Care in a GIS Environment: Synthesis and a Case Study in the Chicago Region**. E*nvironment and Planning B: Planning and Design*, 2003, **30**: 865-884.

Luo L. McLafferty S., and Wang F.H. **Analyzing Spatial Aggregation Error in Statistical Models of Late-Stage Cancer Risk: A Monte Carlo Simulation Approach**. *International Journal of Health Geographics*, 2010, **9**:51.

Luo W. **Using a GIS-Based Floating Catchment Method to Assess Areas with Shortage of Physicians.** *Health and Place*, 2004, **10**: 1-11.

Luo W., and Qi Y. **An Enhanced Two-Step Floating Catchment Area (E2SFCA) method for Measuring Spatial Accessibility to Primary Care Physicians**. *Health & Place*, 2009, **15**: 1100-1107.

MacKinnon J.A., Duncan R.C., Hunag Y., Lee D.J., Fleming L.E., Voti L., Rudolph M., and Wilkinson J.D. **Detecting an Association between Socioeconomic Status and Late Stage Breast Cancer Using Spatial Analysis and Area-Based Measures**. *Cancer Epidemiology Biomakers & Prevention*, 2007, **16**: 756-762.

Mandelblatt J., Andews H., Kao R., Wallace R., and Kerner J. **The Late-Stage Diagnosis of Colorectal Cancer: Demographic and Socioeconomic Factors.** *American Journal of Public Health*, 1996, **86**: 1794-1797.

Marcella S., Miller J.E. **Racial Differences in Colorectal Cancer Mortality: The Importance of Stage and Socioeconomic Status**. *Journal of Clinical Epidemiology*, 2001, **54**: 359-366.

Mayberry R.M., Coates R.J., Hill H.A. Click L.A., Chen V.W., Austin D.F., Redmond C.K., Fenoglio-Preiser C.M., Hunter C.P., Haynes M.A., Muss H.B., Wesley M., N., Greenberg R.S., and Edwards B.K. **Determinants of Black/White Differences in Colon Cancer Survival**. *Journal of National Cancer Institute*, 1995, **87**: 1686-1693.

McLafferty S., and Wang F.H. **Rural Reversal? Rural-Urban Disparities in Late-Stage Cancer Risk in Illinois**. *Cancer*, 2009, **15**: 2755-2764.

McMahon L.F.Jr., Wolfe R.A., Huang S. Tedeschi P., Manning W.Jr., and Edlund M. **Racial and Gender Variation in Use of Diagnostic Colonic Procedures in the Michigan Medicare Population**. *Medical Care*, 1999, **37**: 712-717.

McMahon P.M., and Gazelle G.S. **Colorectal Cancer Screening Issues: A Role for CT Colonography?** *Abdominal Imaging*, 2002, **27(3)**: 235-243.

Mullins D. **An Overview of Cancer Economics** (based on a presentation). *American Journal of Managed Care*, 1999, **5S**: 371-376.

National Cancer Institute. Dictionary of Cancer Terms: Late-Stage Cancer. National Cancer Institute, Bethesda, Maryland, 1999. Available at: http://www.cancer.gov/dictionary?CdrID=561600 (Accessed: April 3rd, 2011).

Nelson R.L., Persky V., and Turyk M. **Carcinoma in Situ of the Colorectum: SEER Trends by Race, Gender, and Total Colorectal Cancer.** *Journal of Surgical Oncology*, 1999, **71**: 123-129.

Nelson D.E., Bolen J., Marcus S., Wells H.E., and Meissner H. **Cancer Screening Estimates for U.S. Metropolitan Areas**. *American Journal of Preventive Medicine*, 2003, **24**: 301-309.

Newman A.M., and Spengler R.F. **Cancer Mortality among Immigrant Populations in Ontario, 1969 through 1973**. *Canadian Medical Association Journal*, 1984, **130(15)**: 399-405.

O'Malley A.S., Forrest C.B., Feng S. B., and Mandelblatt J. **Disparities despite Coverage: Gaps in Colorectal Cancer Screening, among Medicare Beneficiaries**. *Archives of Internal Medicine*, 2005, **165**: 2129-2135.

O'Malley A.S., Beaton E., Yabroff K.R. Abramson R., Mandelblatt J. **Patient and Provider Barriers to Colorectal Cancer Screening in the Primary Care Safety-Net**. *Preventive Medicine*, 2004, **39(1)**: 56-63.

Paskett E.D., Tatum C., Rushing J., Michielutte R., Bell R., Foley K.L., Bittoni M., and Dickinson S. **Racial Differences in Knowledge, Attitudes, and Cancer Screening Practices among A Triracial Rural Population**. *Cancer*, 2004, **101(11)**: 2650-2659.

Pagano I.S., Morita S.Y., Dhakal S., Hundahl S.A., and Maskarinec G. **Time Dependent Ethnic Convergence in Colorectal Cancer Survival in Hawaii.** *BMC Cancer*, 2003, **3**:5.

Palmer R.C., and Schneider E.C. **Social Disparities across the Continuum of Colorectal Cancer: A Systematic Review**. *Cancer Causes and Control*, 2005, **16**: 55-61.

Pandya K.J., McFadden E.T., Kalish L.A., Tormey D.C., Taylor S.G. IV, and Falkson G. **A Retrospective Study of Earliest Indicators of Recurrence in Patients on Eastern Cooperative Oncology Group Adjuvant Chemotherapy Trials for Breast Cancer: a Preliminary Report**. *Cancer*, 1985, **55**: 202-205.

Paquette I., and Finlayson S.R. **Rural versus Urban Colorectal and Long Cancer Patients: Differences in Stage at Presentation**. *Journal of the American College Surgeons*, 2007, **205**: 636-641.

Phillips D.R. **Health and Health Care in the Third World**. Longman Scientific & Technical, Harlow, Essex, England, 1990, 334pp.

Pine M., Jordan H.S., Elixhauser A., Fry D.E., Hoaglin D.C., Jones B., Meimban R., Warner D., and Gonzales J. **Enhancement of Claims Data to Improve Risk Adjustment of Hospital Mortality**. *The Journal of the American Medical Association*, 2007, **297**: 71-76.

Pollack L.A., Gotway C.A., Bates J.H., Parikh-Patel A., Richards T.B., Seeff L.C., Hodges H., and Kassim S. **Use of the Spatial Scan Statistic to Identify Geographic Variations in Late Stage Colorectal Cancer in California (United States).** *Cancer Causes & Control*, 2006, **17**: 449-457.

Polite B.N., Dignam J.J., and Olopade O.I. **Colorectal Cancer Model of Health Disparities: Understanding Mortality Differences in Minority Populations**. *Journal of Clinical Oncology*, 2006, **24**: 2179-2187.

Ries L.A.G., Kosary C.L., Hankey B.F., Miller B.A., Edwards B.K (eds) **SEER Cancer Statistics Review**, **1973-1996**. NIH Publication No., 99-2789. Bethesda, MD: National Cancer Institute, 1999.

Ries L.A.G., Wingo PA., Miller D.S. Howe H.L., Weir H.K., Rosenberg H.M., Vernon S.W., Cronin K., and Edwards B.K. **The Annual Report to the Nation on the Status of Cancer, 1973-1997, With a Special Section on Colorectal Cancer**. *Cancer*, 2000, **88**: 2398-2424.

Roetzheim R.G., Pal N., Tennant C., Voti L., Ayanian J.Z., Schwabe A., and Krischer J.P. **Effects of Health Insurance and Race on Early Detection of Cancer**. *Journal of the National Cancer Institute*, 1999, **91**: 1409-1415.

Roche L.M., Skinner R., and Weinstein R.B. **Use of a Geographic Information System to Identify and Characterize Areas with High Proportions of Distant Stage Breast Cancer**. *Journal of Public Health Management & Practice*, 2002, **8**: 26-32.

Rossi S., Cinini C., Di Pietro C., Lombardi C.P., Crucitti A., Bellantone R., and Crucitti F. **Diagnostic Delay in Breast Cancer: Correlation with Disease Stage and Prognosis**. *Tumori*, 1990, 76(6): 559-562.

Rushton G., Peleg I., Banerjee A., Smith G., and West M. **Analyzing Geographic Patterns of Disease Incidence: Rates of Late-Stage Colorectal Cancer in Iowa**. *Journal of Medical Systems*, 2004, **28(3)**: 223-235.

SAS. **Production GLIMMIX Procedure**. SAS Institute, Cary NC, 2006. Available at: http://support.sas.com/rnd/app/da/glimmix.html (Accessed: Nov 25th, 2010).

SAS. **SAS/STAT® 9.22 User's Guide: The Factor Procedure**. SAS Institute Inc, Cary, NC, 2011. Available at:

http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#factor_toc.htm (Accessed: Nov 2nd, 2010).

Seeff L.C., Nadel M., Blackman C., and Pollack L.A. **Colorectal Cancer Test Use among Persons Aged Greater than Equal to 50 Years-United States, 2001**. *Morbidity and Mortality Weekly Report*, 2003, **52**: 193-195.

Seeff L.C., Nadel M.R., Klabunde C.N. Thompson T., Shapiro J.A., Vernon S.W., and Coates R.J. **Patterns and Predictors of Colorectal Cancer Test Use in the Adult U.S. Population**. *Cancer*, 2004, **100**: 2093-2103.

Schabenberger O. **Introducing the GLIMMIX Procedure for Generalized Linear Models**. SUGI 30, Cary, NC: SAS Institute, 2007.

Schneider E.C. **Chapter 7 Disparities and Colorectal Cancer.** *Toward the Elimination of Cancer Disparities*, H.K. Koh ed., DOI 10.107/978-0-387-89443-0_7. Springer Science+Business Media, LLC, 2009.

Sheehan T.J., DeChello L.M., Kulldorff M., Gregorio D.I., Gershman S., and Mroszczyk M.**The Geographic Distribution of Breast Cancer Incidence in Massachusetts 1988 to 1997, Adjusted for Covariates**. *International Journal of Health Geographics*, 2004, **3**:17.

Shen Q. **Location Characteristics of Inner-City Neighborhoods and Employment Accessibility of Low-Income Workers**. *Environment and Planning B: Planning and Design*, 1998, **25**: 345-365.

Shih Y.C., Zhao L., and Elting L.S. **Does Medicare Coverage of Colonoscopy Reduce Racial/Ethnic Disparities in Cancer Screening among the Elderly?** *Health Affairs* (Millwood), 2006, **25**: 1153-1162.

Singh G., Miller B.A., Hankey B.F., and Edwards B.K. **In NCI Cancer Surveillance Monograph Series**. National Cancer Institute, Bethesda, MD, 2003.

Staszewski J., and Haenszel W. **Cancer Mortality among the Polish-Born in the United States**. *Journal of the National Cancer Institute*, 1965, **35**: 291-297.

Terry P., Ekbom A., Lichtenstein P. Feychiting M., and Wolk A. Long-**Term Tobacco Smoking and Colorectal Cancer in a Prospective Cohort study**. *International Journal of Cancer*, 2001, **91**: 585-587.

The Polish American Association. **The Polish Community in Metro Chicago: A Community Profile of Strengths and Needs**. **A Census 2000 Report.** Published by the Polish American Association, 2004.

Available at: http://www.robparal.com/downloads/Polish%20Community%20in%20Chicago.pdf (Accessed: March 23rd, 2011).

Thouez J.M., Bodson P., and Joseph A.E. **Some Methods for Measuring the Geographic Accessibility of Medical Service in Rural Regions**. *Medical Care*, 1988, **26(1)**: 34-44.

Thun M.J., Calle E., Namboodiri M.M., Flanders W.D., Coates R.J., Byers T., Boffetta P., Garfinkel L., and Heath C.W. **Risk Factors for Fatal Colon Cancer in a Large Prospective Study**. *Journal of National Cancer Institute*, 1992, **84**: 1491-1500.

Thomas A., and Carlin B.P. **Late Detection of Breast and Colorectal Cancer in Minnesota Counties: An Application of Spatial Smoothing and Clustering**. *Statistics in Medicine*, 2003, **22**: 113-127.

Tobler W.R. **Smooth Pycnophalactic Interpolation for Geographic Regions**. *Journal of the American Statistical Association*, 1979, **74**: 519-536.

VanEenwyk J., Campo J.S., and Ossiander E.M. **Socioeconomic and Demographic Disparities in Treatment for Carcinomas of the Colon and Rectum**. *Cancer*, 2002, **9**: 39-46.

U.S. Census Bureau. **1990 Census: Sample Data-Summary Tape File 3**. U.S. Census Bureau, Washington D.C. Available at: http://www.census.gov/main/www/cen1990.html (Accessed; Nov 23rd, 2010).

U.S. Census Bureau. **United States Census 2000: Summary File 3 (SF3).** U.S. Census Bureau, Washington D.C. Available at: http://www.census.gov/census2000/sumfile3.html (Accessed: Dec 15th, 2010).

U.S. Census Bureau. **Census 2000 Tiger/Line Files**. U.S. Census Bureau, Washington D.C. Available at: http://www.census.gov/geo/www/tiger/tiger2k/othertgr.html (Accessed: May 15th, 2010).

U.S. Census Bureau. Census 2000 ZCTAs: Zip Code Tabulation Areas Technical Documentation. U.S. Census Bureau, Washington D.C. Available at:http://www.census.gov/geo/ZCTA/zcta_tech_doc.pdf (Accessed: Nov 4th, 2011).

U.S. Census Bureau. **Cartographic Boundary Files: 1990 Census County Subdivisions in ArcView Shapefile (.shp) Format.** Illinois-cs17_d90_shp.zip. U.S. Census Bureau, Washington D.C. Available at: http://www.census.gov/geo/www/cob/cs1990.html#shp (Accessed: Nov 19th, 2010).

U.S. Census Bureau. **Cartographic Boundary Files: Census 2000 County Subdivisions in ArcView Shapefile (.shp) Format**. Illinois-cs17_d00_shp.zip. U.S. Census Bureau, Washington D.C. Available at: http://www.census.gov/geo/www/cob/cs2000.html#shp (Accessed: Nov 22nd, 2010).

U.S. Department of Health and Human Services. Health Resources and Services Administration. **Find Shortage Areas: HPSA by State & County**. U.S. Department of Health and Human Services, Rockville, MD, 2011. Available at: http://hpsafind.hrsa.gov/ (Accessed: April 24th, 2011).

VanEenwyk J., Campo J.S. and Ossiander E.M. **Socioeconomic and Demographic Disparities in Treatment for Carcinomas of the Colon and Rectum**. *Cancer*, 2002, 95: 39-46,

Wang F., and Minor W.W. **Where the Jobs Are: Employment Access and Crime Patterns in Cleveland**. *Annals of the Association of American Geographers*, 2002, **92**:435-450.

Wang F.H., Luo L. and McLafferty S. **Healthcare Access, Socioeconomic Factors and Late-Stage Cancer Diagnosis: An Exploratory Spatial Analysis and Public Policy Implication**. *International Journal of Public Policy*, 2010, **5(2-3)**: 237-258.

Walsh J., and Terdiman J.P. **Colorectal Cancer Screening: Scientific Review.** *Journal of the American Medical Association*, 2003, **289(10)**: 1288-1296.

Weibull J.M. **An Axiomatic Approach to the Measurement of Accessibility**. *Regional Science and Urban Economics*, 1976, **6**: 357-379.

Willett W.C., Stampfer M.j., Colditz G.A., Rosner B.A., and Speizer F. E. **Relation of Meat, Fat and Fiber Intake to the Risk of Colon Cancer in a Prospective Study among Women**. *The New England Journal of Med*, 1990, **323**: 1664-1672.

Winchester D.P., Sener S.F., Khandekar J.D., Cunningham M.P., Caprino J.A., Burkett F., and Scanlon E. **Symptomatology as an Indicator of Recurrent or Metastatic Breast Cancer**. *Cancer*, 1979, **43**: 956-960.

Wingo P.A., Ries L.A., Parker S.L., and Heath C.W. Jr. **Long-Term Cancer Patient Survival in the United States**. *Cancer Epidemiological Biomarkers & Prevention*, 1998, **7**: 271-282.

Wu C.X., Chen V.W., Steele B., Ruiz B., Fulton J., Liu L., Carozza S.E., and Greenlee R. **Subsite-specific Incidence Rate and Stage of Disease in Colorectal Cancer by Race, Gender, and Age Group in the United States, 1992-1997**. *Cancer*, 2001, **92(10)**: 2547-2554.

Yang D., Goerge R., and Mullner R. **Comparing GIS-Based Methods of Measuring Spatial Accessibility to Health Services**. *Journal of Medical Systems*, 2006, **30(1)**: 23-32.

Young J.L., Jr., Roffers S.D., Ries L.A.G., Fritz A.G., and Hurlbut A.A.(eds) (2001). **SEER Summary Staging Manual-2000: Codes and Coding Instructions**. NIH Pub. No. 01-4969. Bethesda, MD: National Cancer Institute, 2001.

Zenk S.N., Tarlov E., and Sun J. **Spatial Equity in Facilities Providing Low- or No-Fee Screening Mammography in Chicago Neighborhoods**. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 2006, **83(2)**: 195-210.

Chapter II[1]

Analyzing Spatial Aggregation Error in Statistical Models of Late-Stage Cancer Risk: A Monte Carlo Simulation Approach

1. Introduction

Detecting and analyzing spatial aggregation error in large spatial data sets is an increasingly important topic in GIS and public health research (Gregorio et al., 2005, Hewko et al., 2002, Hillsman and Rhoda, 1978, Hodgson et al., 1997, Shi 2007). Spatial aggregation error arises because of the agglomeration of individual, georeferenced observations into larger spatial zones. The spatial aggregation process smoothes local variation, leading to errors in measurement of geographical variables. This error in turn affects the estimation of statistical models that incorporate spatially-aggregated variables. Spatial aggregation error is particularly important in cancer research, given that cancer data sets are often only released publicly at the ZIP code level due to privacy and confidentiality issues (Rushton et al., 2006). Thus, studies that use ZIP code-level data to examine the associations between geographical and environmental variables and cancer incidence may be adversely affected by spatial aggregation error. Although spatial aggregation error has been widely investigated, few studies have examined how spatial aggregation error affects the statistical analysis of cancer data at the ZIP code level. This study estimates the potential impact of spatial aggregation error on the parameter values of multilevel statistical models which analyze the association between spatial accessibility to mammography facilities and late-stage breast cancer risk. This study focuses on breast cancer based on the fact that it is the most common cancer among women and an important cause of cancer mortality in Illinois (Wang et al., 2010).

This study develops a Monte Carlo simulation procedure for disaggregating cancer cases from larger to smaller study units in empirical simulations, and uses that procedure to examine the implications of spatial aggregation error for multilevel model coefficients. The context sensitivity of spatial

[1] This paper was already published in *International Journal of Health Geographics*, 2010, 9:51.

Luo L., McLafferty S., and Wang F.H. Analyzing Spatial Aggregation Error in Statistical Models of Late-Stage Cancer Risk: A Monte Carlo Simulation Approach. *International Journal of Health Geographics*, 2010, 9:51.

aggregation error is also examined by comparing two study areas. This paper is divided into the following sections: literature background; data pre-processing and analytical methodology; description and analysis of results; and conclusion.

2. Background

In many scientific disciplines, data are collected at a spatial scale appropriate to the research question of interest. However, in geography and public health, much data is publicly available to researchers for analysis in predefined areas (zones) with an arbitrary and modifiable boundary. These zones were not optimally designed to answer the research question, thus introducing geographical bias which affects subsequent statistical analyses based on such data. This is the well-known Modifiable Area Unit Problem (MAUP). One of the most common consequences of this problem is the ecological fallacy.

The ecological fallacy arises when making inferences from higher to lower levels of analysis (Johnston, 2000). The model coefficients estimated based on aggregated data differ from those at the individual level, leading to errors of interpretation (Gehlke and Biehl, 1934, Openshaw and Taylor, 1979, Robinson, 1950, Yule and Kendall, 1950). Gehlke and Biehl (1934) found that the magnitude of the correlation coefficient increased with aggregation. Openshaw and Taylor (1979) demonstrated the impact on correlation coefficients of spatial aggregation of data from smaller to larger geographic zones. As in earlier work on the ecological fallacy, they found that spatial aggregation tends to increase the magnitude of correlation coefficients, confirming that spatial aggregation error has an impact on statistical analysis. Spatial aggregation error is an example of biased inference caused by the mismatch between spatial units and the research question of interest. It particularly occurs when a large area or a single point is employed to represent spatially distributed individuals (Hodgson et al., 1997). Hillsman and Rhoda (1978) identified three types of the spatial aggregation error that arise when estimating a population's average distance to the nearest service facility. The three types of error are based on different geographical characteristics of origins and destinations and can result in under- or over-estimation of individuals' actual travel distances.

Recently, with the rapid expansion of computational resources and GIS, spatial aggregation error has been studied more thoroughly. Researchers have adopted different approaches to evaluate the influence of spatial aggregation error in large study areas. Hewko et al., (2002) analyzed the spatial aggregation error associated with the measurement of neighborhood spatial accessibility (NSA). Neighborhood spatial accessibility describes the ease with which residents can travel to service facilities, and it can be approximated by the network distance from home to the closest facility. Because population data are typically aggregated to zones (census tracts, zip codes), distance is calculated from zonal

centroids to facilities resulting in spatial aggregation error. Hewko et al. (2002) compared three methods for estimating distance: one involves the use of unweighted (geographic centroids) while the others incorporate finer-scale, block-level population data thus reducing spatial aggregation error. Comparing the NSA values based on these three methods, the authors concluded that spatial aggregation error does create bias, but the impact varies with the type of centroids and the number and locations of service destinations. Spatial autocorrelation tests were also affected. Fortney, Rost & Warren (2000) studied the impact of spatial aggregation error on measures of spatial accessibility to physicians. Their results showed substantial differences between area centroid-based estimates of distance to physicians and distances calculated from individual residences, confirming that spatial aggregation error leads to significant "errors in variables" in measuring spatial accessibility.

Gregorio et al., (2005) studied the impact of spatial aggregation on tests of spatial clustering. They compared the analysis of spatial clustering of late-stage cancer in Connecticut using cancer data at different geographic scales – block group, census tract and town. Results showed little difference in the outcomes of spatial clustering tests using data at different scales. In this example, the impacts of spatial aggregation error were minimal, contradicting the aforementioned literature and suggesting the need for further analysis of the issue.

Examining spatial aggregation error requires the use of high resolution data; however, such high resolution data is often not available due to privacy and confidentiality restrictions (Rushton et al., 2006). Although with proper approvals, some health departments do provide access to high resolution data; in many cases it is only possible to obtain cancer data at a low spatial resolution such as county or ZIP code. ZIP codes are devised by the U.S. Postal Service to facilitate mail delivery, and each ZIP code comprises a set of mail distribution points which can be joined to create ZIP code areas. ZIP codes vary greatly in geographic and population size, with an average population size of 30,000 in 2000 (Krieger et al., 2002). The large and variable sizes of ZIP codes, and the fact that they are not well-defined geographic zones, pose challenges for spatial analysis of health data.

Using large-area data increases the risk of spatial aggregation error. Recently, some authors have used Monte Carlo methods to analyze spatial aggregation error by assigning data from larger to smaller zones based on the demographic characteristics of individual cancer cases (Henry and Boscoe, 2001, Shi, 2007). To obtain cancer data with a high resolution and reduce spatial aggregation error, Henry and Boscoe (2008) used demographically-based geo-imputation to assign cancer cases from ZIP codes to census tracts. Cases were assigned to tracts based on their age, gender and racial characteristics, and cases were more likely to be assigned to tracts whose populations have similar demographic characteristics. To

test the geographic accuracy of the assignment, the authors obtained data on the actual residential locations of cancer cases. The actual census tract of residence was compared to the tract assigned via geo-imputation. They found that the validity and reliability of the geo-imputation outcomes were dependent on demographic variables; that is, using race/ethnicity in geo-imputation provided a more accurate disaggregation than the one utilizing population only. The authors also detected that the geo-imputation performed differently within different census tracts. Homogeneous census tracts were more likely to have a low match rate than more heterogeneous ones.

Spatial aggregation error can also arise when using hybrid data with point- and polygon-levels (Bonner et al., 2003, Krieger et al., 2001, McElroy et al., 2003). Some methods of analysis require a consistent set of geographic units, so that hybrid data require conversion of data from either point to polygon or vice versa. If points are aggregated to corresponding polygons, however, localized information from point-level data is lost (Jacquez and Waller, 1999). An alternative approach is to convert polygon data to point data. For example, one can assign random locations to observations within polygons, and repeat the process many times using Monte Carlo simulation to estimate uncertainty (Jacquez and Jacquez, 1999). Shi (2009) devised a restricted Monte Carlo method to assign polygon-level addresses into suitable random point locations in investigating spatial variation in lung cancer incidence in New Hampshire. The method was employed to detect spatial clusters of high cancer incidence while incorporating spatial uncertainty associated with imprecise address locations. By quantifying uncertainty, this approach provides an indication of the error associated with spatial aggregation.

Although previous studies have emphasized the importance of spatial aggregation error and developed methods to reduce its effects, less is known about the impacts of spatial aggregation error on statistical estimates of model coefficients. Two recent studies investigate this issue with respect to positional error – a form of geographic error-in-variables that is similar to spatial aggregation error. Positional error occurs when residences are placed at incorrect locations due to errors in geocoding and inaccuracies in street network information. Griffith et al., (2007) studied the impacts of positional error on spatial regression analysis by comparing analytical results using datasets with different geocoding accuracies. They found that positional error had a noticeable influence on parameter estimates obtained through spatial statistical analysis. Mazumdar et al., (2008) examined a similar question using somewhat different methods. They found that the observed strength of association between environmental exposure and disease incidence decreased as positional error increased. The implication is that it is more difficult to uncover the true association between environmental exposures and disease using less accurate spatial data. These studies suggest that geographic error-in-variables can lead to errors in statistical estimates of model coefficients. Spatial aggregation error results in a similar kind of error-in-variables and is likely to have

similar kinds of impacts on model coefficients. The only difference is that spatial aggregation error has an explicit spatial structure rooted in the zones to which data are aggregated. In contrast, positional error does not have an explicit spatial structure and can be associated with very large displacements of points from their true locations.

In this paper, we examine the impact of spatial aggregation error on the coefficients of multilevel statistical models which analyze the associations between late-stage breast cancer, demographic variables and distance to mammography facilities. Using Monte Carlo simulation methods similar to those adopted by Henry and Boscoe (2008) and Shi (2007), we generate a large number of 'disaggregations' of breast cancer cases from the ZIP code to the census block level. The assignment of individual breast cancer cases from ZIP codes to blocks is proportional to the age/racial composition of block populations as in Henry and Boscoe (2008). We estimate a multilevel statistical model of late-stage breast cancer risk that includes a spatial variable, distance to the nearest mammography facility, as a predictor of late-stage risk. Models are estimated at the ZIP code and census block levels, and differences between model coefficients at the two levels reveal the impacts of spatial aggregation error.

3. Methods

Two geographically and demographically diverse study areas are chosen for analysis: Kane and Peoria counties in Illinois. Kane County is located in the southwest section of the Chicago Metropolitan area. The eastern part of this county is highly populated, while the western part is mainly farmland with a few residential areas. The population is predominantly Caucasian, with concentrations in the young and middle age groups. African-Americans make up around 6 percent of the county's population. Peoria County is located in central Illinois. Its population characteristics are similar to those of Kane County, except for a higher representation in the elderly age group (>65 years). Because ZIP code boundaries sometimes cut across county borders, all contiguous ZIP codes are included in the study areas as long as the ZIP code centroids fall within county boundaries. The two study areas are illustrated in Figures 20 and 21.

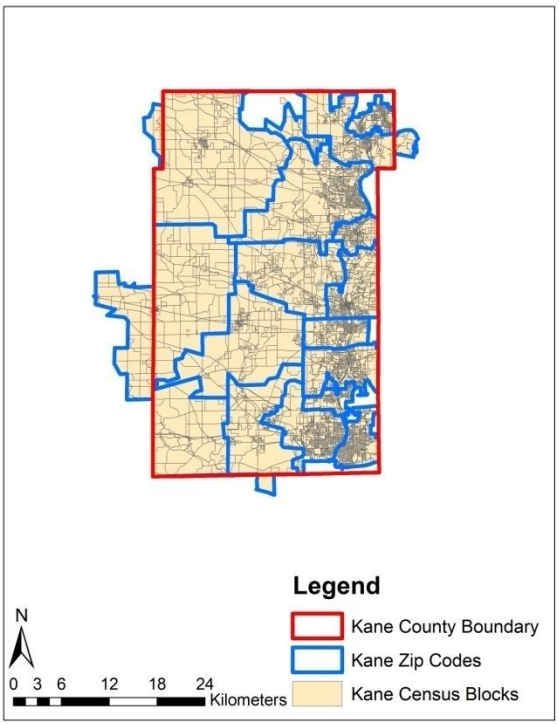

Figure 20. Census Blocks and ZIP Codes in Kane Study Area

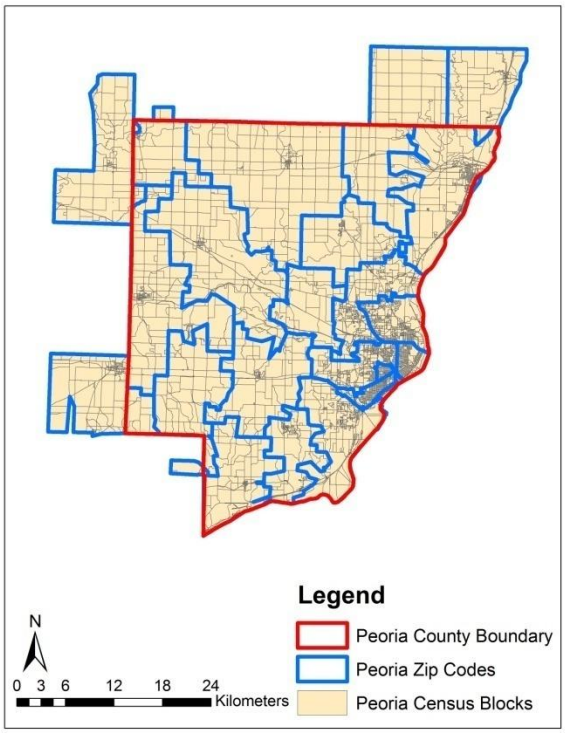

Figure 21. Census Blocks and ZIP Codes in Peoria Study Area

Breast cancer cases in Illinois were obtained from the Illinois State Cancer Registry (ISCR). The dataset contains demographic and epidemiologic records at the individual level and each record is geocoded to the residential ZIP code. Variables include age group, sex, race, diagnosis stage and year. ISCR utilized a classification scheme parallel with SEER summary stage to measure cancer stage at diagnosis (Young et al., 2001). Cancer cases at stages 0 and 1 were considered as early stage, and cases staged from 2 to 7 were regarded as late stage (Wang et al., 2010). Cases with unknown stage were excluded from this study. For both study areas, female breast cancer cases from 1998 to 2002 were selected. The percent of cases at different stages for the two study areas is shown in Table 14.

Table 14. Breast Cancer Cases by Stage in Kane and Peoria, 1998-2002

Study Area	# Cases	Unstaged		Late-Stage	
		# Cases	Percent (%)	# Cases	Percent (%)
Kane	1102	65	5.90	406	39.2
Peoria	804	38	4.73	245	32.0

The Monte Carlo Simulation procedure involves assigning cancer cases from a ZIP code area to the census blocks within that ZIP code. The probability of assignment is proportional to the age-race composition of the block population; so, for example, a cancer case in a black woman aged 50-69 has a higher probability of assignment to a census block that has a large population in the same demographic group. To facilitate this assignment procedure, we divided cancer cases into 6 categories based on age-race combinations. To differentiate the age categories, three age groups were used: less than 50-years old, between 50-and 70-years old, and more than 70-years old. Research shows that the risk of late-stage diagnosis varies according to age, and young patients have a higher risk of late diagnosis (Joslyn et al., 2005). Cases also were divided into 'black' and 'non-black' racial categories, given that late-stage breast cancer risk is high among blacks (Eley et al., 1994, Hunter et al., 1993, Lannin et al., 2002, McCarthy et al., 1998, Yost et al., 2001). The numbers of breast cancer cases in each county in the six categories are listed in Table 15.

Table 15. Summary of Breast Cancer Cases by Demographic Subgroup, Kane and Peoria

Kane		Peoria	
Population Subset	Cases	Population Subset	Cases
Total Population	1037	Total Population	766
Non-Black		Non-Black	
Female <50 years	243	Female <50 years	131
Female 50~70 years	420	Female 50~70 years	294
Female >70 years	334	Female >70 years	276
Black		Black	
Female <50 years	17	Female <50 years	24
Female 50~70 years	14	Female 50~70 years	30
Female >70 years	9	Female >70 years	11

Demographic information for the year 2000 at the census-block level was obtained from the U.S. Census for all the census blocks in the two study areas. There were a total of 7,619 census blocks in the Kane study area and 5,689 in Peoria. The census block female populations were divided into the six subgroups described above to match the breast cancer data.

3.1. Shortest Travel Distance Calculation

The spatial variable examined in this study is travel distance to the nearest mammography facility. Some research suggests that poor spatial accessibility to mammography screening facilities is associated with late-stage diagnosis. There are many ways to measure spatial accessibility, including provider-to-population ratio, and travel impedance to nearest provider (Guagliardo, 2004). We estimated spatial

accessibility based on shortest travel distance --the road network distance from the ZIP code or block centroid to the nearest provider. Many studies have used shortest travel distance to evaluate spatial accessibility at neighborhood level (Athas, 2000, Chen et al., 2008, Hyndman et al., 2000, Maheswaren et al., 2006, Nattinger et al., 2001). Within each ZIP code, population-weighted centroids were used to better reflect the uneven distribution of population (SAS, 2007). Geographic centroids were used for the block-level analysis. Data on registered mammography screening facilities in Illinois were obtained for 2000, and facilities were geocoded using street address information. Mobile mammography facilities do not operate in either county and thus were not included in the analysis. The shortest travel distance was computed from each centroid to its nearest mammography screening facility through the road network. The block-level shortest distances in the Kane and Peoria study areas are mapped in Figures 22 and 23. In both counties, the shortest distances do not exceed 46 kilometers, suggesting that spatial access to mammography is reasonably good overall.

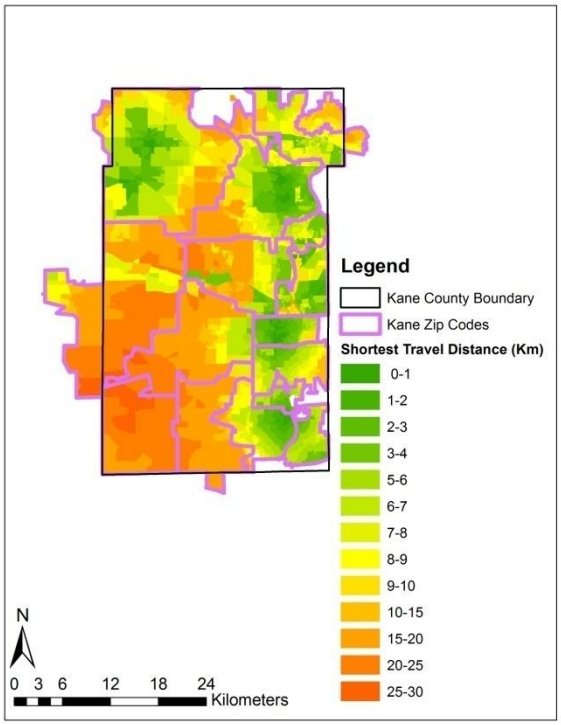

Figure 22. Block-level Shortest Travel Distance Distribution in the Kane County Study Area

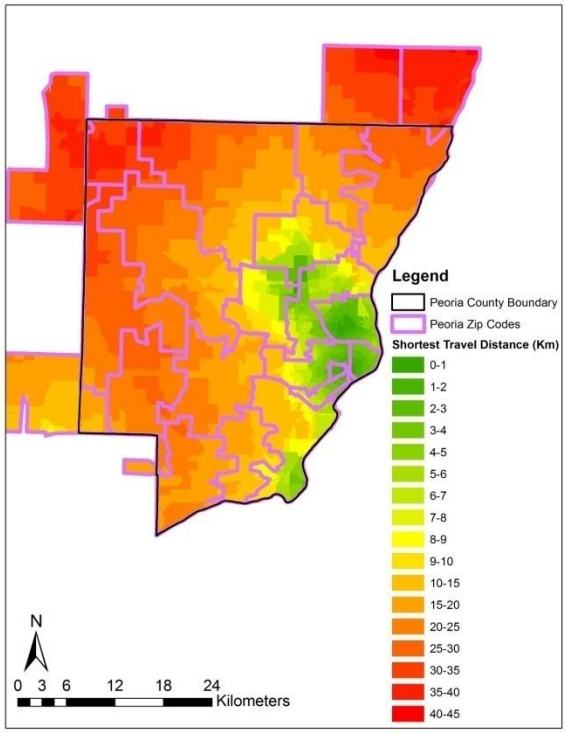

Figure 23. Block-level Shortest Travel Distance Distribution in the Peoria County Study Area

Figures 22 and 23 show that within some ZIP codes, block-level travel distances vary significantly which indicates the potential for spatial aggregation error. Summary statistics for the distance variable at the ZIP code and block levels also reveal substantial disparities, particularly for Peoria County (Table 16). In Peoria County, the average and median travel distances differ by 4 and 9 kilometers respectively for ZIP codes and blocks, whereas in Kane County, the mean and median values are quite similar. This suggests that the impact of spatial aggregation error will be greater in Peoria County where the distance measurements at the two levels are very different.

Table 16. Summary Statistics for Travel Distance to Mammography at Block ZIP Code Levels

Variables(Km)	Kane Study Area			
	Min	Max	Mean	Median
Block-level Distance	0.0315	22.601	5.951	4.995
ZIP-level Distance	0.670	13.149	5.621	4.564
Variables(Km)	Peoria Study Area			
	Min	Max	Mean	Median
Block-level Distance	0.0527	43.255	11.902	8.027
ZIP-level Distance	1.373	36.262	15.569	17.110

3.2. Disaggregation of Breast Cancer Data Using Monte Carlo Simulation

The purpose of the Monte Carlo simulation is to investigate the impact of the spatial aggregation error by comparing ZIP code-level model coefficient estimates with a reference distribution of values based on small-area (block level) data. Ideally, one would want to compare the ZIP-code values with those based on actual patient residential locations. However, because of privacy and confidentiality issues, we were unable to obtain breast cancer data below the ZIP code level. Therefore, simulation was used to create 'reasonably' distributed cancer cases at the census block level, building upon the work of Shi (2007) and Henry and Boscoe (2008). Each case was randomly assigned to a block within its ZIP code, and the likelihood of assignment depended on the age-race composition of the block population defined according to the six subgroups mentioned earlier.

To implement the Monte Carlo simulation, the block-level population in each demographic subgroup was accumulated and summed. The output was then normalized so that each subgroup's population ranged from 0 to 1, with intermediate values representing the cumulative share of that subgroup's population located in each block. This process was repeated for each ZIP code and each subgroup. Based on this data, the Monte Carlo simulation was implemented.

The Monte Carlo simulation involved several steps. First, for each cancer case, an array of 1,000 uniform random numbers was generated. A nested-structure of generating seeds was used to ensure the independence of each random number. Specifically, a series of random numbers was generated using system time as the generating seed. Then this series was employed as the secondary generating seed to produce final numbers. The end result was an 'n' by 1000 matrix in which n is the number of cancer cases. Rows represent individual cancer cases and columns represent random numbers. Second, we used the random numbers to assign a case from ZIP code to a block, with each random number representing a simulated block assignment. Each block assignment was based on the following principle: a case was assigned to a census block if the block-level normalized range of values contained that specific random number. Hence, the assignment was not based on a uniform distribution, but was proportional to the block population falling in the same demographic category as the cancer case. Assignments were made sequentially within each column of the matrix. Third, within a specific column, once a block received a cancer case, the block population in that demographic category was reduced by 1, because one person cannot be diagnosed with cancer twice simultaneously. If a population subgroup within a block went down to zero, the block was taken out from the remaining candidates for subsequent assignments in the same demographic category. We iterated the second and third steps, disaggregating cases from ZIP codes to blocks, and thus generated 1,000 disaggregated patterns of cases. As a result, a final matrix was produced in which rows represented cancer cases and columns denoted different assignments of census blocks for each cancer case. The matrix was diagramed as 1,037 rows by 1,000 columns for cases in Kane, and 766 rows by 1,000 columns for cases in Peoria. We wrote the Monte Carlo simulation procedure using Javascript1.5 and used Eclipse 3.4.0 as the software interface.

3.3. Analysis of Spatial Aggregation Error Using Hierarchical Logistic Regression

Hierarchical (multilevel) logistic regression was utilized to evaluate the impact of spatial aggregation error on statistical models of late-stage breast cancer risk. We used a two-level hierarchical modeling approach in which individual cancer patients (level 1) are nested within either ZIP codes or blocks (level 2). First, hierarchical models were estimated with the ZIP code as level 2; then, after Monte Carlo simulation, models were estimated at the block scale, with blocks representing level 2. The

dependent variable in the hierarchical regression models is late-stage diagnosis. Only a limited set of independent variables is included in the models so that the effects of spatial aggregation error can easily be observed. Individual variables include the patient's age and race categories, defined according to the categories used earlier. Race is represented by a dummy variable (BLACK) in which 'non-black' is the reference category. Age is represented by two dummy variables (AGE<50, AGE 50-70), and the reference category is the oldest age group (>=70). The level 2 independent variable is shortest travel distance (in meters) to the closest mammography facility, measured based on ZIP code centroid for the ZIP code model and block centroid for the block-level disaggregations. The formulations of the hierarchical logistic regression are shown below:

The micro specification (level 1) is:

$$Logit(\Pr ob(Y_{ij} = latestage)) = \beta_{0j} + \beta_{1j}(Race)_{ij} + \beta_{2j}(Age)_{ij} + R_{ij} \qquad (1.1)$$

where the βs denote the constant (intercept) and regression coefficients of the independent variables, i=1,…, n_j denotes individuals within different ZIP code or census block areas, and j=1,…, J denotes ZIP code or census block areas. The R_{ij} are micro errors with independent normal distributions, $R_{ij} \sim N(0, \sigma^2)$.

The macro stage (level 2) model is:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(ShortestTravelDistance)_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} \\ \beta_{2j} &= \gamma_{20} \end{aligned} \qquad (1.2)$$

where Us are macro errors, $U_{oj} \sim N(0, \tau_0^2)$ and they are independent over j and with R_{ij}. Equations (1.1) and (1.2) define a hierarchical logistic model that can be written equivalently as a combined single-equation model by substituting (1.1) – (1.2) into (2):

$$\begin{aligned} Logit(\Pr ob(Y_{ij} = latestage)) &= \gamma_{00} + \gamma_{01}(ShortestTravelDistance)_j \\ &+ \gamma_{10}(Race)_{ij} + \gamma_{20}(Age)_{ij} + U_{0j} + R_{ij} \end{aligned} \qquad (2)$$

The variable most likely to be affected by spatial aggregation error is shortest travel distance, so any change in model coefficients between the ZIP code and block levels is mainly due to changes in measurement of this variable resulting from spatial aggregation.

All models were estimated using 'proc glimmix' in SAS 9.1 (SAS, 2007). Given that there are 1,000 randomized patterns of cases at the block level, macro-level SAS syntax was used to automatically estimate the block-level hierarchical regression analyses. The coefficient estimates for the block level models were displayed as histograms and compared with the respective values for the ZIP code level coefficients.

4. Results and Discussion

The comparison of model coefficients at the ZIP code and block levels for Kane County is shown in Table 17 and Figure 24. The results for Kane show only a small impact of spatial aggregation error on model coefficients. The means of the block coefficients are very close to the corresponding ZIP code values except in the case of shortest travel distance. In addition, the ranges of block-level coefficients include the corresponding ZIP code parameters for all independent variables. The similarity of ZIP code and block level coefficients is also evident in Figure 5 which shows, for each independent variable, a histogram of the block-level coefficients and a dotted line representing the ZIP code coefficient. All of the ZIP code coefficients are located near the peak of their corresponding block-level histograms. Moreover, at both levels, the coefficient for distance indicates no statistically significant association between shortest travel distance to mammography and late-stage diagnosis, so the overall findings are consistent. Therefore, for the Kane study area, the closeness of the means and the fact that the ZIP code values fall within their respective block-level ranges show that spatial aggregation error does not have much influence on inferences made based on statistical analysis at the ZIP code level.

Table 17. Model Coefficients at the Block and ZIP Code Levels for Kane County

Variables	Census Block Level			ZIP Code Level			
	Mean Coefficient	Min	Max	Coefficient	Std Error	p-value	95% Confidence Interval
Age < 50	0.536	0.503	0.575	0.537	0.171	0.00175	(0.201, 0.873)
Age 50~70	0.326	0.273	0.365	0.328	0.152	0.0307	(0.0305, 0.626)
Black	0.396	0.332	0.468	0.391	0.326	0.230	(-0.248, 1.031)
Shortest Travel Distance (m)	5.270E-6	-4.155E-5	5.394E-5	2.620E-6	2.489E-5	0.916	(-4.623E-5, 5.147E-5)

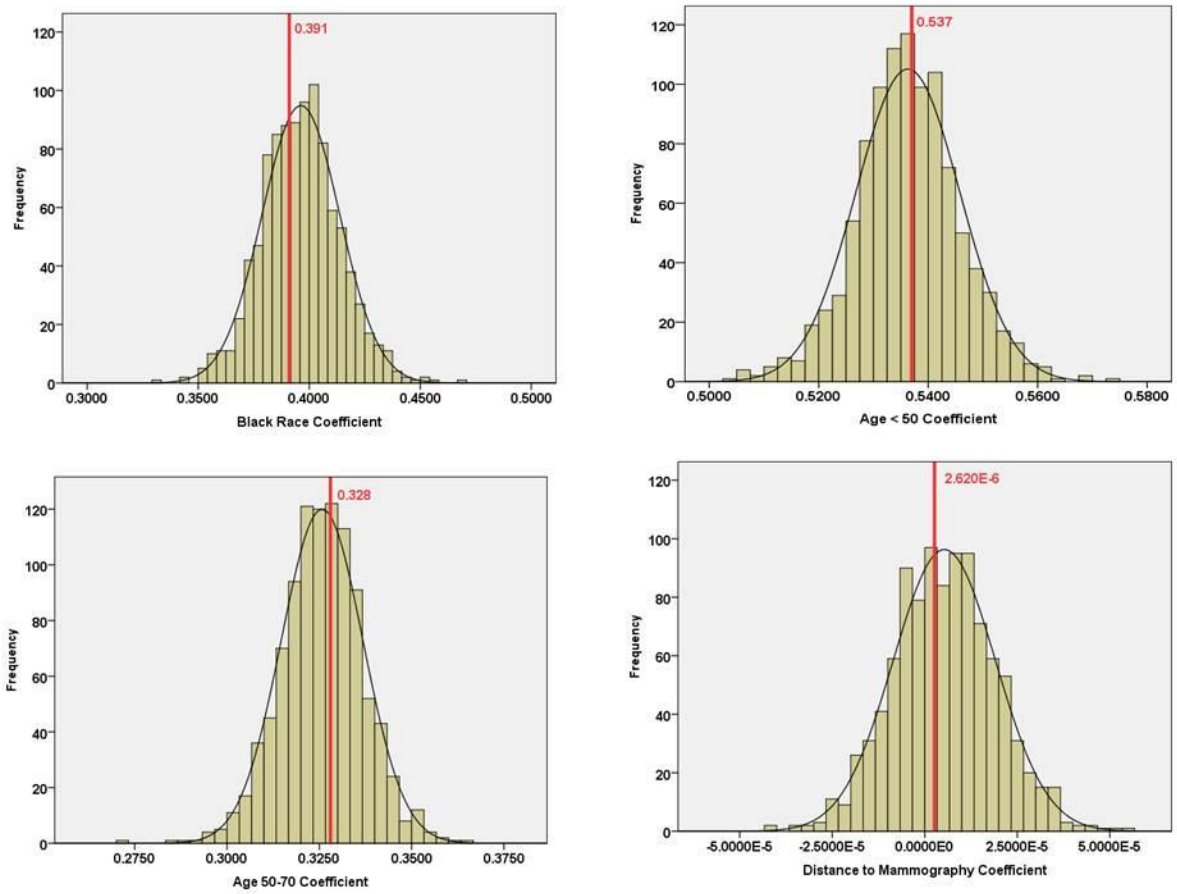

Figure 24. ZIP code-level Coefficient (Red, Bold line) and Histogram of Block-level Coefficients for Each Independent Variable, Kane County

The findings are very different for the Peoria study area. Large differences are evident between ZIP code- and block- coefficients. As shown in Tables 18, each of the ZIP code-level coefficients falls outside the range of the respective block-level coefficients. Also, the ZIP code coefficients differ much more from their respective block means than was the case in the Kane study area. This is especially true for shortest travel distance, in which the coefficient signs for models at the two levels are different. Specifically, for shortest travel distance, the ZIP code-level parameter is negative and an order of magnitude less (in the negative direction) than the mean of the block-level values which has a positive sign.

Table 18. Model Coefficients at the Block and ZIP Code Levels for Peoria County

Variables	Census Block Level			ZIP Code Level			
	Mean Coefficient	Min	Max	Coefficient	Std Error	p-value	95% Confidence Interval
Age < 50	0.673	0.661	0.683	0.714	0.219	0.0012	(0.283, 1.145)
Age 50~70	0.445	0.434	0.454	0.482	0.184	0.0089	(0.121, 0.842)
Black	1.082	1.040	1.128	0.943	0.271	0.00054	(0.411, 1.475)
Shortest Travel Distance (m)	1.419E-5	5.950E-6	2.250E-5	-3.51 E-4	2.08 E-4	0.091	(-7.596E-4, 5.678E-5)

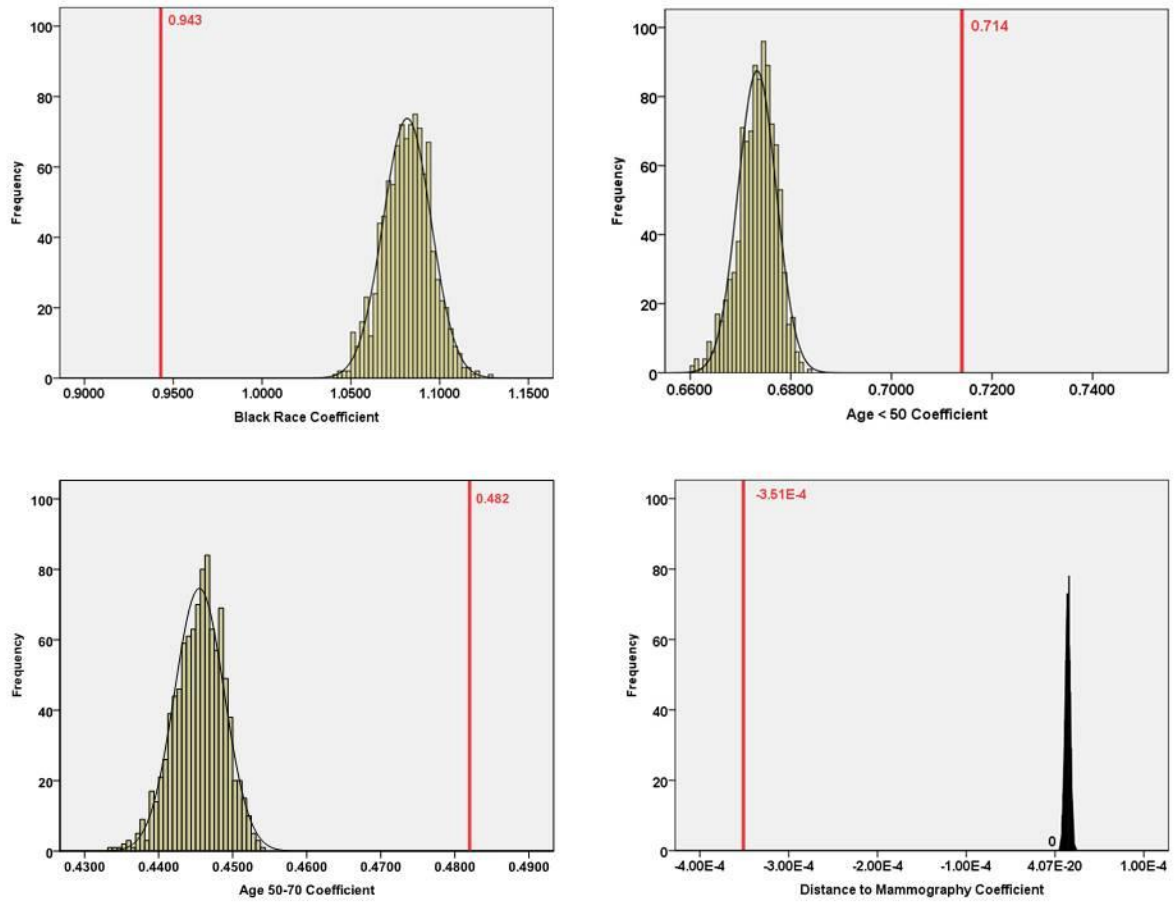

Figure 25. ZIP code-level Coefficient (Red, Bold line) and Histogram of Block-level Coefficients for Each Independent Variable, Peoria County

For the Peoria case, we calculated the impact of these differences in model coefficients on model predictions by plugging in values for a "reference person" (non-black, age >70) located at distances of 0 and 20 km from the closest mammography facility. At zero kilometers, the predicted late-stage risks are very similar for the ZIP code and block (average) models – 0.250 and 0.266 respectively. However, at 20 kilometers, differences are extraordinarily large because the effects of the different distance coefficients are magnified. The ZIP code model gives a predicted late-stage risk of less than 1 percent, a nonsensical value; whereas the block (average) model yields a predicted risk of 24 percent. Thus, using the ZIP code model for predictive purposes does not give meaningful results.

For Peoria County, impacts of spatial aggregation error are also apparent in the plots comparing model coefficients at the ZIP code and block levels (Figure 25). For each independent variable, the ZIP code-level coefficient falls substantially outside the range of the block-level values. For the distance variable, the ZIP code level parameter estimate is completely isolated from the block-level values, differing greatly in magnitude and with the opposite sign as noted above. This suggests that the association between distance and late-stage cancer risk is completely different from that observed based on ZIP code data. Among the remaining demographic variables, the coefficient for black race changed more than those for the two age variables. Coefficients for the two age variables move towards zero when we shift from the ZIP code to block scale, whereas the coefficient for black race increases. Block-level models indicate that black race is more strongly associated with late-stage breast cancer risk than was evident in the ZIP code-level model. Thus, spatial aggregation error affects not only the coefficient for the spatial variable in the model, distance to mammography, but also the coefficients for the other socio-demographic variables, age and race, which were incorporated in the Monte Carlo simulation procedure.

These results indicate that in some geographic contexts, spatial aggregation error results in significant bias in model coefficients, bias that can lead to inaccurate conclusions and inappropriate statistical inferences. Results for Peoria County suggest that if cancer data were available at the block level, the resulting model coefficients for all independent variables would most likely be quite different from the values observed based on ZIP code data. For the distance variable, the impact of spatial aggregation error is substantial enough to affect statistical inference. In particular, a significance test (one-sided, α=0.1) indicates that the ZIP code-level coefficient for the distance variable is significantly different from zero, suggesting that distance to mammography is significantly and negatively associated with late-stage breast cancer risk. This is an unexpected finding implying that late-stage risk decreases with increasing distance. Yet our simulations indicate that this conclusion is most likely a spurious result of spatial aggregation error. The block-level coefficients are all close to zero, suggesting a lack of statistical association. Without address-level data, we cannot know the true association between distance

and late-stage breast cancer risk; however, the simulated block-level values overwhelmingly suggest no association.

Although spatial aggregation error is important, the differences between the two study areas reveal that the influence of spatial aggregation error is highly case-sensitive. In Kane County, spatial aggregation has a minimal impact on model coefficients; whereas in Peoria County, the impact is substantial. We believe that these differences are linked to differences in the underlying spatial distributions of socio-demographic groups and differences in the sizes and configurations of ZIP codes and blocks which are superimposed on each county's demographic landscape. We can only speculate about the causes of differences observed between these two counties. Located on the fringe of the Chicago metropolitan region, Kane County has a higher population density than Peoria County, and Kane's population appears to be more uniformly distributed, although with an east-west gradient. Mammography facilities are well-distributed throughout the more populated areas of the county. In comparison, Peoria County contains a more bifurcated rural-urban pattern, with a single, densely populated city (Peoria) surrounded by low density suburban and rural zones. The few mammography facilities are concentrated in Peoria city. In this bifurcated landscape, disaggregation of cases to the block level via Monte Carlo simulation results in heterogeneous assignments that greatly influence model coefficients.

Another important finding is that the statistical impacts of spatial aggregation error are not confined to coefficients for spatial variables. In Peoria County, coefficients for all variables are affected. These interconnected impacts most likely reflect the correlations between race, age and residential location. Residential segregation by race is a strong feature of both study areas, and it implies that the 'black' and 'non-black' racial categories have distinct residential geographies at the block scale. Disaggregating data from ZIP codes to blocks on the basis of racially- and demographically-based probabilities incorporates these localized, segregated geographies. Although we used population-weighted centroids in calculating shortest distance to mammography, using race- and age-specific population centroids may be more effective in minimizing spatial aggregation error associated with residential segregation. Still, these more finely-tuned centroids can be problematic when racial groups are both segregated and unevenly distributed within ZIP code boundaries as is often the case.

5. Conclusion

The paper analyzed the impact of the spatial aggregation error on ZIP code level statistical analysis of the associations between spatial and non-spatial variables and late-stage breast cancer risk in two study areas in Illinois. Given the difficulties in obtaining cancer cases below the ZIP code level, we

designed a Monte Carlo simulation procedure to disaggregate cancer cases from ZIP codes to census blocks on the basis of the demographic characteristics of cancer case and block populations. Spatial aggregation error significantly affected the coefficients of statistical models in the Peoria study area, leading to inaccurate inference, whereas in Kane County the impact was minimal. The distinctive outputs for Kane and Peoria counties illustrate that the impacts of spatial aggregation error are context-dependent. Impacts appear to be most pronounced in areas like Peoria County, which have both highly uneven and segregated residential geographies. The spatial autocorrelation of age- and racially-categorized population groups by block may be important in affecting spatial aggregation error. Error also depends on the configuration of zones overlaying those geographies. Many studies have demonstrated that large zones are associated with high levels of spatial aggregation error, but clearly the residential geographies within the zones are also important. Other factors affecting spatial aggregation error in analyzing distance to health services are the number and spatial configuration of service facilities (Hewko et al., 2002). In general, the potential for error will be greater in places with fewer facilities and where facilities are spatially clustered. Compared to Peoria, Kane County has more mammography facilities, and facilities are more spatially dispersed, perhaps reducing the scope for spatial aggregation error.

Given the range and complexity of factors involved in spatial aggregation error, the specific nature of these associations requires further investigation using a much wider range of study areas representing varied social and geographical characteristics. The Monte Carlo simulation procedure implemented here is very useful in these efforts. Moreover, analyzing how spatial aggregation error compares with other kinds of uncertainty such as sampling error in statistical modeling is also critically important.

Our findings highlight the need to develop methods and procedures for minimizing spatial aggregation error in statistical models that rely on zonal health data. Monte Carlo simulation provides a way to generate the highly likely distribution of block-level coefficients associated with a particular dataset, but the method is both data- and computationally-intensive. Much simpler procedures, like using age- and race-specific ZIP code centroids offer a feasible, low-tech alternative, but these methods may not be effective in areas where the spatial distributions of population groups are highly uneven (Langford and Higgs, 2006). Shi and Berke (2009) discuss promising methods which utilize area-based representations of population. Another option is explicit modeling of aggregation effects through the use of variograms and other indicators of spatial autocorrelation. Promising methods have been developed for use with environmental and population data (Kyriakidis, 2004), and the methods have great potential value for health studies (Goovaerts, 2009).

Although we have demonstrated the importance of spatial aggregation error, our study has several limitations. Because we do not have access to data on actual breast cancer cases locations, we do not know the real extent and impact of spatial aggregation error in the two case study areas. The simulations delineate the likely distribution of possible coefficient values, but do not quantify the true spatial aggregation error. Still knowing the likely extent of error is important in signaling the need for more advanced methodologies that explicitly address spatial aggregation effects. Another limitation is that by relying on actual cancer case data we have no knowledge of, the underlying 'true' risk model for late-stage breast cancer, and we are unable to control or manipulate that model in the process of Monte Carlo simulation. A more experimental approach based on hypothetical data would enable researchers to assess the relative magnitude of spatial aggregation error compared to other sources of error in statistical models of cancer risk factors. Despite these limitations, this research demonstrates that spatial aggregation error has substantial effects in some geographic contexts on the results of statistical modeling of the association between cancer and spatial and non-spatial risk factors. Understanding how and why these effects vary stands as a key topic for future research investigations.

6. References

Athas W, Adams-Cameron M, Hunt W, Amir-Fazli A, Key C: **Travel Distance to Radiation Therapy and Receipt of Radio-Therapy Following Breast-Conserving Surgery.** *Journal of the National Cancer Institute,* 2000, **92**: 269-271.

Bonner M, Han D, Nie J, Rogerson P, Vena J, Freudenheim J: **Positional Accuracy of Geocoded Addresses in Epidemiologic Research.** *Epidemiology*, 2003, **14**: 408-412.

Chen AY, Halpern MT, Schrag NM, Stewart A, Leitch M, Ward E: **Disparities and Trends in Sentinel Lymph Node Biopsy Among Early-Stage Breast Cancer Patients (1998-2005).** *Journal of the National Cancer Institute*, 2008, **100(7):** 462-474.

Eley JW, Holly AH, Chen VW, Austin DF, Wesley MN, Muss HB, Greenberg RS, Coates RJ, Correa P., Redmond CK, Hunter CP, Herman AA, Kurman R, Black R, Shapiro S, Edwards BK: **Racial Differences in Survival From Breast Cancer Results of the National Cancer Institute Black/White Cancer Survival Study.** *Journal of the American Medical Association*, 1994, **272(12):** 947-954.

Fortney J, Kathryn R, Warren J: **Comparing Alternative Methods of Measuring Geographic Access to Health Services**. *Health Services & Outcomes Research Methodology*, 2000, **1(2):** 173-184.

Gehlke CE, Biehl H: **Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material.** *Journal of the American Statistical Association, Supplement,* 1934, **29**: 169-170.

Goovaerts, P: **Medical Geography: A Promising Field of Application for Geostatistics**. *Mathematical Geosciences,* 2009, **41**:243-264.

Gregorio D, DeChello L, Samociuk H, Kulldorff M: **Lumping or Splitting: Seeking the Preferred Areal Unit for Health Geography Studies.** *International Journal of Health Geographics,* 2005, **4**: 6.

Griffth DA, Millones M, Vincent M, Johnson DL, Hunt A: **Impacts of Positional Error on Spatial Regression Analysis: A Case Study of Address Locations in Syracuse, New York.** *Transactions in GIS,* 2007, **11(5):** 655-679.

Guagliardo MF: **Spatial Accessibility of Primary Care: Concepts, Methods and Challenges.** *International Journal of Health Geographics,* 2004, **3(3):** 1-13.

Henry KA, Boscoe FP: **Estimating the Accuracy of Geographical Imputation.** *International Journal of Health Geographics*, 2008, **7(3):** 1-10.

Hewko J, Smoyer-Tomic KE, Hodgson MJ: **Measuring Neighborhood Spatial Accessibility to Urban Amenities: Does Aggregation Error Matter?** *Environment and Planning A,* 2002, **34**: 1185-1206.

Hillsman E, Rhoda R: **Errors in Measuring Distances from Populations to Services Centers**. *Annals of Regional Science,* 1978, **12**: 74-88.

Hodgson MJ, Shmulevitz F, Kőrkel M: **Aggregation Error Effects on the Discrete-Space *P*-Median Model: The Case of Edmonton, Canada**. *The Canadian Geographer,* 1997, **41**: 415-428.

Hunter CP, Redmond CK, Chen VW, Austin DF, Greenberg RS, Correa P, Muss HB, Forman MR, Wesley MN, Blacklow RS: **Breast Cancer: Factors Associated with Stage at Diagnosis in Black and White Women. Black/White Cancer Survival Study Group.** *Journal of the National Cancer Institute*, 1993, **85(14):** 1129-1137.

Hyndman JCG, Holman CFJ, Dawes VP: **Effects of Distance and Social Disadvantage on the Response to Invitations to Attend Mammography Screening.** *Journal of Medical Screening,* 2000, **7(3):** 141-145.

Jacquez GM, Waller LA: **The Effect of Uncertain Locations on Disease Cluster Statistics. In Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing**. Mowrer HT, Congalton RG eds., *Chelsea MI: Sleeping Bear Press,* 1999, 53-64.

Jacquez GM, Jacquez JA: **Disease Clustering for Uncertain Locations**. *In Disease Mapping and Risk Assessment for Public Health Decision Making*. Lawson A, Biggeri D, Bőhning E, Lesaffre J-F V, Bertollini R (eds), 1999, 151-168, Wiley, London.

Johnston RJ: **Ecological Fallacy**. In Johnston RJ, Gregory D, Pratt G, Watts M, eds., *The Dictionary of Human Geography*, 2000, 190-191. Oxford: Blackwell.

Joslyn SA, Foote ML, Nasseri K, Coughlin SS, Howe HL: **Racial and Ethnic Disparities in Breast Cancer Rates by Age: NAACCR Breast Cancer Project**. *Breast Cancer Research and Treatment* 2005, **92(2):** 97-105.

Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW: **On the Wrong Side of the Tracts? Evaluating the Accuracy of Geocoding in Public Health Research.** *American Journal of Public Health,* 2001, **91**: 1114-1116.

Krieger N, Waterman P, Chen J, Soobader M, Subramanian S, Carson R: **Zip Code Caveat: Bias Due to Spatiotemporal Mismatches Between Zip Codes and US Census-Defined Geographic Areas – The Public Health Disparities Geocoding Project**. *American Journal of Public Health,* 2002, 92: 1100-1102.

Kyriakidis P: **A Geostatistical Framework for Area-to-Point Spatial Interpolation**. *Geographical Analysis*, 2004, **36(3):** 259-289.

Lannin DR, Mathews HF, Mitchell J, Swanson MS: **Impacting Cultural Attitudes in African-American Women to Decrease Breast Cancer Mortality.** *The American Journal of Surgery*, 2002, **184(5):** 418-423.

Langford, M., & Higgs, G. **Measuring potential access to primary healthcare services: The influence of alternative spatial representations of population**. *Professional Geographer*, 2006, **58(3):** 294-306.

Maheswaran R, Pearson T, Jordan H, Black D: **Socioeconomic Deprivation, Travel Distance, Location of Sevice, and Uptake of Breast Cancer Screening in North Derbyshire, UK.** *Journal of Epidemiology and Community Health,* 2006, **60**:208-212.

Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ: **Geocoding Accuracy and the Recovery of Relationships between Environmental Exposures and Health**. *International Journal of Health Geographics*, 2008, **7(13):** 1-18.

McCarthy EP, Burns RB, Coughlin SS, Freund KM, Rice J, Marwill SL, Ash A, Shwartz A, Moskowitz MA: **Mammography Use Helps To Explain Differences in Breast Cancer Stage at Diagnosis between Older Black and White Women**. *Annals of Internal Medicine*, 1998, **128(9):** 729-736.

McElroy JL, Remington P, Trentham-Dietz A, Robert SA, Newcomb PA: **Geocoding Addresses from A Large Population-Based Study: Lessons Learned.** *Epidemiology*, 2003, **14**: 399-407.

Nattinger AB, Kneusel RT, Hoffmann RG, Gilligan MA: **Relationship of Distance from A Radiography Facility and Initial Breast Cancer Treatment.** *Journal of the National Cancer Institute*, 2001, **93(17):** 1344-1346.

Openshaw S, Taylor PJ: **A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem.** *Statistical Methods in the Spatial Science*, Wrigley N eds., London, 1979, 127-144.

Robinson AH: **Ecological Correlation and the Behavior of Individuals.** *American Sociological Review*, 1950, **15**: 351-357.

Rushton G, Armstrong P, Gittler J, Greene B, Pavlik CE, West MM, Zimmerman DL: **Geocoding in Cancer Research A Review**. *American Journal of Preventive Medicine*, 2006, **30(25):** 16-24.

Shi X: **Evaluating the Uncertainty Caused by Post Office Box Addresses in Environmental Health Studies: A Restricted Monte Carlo Approach.** *International Journal of Geographical Information Science*, 2007, **21(3):** 325-340.

Shi, X: **A Geocomputational Process for Characterizing the Spatial Pattern of Lung Cancer Incidence in New Hampshire**. *Annals of the Association of American Geographers*, 2009, **99(3):** 521-533.

Shi X., & Berke, E. **Computing travel time when the exact address is unknown: A comparison of point and polygon ZIP code approximation methods**. *International Journal of Health Geographics*, 2009, 8, 23.

Statistical Analysis Software: **Proc Glimmix Procedure**. *SAS Institute* 2007, Cary, North Carolina.

http://support.sas.com/rnd/app/da/glimmix.html (Accessed: April 8th, 2009).

Wang F, Luo L, McLafferty S: **Health Access, Socioeconomic Factors and Late-Stage Cancer Diagnosis: An Exploratory Spatial Analysis and Public Policy Implication.** *International Journal of Public Policy*, 2010, **5(2/3):** 237-258.

Wang F, McLafferty S, Escamilla V, Luo L. **Late-Stage Breast Cancer Diagnosis and Health Access in Illinois.** *The Professional Geographer*, 2008, **60**: 54-69.

Young JL, Roffers SD, Ries LAG, Fritz AG, Hurlbut AA(eds): **SEER Summary Staging Manual-2000: Codes and Coding Instructions.** Bethesda MD: National Cancer Institute, NIH Pub, 2001. No. 01-4969.

Yost K, Perkins C, Cohen R, Morris C, Wright W: **Socioeconomic Status and Breast Cancer Incidence in California for Different Race/Ethnic Groups.** *Cancer Causes and Control*, 2001, **12(8):** 703-711.

Yule GU, Kendall MG: **An Introduction to the Theory of Statistics.** *Griffin*: London, 1950.

Chapter III

The Impact of Spatial Aggregation Error on the Spatial Scan Analysis:

A Case Study of Colorectal Cancer

1. Introduction

The choice of geographic unit plays a very important role in analyzing the uneven distribution of cancer cases and designing appropriate policies for disease control and prevention (Rushton, 1995). To protect privacy and confidentiality, cancer data obtained from surveillance systems are usually released only for predefined areal units, for example, counties or ZIP codes, which may have arbitrary and unstable characteristics. Because these predefined areas were not originally designed for cancer research, true patterns of cancer incidence can be distorted, producing misleading results. This problem has been well described as the 'modifiable areal unit problem' (MAUP) (Amrhein, 1994, Openshaw and Alvandies, 1999). One of the important components of the MAUP is spatial aggregation error, error that is caused by using data at an aggregated, large-area level, to generate inferences about data at lower (small-area) levels (Hodgson et al., 1997). Although the biases brought about by spatial aggregation error have been widely analyzed (Fortney et al., 2000, Hewko et al., 2002, Hillsman and Rhoda, 1978, Hodgson et al., 1997, Luo et al., 2010), its impacts on detection of spatial clusters of cancer cases has rarely been studied.

In analyzing spatial clustering of cancer, many researchers have used the spatial scan statistic to detect cluster locations (Gregorio et al., 2002, Gregorio and Samociuk, 2003, Gregorio et al., 2004, Jemal et al., 2002, Kulldorff et al., 1997, Pollack et al., 2006, Rushton et al., 2004, Roche et al., 2002, Seeff et al., 2003, Thomas and Carlin, 2003). The spatial scan statistic is a "local" spatial clustering test that identifies the locations and characteristics of statistically significant clusters of cases within a study area (Kulldorff, 1997). Studies have utilized spatial scan statistics to examine spatial disparities in cancer incidence and mortality (Gregorio et al., 2002, Gregorio et al., 2004, Roche et al., 2002, Kulldorff et al., 1997, Jemal et al., 2002). Several studies have applied spatial scan statistics to identify areas with high or low incidence rates of breast cancer (Gregorio and Samociuk, 2003) and to detect areas with an elevated proportion of late-stage breast cancer cases (Roche et al., 2002). Other studies have evaluated geographical patterns of colorectal cancer (CRC) in California, Iowa, Minnesota and New York (Pollack et al., 2006, Rushton et al., 2004, Seeff et al., 2003, Thomas and Carlin, 2003). Different geographical units were used in these studies, such as towns, ZIP code areas, counties, census blocks and census tracts. Each of these study units has pros and cons. Small areal units can depict local variations more clearly than

larger areal units, while larger units produce more reliable and stable estimates of disease incidence or risk across a large region. Lacking a 'gold standard', it is very difficult to select the optimum areal unit, and the optimum may vary from one case to another. Given these challenges, an important question is: How sensitive are the results of the spatial scan statistic to the choice of areal units?

This paper aims to evaluate the influence of spatial aggregation error on one of the most widely-used spatial scan statistics, the Bernoulli-based spatial scan statistic. The paper mainly addresses one question: How does spatial aggregation error affect spatial cluster detection using the well-known spatial scan statistic method? I examine the impact of spatial aggregation error on the ZIP code-level spatial scan statistic which is used to identify significant clusters of late-stage CRC cases based on cancer data at the ZIP code level. Following the second paper, Monte Carlo simulation methods are used to disaggregate cancer cases from the ZIP code level to the census tract, block group and block levels. Results of the spatial scan statistic are compared at each level to evaluate the sensitivity of results to the geographic scale of cancer data. This paper is divided into the following sections: literature background; data pre-processing and methodology; results and discussion and conclusions.

2. Background

A few studies have compared geographic outcomes in spatial analyses of cancer using cancer incidence data at different geographic scales (Gregorio et al., 2005, Krieger et al., 2002, Sheehan et al., 2000). Specifically, Sheehan et al., (2000) utilized the spatial scan statistic to detect significant spatial clusters of late-stage breast cancer diagnoses across Massachusetts, using towns, ZIP code areas, and census tracts. They observed that differences exist among the three geographic levels, in terms of the cluster sizes and the number of cases included in each cluster. However, they found that fluctuations in cluster characteristics were caused by geocoding problems, and the fluctuations had little association with the sizes and the boundaries of study units. Krieger et al. (2002) examined all-cause and some particular cause-specific mortality rates, and all-cause and site-specific cancer incidence rates within block groups, census tracts and ZIP code regions, across Massachusetts and Rhode Island. They concluded that analyses by block group and census tract performed comparably, but results at the ZIP code level were contradictory. Gregorio et al., (2005) applied the spatial scan statistic to compare geographical variation in late-stage prostate and breast cancers across Connecticut, using block groups, tracts, and towns. They reported that the local clusters identified at each scale were similar in terms of locations, populations at risk and other estimated parameters (centroid coordinates, p-values, and the ratios of observed-to-expected). Only a few differences were found in analytical results across the three study levels. All of

these studies began with cancer case data for the smallest areas and then aggregated the cases into larger areas for comparative analysis.

These aforementioned studies provided good strategies for comparing different spatial cancer clusters among areal units. Particularly, Gregorio et al., (2005) summarized all the clustering parameters from the spatial scan statistic results in a straightforward table for clear comparison. They used the block-level clusters as reference points, and compared clusters at the tract and town levels to the block-level clusters. One metric used was the average distance between the geographic coordinates of block-level centroids and the ones at town and census-tract levels. Clusters comparisons were also illustrated by a nested-structure format which displays cluster locations, cluster sizes, and the shared sections (overlap) among clusters on the same map.

These studies have found little difference in cluster results using data at different scales ranging from blocks to towns, indicating that the spatial aggregation error has a minimal effect on cluster detection. The approach taken in these studies is to begin with data for small areas and aggregate the data into larger areas. In this approach, there is only one outcome at each level, and the effect of spatial aggregation error is exactly known. In many cases, however, researchers do not have access to data for small reference units, so it is important to know how much error might exist as a result of the need to work with data that are highly spatially aggregated. For example: how reliable are clusters detected based on large-area data? How likely is it that those clusters would also be detected if small-area data were analyzed? According to my previous study (Luo et al., 2010), spatial aggregation error can be highly context-dependent. It is expected that spatial aggregation error has a larger impact on cluster detection when the distribution of disease cases and at-risk population vary across the study area. Therefore, this paper aims to enumerate possible distribution patterns of cases within ZIP code areas using a Monte Carlo simulation approach and to examine the effects of spatial aggregation error at different reference levels (census tract, block group and block).

3. Methods

3.1. Data and Pre-Processing

To analyze the impact of spatial data aggregation on the results of the spatial scan statistic, data on CRC cases in Cook County were used. The health outcome analyzed is the binary variable, late-stage CRC at diagnosis. CRC is classified as 'late-stage' if the tumor is large and/or the disease has spread beyond the initial site when first diagnosed. People diagnosed with late-stage CRC have a higher risk of mortality and morbidity than those whose cancer is diagnosed early. Clusters of late-stage CRC were

detected via SaTScan based on data at four geographic scales, from ZIP code to census block, and results are compared.

The data were obtained from Illinois State Cancer Registry (ISCR), and include all CRC cases diagnosed in Illinois residents between 1998 and 2002. Records in the data set represent individual cancer cases, with variables including age group, sex, race, diagnosis stage, year, and ZIP code of residence. This study focuses on advanced-stage (late-stage) cases so that the CRC cases were divided into early-stage (stages 0 and 1) and late-stage (stages 2 to 7) groups. Following the second paper, examining the influence of spatial aggregation error involved allocating the same CRC cases to smaller geographic units. The Monte Carlo simulation method used in the second chapter was applied in this study to accomplish this. To prepare the demographic link for the disaggregation, the CRC cases were divided into 12 categories by combinations of race by age by gender. Specifically, CRC cases were aggregated into black and non-black groups based on race information; the original 5-year age groups were classified into 3 main groups (< 50 years-old, 50-70-years old, and > 70 years-old), and gender is categorized as male and female. Population-level data for census areal units (tracts, block groups and blocks) in 2000 was derived from the Summary File 1 (SF1) data from the U.S. Census Bureau (US Census, 2000), and was categorized into 12 groups based on the same race-age-sex categories.

3.2. Areal Units and Study Region

For comparison with the ZIP code level, three smaller geographical levels were selected as reference units: census tracts, census block groups, and census blocks. These areal units are hierarchically structured and defined by the United States Census Bureau. According to the definition from US Census Bureau (US Census, 2000), census tracts are "designated to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions", and average 4,000 inhabitants in each area. Census tracts can be subdivided into block groups and blocks, with blocks being the smallest areal units, and block groups intermediate in size between blocks and tracts. The census block is the smallest geographic unit designed by the US Census Bureau for tabulating population and housing data (the complete data collected from all houses) (US Census, 2000). On average, 39 blocks form a block group with some small variations across the country. These three census areal units make appropriate choices because of their well-established association with demographic information, their nested structure, and their relatively stable boundaries over time.

Cook County was chosen as the study region, mainly because the spatial relationships between the four geographic levels are well-defined. Cook County is the most populated area in Illinois, and the high population density means that the three census areal units are typically smaller than ZIP codes. Thus,

the spatial relation between census tracts and ZIP codes can be easily defined as 'within' or 'outside'. Outside Cook County, most census tracts, and even some block groups, are comparable in size to ZIP code areas which mean that there is no spatial aggregation effect when comparing ZIP codes and tracts. Outside Cook County, one census tract or block group sometimes comprises several ZIP code areas. Without knowing the localized distribution of CRC cases in each ZIP code area, it would have been difficult to assign CRC cases from ZIP codes to the overlapping sections of different census tracts, and such a process would result in a different kind of spatial error. Consequently, I selected Cook County as the only study region, to provide clear hierarchical spatial relationships between the ZIP code level and smaller census area units. In addition, Cook County contains a large sample size of CRC cases: 3,608 total cases and 2,353 falling into the late-stage category. Also, my previous work (Paper 1) shows that Cook County contains a spatial cluster of late-stage CRC that is close to statistically significant(α=0.10), making it possible to evaluate the sensitivity of cluster detection to the geographic scale of cancer data.

3.3. Disaggregation of Cancer Cases

Because the cancer data is unobtainable at a level below the ZIP code scale, the Monte Carlo simulation approach designed in the second paper was applied here to disaggregate cancer data from the ZIP code level to each smaller reference unit. The most critical step in the disaggregation process was to define which ZIP code contains each census tract, so that tracts were not shared by neighboring ZIP codes. In the ISCR cancer dataset, CRC patients lived in 152 out of 161 ZIP code areas, covering most sections of Cook County. As the smallest reference unit, census blocks are mostly completely inside of each ZIP code area. If a block overlapped a ZIP code boundary, the block was treated as within a ZIP code area if the block centroid fell within the ZIP code. As a result, 64,231 blocks were chosen within the 152 ZIP code areas. Linking census block groups with ZIP codes was more complicated, because the larger size of a block group increases the chances of it overlapping multiple ZIP codes. Several steps were implemented to specify the spatial relation between ZIP codes and block groups. First, the population-weighted centroid of each block group was generated based on block-level population information; then each block group was regarded to be within a ZIP code if its population-weighted centroid was located inside of that ZIP code. Seven block groups were deleted on the edge of Cook County, given that each of them only shared a small tip with one of the 152 ZIP code areas and their population-weighted centroids were outside the study area. Two other block groups were merged with their neighbors, because their small sizes were completely within a ZIP code and they shared that ZIP code with their neighboring block groups. Finally, 4,260 block group were selected to comprise this reference level. Similar strategies were implemented to assign census tracts to ZIP codes. Only 9 of the Cook County census tracts were excluded, leaving 1,365 tracts.

After establishing the areal units at the three reference levels, the Monte Carlo simulation procedure was applied to disaggregate CRC cases from the ZIP code level to each reference level. The demographic link between the ZIP code level and each smaller geographic level was the 12 race-age-gender categories. Because of the intensive computation time for re-running SaTScan, the number of Monte Carlo simulations was set at 100. Consequently, at each reference level, the spatial scan algorithm was run 100 times, each time on a separate simulated CRC dataset.

3.4. Automation of SaTScan

The spatial scan statistic was utilized to analyze spatial clustering patterns of high late-to-early CRC cancer cases at the ZIP code and the three reference levels. SaTScan was chosen over other spatial clustering methods like Local Indicators of Spatial Autocorrelation (LISA) and Getis-Ord G*, because it uses a varying scanning window and an appropriate maximum likelihood test to detect clusters accurately. Due to these advantages, SaTScan is widely used in cancer research. The specific spatial scan statistic applied in this study was the Bernoulli-based model to address the binary characteristic of late-to-early cancer data. The Bernoulli model has been explicitly described in Kulldorff (1997). A scanning window is passed over the study area. In any scanning window, the number of cases is computed within and outside the window. The likelihood ratio test is utilized to compare the null hypothesis of constant risk within and outside the window with the alternative hypothesis of non-equal risk. The outcomes of the maximum likelihood ratio test provide an indication of the most likely clusters. The formulation of the Bernoulli-based spatial scan statistic is provided below.

$$\lambda = \max_z \left(\frac{c_z}{n_z}\right)^{c_z} \left(1-\frac{c_z}{n_z}\right)^{n_z-c_z} \left(\frac{C-c_z}{N-n_z}\right)^{C-c_z} \left(1-\frac{C-c_z}{N-n_z}\right)^{(N-n_z)-(C-c_z)} \times I\left(\frac{c_z}{n_z} > \frac{C-c_z}{N-n_z}\right) \qquad (1)$$

where c_z is defined as the total number of late-stage CRC cases and n_z the total number of CRC cases within a circular area (Z). Let C be the total number of late-stage CRC cases and N be the total number of CRC cases in the whole study area G. I denotes the indicator function (this formula only maximizes the likelihood function for windows where the observed probability inside the window is larger than the one outside the window).

To implement this procedure, SaTScan uses a coordinate file (the geocoded addresses) to assign each case (a late-stage CRC case) and each control (an early-stage CRC case) into the study area. Then it generates numerous (over 100,000) circular windows, whose centroids are the coordinates of cases, across the whole study area. The radii of these circular windows vary from the smallest observed distance between a pair of cases to a user-defined threshold (Waller and Gotway, 2004). I set the threshold as the

radius containing up to 33% of the entire population of the study area. In each circle, the ratio of late-stage compared to early-stage CRC cases is evaluated and compared with that in the rest of study area. The likelihood ratio statistic is applied to test the null hypothesis of constant risk versus the alternative hypothesis that the late-stage rate within the scanning window is greater than that outside the window. The whole study region is scanned to identify the areas where a significantly higher late-to-early ratio exists based on the observed results of the maximum likelihood ratio statistic. The corresponding parameters of these clusters are calculated, such as the radius of window, population at risk inside the cluster, cases in cluster, the observed-to-expected ratio, and cluster area. The statistical significance of clusters was tested by Monte Carlo simulation with 999 replications. The Bernoulli-based spatial scan statistic normally generates a number of spatial clusters with different p-values. This study focused on primary, statistically significant clusters which had the smallest p-values.

SaTScan is a widely-used software designed to implement the spatial scan statistic. It is freely-accessible and has a straightforward graphical user interface (GUI) (Kulldorff and Information Management Services Inc., 2005). The main components needed to run spatial scan statistics in SaTScan are the input files (containing case file, control file/population file, and coordinate file/grid file), and the parameter file which includes the input interface (specifying input files), analysis interface (containing options for different spatial scan statistics and related parameter settings) and output interface (consisting of options for outputting the results). In this study, the input file section includes the case file (late-stage CRC cases), control file (early-stage CRC cases), and coordinate file (the geographical coordinates of all CRC cases). The analysis interface sets up the settings for the Bernoulli-based spatial scan statistic model, and the output interface specifies two destination files that contain cluster statistics and location information.

For a single run of SaTScan, the procedure is quite simple involving menu-driven data input and setting of parameters. However, using SaTScan repeatedly on different data sets is more difficult: input files in SaTScan are independent and they require separate importing steps. Moreover, the destination files require a non-duplicated name in each run. In comparing geographic clusters of cancer using data for different areal units, SaTScan needs to be run multiple times. Normally, when running SaTScan multiple times, the majority of options in the parameter files remain the same, and only the input and destination files change. Inputting different data sets and specifying unique destination files in every run are cumbersome tasks in running SaTScan repeatedly through its GUI. The situation is exacerbated if we need to run SaTScan hundreds of times. Thus it is very important to make SaTScan run automatically, especially for the steps related to importing input files and naming destination files. However, information and literature on auto-running SaTScan is extremely scarce. The developer of SaTScan (Kulldorff and

Information Management Services Inc., 2005) only provides "Batch File" mode to implement SaTScan automatically, after users manually set up input and parameter files for each run. Abrams and Kleinman (2007) designed the SaTScan Macro Accessory for Cartography (SMAC) package, comprising four SAS macros, to fully automate SaTScan. Nevertheless, SMAC is only available for the Poisson-based spatial scan statistic. Additionally, the macro-syntaxes in SAS are lengthy and complicated, so that it is quite challenging for users to customize or apply the SMAC package, especially for those who are not familiar with macro-level programming in SAS. Therefore, a need exists to automate the whole SaTScan procedure, and it needs to be fulfilled in a cost-effective and efficient way. For this research, a new and much simpler strategy was created by the author to automate the generation of input-and parameter-files, as described below.

The three input files (case, control, and coordinate) were all tab-formatted text files with respective extensions and formats. Among them, the case file with extension 'cas', contained two columns: one was location id that worked as a geographic link between case file and coordinate file; another column described individual late-stage cases, with 1 representing late-stage and 0 early-stage. Similar to the case file, the control file with extension as 'ctl', included the id column and another column for early-stage cases in which an early-stage case was presented as 1 and a late-stage case was coded by 0. The coordinate file consisted of three columns: location id, and x- and y-coordinates to provide geographic information for each case and control, as well as the extension name 'geo'. The simple structures of the case, control and coordinate files inspired the author to auto-generate these input files by macro-level programming in SAS. In each disaggregation outcome, the location id, late-and early-stage cases were stored as separate columns at each reference level. Among these attributes, the location id is the key link to obtain the x-y coordinates from the coordinate files. Therefore, the whole process in SAS was separated into 4 sequential steps: (1) import the disaggregated files and the coordinate file into SAS; (2) merge each imported disaggregated file with the coordinate file, to ensure that every location id in each disaggregated file has x-y coordinates; (3) create the four input files from each disaggregated file, and re-order the variables in each input file to meet the requirements of SaTScan input files; (4) export input files in text format with particular extensions for SaTScan usage. Given that the process in SAS only utilizes basic data-step syntaxes, the macro-level program is cost-effective (syntaxes only around 100 lines), efficient, and has the flexibility to be adjusted for generating input files for different spatial scan models. The detailed Macro-level SAS syntaxes are presented in Appendix A.

As discussed earlier, when SaTScan runs the same spatial scan statistic model repeatedly, the parameter settings remain the same, except for importing and exporting different input-and destination-files. Given the consistency in the parameter file, Java programming was used to automatically produce

different parameter files by changing names of input-and output-files at each reference level. To make this step efficient, the names of input files were varied by consecutive numbers. The names of each parameter file included a constant part plus the consecutive number. Therefore, for every SaTScan run at each reference level, the Java program only needed to set a loop to automatically change the number in the names of input- and destination-files in each parameter file. Afterwards, each parameter file with extension 'prm' was exported for each SaTScan implementation. The programming details in Java are listed in Appendix B. After the two automatic steps of generating input-and parameter-files, the Batch-mode in SaTScan was applied to auto-run the Bernoulli-based spatial scan statistic at the three reference levels.

3.5. Analyzing SaTScan Outcomes

To compare SaTScan outcomes at different geographic scales, the locations, sizes and other characteristics of statistically significant spatial clusters were compared. SaTScan outcomes included the primary cluster at the ZIP code level, and the 100 primary clusters at census tract, census block group and census block levels. However, many of the primary clusters did not achieve statistical significance (p-value<0.1). Only the statistically significant clusters at each level were compared with the ZIP code-level primary cluster. The primary clusters with statistical significance at each reference level were displayed on a map with the ZIP code-level cluster to show the geographic similarity or difference between the results at the two levels. Additionally, the parameters of the statistically significant clusters at each reference level were compared with those at the ZIP code level. The geographical and statistical comparisons between ZIP code level and reference levels reveal the impact of spatial aggregation error on the Bernoulli-based spatial scan statistic results.

4. Results and Discussion

Overall more than half of CRC cases in Cook County in 1998-2002 were diagnosed at a late-stage. The late-stage percentage varies among age, gender, and race groups. Generally, the ratio of late-stage to early-stage cases is 1.5 and 2 in each demographic category (Table 19). The most dramatic excess of late-stage CRC cases compared to early-stage happens in the youngest group. Specifically, the number of late-stage cases is 200% higher than early-stage cases among Black females <50 years old and 225% for Black males < 50 years old. In the Non-Black group, for females < 50 years old, the percent of late-stage cases is more than double that of early-stage cases – 75.23% compared with 24.77% – and the respective percentages are 69.49% and 30.51% in the corresponding male group. The largest numbers of early-and late-stage CRC cases are observed in the oldest age group for every race/gender group.

Table 19. Demographic and Epidemiological Summary of Colorectal Cancer Cases in Cook County from 1998 to 2002

Early-Staged Black		Late-Staged Black		Early-Staged Non-Black		Late-Staged Non-Black	
Female < 50	6 (25%)	Female < 50	18 (75.00%)	Female < 50	27 (24.77%)	Female < 50	82 (75.23%)
Female 50 ~ 70	23 (31.94%)	Female 50 ~ 70	49 (68.06%)	Female 50 ~ 70	163 (36.14%)	Female 50 ~ 70	288 (63.86%)
Female > 70	45 (34.09%)	Female > 70	87 (65.91%)	Female > 70	349 (33.85%)	Female > 70	682 (66.15%)
Male < 50	4 (23.53%)	Male < 50	13 (76.47%)	Male < 50	36 (30.51%)	Male < 50	82 (69.49%)
Male 50 ~ 70	39 (38.61%)	Male 50 ~ 70	62 (61.39%)	Male 50 ~ 70	212 (36.74%)	Male 50 ~ 70	365 (63.26%)
Male > 70	20 (28.57%)	Male > 70	50 (71.43%)	Male > 70	317 (36.02%)	Male > 70	563 (63.98%)

Running SaTScan at the ZIP code level identifies one primary cluster. This cluster occurs in the northwestern section of Cook County, covering the northwestern edge of Chicago city. The radius of the cluster is approximately 6 km, and the cluster covers almost 113 km^2 area. The number of late-stage cases in this cluster is 288, and the relative-risk is 1.141, indicating that CRC patients living in the cluster are approximately 14 percent more likely to be diagnosed with late-stage CRC than those residing outside the zone. In terms of p-value (0.119), the ZIP code-level cluster is not statistically significant according to standard significance levels. However, ZIP codes may be oversized areal units for studying the local patterns of late-stage CRC, and one might suspect that the ZIP code analysis will miss some significant clusters that would be detected based on small-area data. The ZIP code cluster is used as a benchmark for comparison: the clusters with statistically significant p-values at each reference level were selected to compare with the ZIP code cluster. This comparison suggests the validity of ZIP code-level cluster, and the types of clusters it might miss. These comparisons are discussed in the following sections.

At the census tract level, 14 of 100 simulations resulted in statistically significant spatial clusters containing significantly high ratios of late-to-early stage CRC cases. Table 20 displays characteristics of these clusters, including centroid coordinates, the radius and covering area of circular windows, numbers of observed late-stage cases within each cluster, the ratio of observed-to-expected cases, p-values, relative risks, the distance from the ZIP code-level centroid, and the percent of ZIP code cluster area that overlaps with the tract cluster. This table also lists the ZIP code-level parameters in the last row for comparison. Compared to the ZIP code cluster, the significant census tract clusters all have higher relative-risks and ratios of observed-to-expected cases. This localized clustering indicates that in a highly populated region, spatial clusters of CRC cases are more likely to be detected using data for smaller areal units than at the

ZIP code scale. Nine census-tract level clusters overlap the ZIP code cluster, and the overlap percentages vary from 2.74% to 100.0%.

The census tract clusters are shown in Figures 26 and 27. Census tract clusters are illustrated as hollow circles with purple borders in Figure 26, and the ZIP code cluster is displayed as a green circle. The centroids of each cluster are mapped in Figure 27. In Figure 1, two clusters (12 and 13 in Table 20) at the census-tract level have very similar covering areas as the one at the ZIP code level, and the 12th cluster can almost be treated as a replica of the ZIP code-level one except for a small curved area outside of the ZIP code cluster zone. The 13th cluster includes more area than the ZIP code one, including a crescent-shaped buffer surrounding the ZIP code cluster. The 6th cluster also highly overlaps the ZIP code cluster, covering 76.52% of its area. At the southeastern edge of the ZIP code cluster, four census tract clusters are completely within the ZIP code cluster, and another cluster mainly falls into the ZIP code cluster except for a small tip outside.

On the other hand, five clusters at the census-tract level are completely outside the ZIP code-level cluster: one is close to the northern border of Cook County, two are southeast of the ZIP code-level cluster, and other two are located at the southern border of Chicago city. However, these clusters are small and the numbers of observed cases within these clusters are no larger than 35. Figure 27 also shows the location of each tract cluster centroid in relation to the one at the ZIP code level. The centroid of the 12th cluster is almost identical to the ZIP code one. The centroids of eight other clusters also closely surround the ZIP code centroid, with distances ranging from 0.92 km to 6.7 km (Table 20). On the other hand, four clusters have centroids located more than 10 km from the ZIP code cluster centroid. In summary, generally the tract-level clusters correspond quite well geographically to the ZIP code cluster, although the ZIP code cluster fails to represent some smaller, distant clusters that are detected with tract-level data.

Table 20. Results of Bernoulli-based Spatial Scan Statistic at ZIP Code and Census Tract Levels

Cluster	Centroid Coordinates	Radius (km)	Area(km^2)	Cases in Cluster	O/E	p-value	Relative-Risk	Distance (km)	Overlap Area (% of ZIP code level cluster)
1	39.773; -85.044	1.326	5.523	23	1.530	0.0420	1.535	10.827	0.00
2	39.734; -84.990	1.223	13.396	26	1.530	0.0170	1.536	4.617	4.89
3	39.809; -85.252	2.065	5.523	23	1.530	0.0550	1.535	27.681	0.00
4	39.708; -85.252	0.984	3.041	21	1.530	0.0900	1.535	5.269	2.47
5	39.697; -85.023	3.719	43.450	126	1.228	0.0949	1.241	3.448	33.12
6	39.703; -84.955	5.274	87.381	204	1.178	0.0874	1.195	0.922	76.52
7	39.798; -85.327	1.834	10.566	23	1.530	0.0590	1.535	33.137	0.00
8	39.781; -85.052	1.012	3.217	35	1.488	0.0116	1.495	11.982	0.00
9	39.731; -85.028	3.842	46.371	144	1.231	0.0290	1.246	6.706	13.26

Table 20. (cont.)

10	39.586; -84.808	2.009	12.679	21	1.530	0.0900	1.535	18.266	0.00
11	39.712; -85.008	1.472	6.806	57	1.363	0.0658	1.372	4.193	6.03
12	39.699; -84.965	6.248	122.637	278	1.150	0.0900	1.170	0.218	100.00
13	39.703; -84.955	7.249	165.081	336	1.137	0.0648	1.160	0.922	100.00
14	39.704; -85.002	0.993	3.097	22	1.530	0.0698	1.535	3.455	2.74
ZIP Code Level	39.698; -84.963	5.994	112.878	288	1.124	0.119	1.141		

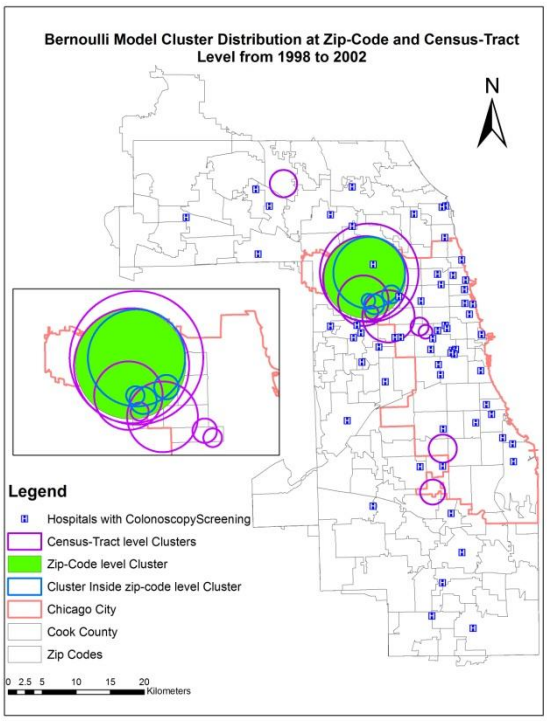

Figure 26. The Distribution of Clusters at Census Tract and ZIP Code Levels in Cook County

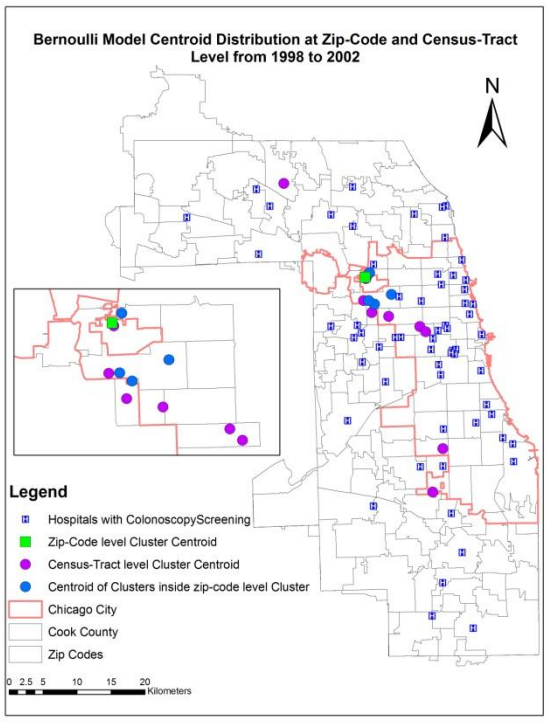

Figure 27. The Distribution of Cluster Centroids at Census Tract and ZIP Code Levels in Cook County

The block group-level spatial scan statistic generated 18 clusters with significantly high late-to-early ratios, more than were found at either of the other two levels (Table 21). Similar to the clusters at the census-tract level, these block-group clusters present higher ratios of observed-to-expected cases and larger relative risks than the one at the ZIP code level. Block group clusters tend to be smaller than those at the tract and ZIP code levels. Only one block group level cluster overlaps greatly (85.24% overlap area) with the ZIP code cluster. Nine other clusters overlap with small sections of the ZIP code cluster: overlap

percentages range from 2.31% to 14.52% of the ZIP code cluster area. The number of block-group level clusters that are completely outside the ZIP code clustering zone is eight, compared with only five at the census-tract level. These 'outside' clusters appear to have larger radii and cover more area than the 'outside' ones at the census tract level. Furthermore, the p-values at block-group level are generally smaller than the ones at census-tract level, indicating that this smaller area level is capable of detecting more distinctive patterns of late-stage CRC clustering.

Figure 28 describes the spatial distribution of block-group level clusters, illustrated by hollow circles with yellow boundaries. Similar to the census-tract results, the block-group level clusters that overlap with the ZIP code cluster are often located along the southern part of the ZIP code cluster, indicating a tendency for clusters to be focused in this area. Among the ten clusters that overlap with the ZIP code cluster, only two clusters (highlighted by blue boundary) lie completely inside, respectively occupying 14.52% and 6.24% of the ZIP code cluster area. Among 'outside' clusters, several appear southeast of the ZIP code cluster, in locations similar to those detected with census tract data. The other 'outside' clusters are also located in areas similar to clusters at the census-tract level. One appears in the northern part of Cook County and two others around the southern border of Chicago city. Furthermore, their radii and covering areas are generally larger than those for the corresponding clusters at the census tract level. These 'outside' clusters reveal that using data at the block group level enhances the possibility of detecting late-stage CRC clusters outside the dominant clustering area compared to using data at the tract or ZIP code levels. Examining block group cluster centroid locations in Figure 29, shows a concentration of centroids near the ZIP code centroid and along the southeastern edge of the ZIP code cluster – a pattern similar to that observed based on tract-level data. Fourteen of the eighteen block group cluster centroids lie within the 11km buffering zone of the ZIP code-level centroid, indicating a relatively good geographic correspondence between clusters at both levels. However, four cluster centroids fall far outside, with centroid distances ranging from 17 to 37 kilometers (Table 21). The maximum value of the block group-level distances to the ZIP code centroid is 37 km, compared with 33 km at census tract level.

Table 21. Results of Bernoulli-based Spatial Scan Statistic at ZIP Code and Block Group Levels

Cluster	Centroid Coordinates	Radius (km)	Area(km^2)	Cases in Cluster	O/E	p-value	Relative-Risk	Distance (km)	Overlap Area (% of ZIP code level cluster)
1	39.675; -84.989	2.284	16.388	25	1.530	0.0440	1.536	3.393	14.52
2	39.717; -85.017	1.151	4.162	25	1.530	0.0480	1.536	5.183	3.32
3	39.712; -85.008	1.498	7.049	40	1.493	0.00343	1.501	4.193	6.24
4	39.742; -85.049	2.490	19.477	47	1.410	0.0694	1.419	8.826	0.00
5	39.811;-85.204	1.368	5.879	24	1.530	0.0634	1.536	23.764	0.00
6	39.712; -85.017	1.615	8.193	46	1.408	0.0679	1.416	4.864	6.49

Table 21. (cont.)

7	39.713; -85.021	0.949	2.829	25	1.530	0.0530	1.536	5.288	2.31
8	39.695; -84.973	5.733	103.253	244	1.178	0.0410	1.198	0.976	85.24
9	39.611; -84.799	3.101	30.209	26	1.530	0.0390	1.536	17.086	0.00
10	39.723; -85.025	2.034	12.996	59	1.368	0.0513	1.378	6.056	5.12
11	39.729; -85.040	3.349	7.942	107	1.289	0.0190	1.303	7.451	6.09
12	39.774; -85.042	1.590	3.684	27	1.530	0.0240	1.536	10.804	0.00
13	39.782; -85.045	1.082	15.329	31	1.483	0.0714	1.489	11.677	0.00
14	39.878; -85.291	2.209	18.064	25	1.530	0.0470	1.536	34.514	0.00
15	39.723; -85.040	2.398	68.895	54	1.378	0.0890	1.386	7.197	2.73
16	39.721; -85.013	1.448	6.559	32	1.484	0.0543	1.490	4.976	5.21
17	39.837; -85.358	4.683	14.065	92	1.291	0.0826	1.303	37.271	0.00
18	39.775; -85.042	2.116	35.234	48	1.412	0.0490	1.421	10.968	0.00
ZIP Code Level	39.698; -84.963	5.994	112.878	288	1.124	0.119	1.141		

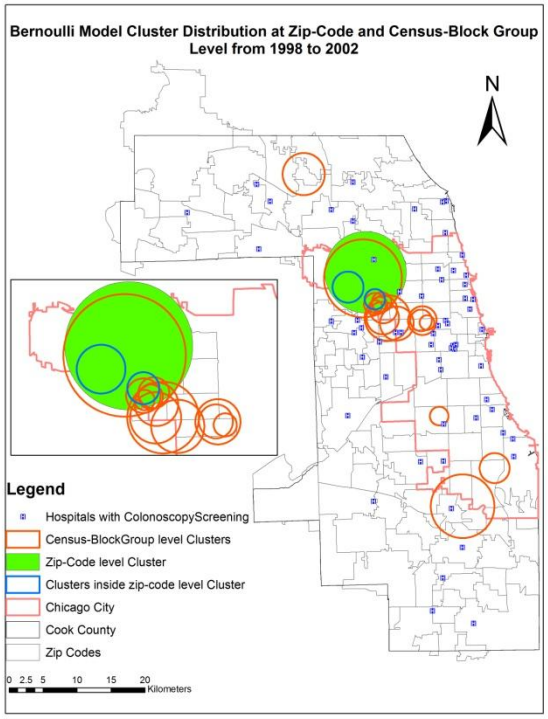

Figure 28. The Distribution of Clusters at Census Block Group and ZIP Code Levels in Cook County

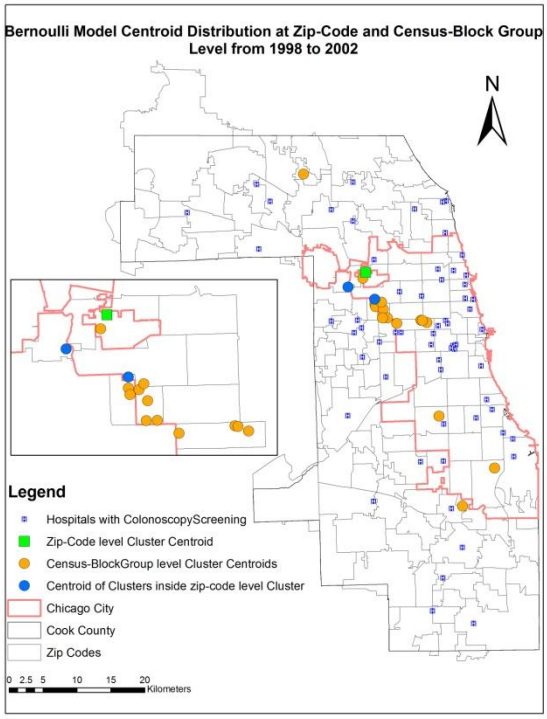

Figure 29. The Distribution of Cluster Centroid at Census Block Group and ZIP Code Levels in Cook County

Fifteen clusters have significantly high late-to-early ratios based on block-level data. These clusters tend to be smaller in size than those at the tract or block group level (Table 22): their radii and covering areas are generally smaller than the ones at census tract and block group levels. As the smallest reference unit, blocks provide the most localized detail about the spatial clustering patterns of late-stage CRC cases. The majority of block-level clusters present high ratios of observed-to-expected late-stage cases and relative risks, indicating that more localized variation in late-stage CRC cases can be detected using data for the smallest reference unit. Numbers of observed cases in each block-level cluster are generally less than those in clusters at other levels, so the block data uncover small, localized clusters of late-stage CRC. The percentages of ZIP code cluster area that overlap with the block-level clusters are much less than the ones for clusters at the other two scales, because of the small sizes of block-level clusters.

Figure 30 displays the clusters with statistically significant p-values at the block level as hollow circles with red boundaries. These clusters clearly reveal concentrations of high late-to-early ratios around the eastern and southeastern sections of the ZIP code-level cluster. Three clusters (highlighted by blue boundary) completely fall inside the ZIP code-level cluster, respectively covering 15.66%, 7.35%, and 5.22% of the ZIP code-level cluster area. Five clusters at the block level are located completely outside the ZIP code-level cluster: four are located southeast of the ZIP code-level cluster, and another one is placed in the southern part of Cook County. Clusters in the northern part of Cook County and around the southwestern border of Chicago city that emerged in the tract and block group analyses do not appear in the block-level analysis. The reason may be that the simulated cancer cases at the block level are more evenly distributed than those at the tract and block group levels, resulting in less tendency towards clustering. Of course, the Monte Carlo simulation involves a random assignment procedure in which the spatial disaggregation of cases within ZIP codes is only based on demographic information and is otherwise spatially random. In areas with few CRC cases, disaggregation of cases to the block level may result in more dispersed geographic patterns.

In terms of distances between centroids, only the cluster located in the southern part of Chicago city presents a relative long distance, 34 km (Figure 31). Centroids of the three ‘inside’ clusters are located near the ZIP code centroid, with centroid distance of 4.1 km or less. The other clusters have distances varying from 3.3 km to 12.4 km. Compared to the clusters at census-tract and block-group levels, the block-level clusters reveal more detailed spatial aggregations of late-stage CRC cases in areas containing large numbers of late-stage CRC cases. However, in regions with fewer late-stage CRC cases, such as the northern part of Cook County and southeastern section of Chicago city, the block level failed to identify clusters with significantly high ratios.

Table 22. Results of Bernoulli-based Spatial Scan Statistic at ZIP Code and Census Block Levels

Cluster	Centroid Coordinates	Radius (km)	Area(km^2)	Cases in Cluster	O/E	p-value	Relative-Risk	Distance (km)	Overlap Area (% of ZIP code level cluster)
1	39.722; -85.021	1.646	8.511	42	1.428	0.0898	1.436	5.644	4.57
2	39.785; -85.053	1.970	12.191	45	1.434	0.0390	1.443	12.382	0.00
3	39.786; -85.040	1.205	4.561	33	1.530	0.00311	1.538	11.823	0.00
4	39.733; -84.972	2.378	17.764	71	1.341	0.0524	1.352	3.971	15.08
5	39.774; -85.046	1.671	8.771	33	1.485	0.0714	1.492	11.080	0.00
6	39.724; -85.018	1.271	5.075	27	1.530	0.0420	1.536	5.546	3.15
7	39.872; -85.290	1.854	10.798	24	1.530	0.0849	1.536	34.111	0.00
8	39.673; -84.981	2.371	17.65	24	1.530	0.0867	1.536	3.160	15.66
9	39.693; -84.954	1.625	8.295	25	1.530	0.0612	1.536	0.968	7.35
10	39.721; -85.018	2.435	18.626	81	1.319	0.0669	1.330	5.403	10.07
11	39.772; -85.050	1.451	6.614	27	1.530	0.0400	1.536	11.155	0.00
12	39.672; -84.983	2.788	24.418	25	1.530	0.0545	1.536	3.387	21.35
13	39.662; -84.978	3.284	33.880	28	1.530	0.0220	1.536	4.163	23.76
14	39.668; -84.966	2.979	27.879	26	1.530	0.0380	1.536	3.324	24.04
15	39.716; -85.005	1.370	5.896	31	1.530	0.00680	1.537	4.098	5.22
ZIP Code Level	39.698; -84.963	5.994	112.878	288	1.124	0.119	1.141		

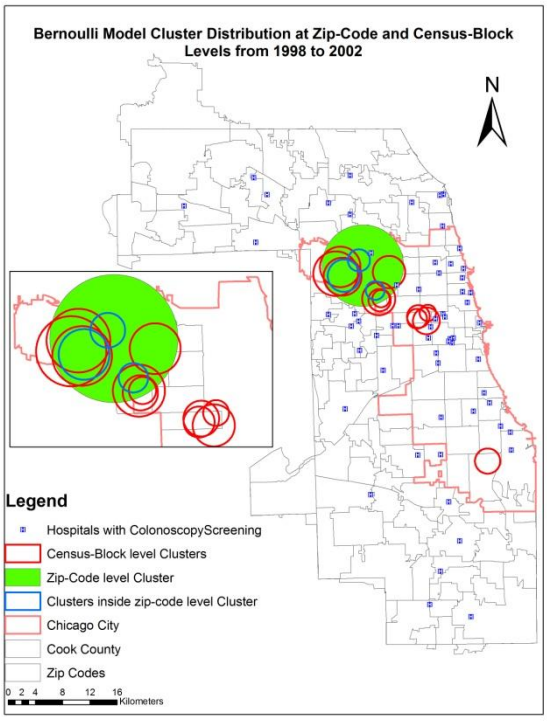

Figure 30. The Distribution of Clusters at Census Block and ZIP Code Levels in Cook County

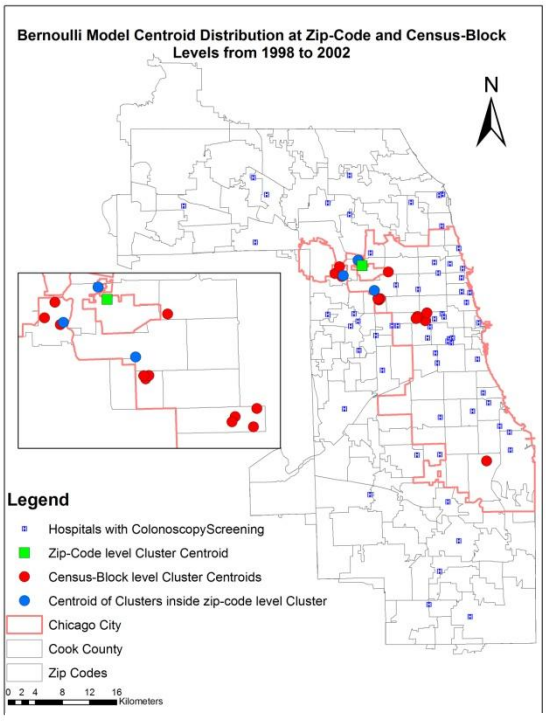

Figure 31. The Distribution of Cluster Centroids at Census Block and ZIP Code Levels in Cook County

Comparing the ZIP code-level cluster with the clusters at the three reference levels indicates strengths and weaknesses of using the ZIP code level as the study unit. Specifically, the Bernoulli-based spatial scan statistic at the ZIP code level can detect clusters in areas with large concentrations of cases. However, even in these concentrated settings, the ZIP code cluster is at the global-level, which means it gives general clustering information without much local detail. In other areas with fewer CRC cases, the ZIP code level is too large to detect 'local-level' clusters. Thus, spatial aggregation error may have more influence in areas where the sample size is small, compared to areas with many cases.

5. Conclusion

This study compared the results of the Bernoulli-based spatial scan statistic at the ZIP code level with the outcomes at three reference census units (census tract, census block group, and census block) to examine if reliable and accurate spatial analysis results can be generated using ZIP code-level data.

Lacking actual data on patient locations by census tract, block group and block, a Monte Carlo simulation procedure was used to disaggregate cancer cases from the ZIP code level to smaller census geographic units. Thus, the research focused on possible geographic patterns of CRC cases that conform to the demographic and geographic characteristics of cases at the ZIP code level. The number of simulated results was 100 at each reference level, and every result was tested for spatial clustering using the Bernoulli-based spatial scan statistic in SaTScan. Because the steps of importing input files and providing non-duplicated output names in each SaTScan run were tedious and time-consuming to perform manually, I designed a cost-effective procedure to automate the running of SaTScan. This procedure mainly consisted of a Macro-level SAS program to automatically generate input files and a Java program to automate the parameter file generation. Compared with the SMAC package created by Abrams and Kleinman (2007), my procedure is simpler, more efficient and highly adaptive to other spatial scan statistics in SaTScan, because it only comprises two small programs and there is no need to build the major part of a parameter file.

Comparing geographic clusters with statistically significant p-values at each reference level with the ZIP code cluster yielded several innovative results. One important observation was that only a small number (14 – 18) of the simulated data patterns at each reference unit produced statistically significant clusters. Thus, the fact that the ZIP code-level cluster had a p-value of 0.12 seems appropriate, given that 80-85% of the clusters generated based on simulated data at each reference level were not statistically significant.

The spatial scan statistic at the ZIP code level did well at identifying a primary cluster in an area with a high density of cases. However, the spatial scan analysis at this level lost the power to detect more localized clusters. In some instances, the simulated datasets contained statistically significant clusters located in areas with smaller numbers of late-stage cases. Even in the areas with a large sample size of cases, using ZIP code level data fails to detect statistically significant clusters that appear at the census tract, block group and block levels. At these levels, clusters often were detected along an axis extending southeast of the ZIP code level cluster. Some of these clusters partially overlapped with the ZIP code cluster, while others did not.

In areas containing fewer CRC cases, utilizing ZIP code level data misses statistically significant clusters that are detected based on small-area data. Clusters located near the northern border of Cook County and southern border of Chicago city could not be detected at the ZIP code level. At the block level, simulated data contained significant clusters located in the eastern, western, southeastern parts of the ZIP code-level cluster and some surrounding areas. Although some of these clusters overlapped the ZIP code cluster, others were more geographically distinct. Thus, the spatial scan statistic at the ZIP code scale can produce reliable and stable 'global-level' results, however it has difficulty in identifying clusters at a smaller and more localized level. If cancer data for small areas is not available, applying the spatial scan statistic at the ZIP code level can detect the dominant cluster(s) in areas where the sample size is large.

Additionally, the influence of spatial aggregation error on the spatial scan analysis may vary across the study area, depending on the densities of cases within different local areas. The influence is typically greater in areas with a low density of cases, where the combination of low statistical power and spatial aggregation of cases makes it difficult to detect localized clusters. Although utilizing ZIP code-level data made it possible to detect a stable and large cluster in Cook County, the ZIP code-level spatial scan analysis was less appropriate for detecting clusters in areas with a lower density of CRC cases. Thus, to detect spatial clusters using the spatial scan statistic, a trade-off strategy needs to be applied in selecting the study unit. In areas with large numbers of cases, using the smallest unit, such as census block, can reveal localized clusters in great detail. However, in areas with fewer cases, using a 'middle-size' study level which can contain enough sample size of cases without oversized concern, such as census tract or census block-group, the clustering patterns can be identified better than using the smallest areal unit. Thus, the optimum study size of the spatial scan statistic needs to be varied based on the distribution of cases in different regions across the whole study area.

Given that the resulting quality of spatial scan statistic highly depends on the spatial displacement and density of cases in a specific area, using a uniform policy to release cancer data for research in

different study areas is not very appropriate. In areas with high density of cases, cancer data can be released at a smaller areal unit without violating the privacy issue. In areas including fewer cases, cancer data can be published for medium-sized areal units in order to detect significant clusters as well as conform to confidentiality regulations.

Several limitations and drawbacks need to be pointed out in this study. The distributions of CRC cases at the three reference levels were computed by simulation, and they do not represent actual CRC case locations. This study also constrained the study unit to Cook County, a highly-populated and urbanized area, and the corresponding results may not be applicable to suburban or rural areas. The edge effect may add some bias to the results of the spatial scan statistics at all levels of analysis, especially in locations along the boundary of the study area. The findings also may be limited by errors in assigning tracts, block groups and block to their respective ZIP code areas. Additionally, the number of simulated datasets at each reference level was constrained as 100, given the very long processing time of each run in SaTScan. The simulated data patterns may not capture all the possible distributions of cancer cases at each reference level and bring potential bias with the inadequate possibilities. With the rapid development of super computing, speeding up the application of SaTScan in the super computing environment may become a reality in the near future. Then the number of simulated datasets can be increased to include large numbers of distribution possibilities to provide a much more unbiased analysis.

The main tasks for future research are to overcome data limitations and design more appropriate spatial relationships in the disaggregation method to deal with the problem of multiple ZIP code areas overlapping a single census tract/block group area. Introducing ZIP code tabulation areas (ZCTAs) may be a good idea in terms of using their internal populations to compute the weights for assigning cancer cases from one ZIP code area to its shared multiple census tracts (US Census Bureau, 2001). In future research, it is also important to use an enlarged study area – a buffer zone – to deal with the edge effect. Similar to the influence of the spatial aggregation error on ZIP code level statistical analysis (Luo et al., 2010), the impact of this error on spatial scan analysis has also been found to be case-sensitive and to vary with the number of cases across the study area. This empirical evaluation of spatial aggregation error was limited to an urban setting and may not apply to suburban-and rural-areas. More diverse study areas should be studied to obtain detailed information about the impact of spatial aggregation error on ZIP code-level spatial scan statistics.

6. References

Abrams and Kleinman. **A SaTScan™ Macro Accessory for Cartography (SMAC) Package Implemented with SAS Software**. *International Journal of Health Geographics*, 2007, **6**:6.

Armhein C. **Searching for the Elusive Aggregation Effect: Evidence from Statistical Simulations**. *Environment & Planning A*, 1994, **27**: 105-109.

Fortney J., Kathryn R., Warren J. **Comparing Alternative Methods of Measuring Geographic Access to Health Services**. *Health Services & Outcomes Research Methodology*, 2000, **1(2)**: 173-184.

Gregorio D.I., Kulldorff M., Barry L., and Samociuk H. **Geographic Differences in Invasive and in Situ Breast Cancer Incidence according to Precise Geographic Coordinates, Connecticut, 1991-1995**. *International Journal of Cancer*, 2002, 100: 194-198.

Gregorio D.I., and Samociuk H. **Breast Cancer Surveillance Using Gridded Population Units, Connecticut, 1992 to 1995**. *Annals of Epidemiology*, 2003, **13**: 42-49.

Gregorio D.I., Kulldorff M., Sheehan T.J., and Samociuk H. **Geographic Distribution of Prostate Cancer Incidence in the Era of PSA Testing, Connecticut, 1984 to 1998**. *Urology*, 2004, 63: 78-82.

Gregorio D.I., DeChello L.M., Samociuk H., and Kulldorff M. **Lumping or Splitting: Seeking the Preferred Areal Unit for Health Geography Studies**. *International Journal of Health Geographics*, 2005, **4**:6.

Hewko J., Smoyer-Tomic K.E., Hodgson M.J. **Measuring Neighborhood Spatial Accessbility to Urban Amenities: Does Aggregation Error Matter?** *Environment and Planning A*, 2002, 34: 1185-1206.

Hillsman E., and Rhoda R. **Errors in Measuring Distances from Populations to Services Centers**. *Annals of Regional Science*, 1978, 12: 74-88.

Hodgson M.J., Shmulevitz F., Körkel M. **Aggregation Error Effects on the Discrete-Space P-Median Model: The Case of Edmonton, Canada**. *The Canadian Geographer*, 1997, **41**: 415-428.

Jemal A., Kulldorff M., Devesa S.S. Hayes R.B., and Fraumeni J.F.Jr. **A Geographic Analysis of Prostate Cancer Mortality in the United States, 1970-89**. *International Journal of Cancer*, 2002, **101**: 168-174.

Krieger N., Chen J.T., Waterman P.D., Soobader M.J., Subranmanian S.V., and Carson R. **Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-Based Measure and Geographic Level Matter?** *American Journal of Epidemiology*, 2002, **156**: 471-482.

Kulldorff M. **A Spatial Scan Statistic**. *Communications in Statistics-Theory and Methods*, 1997, **26(6):** 1481-1496.

Kulldorff M., Feuer E.J., Miller B.A., and Freedman L.S. **Breast Cancer Clusters in the Northeast United States: a Geographic Analysis.** *American Journal of Epidemiology*, 1997, **146**: 161-170.

Kulldorff M., and Information Mangement Services, Inc.: **SatScanTM v5.1: Software for the Spatial and Space-Time Scan Statistics**. 2005. Available at: http://www.satscan.org (Accessed: Dec 12th, 2010).

Luo L., McLafferty S., and Wang F. **Analyzing Spatial Aggregation Error in Statistical Models of Late-Stage Cancer Risk: a Monte Carlo Simulation Approach**. *International Journal of Health Geographics*, 2010, 9:51.

Openshaw S., and Alvandies S. **Applying Geocomputing to the Analysis of Spatial Distributions**. *In Geographic Information Ssystems: Principles and Technical Issues Volume I.* 2nd edition. Edited by: Longley P., Goodchild M., Maguire D., Rhind D., New York: John Wiley and Sons, Inc; 1999.

Pollack L.A., Gotway C.A., Bates J.H., Parikh-Patel A., Richards T.B., Seeff L.C., Hodges H., and Kassim S. **Use of the Spatial Scan Statistic to Identify Geographic Variations in Late Stage Colorectal Cancer in California (United States).** *Cancer Causes Control*, 2006, **17**: 449-457.

Roche L.M., Skinner R., and Weinstein R.B. **Use of a Geographic Information Systems to Identify and Characterize Areas with High Proportions of Distant Stage Breast Cancer**. *Journal of Public Health Management and Practice*, 2002, 8: 26-32.

Rushton G. **Methods to Evaluate Geographic Access to Health Services**. *Journal of Public Health Management & Practice*, 1999, 5:93-100.

Rushton G., Peleg I., Banerjee A., Smith G., and West M. **Analyzing Geographic Patterns of Disease Incidence: Rates of Late-Stage Colorectal Cancer in Iowa**. *Journal of Medical Systems*, 2004, **28**: 223-236.

Seeff L.C., Nadel M., Blackman C., and Pollack L.A. **Colorectal Cancer Test Use among Persons Aged Greater than or Equal to 50 Years-United States, 2001**. *Morbidity and Mortality Weekly Report*, 2003, **52**: 193-195.

Sheehan T.J., Gershman S.T., McDougal L., Danley R.A., Mroszczyk M., Sorensen A.M., and Kulldorff M. **Geographic Surveillance of Breast Cancer Screening by Tracts, Towns and Zip Codes.** *Journal of Public Health Management and Practice*, 2000, **6**: 48-57.

Thomas A., and Carlin B.P. **Late Detection of Breast and Colorectal Cancer in Minnesota Counties: An Application of Spatial Smoothing and Clustering**. *Statistics in Medicine*, 2003, **22**: 113-127.

U.S. Census Bureau. **United States Census 2000 Summary File 1(SF1)**. U.S. Census Bureau, Washington D.C. Available at: http://www.census.gov/census2000/sumfile1.html (Accessed: January 21st, 2011).

U.S. Census Bureau. **Census Tracts and Block Numbering Areas**. U.S. Census Bureau, Washington D.C., 2000. Available at: http://www.census.gov/geo/www/cen_tract.html. Retrieved 2007-12-05 (Accessed: January 3rd, 2011).

U.S. Census Bureau. **American FactFinder Help-Glossary**. U.S. Census Bureau, Washington D.C., 2000. Available at: http://factfinder.census.gov/home/en/epss/glossary_b.html (Accessed: February 3nd, 2011).

U.S. Census Bureau. **ZIP Code Tabulation Areas (ZCTAs™).** U.S. Census Bureau, Washington D.C., 2001. Available at: http://www.census.gov/geo/ZCTA/zcta.html (Accessed: March 12th, 2011).

Waller L.A., and Gotway C.A. **Applied Spatial Statistics for Public Health Data**, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2004.

Appendix A. Macro-level SAS Syntaxes to Generate Input Files for SaTScan

```
*Import the centroids' x y coordinates at a census unit;
PROC IMPORT OUT= WORK.xy
            DATAFILE= "Directory\censusunit_Cook_XYCoord.dbf"
            DBMS=DBF REPLACE;
     GETDELETED=NO;
RUN;
*Import disaggregated results into SAS, from res_001 to res_100;
%macro loop;
%do i=1 %to 100;

PROC IMPORT OUT= WORK.res%sysfunc(putn(&i,z3.))
          DATAFILE= "Directory\res_%sysfunc(putn(&i,z3.)).xls"
          DBMS=EXCEL REPLACE;
   RANGE="res$";
   GETNAMES=YES;
   MIXED=NO;
   SCANTEXT=YES;
   USEDATE=YES;
   SCANTIME=YES;
RUN;

*Merge centroids and each result file (from res_001 to res_100);
proc sort data=xy; by STFID; run;
proc sort data=res%sysfunc(putn(&i,z3.)); by STFID; run;
data res%sysfunc(putn(&i,z3.))_xy;
      set res%sysfunc(putn(&i,z3.)) xy;
      merge res%sysfunc(putn(&i,z3.)) xy;
      by STFID;
run;

data res%sysfunc(putn(&i,z3.))_xy;
      set res%sysfunc(putn(&i,z3.))_xy;
      if ID^=.;
run;

*Separate into three SAS datasets;
/* Case File;*/
data case_%sysfunc(putn(&i,z3.));
      set res%sysfunc(putn(&i,z3.))_xy;
      keep STFID late;
run;
data case_%sysfunc(putn(&i,z3.));
      set case_%sysfunc(putn(&i,z3.));
      count=late;
      drop late;
run;

/*Control File;*/
data cont_%sysfunc(putn(&i,z3.));
      set res%sysfunc(putn(&i,z3.))_xy;
      keep STFID early;
run;
data cont_%sysfunc(putn(&i,z3.));
      set cont_%sysfunc(putn(&i,z3.));
```

```
      control=early;
      drop early;
run;

/*Geographic File;*/
data coor_%sysfunc(putn(&i,z3.));
      set res%sysfunc(putn(&i,z3.))_xy;
      keep STFID x y;
run;

*Re-order the variables in coordinates files;
data cnew_%sysfunc(putn(&i,z3.));
      retain STFID y x;
      set coor_%sysfunc(putn(&i,z3.));
run;

*Export text files (case, control, and coordinate) for the SaTScan use;
data _null_;
      set case_%sysfunc(putn(&i,z3.));
      file "Directory\SaTScan_Files\case_%sysfunc(putn(&i,z3.)).cas"
dlm="09"X;
put STFID
    count;
run;

data _null_;
set cont_%sysfunc(putn(&i,z3.));
file "Directory\SaTScan_Files\cont_%sysfunc(putn(&i,z3.)).ctl" dlm="09"X;
put STFID
   control;
run;

data _null_;
set cnew_%sysfunc(putn(&i,z3.));
file "Directory\SaTScan_Files\coor_%sysfunc(putn(&i,z3.)).geo" dlm="09"X;
put STFID
    y
    x;
run;

%end;
%mend;

options mprint mlogic symbolgen;
%loop
```

Appendix B. The Java Program to Automatically Generate Parameter Files

```
import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;

public class FileGen{
	public static void main(String[] args) throws IOException{
		String strLine, strLine1 = "", strLine2 = "", strLine3 = "", strLine4 = "";
		String newName = "";
		String outName = "";
		BufferedReader input;
		BufferedWriter output;
		for(int i=1;i<=100;i++){
			try {
			 //Auto-name the parameter files by the consecutive numbers
				if(i<10)
					outName = "output/prm_00"+i+".prm";
				else if(i<100)
					outName = "output/prm_0"+i+".prm";
					output = new BufferedWriter(new FileWriter(outName));
				input = new BufferedReader(new FileReader("input/parameterfile.txt"));
				int j=0;

			//Auto-change the names of case files to input each case file
```

```
while ((strLine = input.readLine()) != null){
                    if (j==2){
                            if(i<10)
                                    newName = "case_00"+i+".cas";
                            else if(i<100)
                                    newName = "case_0"+i+".cas";
                            strLine1 = strLine.replace("case_001.cas", newName);
                    }
//Auto-change the names of control files to input each control file
                    if (j==4){
                            if(i<10)
                                    newName = "cont_00"+i+".ctl";
                            else if(i<100)
                                    newName = "cont_0"+i+".ctl";
                            strLine2 = strLine.replace("cont_001.ctl", newName);
                    }
 //Auto-change the names of coordinate files to input each coordiante file
                    if (j==8){
                            if(i<10)
                                    newName = "coor_00"+i+".geo";
                            else if(i<100)
                                    newName = "coor_0"+i+".geo";
                            strLine3 = strLine.replace("coor_001.geo", newName);
                    }
    //Auto-change the names of output files to generate each destination file
                    if (j==36){
                            if(i<10)
                                    newName = "output"+i;
```

```
                    else if(i<100)
                            newName = "output"+i;
                    strLine4 = strLine.replace("output1", newName);
                    //System.out.println(strLine);
            }
            j++;
        }
        input.close();
        input = new BufferedReader(new FileReader("input/paramterfile.txt"));
//Auto-generate each parameter file by changing the input-and output files
        j=0;
        while ((strLine = input.readLine()) != null){
            if (j==2){
                strLine = strLine1;
            }
            if (j==4){
                strLine = strLine2;
            }
            if (j==8){
                strLine = strLine3;
            }
            if (j==36){
                strLine = strLine4;
            }
            output.write(strLine + "\n");
            j++;
        }
```

```
                output.close();
                input.close();
            } catch (FileNotFoundException e1) {
                e1.printStackTrace();
            } catch (IOException e) {
                System.err.println("Error: " + e.getMessage());
                e.printStackTrace();
            }
        }
    }

    private static String getFileNumber(int i){
        if (i<10)
            return "00"+i;
        else if(i<100)
            return "0" +i;
        else
            return String.valueOf(i);
    }
}
```