12-2011

Spatio-temporal modeling of arthropod-borne zoonotic diseases: a proposed methodology to enhance macro-scale analyses

Stephen Jones
Clemson University, stephen_jones@bcbst.com

SPATIO-TEMPORAL MODELING OF ARTHROPOD-BORNE ZOONOTIC DISEASES: A PROPOSED METHODOLOGY TO ENHANCE MACRO-SCALE ANALYSES

A Dissertation
Presented to
the Graduate School of
Clemson University

In partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Wildlife Biology

by
Stephen G. Jones
December 2011

Accepted by:
Dr. William Conner, Chair
Dr. Bo Song
Dr. David Gordon
Dr. Anand Jayakaran

ABSTRACT

Zoonotic diseases are infectious diseases that can be transmitted from or through animals to humans, and arthropods often act as vectors for transmission. Emerging infectious diseases have been increasing both in prevalence and geographic range at alarming rates the last 30 years, and the majority of these diseases are zoonotic in nature. Many zoonotic diseases are considered notifiable by the Centers for Disease Control and Prevention (CDC). However, though state regulations or contractual obligations may require the reporting of certain diseases, significant underreporting is known to exist. Because of the rich volume of information captured in health insurance plan databases, administrative medical claims data could supplement the current reporting systems and allow for more comprehensive spatio-temporal analyses of zoonotic infections.

The purpose of this dissertation is to introduce the use of electronic administrative medical claims data as a potential new source that could be leveraged in ecological field studies in the surveillance of arthropod-borne zoonotic diseases. If using medical claims data to study zoonoses is a viable approach, it could be used to improve both the temporal and spatial scale of study through the use of long-term longitudinal data covering large geographic expansions and more geographically refined ZIP code scales. Additionally, claims data could supplement the current reporting of notifiable diseases to the CDC. This effort may help bridge the disease incidence gap created by health care providers' underreporting and thus allow for more effective tracking and monitoring of infectious zoonotic diseases across time and space.

I specifically examined 5 tick-borne (Lyme disease [LD], babesiosis, ehrlichiosis, Rocky Mountain spotted fever [RMSF], and tularemia) and 2 mosquito-borne (West Nile virus, La Crosse viral encephalitis) diseases known to occur in the southeastern US. I first compared disease incidence rates from cases reported to the Tennessee Department of Health (TDH) state registry system with medically diagnosed cases captured in a southeastern managed care organization (MCO) claims data warehouse. I determined that LD and RMSF are significantly underreported in Tennessee. Three (3) cases of babesiosis were discovered in the claims data, a significant finding as this disease has never been reported in Tennessee. Next, I used a cluster scan statistic to statistically validate when (temporal) and where (spatial) these data sources differ. Findings highlight how the data sources do not overlap in their significant cluster results, supporting the need to integrate administrative and state registry data sources in order to provide a more comprehensive set of case information. Once the usefulness of administrative data was demonstrated, I focused on how these data could improve spatio-temporal macro-scale analyses by examining information at the ZIP code level as opposed to traditional county level assessments. I expanded on the current literature related to spatially explicit modeling by employing more advanced data mining modeling techniques. Four separate modeling techniques were compared (stepwise logistic regression, classification and regression tree, gradient boosted tree, and neural network) to describe the occurrence of tick-borne diseases as they relate to socio-demographic, geographic, and habitat characteristics. Covariates most useful in explaining LD and RMSF were similar and included co-occurrences of RMSF and LD, respectively, amount of forested and non-forested wetlands, pasture/grasslands, and urbanized/developed lands, population counts, and median income levels. Finally, I conclude with a ZIP code level spatio-temporal modeling exercise to determine areas and time periods in Tennessee where significant clusters of the studied diseases occurred. ZIP code level clusters were compared to the previously defined county-level clusters to discuss the importance of spatial scale. The findings suggest that focused disease/vector prevention efforts in non-endemic areas are warranted.

Very little work exists using administrative claims data in the study of zoonotic diseases. This body of work thus adds to an area void of much knowledge. Administrative medical claims data are relatively easy to access given the appropriate permissions, have relatively no cost once access is granted, and provides the researcher with a volume rich dataset from which to study. This data source should be properly considered in the wildlife and biological sciences fields of research.

DEDICATION

I dedicate this work to my family, especially my beautiful wife Lindsey, my son Porter, and my daughters Addison and Ivy Elizabeth…I can only hope that they may one day read it and desire to write their own. And to my mom, who gave me the academic guidance to get me here.

ACKNOWLEDGEMENTS

I would first and foremost like to acknowledge my graduate committee for allowing me this unique opportunity. I especially appreciate their willingness, support and belief in me to accomplish this tremendous task. I thank Dr. William Conner, Dr. Bo Song, Dr. David Gordon, Dr. Jim Rieck, and Dr. Anand Jayakaran for all of their guidance and support. I cannot thank them enough for taking this project on. I greatly appreciate the staff and faculty at Clemson who helped me along the way, especially Dr. Pat Layton, Dr. Mary Beck, Dr. William Bowerman, Dr. Rob Baldwin, Gayle Estes, and Shannon Coker.

I especially want to acknowledge BlueCross BlueShield of Tennessee and their support throughout this project. They made it possible for me to explore this concept and realize the importance of sound research and proactive movements. I especially thank Soyal Momin, our director of research and development, and Dr. Inga Himelright, our infectious disease specialist, for their support throughout.

I would also like to acknowledge and thank the Tennessee Department of Health and their astute epidemiologist, Dr. Bob Taylor. His promptness and willingness to answer my countless data questions made this project possible. Finally, I want to thank Dr. Martin Kulldorff from the Department of Population Medicine at Harvard Medical School and Harvard Pilgrim Health Care Institute…his insightful input and communication related to spatio-temporal modeling for this project were invaluable.

TABLE OF CONTENTS

Table of Contents (Continued)

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

List of Figures (Continued)

Figure Page

List of Figures (Continued)

Figure Page

List of Figures (Continued)

Figure Page

CHAPTER 1

IMPORTANCE OF STUDYING ZOONOTIC DISEASES

INTRODUCTION

Zoonotic diseases, also termed zoonoses, are infectious diseases that can be transmitted from or through animals to humans, and arthropods often act as vectors for transmission. Emerging infectious diseases have been increasing both in prevalence and geographic range at alarming rates the last 30 years, and the majority of these diseases are zoonotic in nature (Jones et al. 2008). Zoonotic diseases are of significant concern to public health and account for approximately 75% of recently emerging infectious diseases, and approximately 60% of all human pathogens originate from animals (CDC NCEZID 2010). Zoonoses can be in the form of viral (*e.g.,* West Nile virus), bacterial (*e.g.,* Lyme disease), fungal (*e.g.,* Histoplasmosis), protozoan (*e.g.,* babesiosis) or parasitic (*e.g.,* Filariasis) infections. These zoonotic diseases can pose a serious public health threat as some diseases such as rabies, though rare, can be fatal, while others (*e.g.,* ringworm) are minor health concerns. Even when incidence rates are relatively low, large-scale public health scares can emerge (*e.g.,* West Nile and H1N1 viruses).

Arthropods (*e.g.,* mosquitoes, ticks, mites, spiders) make up over 80 percent of all animal species but only a very small percentage of the over 1 million described species are potentially dangerous to humans (Goddard 2008).

However, arthropods can serve as vectors for the transmission of zoonotic diseases from the infected animal (reservoir) to a susceptible human host. Diseases transmitted by arthropods are thus termed arthropod-borne diseases (Eisen and Eisen 2007). The ability for vectors to successfully transmit disease from reservoir to host depends on many factors including vector physiology, morphology, reproductive capacity, and genetics. Additionally, the occurrence, extent, and suitability of arthropod habitats depend on multiple factors such as temperature, topography, moisture, rainfall, soil pH, weather, and geographical location. Human induced factors may also contribute to the rise in disease prevalence because of climate change, public health policy, lack of prevention and control, and increasing urbanization (Goddard 2008).

Many zoonotic diseases are considered notifiable by the Centers for Disease Control and Prevention (CDC). This means that when a case is diagnosed or suspected, the diagnosing clinician (*i.e.*, health care provider) should report this information to their local or state health department. In addition to diagnosing and treating individual patients, health care providers play an important role in protecting the public health through the identification and reporting of infectious diseases. Health care providers are typically the first health officials to encounter cases of infectious zoonotic diseases, and therefore play an important role in disease surveillance activities (GAO 2004). Health insurance plans, sometimes referred to as managed care organizations (MCO), contract with providers to deliver health care to their members. Therefore, there is both a direct (self-reporting) and indirect (provider-reporting) responsibility of health plans in the reporting of infectious diseases. Indirectly within their legally binding contract with providers, health plans could require providers to report 100% of diagnosed cases. However, though state regulations or contractual obligations may require the reporting of certain diseases, underreporting still exists as not all diagnosed or suspected cases are reported by health care providers (Marier 1977; Meek et al. 1996; Young 1998; Koo and Caldwell 1999), and can vary by physician specialty (Campos-Outcalt et al. 1991).

Medical claims data are recorded within the healthcare system every time a patient visits their doctor or hospital for a medical service, fills a prescription medicine, or seeks consultation from a clinician. Of particular interest is the amount of available data from MCOs, as well as the temporal and spatial granularity of captured data elements from each medical encounter. Medical claims data contain, among other things, the patient's ZIP code at the time of service, date of medical service, and medical diagnosis codes which describe the reason why the patient is seeking medical care. The geographic element of a patient's residence location combined with the date of diagnosis provides both a spatial and temporal "stamp" of what the patient was exposed to, and potentially when and where the exposure may have occurred.

Medical claims data may aid in the study and tracking zoonoses. If using medical claims data is a viable approach, it could be used to improve both the temporal and spatial scale of study through the use of long-term longitudinal data covering a large geographic expansion and at a more geographically refined scale. Additionally, claims data could supplement the current reporting of notifiable diseases to the CDC. This effort may help bridge the disease incidence gap created by health care providers' underreporting and thus allow for more effective tracking and monitoring of infectious zoonotic diseases across time and space.

BACKGROUND

Southeastern Arthropod-borne Diseases

Arthropod-borne diseases are infectious diseases in which arthropods are considered a transmitting vector or intermediate host. These diseases can be separated into categories based upon vector phylogeny. The most prevalent vectors in the southeastern US and the primary focus of this study are ticks and mosquitoes, and thus this study proposes to categorize diseases as tick-borne and mosquito-borne, respectively. Specifically, we examine five tick-borne (Lyme disease, babesiosis, ehrlichiosis, Rocky Mountain spotted fever, and tularemia) and 2 mosquito-borne (West Nile virus and La Crosse viral encephalitis) diseases known to occur in the southeastern US (Table 1-1). For the purposes of this study, the southeast was considered South Carolina, North Carolina, Georgia, Tennessee, Florida, Alabama, Mississippi, Louisiana, and Arkansas (Figure 1-1).

Tick-borne Diseases (Table 1-1)

Borreliosis - Lyme disease (ICD-9 Diagnosis Code: 088.81*)

Lyme disease is the most frequently reported vector-borne disease in the US (Varela et al. 2004) with 29,780 cases reported nationwide in 2009, with 32 occurring in Tennessee (CDC MMWR 2010). Lyme disease is caused by the bacterium *Borrelia burgdorferi,* which is transmitted to humans via the Blacklegged or deer tick (*Ixodes scapularis*), the same tick responsible for transmitting babesiosis and certain forms of ehrlichiosis. Symptoms include a characteristic "bulls-eye rash" within 2 weeks after exposure, fever, headache, and fatigue. If left untreated, infection can spread to joints, the heart, and the nervous system. Most cases of Lyme disease can be treated successfully with a few weeks of antibiotics. The majority of diagnosed cases occur in the New England area, upper Mid-West, southeastern US, and Pacific Coast states.

Babesiosis (ICD-9 Diagnosis Code: 088.82)

Babesiosis is an uncommon tick-borne malaria-like illness caused by the *Babesia microti* organism, which usually infects white-footed mice and other small mammals. This organism is then transferred to humans by *Ixodes scapularis* (CDC NCEZID 2010). Most cases of babesia infection are asymptomatic or include mild fevers and anemia while more severe cases carry symptoms similar to malaria and can be life-threatening. Reported cases are on

* See below for a detailed description of the ICD-9 medical coding system

the rise, perhaps because of expanded medical awareness (Hunfeld et al. 2008). In North America, the disease is most commonly found in the Northeast and upper Midwest, particularly in parts of New England, New York, New Jersey, Wisconsin, and Minnesota. According to the Tennessee Department of Health Communicable Disease Interactive Data Site, there have been no reported cases within Tennessee for the 1995-2009 time period. However, a recent 2010 report indicates what they believe to be the first zoonotic babesiosis case documented in Tennessee (Mosites et al. 2010).

Rickettsiosis – Rocky Mountain spotted fever (ICD-9 Diagnosis Code: 082.0)

Rocky Mountain spotted fever is the most severe tick-borne rickettsial illness in the US and is caused by the *Rickettsia rickettsii* bacterial organism. Infections occur most commonly in the southeastern and south central US and are typically transmitted from the bite of an infected American Dog tick (*Dermacentor variabilis*). Symptoms include the development of a rash within 2 to 4 days after the onset of fever, and can be non-descript or mimic other illnesses with headache, muscle pain, nausea, and lack of appetite. In 2009 there were 1,393 cases reported nationwide, with 184 occurring in Tennessee (CDC MMWR 2010).

Ehrlichiosis – human monocytic ehrlichiosis, Ehrlichia chaffeensis (ICD-9 Diagnosis Code: 082.41)

Ehrlichia chaffeensis can refer to both the disease name and the responsible bacterial pathogen. Ehrlichiosis caused by *E. chaffeensis* is also referred to as human monocytic ehrlichiosis (HME). As with tularemia, HME is associated with bites from the Lone Star tick (*Amblyomma americanu*) and is characterized by acute onset of fever and headache, malaise, anemia, nausea, vomiting, and/or a rash. This disease occurs most often in the Southeastern and Midwestern US, and the number of diagnosed cases have risen steadily from 1999 – 2006 (CDC NCEZID 2010). In 2009 there were 122 cases reported nationwide, with 16 occurring in Tennessee (CDC MMWR 2010).

Tularemia - (ICD-9 Diagnosis Code: 021)

Tularemia is a relatively uncommon but potentially fatal infectious disease most common in the south central US, Pacific Northwest, and Massachusetts. It is caused by the bacterium *Francisella tularensis* which is transmitted to humans through the bite of 2 different ticks, the American dog tick and the Lone Star tick. Tularemia can also be transmitted through the handling of infected animal carcasses, consuming contaminated food or water, or breathing in the bacteria. Because of the latter aerosol transmission capability, this disease is considered a possible bioterrorism indicator and is classified as a category 1B disease, which requires immediate telephonic notification followed by a written report within 1 week of diagnosis (TDH CEDS 2010). Symptoms occur within 5 days of exposure and include a sudden fever onset, chills, headache, diarrhea, muscle and joint pain, dry cough, difficulty breathing, and progressive weakness. In 2009 there were 123 cases of tularemia reported nationwide, with only 2 occurring in Tennessee (CDC MMWR 2010).

Tick Species

Blacklegged tick, deer tick (*Ixodes scapularis*)

This tick is widely distributed in the northeastern and upper midwestern US (Figure 1-2; Figure 1-3) and can transmit *Borrelia burgdorferi* (responsible for Lyme disease) and *Babesia microti* (responsible for babesiosis). Larvae and nymphs feed on small mammals and birds, while adults feed on larger mammals and will bite humans on occasion. It is important to note that the pathogen that causes Lyme disease is maintained by wild rodent and other small mammal reservoirs, and is not transmitted everywhere that the blacklegged tick lives. In some regions, particularly in the southern US, the tick has very different feeding habits that make it an unlikely vector in the spread of human disease (CDC NCEZID 2010).

Lone Star tick (*Amblyomma americanum*)

This tick is primarily found in the southeastern and eastern US (Figure 1-3) and is responsible for the transmission of organisms causing forms of borreliosis (Southern Tick-Associated Rash Illness), ehrlichiosis (human monocytic ehrlichiosis), and tularemia. White-tailed deer are a major host of Lone Star ticks and appear to represent a natural reservoir for *Ehrlichia chaffeensis.* Larvae and nymphs feed on birds and deer (CDC NCEZID 2010).

American Dog tick (*Dermacentor variabilis*)

This tick is the most commonly identified species responsible for transmitting the *Rickettsia rickettsii* bacterial organism and causes Rocky Mountain spotted fever in humans. This tick can also transmit tularemia. It is widely distributed east of the Rocky Mountains and also occurs in limited areas on the Pacific Coast (Figure 1-3). Larvae and nymphs feed on small rodents. Dogs and medium-sized mammals are the preferred hosts of adult ticks, although it feeds readily on other large mammals, including humans (CDC NCEZID 2010).

Mosquito-borne Diseases

Arthropod-borne viruses are primarily transmitted during the summer and fall in the US, with disease incidence peaking in late summer. Presently, there are two more commonly described mosquito-borne viral diseases (La Crosse and West Nile) occurring in the southeastern US.

La Crosse viral encephalitis (ICD-9 Diagnosis Code: 062.5)

La Crosse viral encephalitis (LACV) is a relatively uncommon viral illness transmitted to humans by the bite of an infected *Aedes Triseriatus* mosquito. Most cases occur in the upper Midwestern, mid-Atlantic, and southeastern states. Though most often asymptomatic, if symptoms do occur they include fever, headache, nausea, vomiting, and general malaise. If the infection is more severe (typically in children under 16), encephalitis can form and can include seizures, coma, and paralysis. In rare cases, long-term disability or death can result. In 2009, there were 8 confirmed cases in Tennessee (USGS ArboNet 2009).

West Nile virus (ICD-9 Diagnosis Code: 066.4)

The West Nile virus (WNV) was first detected in the US in 1999 and became notifiable in 2002. WNV is spread to humans through the bite of an infected mosquito, typically thought to be the *Culex pipiens* mosquito, which become infected after feeding on infected birds. Though the virus quickly spread across the US from 1999 through 2001, neuroinvasive disease incidence remained low until 2002 when large outbreaks in the Midwest and Great Plains occurred. Approximately 80 percent of people infected with WNV are asymptomatic. Less than 1% of people infected will have severe life-threatening symptoms, such as high fever, neck stiffness, stupor, disorientation, coma, tremors, convulsions, muscle weakness, vision loss, numbness, and paralysis. There were 329 reported cases of non-neuroinvasive West Nile virus in 2009, 4 of which occurred in Tennessee. Additionally, there were 361 reported cases of neuroinvasive West Nile virus in 2009, 4 of which occurred in Tennessee (CDC MMWR 2010).

Mosquito Species

Eastern tree hole mosquito (*Aedes triseriatus*)

This mosquito species is found in wooded regions of eastern and central North America, particularly in areas with temporary pools of stagnant water, such as tree holes and abandoned tires. It favors pools which contain leaf debris and other organic material to provide food for its larvae. Adults remain in areas near larval habitats throughout their lifespan. *Aedes triseriatus* occurs from Florida, north to Ontario and west to Texas (CDC NCEZID 2010).

Northern house mosquito (*Culex pipiens*)

This mosquito is usually the most common pest mosquito in urban and suburban settings and serves as an indicator of polluted water in the immediate vicinity. It is recognized as the primary vector of St. Louis encephalitis and West Nile virus in the eastern US and is normally considered to be a bird feeder, though some urban strains may prefer mammalian hosts (CDC NCEZID 2010).

Notifiable Diseases

According to the TDH, notifiable diseases are:

> “declared to be communicable and/or dangerous to the public and are to be reported to the local health department by all hospitals, physicians, laboratories, and other persons knowing of or suspecting a case in accordance with the provision of the statutes and regulations governing the control of communicable diseases in Tennessee.”

Improving Macro-scale Analyses by Aggregating to the ZIP Code Level

The reporting and tracking of illness cases is essential to knowing who is infected and where the problems are occurring. A major limitation in the study of such diseases, however, is the ability to comprehensively track disease incidence over space and time at a meaningful geographic scale. Data aggregations and disease incidence is most often presented at the county level (Sugumaran et al. 2009). Unfortunately, county level assessments compared to ZIP code level analyses may mask smaller isolated high risk areas as well as obscure within county variability (Mostashari et al. 2003; Eisen et al. 2006). In 2007, the CDC called for a means to improve data collection methods to determine probable pathogen exposure sites based specifically on patient activity spatial patterns. This suggests geocoding the residential location (street address or ZIP code) of the infected patient and conducting a radial search around that point to examine the underlying landscape. However, data describing possible pathogen exposure sites (i.e., patient's actual residential location information) are limited (Glass et al. 1995; Eisen and Eisen 2007), and means to collect this information can be very costly (e.g., patient surveys). Therefore, studies within the wildlife and ecological sciences are often limited in predictive power due to the inability to generate large sample sizes, either because of costs, data availability or both (Bissonette 1999). Health plans and their associated administrative data may help improve this data deficiency.

Role of Health Plans and Providers in the Monitoring of Infectious Diseases

As previously mentioned, MCOs play a major role in the tracking of infectious diseases. Medical claims data are recorded within the healthcare system every time a patient visits their doctor or hospital for a medical service, fills a prescription medicine, or seeks consultation from a physician. Therefore, all diagnosed zoonotic infections where patients are seeking monetary reimbursement from their health plan would be documented in the plans' claims data warehouse. If the services rendered are from an actual person (*e.g.,* physician), a HCFA-1500 form is completed and submitted to the health plan covering the patient. If the services rendered are billed from a facility (*e.g.,* hospital), a UB-92 form is completed. These forms are very similar and capture, among other things, patient information (name, date of birth, address, ZIP code), date of service, services rendered, and diagnosis information (described below in detail in the ICD-9 coding system). The term “electronic” refers to data that is stored electronically in a data warehouse. The term “administrative” refers to data that is transferred from the claim form to the health plan’s data warehouse. Administrative data does not typically include elements like lab results (*e.g.,* blood pressure, white blood cell count).

International Classification of Diseases (ICD) Medical Coding System

The ICD coding system is used throughout the healthcare industry to describe diseases, injuries, symptoms, complaints, and conditions encountered when patients visit a health care provider. Under this coding system, similar health conditions can be categorized together, and each condition/diagnosis is assigned a unique code, up to six characters long in a hierarchical listing. The ICD codes are revised periodically, and the majority of the US currently uses the 9th edition (ICD-9). For example, a patient diagnosed as having West Nile virus could be given a 3-digit ICD-9 code of “066” indicating a diagnosis of an “arthropod-borne viral disease.” More specifically, the patient would be given a 4-digit ICD-9 code of “066.4,” which indicates “West Nile Fever.” And even further, the health care provider could be more specific with the coding if certain symptoms were present, or certain tests confirmed the presence of something. For example, the ICD-9 hierarchy of diagnosis code “066” is:

066 Other arthropod-borne viral diseases

- *066.4* West Nile Fever
  - *066.40* West Nile fever, unspecified
  - *066.41* West Nile fever with encephalitis
  - *066.42* West Nile fever with other neurological manifestation
  - *066.49* West Nile fever with other complications

Another important aspect of diagnosis coding on a claim form is that a medical encounter can have more than one diagnosis code. The initial most important diagnosis (as deemed by the health care provider) is the primary diagnosis, and other diagnoses would be considered secondary, tertiary, and so on. For example, a physician could see a patient about their illness and determine that West Nile fever is the primary diagnosis (066.4). The physician may also code, on the same claim form, another secondary diagnosis for a headache (784.0) and tertiary diagnosis for nausea (787.02). Data in the BlueCross BlueShield of Tennessee (BCBST) data warehouse capture up to 8 diagnosis codes.

Of particular interest is the amount of available data from health plans, as well as the temporal and spatial granularity of captured data elements from each medical encounter. Medical claims data contain, among other things, the patient's ZIP code at the time of service, date of medical service, and medical diagnosis codes which describe the reason why the patient is seeking medical care. The geographic element of a patient's residence location combined with the date of diagnosis provides both a spatial and temporal "stamp" of what the patient was exposed to, and potentially when and where the exposure may have occurred. Therefore, this administrative data source could supplement the current reporting and tracking structure and provide a better estimate of the true incidence rate.

OBJECTIVES

A major limitation in spatial epidemiology is the collection of relevant longitudinal data at the appropriate geographic scale. Research often relies on drawing conclusions from only limited sample sizes usually taken either at a static point in time, or some periodic time interval convenient for sampling, which is further constrained by sampling cost. Additionally, diseases must be reported to the CDC or health departments in order to be recorded in the database, and clinicians or infected patients may not always manually report these.

This study proposes to introduce the use of electronic administrative medical claims data as a potential new source that could be leveraged in ecological field studies in the analyses and monitoring of arthropod-borne zoonotic diseases. If using medical claims data to study zoonoses is a viable approach, it could be used to improve both the temporal and spatial scale of study through the use of long-term longitudinal data covering a large geographic expansion and at a more geographically refined ZIP code scale. Additionally, claims data could supplement the current reporting of notifiable diseases to the CDC. This effort may help bridge the disease incidence gap created by health care providers' underreporting and thus allow for more effective tracking and monitoring of infectious zoonotic diseases across time and space.

Specifically, the 4 main objectives of the study are:

1. To determine if certain notifiable diseases are underreported based on a comparison of MCO administrative claims data and the TN State Health Department (TDH)
2. To determine how MCO and TDH data compare/differ in the context of spatio-temporal cluster analyses at the county level
3. To determine what geographic, habitat, and socio-economic characteristics may be useful in explaining the occurrence of zoonotic diseases
4. To determine where and when (if any) significant spatial and temporal clusters of selected diseases occurred across the state of Tennessee for the 2000-09 time period using MCO data at the ZIP code level

To my knowledge, this project is one of only a very small number of projects that attempt to use administrative data from a MCO to study zoonotic diseases. If successful, this could provide quantifiable evidence of more accurate estimates of disease prevalence. Additionally, there is a multi-state initiative within the

BlueCross BlueShield Association (BCBSA) to combine multiple BlueCross plans' claims data into one centralized data warehouse. This combined data source will contain information on approximately 100 million people across the entire US and thus serve as a potentially powerful data source for mapping and monitoring of zoonotic diseases.

DISSERTATION OVERVIEW

The dissertation is arranged in 5 chapters, where the first serves as a basic introduction to the work, and the last 4 chapters are written as independent papers. Because each project is undertaken individually with the explicit purpose of publication, some information/verbiage contained within the chapters may overlap.

Chapter 2 addresses the feasibility of using administrative medical claims data in the analysis and tracking of infectious zoonotic diseases. The objective is to determine if notifiable diseases are underreported. This is done by comparing the TDH data with administrative claims data extracted from the BlueCross BlueShield of Tennessee data warehouse, later defined as the MCO. The general hypothesis is there is no difference between the state reported incidence rates for a selected disease and the MCO claims derived incidence rate.

Chapter 3 builds on Chapter 2 and compares zoonotic case information derived from the TDH state registry system with MCO administrative medical claims information to statistically validate when (temporal) and where (spatial) these data sources differ. The general research hypothesis is that no differences in clusters exist between the two data sources.

Once MCO data is determined to be useful, Chapter 4 addresses the need to study potential site characteristics at a finer scale (*i.e.,* ZIP code) as opposed to traditional county level analyses using administrative data. Specifically, the objective is to determine what, if any, site level characteristics may be influential in explaining disease occurrence. The general research hypothesis is geographical/habitat characteristics do not influence the presence of zoonotic diseases.

Chapter 5 takes information learned from Chapters 2 – 4 and attempts to address the feasibility of using MCO data in the tracking of zoonotic diseases at the ZIP code level. Specifically, the objective is to determine where and when significant spatial and temporal clusters of selected diseases occurred, and compare these findings to county-level outcomes. The general research hypothesis is there are no significant spatial or temporal clusters of disease incidence across Tennessee for the study period.

REFERENCES

Bissonette, J. 1999. Small sample size problems in wildlife ecology: a contingent analytical approach. *Wildlife Biology*. 5:65-71.

Campos-Outcalt, D., R. England, and B. Porter. 1991. Reporting of communicable diseases by university physicians. *Public Health Reports*. 106(5):579-583.

Centers for Disease Control and Prevention (CDC NCEZID). 2010. National Center for Emerging and Zoonotic Infectious Diseases (http://www.cdc.gov/ncezid/) accessed September 21, 2010

Centers for Disease Control and Prevention (CDC MMWR). 2010. Morbidity Mortality Weekly Report (MMWR): Provisional cases of selected notifiable diseases, United States, week ending January 2, 2010. http://www.cdc.gov/mmwr/index.html accessed October 4, 2010

Eisen, R., R. Lane, C. Fritz, and L. Eisen. 2006. Spatial patterns of Lyme disease risk in California based on disease incidence data and modeling of vector-tick exposure. *American Journal of Tropical Medicine and Hygiene*. 75(4):669–676.

Eisen, L., and R. Eisen. 2007. Need for improved methods to collect and present spatial epidemiologic data for vectorborne diseases. *Emerging Infectious Diseases*. 13(12):1816-1820.

Glass, G., B. Schwartz, J. Morgan III, D. Johnson, P. Noy, and E. Israel. 1995. Environmental risk factors for Lyme disease identified with geographic information system. *American Journal of Public Health*. 85(7):944-948.

Goddard, J. 2008. Infectious Diseases and Arthropods. Humana Press; 2nd edition. 251p.

Government Accountability Office (GAO). 2004. Emerging infectious diseases review of state and federal disease surveillance efforts: report to the Chairman, Permanent Subcommittee on Investigations, Committee on Governmental Affairs, US Senate. GAO report# GAO-04-877.

Hunfeld, K., A. Hildebrandt, and J. Gray. 2008. Babesiosis: Recent insights into an ancient disease. *International Journal of Parasitology*. 38(11):1219–1237.

Jones, K., N. Patel, M. Levy, A. Storeygard, D. Balk, J. Gittleman, and P. Daszak. 2008. Global trends in emerging infectious diseases. *Nature*. 451(7181): 990-993.

Koo, D., and B. Caldwell. 1999. The role of providers and health plans in infectious disease surveillance. *Effective Clinical Practice*. 2(5):247-252.

Marier, R. 1977. The reporting of communicable diseases. *American Journal of Epidemiology*. 105(6):587-590.

Meek, J., C. Roberts, E. Smith Jr, M. Cartter. 1996. Underreporting of Lyme disease by Connecticut physicians. *Journal of Public Health Management and Practice*. 2(4):61-65.

Mosites, E., P. Wheeler, N. Pieniazek, M. Xayavong, L. Carpenter, T. Jones, B. Herwaldt, and J. Dunn. 2010. Human babesiosis caused by a novel babesia parasite--Tennessee, 2009. Unpublished abstract presented at the International Conference on Emerging Infectious Disease (ICEID) July 2010 Atlanta, GA.

Mostashari, F., M. Kulldorff, J. Hartman, J. Miller, and V. Kulasekera. 2003. Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases*. 9(6):641-646.

Sugumaran, R., S. Larson, and J. Degroote. 2009. Spatio-temporal cluster analysis of county-based human West Nile virus incidence in the continental United States. *International Journal of Health Geographics*. 8:43.

Tennessee Department of Health (TDH CEDS). 2010. Tennessee Department of Health, Communicable and Environmental Disease Services: Reportable Diseases and Events. http://health.state.tn.us/ceds/notifiable.htm accessed October 5, 2010.

US Department of the Interior US Geological Survey (USGS ArboNet) ArboNet surveillance system. Available at: http://diseasemaps.usgs.gov/2009/lac_tn_human.html

Varela, A., M. Luttrell, E. Howerth, V. Moore, W. Davidson, D. Stallknecht, and S. Little. 2004. First culture isolation of *Borrelia lonestari*, putative agent of southern tick-associated rash illness. Journal of Clinical Microbiology. 42(3):1163–1169.

Young, J. 1998. Underreporting of Lyme disease. *The New England Journal of Medicine*. 338(22):1629.

Table 1-1: Arthropod-borne zoonotic diseases known to occur in the southeastern US and selected for study

Disease	Common Name	Pathogen Type	Pathogen	Vector Type	Primary Vector	
Borreliosis	Lyme disease	Bacterial	*Borrelia burgdorferi*	Tick	Blacklegged, or deer tick	*(Ixodes scapularis)*
Babesiosis	Babesiosis	Protozoan	*Babesia microti*	Tick	Blacklegged, or deer tick	*(Ixodes scapularis)*
Rickettsiosis	Rocky Mountain spotted fever	Bacterial	*Rickettsia rickettsii*	Tick	American dog tick	*(Dermacentor variabilis)*
Ehrlichiosis	Human Monocytic Ehrlichiosis	Bacterial	*Ehrlichia chaffeensis*	Tick	Lone star tick	*(Amblyomma americanu)*
Tularemia	Tularemia	Bacterial	*Francisella tularensis*	Tick	American dog tick, Lone star tick	*(Dermacentor variabilis, Amblyomma americanum)*
La Crosse Encephalitis	La Crosse viral encephalitis	Viral	*La Crosse encephalitis virus*	Mosquito	Eastern tree hole mosquito	*(Aedes triseriatus)*
West Nile Fever (Virus)	West Nile virus	Viral	*West Nile Virus*	Mosquito	Northern house mosquito	*(Culex pipiens)*

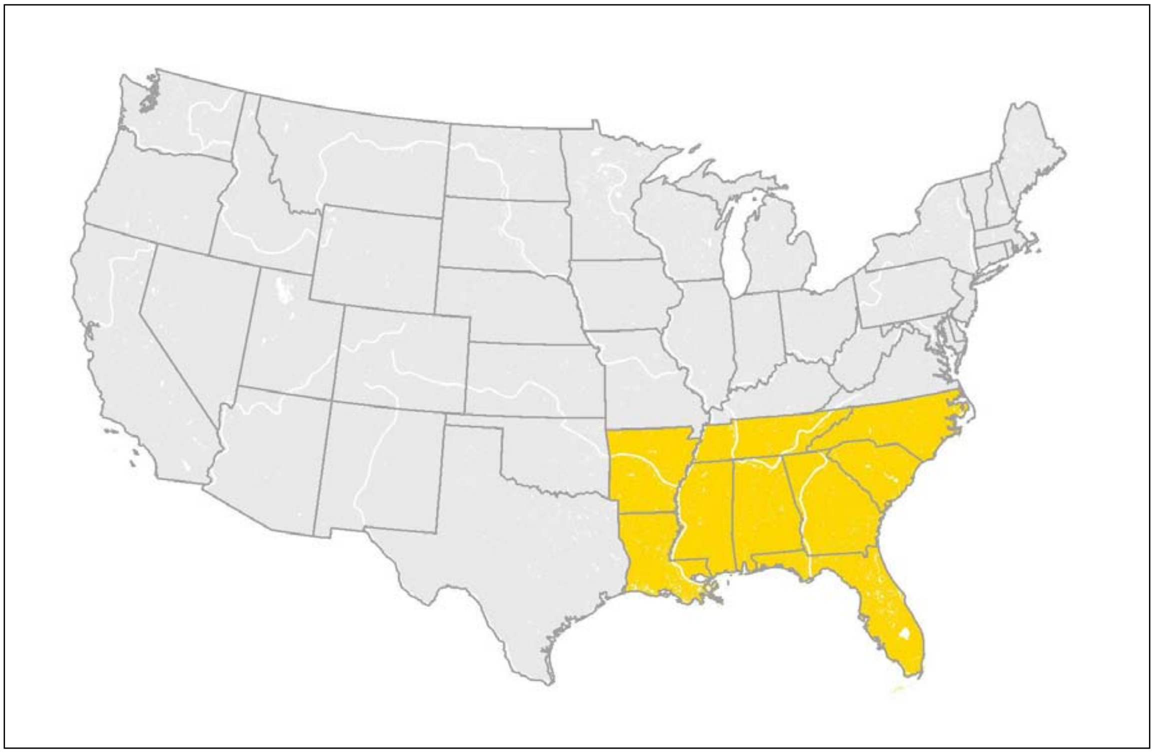


Figure 1-1: Map indicating southeastern states according to this study

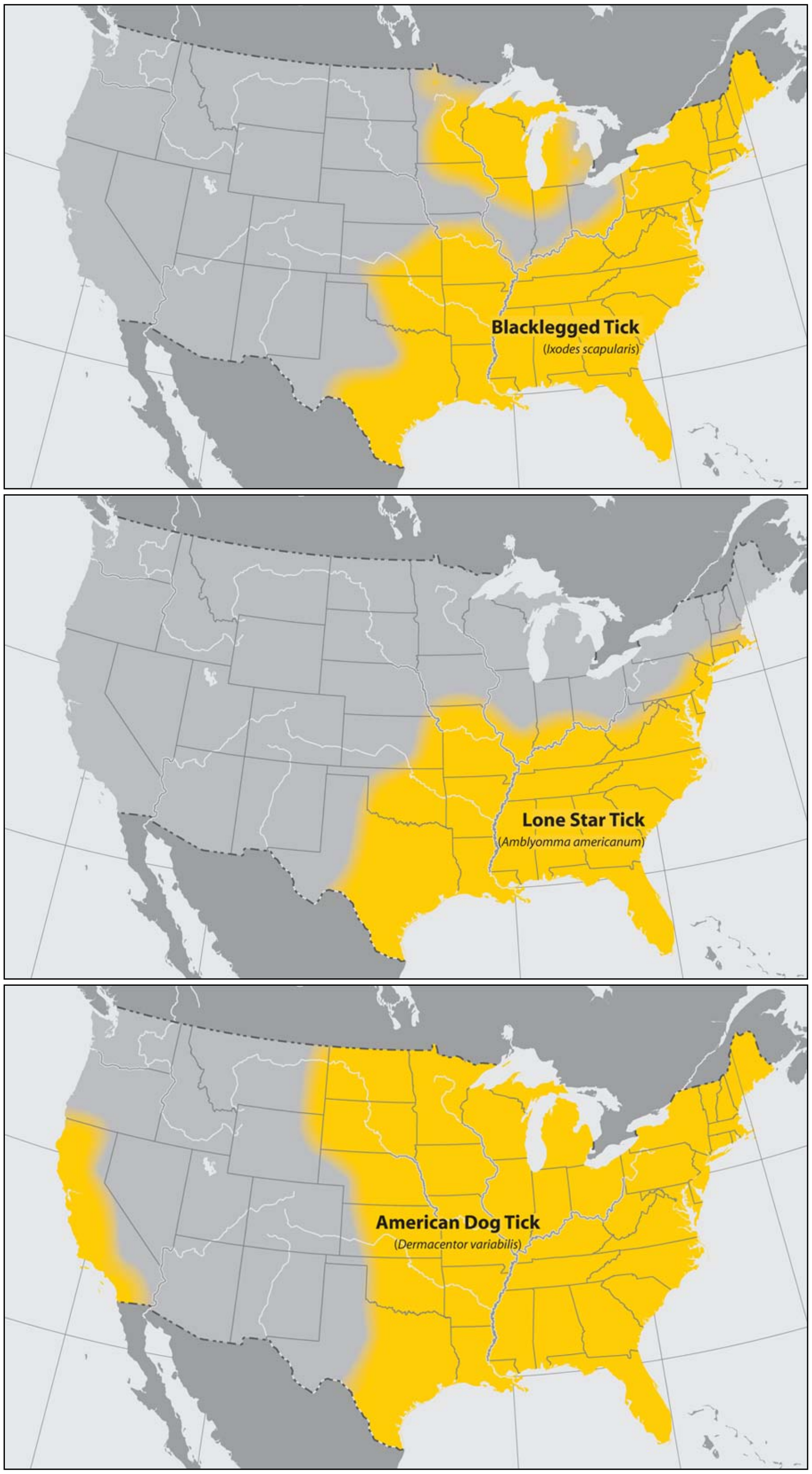


Figure 1-2: Approximate distributions of the Blacklegged Tick (*Ixodes scapularis*), the Lone Star tick (*Amblyomma americanum*) and the American Dog tick (*Dermacentor variabilis*) (images source: CDC)

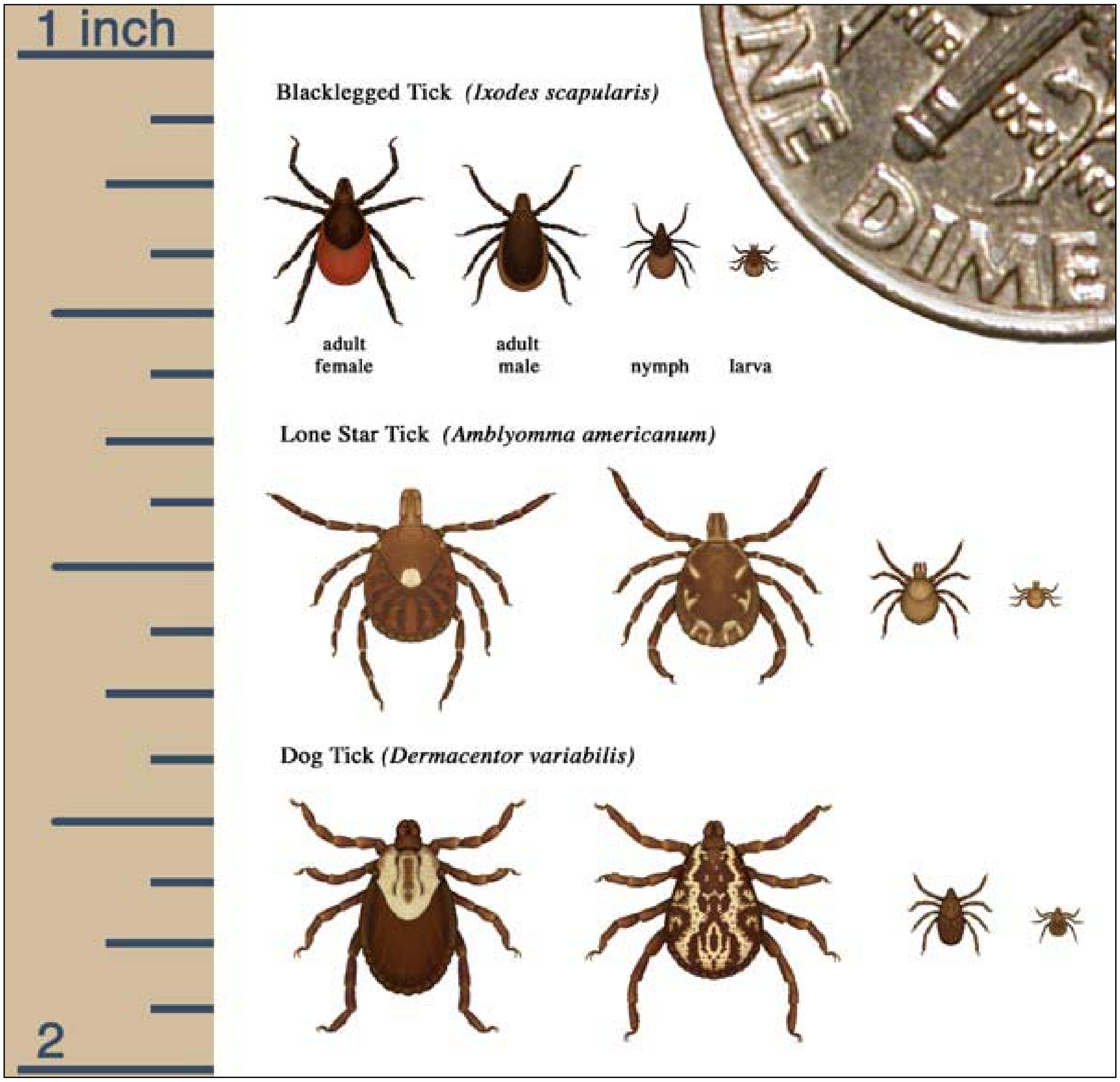

1 inch
2
Blacklegged Tick (Ixodes scapularis)
adult female
adult male
nymph
larva
Lone Star Tick (Amblyomma americanum)
Dog Tick (Dermacentor variabilis)
ONE DIME

Figure 1-3: Life stages and relative sizes of 3 tick species known to be primary vectors for zoonotic diseases (image source: CDC)

CHAPTER 2

USING ADMINISTRATIVE MEDICAL CLAIMS DATA TO ESTIMATE UNDERREPORTING OF INFECTIOUS ZOONOTIC DISEASES

ABSTRACT

Notifiable diseases require regular, frequent, and timely reporting of diagnosed cases to aid in prevention and control. However, manual reporting can be burdensome, incomplete, and delayed. Administrative claims data captured from clinical encounters details the patient's reason for seeking care, service date, and place of residence. To determine if administrative data is useful in the tracking and reporting of diagnosed zoonotic diseases, 5 tick-borne (Lyme disease [LD], babesiosis, ehrlichiosis, Rocky Mountain spotted fever [RMSF], tularemia) and 2 mosquito-borne (West Nile virus, La Crosse viral encephalitis) diseases known to occur in the southeastern US were examined. Disease incidence rates from cases reported to the Tennessee Department of Health (TDH) and medically diagnosed cases captured in a southeastern Managed Care Organization (MCO) claims data warehouse were compared using a complete randomized block design within a general linear mixed model. LD incidence was 7.7 times higher (*P* < 0.001) than what was actually reported to the state, possibly indicating significant underreporting (~196 unreported cases per year in TN). MCO data suggests that about 33 cases of RMSF go unreported per year in TN (*P* < 0.001). Three (3) cases of babesiosis were

discovered using claims data, a significant finding as this disease has never been reported in Tennessee. Significant spatial and temporal variations in disease rates were present ($P < 0.001$). This study successfully demonstrates the usefulness of administrative claims data in the tracking and reporting of zoonotic diseases.

INTRODUCTION

Notifiable diseases are infectious diseases for which regular, frequent, and timely reporting of individual diagnosed cases aids in prevention and control (*e.g.,* Lyme disease, giardiasis, salmonella) (GAO 2004; CDC NNDSS 2010). In 1961, the Centers for Disease Control and Prevention (CDC) was given oversight responsibility for compiling and publishing weekly morbidity statistics for listed notifiable diseases through the Morbidity and Mortality Weekly Report (MMWR). Public health officials from state health departments collaborate annually with the CDC to determine which diseases should be listed. This disease surveillance effort, the National Notifiable Disease Surveillance System (NNDSS), is one of the oldest surveillance systems in the United States (US). Reporting of disease cases by health care providers and laboratories is currently mandated only at the state level and therefore can vary from state to state (Koo and Wetterhall 1996; CDC NNDSS 2010).

In 2009, there were over 7,000 cases of notifiable communicable diseases reported to the Tennessee Department of Health Communicable and Environmental Disease Services (TDH CEDS 2009). This is over twice the number of notifiable diseases reported in 2000 (TDH WebAim 2010). Though state regulations or contractual obligations may require the reporting of certain diseases, traditional passive surveillance initiated by the diagnosing clinician can be burdensome, incomplete, and delayed (Doyle et al. 2002). Thus underreporting of diseases exists as not all diagnosed or suspected cases are reported by health care providers (Marier 1977; Meek et al. 1996; Young 1998; Koo and Caldwell 1999; Figueiras et al. 2004), and can vary by physician specialty (Campos-Outcalt et al. 1991). Many health care providers may not understand the importance of public health surveillance, and generally how, when, why, and what to report (Koo and Caldwell 1999; Figueiras et al. 2004).

Health insurance plans could play a major role in the reporting of infectious diseases (Rutherford 1998; Koo and Caldwell 1999). Medical encounter data are recorded within the healthcare system every time a patient visits their doctor or hospital for a medical service, fills a prescription medicine, or seeks consultation from a clinician. When seeking reimbursement from a health plan for the medical services performed, medical encounter data is captured via an insurance claims form completed by the physician performing the services, and then is submitted to the health plan. Therefore, all claims with medically diagnosed cases being filed to a health plan are captured electronically in the plan's data warehouse. Considering more than 253 million Americans have health insurance and will most likely utilize that service when needed (DeNavas-Walt et al. 2010), the administrative data captured from a medical encounter could serve as a useful source in the tracking of diagnosed infectious diseases. This study examines the feasibility of using administrative claims data in the analysis and tracking of infectious zoonotic diseases by comparing medically diagnosed cases of zoonotic infections extracted from administrative claims data to zoonotic cases reported to the Tennessee State Health Department. The level of underreporting is estimated for five tick-borne (Lyme disease [LD], babesiosis, ehrlichiosis, Rocky Mountain spotted fever [RMSF], tularemia) and 2 mosquito-borne (West Nile virus, La Crosse viral encephalitis) diseases known to occur in the southeastern US. The general research hypothesis is that incidence rates from the state health department are underrepresented when compared to actual diagnosed claims information.

METHODS

Study Area

Disease cases known to occur for residents living within the state boundaries of Tennessee were studied. For the purposes of this study, Tennessee is considered a southeastern state and is approximately bounded within the southernmost west coordinate (-90.309200, 34.995800) to the northern most east coordinate (-81.646900, 36.611900). Approximately 6.3 million people live within the 95 counties, and they have a median age of 37 and median income of $43,600. Ninety-three (93) percent of the state population lived in the same residence or same county as they did one year prior (US Census Bureau 2009). Of the state's approximately 10.9 million hectares, 5% is considered federal lands, 3% water area, 9% non-federal rural, and the remaining 83% is non-federal non-rural lands (USDA 2000). Estimated land cover percentages for the state are as follows: open water (2.7%), forested wetland (3.0%), non-forested wetland (0.4%), grassland/pasture (37.2%), cropland (5.8%), upland deciduous forest (40.6%), upland mixed forest (4.4%), upland coniferous forest (3.6%), urban/developed (1.9%), and non-vegetated (0.2%) (Tennessee Wildlife Resources Agency 1997).

Disease Case Data

Disease occurrence data within the proposed study area of Tennessee were collected from 2 separate data sources and tested for statistical differences. The first data source was electronic administrative medical claims data obtained from a large southeastern managed care organization (MCO) located in Tennessee. This MCO insures approximately 60% of the entire state's population. All medical claims having a primary or secondary arthropod-borne disease diagnosis code of interest (see below) were extracted for the January 1, 2000–December 31, 2009 time period. Allowed claims (*i.e.*, non-voided and approved for payment) having one of the following diagnosis codes were retained for study:

Tick-Borne Diseases:

- Babesiosis (ICD-9 code: 088.82)
- Borreliosis - Lyme disease (LD) (ICD-9 code: 088.81)
- Ehrlichiosis - human monocytic ehrlichiosis (HME) (ICD-9 code: 082.41)
- Rickettsiosis - Rocky Mountain spotted fever (RMSF) (ICD-9 Diagnosis Code: 082.0)
- Tularemia (ICD-9 code: 021)

Mosquito-Borne Diseases:

- La Crosse viral encephalitis (LACV) (ICD-9 code: 062.5)
- West Nile virus (WNV) (ICD-9 code: 066.4)

These diseases were selected because they are known to occur within the southeastern US (CDC NCEZID 2010) and they provide a mix of relatively abundant cases (*e.g.*, borreliosis, rickettsiosis) to very rare cases (*e.g.,* babesiosis, tularemia) to study. Using the MCO database, any patient receiving medical services for one of the selected diseases prior to the start of the study period (January 1, 2000) or after the study period (December 31, 2009) was removed from the analysis. To best ensure the diagnosis for the patient was in fact their first diagnosis, only the first recorded diagnosis date for each patient was retained. Any subsequent claims for the patient were removed and not considered in this study. This analysis utilized the exact diagnosis code for disease identification and served to represent the minimum estimation (*i.e.*, lower limit) of incidence according to MCO data.

The second data source was an extract provided by the TDH, Center for Environmental and Communicable Diseases detailing all notifiable “confirmed” or “probable” disease cases reported to the state of Tennessee during the study period (TDH WebAim 2010). This resulted in a comparison of medically diagnosed cases to CDC confirmed or probable disease cases. For example, a medical diagnosis for LD should be based on an individual’s history of possible exposure to ticks that carry LD, the presence of typical signs and symptoms, and the results of blood tests. CDC case definitions for LD are as follows:

- Confirmed Case – A case of erythema migrans (“bulls-eye” rash) with known exposure, or a case of erythema migrans with laboratory evidence of infection and without a known exposure, or a case with at least one late manifestation that has laboratory evidence of infection.
- Probable Case – Any other case of physician-diagnosed LD that has laboratory evidence of infection.

The comparison between data sources can be made because of the extraction of medical claims with the specific ICD-9 diagnosis code, not a generic root-level code. Thus it is assumed that if a clinician codes at this detail, they have evidence to support the diagnosis beyond suspicion. Furthermore, this very difference and ambiguity is the main intent of the study, which is to compare the difference between state reported incidence rates and medically diagnosed case rates to determine if claims data could support the current surveillance system. Because TDH serves as the compiler of all data sources to the state level, these data represent a theoretically complete set of reported cases for the state. The state data detail what the event was (*e.g.*, West Nile virus), when it occurred (*i.e.*, event date), and the county of residence for the infected person (*i.e.*, county name).

Count Adjustment for MCO Data

Previous unpublished internal work with MCO data indicates that ICD-9 coding errors can exist in the data. For example, a patient had a claim containing exactly 1 line item for Enteric tularemia (ICD-9 code: 0211). This line item was part of a larger medical claim containing many other occurrences of *70211*, which is the ICD-9 code for "Inflamed seborrheic keratosis." This was an obvious miscode in the system and cause for data validity concerns. Further, tularemia cases were evenly distributed throughout the year. This is unexpected as this disease is quite seasonal in nature, occurring in peaks during the summer months (Boyce 1975). Exact quantification of this error type is difficult, if not impossible, as it requires manually reviewing tens-of-thousands of line items of data to check for ICD-9 coding errors, and judgment could in part be subjective. In addition, no known work exists on this subject, so references are unavailable on the number of line items needed to ensure validity. Therefore, an adjustment factor was employed based on the number of line items a patient had for a given disease in the MCO system. Rather than developing an empirical filter upon which to remove these types of claims, it was decided to create a threshold value and apply this to disease cases to remove any cases where the medical claim did not have at least 3 separate line items with the same diagnosis code. This could obviously remove valid cases that contain less than 3 line items, but most likely would remove all cases in error as someone would have to make 3 errors on the same claim record. Thus, the MCO rates may in fact be underrepresented because of this adjustment. All results and analyses use this adjustment applied to the MCO data.

Data Aggregations

For each of the seven diseases, the number of cases was aggregated per county per year separately for TDH and MCO data. Because each data source had different populations from which case information was drawn, raw counts could not be fairly compared. Therefore, the denominator difference was adjusted by including population counts in all models. For TDH data, yearly county level population estimations were provided by the Tennessee Department of Health-Division of Health Statistics. Historical population counts (*i.e.*, plan membership) by county were not known for MCO data and were therefore estimated using an overall monthly adjustment factor. This was done by first calculating total MCO membership enrollment for each month of the study period (this served as the denominator). Next, the total number of medical claims filed for each month of the study period was calculated (this served as the numerator). The adjustment factor was a monthly ratio of medical claims to membership. The next step was to use this monthly adjustment factor to derive a membership estimate per county per month for the study period. Medical claims capture the county of residence of the member, therefore an estimated monthly county level population was calculated by dividing the total number of medical claims filed by MCO members within a given month and county by the overall monthly adjustment factor. For example, if in January 2000 there were 10,000 active members enrolled in the MCO health plan, and 2,500 filed a claim in that month, then the January 2000 adjustment factor is 25% (*i.e.*, 2500 / 10000). If County A had 400 members file claims during January 2000, then the monthly adjusted population count for County A would be 1,600 (*i.e.*, 400 / 0.25).

$MAP_{At} = m_{At} / a_t$ Equation 1: Monthly Adjusted Population Counts

where

MAP_{At} = monthly adjusted population for County A at time t (time t is denoted as the month and year in question)

m_{At} = number of members enrolled in the MCO that reside in County A at time t

a_t = statewide adjustment factor at time t given by $a_{t\,=}\, M_t / C_t$ where

M_t = total number of members enrolled in the MCO at time t

C_t = total number of members that filed a claim during time t

The same monthly adjustment factor was applied to all counties for each respective month, and therefore this approach assumes spatial homogeneity of claims submissions. This process was validated by comparing this method to known current membership levels and showed the median inflated error rate to be approximately 1,768 members, or about a 17% over-adjustment. Therefore, incidence rate estimates from MCO could be underrepresented due to this method because the denominators are inflated.

Statistical Analyses

To estimate if and to what extent underreporting of notifiable diseases exist, a randomized control block design was employed within a generalized linear mixed model (GLMM) approach to compare TDH and MCO case counts. These models are particularly useful in estimating trends in disease rates and where the response variable is not necessarily normally distributed (Salah et al. 2007; SAS® 2008). Input values into the models included a yearly (n = 10) county (n = 95) case total, which produced 950 observations for each data source. Separate models were built for each disease, and the response variable of interest was disease counts assumed to be Poisson distributed with a log-transformed population count as an exposure offset. Disease counts were expected to vary by county (*i.e.,* spatial heterogeneity) due to varying population denominators, socio-economic factors, and varying geographic and habitat characteristics (Kalluri et al. 2007; Wimberly et al. 2008; Winters et al. 2008; Yang et al. 2009). Therefore, county was used as a blocking factor to remove the expected county-to-county variability when comparing TDH to MCO values. Space (county) was considered a random effect, while time (year), data source (MCO vs. TDH), and a time*data source interaction were considered fixed effects. Fixed effects were examined for statistical significance using the F-test with an alpha level of 0.05. Variability in case counts across counties was tested using a covariance test within the GLMM procedure. SAS® Enterprise Guide version 4.2 and SAS/STAT version 9.3 were used for all analyses (SAS® 2008).

Seasonality profiles were created for each disease and visually compared between data sources. These profiles detail the percentage of all recorded cases by month (January-December) for the entire study period to illustrate in which months the disease is most prevalent. This was done for exploratory purposes to see the relationship between recorded event dates from the state and the date of service that patients seek medical care.

RESULTS

Overview

Approximately 58,385,858 medical claims were filed during the 2000-2009 study period. Of these, 6,638 patients had a medical claim with a primary or secondary diagnosis for one of the 7 described arthropod-borne diseases. After removing invalid claims (patients without at least 3 separate ICD-9 entries for the disease, patients with claim dates starting or ending outside of the time period, duplicate patient entries, and patients having non-unique disease coding issues [*e.g.,* code for RMSF and LD on the same claim]), 1,654 unique cases were distributed across the 7 diseases of interest and remained for study. The average age of patients having one of the described diseases was 37.3 (SD: 19.84; SE: 0.49), and 53.2% were female. Proportion of female patients was higher in the mid- to late-age groups (aged 15-60 years). The age/gender distribution for patients with a disease is comparatively different from the population as a whole. In the overall MCO member population, irrespective of disease, the distribution of males is higher compared to females across all age groups over 15 years old (Figure 2-1).

The majority of disease cases were LD (n=903; 55%), followed by RMSF (n=661; 40%). The remaining 5 diseases made up the residual 5% of disease cases (Figure 2-2). Three (3) cases of babesiosis were found within the MCO claims data, specifically within Davidson, Lincoln, and Washington Counties. Average ages varied within each disease type. On average, patients diagnosed with LACV were much younger than the other diseases (Figure 2-3). Gender distributions varied by disease. Lyme disease appears to be diagnosed more in females, while LACV was diagnosed more often in males (Figure 2-4).

Comparison of Medical Claims Case Data to State Reported Data

To determine if and to what extent possible underreporting occurs, MCO case data was compared to the TDH data set for the entire study period (*i.e.*, "data source" comparison). Raw per100k disease rates using the MCO data source appear higher for LD, babesiosis, RMSF, and tularemia. HME and LACV rates are higher from the TDH data source, and WNV rates are equal. However, results from the general linear mixed model suggest that only LD and RMSF values are statistically different, as all other models did not converge (Table 2-1).

The average yearly number of medically diagnosed cases of LD from MCO data were 7.7 times higher those reported to the state (F = 835.44; *P* < 0.0001). LD rates significantly varied over the 10 year study period (F = 2.08; *P* = 0.0283) and there was a significant temporal interaction with year*data source (F = 2.84; *P* = 0.0026). Based on the residual pseudo-likelihood, a tests of covariance suggests there is significant spatial variation of LD cases across the state ($\chi^2 = 84.8$; *P* < 0.0001). The average yearly number of medically diagnosed cases of RMSF from MCO data were 1.24 times higher than those reported to the state (F = 14.45; *P* = 0.0001). RMSF disease rates significantly varied over the 10 year study period (F = 14.82; *P* < 0.0001), and there was a significant temporal interaction with year*data source (F = 10.14; *P* < 0.0001). There is also significant spatial variation of RMSF case across the state ($\chi^2 = 1135.01$; *P* < 0.0001).

Temporal trending indicates the aforementioned per 100k rate differences varied from year to year, and MCO rates were not consistently higher throughout the entire study period. LD rates from MCO were consistently higher than TDH rates (Figure 2-5). TDH indicated no evidence of babesiosis but MCO data indicates 3 separate cases in years 2004, 2005, and 2009 (Figure 2-6). RMSF rates from TDH were lower than MCO rates from 2000-2005, but increased beyond MCO for all years afterwards except 2008 (Figure 2-7). TDH rates for HME were consistently higher than MCO (Figure 2-8). Tularemia rates were much higher in MCO data for years 2000-2004, and then rates became approximately equal from 2005-2009 (Figure 2-9). LACV rates for TDH were higher from 2000-2004, then fluctuate afterwards (Figure 2-10). TDH and MCO rates for WNV follow similar patterns, with TDH rates being slightly higher (Figure 2-11).

Monthly aggregated data show the seasonality of these diseases (Figure 2-12; Figure 2-13; Figure 2-14; Figure 2-15; Figure 2-16; Figure 2-17) (NOTE: babesiosis not shown due to non-representation in TDH data). Overall, the tick-borne diseases were more prevalent during May – August, whereas the mosquito-borne illnesses were most prevalent during August – October. With the exception of tularemia, the seasonal data was relatively consistent between the two data sources. A time lag is evident throughout the seasonal graphs, where MCO data is lagging behind the TDH data. The seasonality relationship between LD and RMSF across the entire study period for MCO data shows RMSF has sharper peaks suggesting cases are relatively more concentrated in the summer months (Figure 2-18).

DISCUSSION

Administrative medical claims data is an important resource for research and surveillance of chronic diseases (Yiannakoulias et al. 2009). Results suggest administrative data could be a valuable resource in the tracking and reporting of infectious zoonotic diseases. The overwhelming majority of cases of LD and RMSF cases within the MCO data source was expected and supports the findings of others (GER 2004; CDC NCEZID 2010). This study suggests LD rates reported to the state are well below that of MCO administrative data, and the actual statewide prevalence rate over the study period may be 3.8 per 100k, rather than 0.49 as reported by TDH statistics. This equates to an approximate 7.7 fold difference over the entire study period, resulting in an additional 1,956 cases above the 292 reported to TDH. This suggests that, on average, about 196 cases of LD go unreported each year in Tennessee. This supports the body of evidence suggesting LD is underreported, possibly up to 12-fold in some areas (Meek et al 1992; Coyle et al. 1996). Though these diseases are required to be reported to the health department through the National Notifiable Disease Surveillance System, reporting is a voluntary process. It is known that many LD cases are incomplete, unavailable, and not reported to the CDC (Bacon et al. 2008). Fines associated with non-reporting of diagnosed cases are relatively low and therefore provide little incentive to do so (MCO internal communication). However, the process of estimating a true prevalence rate is difficult, because there is also evidence suggesting LD cases are over-reported in areas that are not endemic for the disease (Rosen 2009), possibly due to misdiagnoses (Steere et al. 1993; Svenungsson and Lindh 1997) and having similar clinical symptoms as other diseases such as Southern Tick Associated Rash Illness (STARI) (Moncayo 2006; Rosen 2009). Additionally the deer tick (*Ixodes scapularis*), which is the primary vector of Lyme disease, is rarely found in Tennessee (Moncayo 2006; Rosen 2009), thus providing biological evidence of over-estimating the disease. This conflicting evidence supports the need for further investigation into integrating data sources.

RMSF has been a reportable illness since the 1920s. RMSF rates were slightly higher according to the MCO data, and suggest the actual number of cases in the state could have been 3.1 per 100k rather than 2.5 (an average difference of approximately 33 more cases per year). RMSF is the most severe and frequently reported tick rickettsial disease in the US (CDC NCEZID 2010). Tennessee is one of the top 5 states for RMSF transmission, accounting for approximately 12% of cases nationwide. As with LD, the number of RMSF cases may be underreported due to vague and/or asymptomatic infections (Lacz et al. 2010), and despite frequent laboratory testing and reports of RMSF, the true incidence in Tennessee is unknown (Moncayo et al 2010). Indirect immunofluorescence assay (IFA) serologic testing is used by the CDC and most state laboratories, though this test commonly produces false positive and false negative results (GER 2004) and therefore cannot always provide definitive proof of RMSF in the early symptomatic phase. Additionally, diagnostic levels of antibodies do not appear until a week or more after onset of symptoms, thus making early detection difficult. Prospective active surveillance for RMSF in regions where the disease is hyperendemic suggests that as many as 50% of all cases (including confirmed but unreported deaths due to RMSF) are missed by passive surveillance mechanisms (Wilfert et al. 1981).

Overall, disease incidence rates were higher using administrative data for all tick-borne diseases except HME (ehrlichiosis). The state reported a nearly 3-fold increase in HME cases from 2007 to 2008. Cases in the neighboring state of Georgia have increased dramatically since being reportable in 1999 (GER 2004) and recent expansion of the lone star tick has increased cases in the New England area (CDC MMWR 1998; ALDF 2006). HME rates may be comparatively lower in MCO data because clinical diagnosis is difficult due to misdiagnoses and limitations of confirmatory testing. Diagnoses are often made before laboratory confirmation is available. HME in Tennessee is under recognized and not routinely tested (Moncayo 2006). Patients will usually seek medical care when initially experiencing vague and possibly mild flu-like symptoms, prior to the presence of classical diagnostic signs and symptoms (GER 2004). The large deviation in HME cases between MCO and TDH data warrant further investigation into diagnosing patterns because it is apparent the disease is being reported to the state, but not necessarily recorded in the ICD-9 medical claims system as the specified coding level examined during this study.

Though statistical testing of babesiosis was inconclusive due to the small sample size, MCO data indicated at least 3 cases of babesiosis were diagnosed in Tennessee during the 2000-2009 study period. This is of interest because babesiosis has never been reported in the state, and was only recently discussed at the 2010 International Conference on Emerging Infectious Diseases Conference, whereby the authors suggested they had discovered the first diagnosed case in Tennessee in 2009 (Mosites et al. 2010). These authors are now attempting to identify animal reservoir hosts and tick vectors. Data from the MCO could aid in this effort, and suggest that at least 2 other cases occurred prior to this finding.

The large noted differences in tularemia cases for the 2000-2004 time period and then convergence of values for the remaining study period was unexpected. An outbreak of tularemia has occurred in Tennessee in the distant past (Warring and Ruffin 1946), therefore it is possible that an isolated acute outbreak occurred but went unreported (NIAID 2008). It is also plausible this increase was related to bioterrorism because *F. tularensis*, the causative agent of tularemia, can be spread via aerosol transmission. Since the 2001 terrorist attacks at the World Trade Center, there is heightened awareness of this disease (Altman 2002; Palmore et al. 2002) and therefore may explain the spikes in diagnosed cases.

Patients diagnosed with La Crosse viral encephalitis were much younger than all others with a diagnosed zoonotic disease (median age: 8), and is consistent with the findings of others (Erwin et al. 2002). Due to sample size, no definitive conclusion for comparing the TDH with MCO data was reached. However, the data suggest more cases were reported than diagnosed for years 2000-2003. The relationship reversed from 2006-2009, suggesting more cases were diagnosed than reported. LACV is the predominant virus of the California serogroup, which is made up of other viral infections including St. Louis encephalitis, Eastern/Western equine encephalitis, and other unspecified mosquito-borne viral encephalitis. Diseases in this serogroup can present with similarities. Thus, uncertainty around an exact diagnosis of LACV is possible because non-diagnosed cases may have failed to develop antibodies or available testing procedures are not sufficiently sensitive (Erwin et al. 2002). As previously mentioned, examination of specific ICD-9 codes shows it is possible some cases went undetected due to variation in clinician coding practices.

West Nile virus rates from the MCO data followed a similar temporal pattern to TDH reported cases, though actual MCO numbers were slightly lower and statistical testing was inconclusive due to sample size. As with LACV, the clinical diagnosis of WNV (ICD-9: 066.4) falls within a larger more generic root ICD-9 category of “066: Other arthropod-borne viral diseases.” This phase of the study examined only those cases with specific ICD-9 codes, and it is therefore possible that many cases reported to the state were diagnosed at the root level, rather than the actual ICD-9 code. To better understand the impact this may have on rates, I post-hoc examined per 100k rates at the root ICD-9 level for WNV and determined rates averaged over the study period were 52 times higher (4.37 vs. 0.08) when using the root diagnosis code compared to the specific ICD-9 diagnosis. This suggests physicians are far more likely to code at the root level, versus the more specific code level. Further work in this area is needed.

Zoonotic diseases are very seasonal in nature, mainly due to the temporal dynamics of the vectors’ life cycles and population densities of intermediate hosts.

For example, mid-summer peaks in Lyme disease incidence suggest tick nymphs are the life stage most responsible for transmitting infections to humans (Killilea et al. 2008). Additionally, infection rates may be higher in the summer months because the general public spends more time out of doors during this time, thus increasing the likelihood of exposure. Results from this study suggest administrative data and actual reported cases from TDH follow similar seasonal distributions patterns. TDH data cannot be compared directly to MCO on a case by case basis (*i.e.*, cannot match a patient record to a state reported case). MCO prevalence data slightly lagged behind TDH by approximately a month or two. This is expected because the TDH event date represents the estimated date of exposure, whereas MCO data represents the date when the infected individual sought medical care. It is known that symptoms of a zoonotic disease can develop days or even months after a bite (CDC NCEZID 2010). This lag phenomenon further confirms the usefulness of MCO data in tracking zoonotic infections, as it provides an estimate of the time from exposure to treatment. This seasonal overlapping of data serves as both visual and quantitative confirmation that MCO data are in fact a viable source for detecting zoonotic infections. That is, this correlation between data sources can serve as an indication that each data source is measuring similar events.

Limitations in study include the inability to empirically filter out claims in error. Even though the data were filtered to include only cases with at least 3 line items, there still exists the possibility of claims coding errors. This filtering also limits the ability to potentially estimate the true rate, because valid claims might be removed using this filtering process. Treatment for these diseases is also not necessarily consistent between health care providers, so this introduces complexity in any attempts to develop an empirically based claim line item count algorithm. Uncertainties surrounding the diagnosis and reporting of these cases suggest such trends must be interpreted with caution. Historical population counts by county were not known for MCO data and were therefore estimated, and this estimation may not be without error. There was no control for patients' enrollment time in the plan, so it is possible that a patient's claims records are incomplete. For five of the seven studied diseases, statistical testing was not possible due to sample sizes. However, it can be argued that statistical testing is not necessarily required because data for both sources were not drawn from a sample and represent the population. Therefore, any noted differences are in fact real differences. We are further limited by the inability to relate, via a patient identifier, a medically diagnosed case to a CDC defined “confirmed” or “probable” case. A patient could be coded with LD in the MCO claim system without necessarily having a laboratory confirmed diagnosis, or a physician could report a confirmed case without laboratory confirmation if the patient presented with erythema migrans and was recently in an endemic county (CDC 1995). Only one other study has examined the differences between administrative data and notifiable disease data (Yiannakoulias and Svenson 2009), and they arrived at the same conclusion that administrative health data may be insufficiently precise without laboratory confirmation. Both studies conclude that administrative data could enhance the current passive surveillance registry system.

CONCLUSIONS

Zoonotic diseases in Tennessee, particularly LD and RMSF, may be significantly underreported to the state health department within the current passive system. Administrative medical claims data suggest that approximately 200 cases of Lyme disease and 30 cases of Rocky Mountain spotted fever go unreported each year in Tennessee. Medical claims data show babesiosis may have been present in the state 8 years prior to what is currently thought to be the first reported incident. This study successfully demonstrates the usefulness of administrative claims data in the tracking and reporting of zoonotic diseases.

In the past 10 years, the number of officially reported cases of tick-borne diseases in Tennessee has increased (*e.g.*, from 0 officially reported cases of ehrlichiosis in 1995 to 74 in 2008, and from 0 reported cases of RMSF in 1995 to 232 in 2008) (TDH WebAim 2010). State and local public health officials rely on health care providers, laboratories, and other public health personnel to report the occurrence of notifiable diseases to state and local health departments (CDC 1997). Missing from this statement is health plans and the data they could provide to state and national surveillance efforts. Without such data, trends cannot be accurately monitored, unusual occurrences of diseases might not be detected, and the effectiveness of intervention activities cannot be easily evaluated. Reporting methods using administrative data and CDC surveillance are similar in that both represent the location of the disease based on the resident county of the infected individual, not exposure. This should help control for differences in data gathering methodologies. Though dates may differ, both data sources also capture a temporal component, where TDH reports the estimated exposure date and MCO data reports the data the infected individual sought medical treatment. However, specific spatio-temporal differences of these two data sources are not known and further work examining these attributes is warranted.

REFERENCES

Altman, G. 2002. Bioterrorism's invisible threats: heightened awareness will help nurses identify real and suspected bioterrorism. *Nursing Management*. 33(1):43, 45-47.

American Lyme Disease Foundation (ALDF). 2006. Ehrlichiosis fact sheet. http://www.aldf.com/Ehrlichiosis.shtml; Accessed November 25, 2010

Bacon, R., K. Kugeler, and P. Mead. 2008. Surveillance for Lyme disease -- United States, 1992 - 2006. MMWR Surveillance Summaries. 57(SS10):1-9.

Boyce, J. 1975. Recent trends in the epidemiology of tularemia in the United States. *The Journal of Infectious Diseases*. 131(2):197-199.

Campos-Outcalt, D., R. England, and B. Porter. 1991. Reporting of communicable diseases by university physicians. *Public Health Reports*. 106(5):579-583.

Centers for Disease Control and Prevention (CDC). 1995. Notice to Readers: Recommendations for Test Performance and Interpretation from the Second National Conference on Serologic Diagnosis of Lyme disease. MMWR Weekly 44(31): 590-591.

Centers for Disease Control and Prevention (CDC). 1997. Case definitions for infectious conditions under public health surveillance: Recommendations and reports. 46(RR10):1-55.

Centers for Disease Control and Prevention (CDC MMWR). 1998. Statewide surveillance for ehrlichiosis -- Connecticut and New York, 1994-1997. 47(23):476-480.

Centers for Disease Control and Prevention (CDC NCEZID). 2010. National Center for Emerging and Zoonotic Infectious Diseases (http://www.cdc.gov/ncezid/) accessed September 21, 2010

Centers for Disease Control and Prevention (CDC NNDSS). 2010. National Notifiable Diseases Surveillance System (http://www.cdc.gov/ncphi/disss/nndss/nndsshis.htm) accessed September 21, 2010.

Coyle, B., G. Strickland, Y. Liang, C. Pena, R. McCarter, and E. Israel. 1996. The public health impact of Lyme disease in Maryland. *Journal of Infectious Diseases*. 173(5):1260-1262.

DeNavas-Walt, C., B. Proctor, and J. Smith. 2010. US Census Bureau, Current population reports, P60-238, Income, poverty, and health insurance coverage in the United States: 2009. US Government Printing Office, Washington, DC

Doyle, T., M. Glynn, S. Groseclose. 2002. Completeness of notifiable infectious disease reporting in the United States: an analytical literature review. *American Journal of Epidemiology*. 155(9):866-874.

Erwin, P., T. Jones, R. Gerhardt, S. Halford, A. Smith, L. Patterson, K. Gottfried, and W. Schaffner. 2002. La Crosse encephalitis in eastern Tennessee: Clinical, environmental, and entomological characteristics from a blinded cohort study. *American Journal of Epidemiology*. 155(11):1060-1065.

Figueiras, A., E. Lado, S. Fernández, and X. Hervada. 2004. Influence of physicians' attitude on under-notifying infectious diseases: a longitudinal study. *Public Health*. 118(7):521-526.

Georgia Epidemiology Report (GER). 2004. Common tickborne diseases in Georgia: Rocky Mountain spotted fever and Human Monocytic Ehrlichiosis. 20:5.

Government Accountability Office (GAO). 2004. Emerging infectious diseases review of state and federal disease surveillance efforts: report to the Chairman, Permanent Subcommittee on Investigations, Committee on Governmental Affairs, US Senate. GAO report# GAO-04-877.

Kalluri, S., P. Gilruth, D. Rogers, and M. Szczur. 2007. Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: A review. *PLoS Pathogens*. 3(10):1361-1371.

Killilea, M., A. Swei, R. Lane, C. Briggs, and R. Ostfeld. 2008. Spatial dynamics of Lyme disease: A Review. *EcoHealth*. 5(2):167–195.

Koo, D., and S. Wetterhall. 1996. History and current status of the National Notifiable Diseases System. *Journal of Public Health Management and Practice*. 2(4):4-10.

Koo, D., and B. Caldwell. 1999. The role of providers and health plans in infectious disease surveillance. *Effective Clinical Practice*. 2(5):247-252.

Lacz, N., R. Schwartz, and R. Kapila. Rocky Mountain spotted fever. MedScape eMedicine. Available at: http://emedicine.medscape.com/article/1054826-overview Accessed Dec. 1, 2010

Marier, R. 1977. The reporting of communicable diseases. *American Journal of Epidemiology*. 105(6):587-590.

Meek, J., C. Roberts, E. Smith Jr, M. Cartter. 1996. Underreporting of Lyme disease by Connecticut physicians. *Journal of Public Health Management and Practice*. 2(4):61-65.

Moncayo, A. 2006. Vector-borne diseases in Tennessee. Tennessee Department of Health & Vanderbilt University Medical Center. Twenty Sixth Biennial State Public Health Vector Control Conference. Available at: http://www.cdc.gov/ncidod/dvbid/westnile/conf/26thbiennialVectorControl/pdf/state/Tennessee.pdf

Moncayo, A., S. Cohen, C. Fritzen, E. Huang, M. Yabsley, J. Freye, B. Dunlap, J. Huang, D. Mead, T. Jones, J. Dunn. 2010. Absence of Rickettsia rickettsii and occurrence of other spotted fever group rickettsiae in ticks from Tennessee. American Journal of Tropical Medicine and Hygiene. 83(3):653-657.

Mosites, E., P. Wheeler, N. Pieniazek, M. Xayavong, L. Carpenter, T. Jones, B. Herwaldt, and J. Dunn. 2010. Human babesiosis caused by a novel babesia parasite--Tennessee, 2009. Unpublished abstract presented at the International Conference on Emerging Infectious Disease (ICEID) July 2010 Atlanta, GA.

National Institute of Allergy and Infectious Diseases (NIAID). 2008. Tularemia fact sheet. Available at: http://www.niaid.nih.gov/topics/tularemia/pages/default.aspx

Palmore, T., G. Folkers, C. Heilman, J. La Montagne, and A. Fauci. 2002. The NIAID Research Agenda on Biodefense: The National Institute of Allergy and Infectious Diseases faces new challenges in fighting the war on bioterrorism. *American Society for Microbiology.* 68(8):375-382.

Rosen, M. 2009. Investigating the maintenance of the Lyme disease pathogen, *Borrelia burgdorferi*, and its Vector, *Ixodes scapularis*, in Tennessee. Master's Thesis, University of Tennessee. Available at: http://trace.tennessee.edu/utk_gradthes/554

Rutherford, G. 1998. Public health, communicable diseases, and managed care: will managed care improve or weaken communicable disease control? *American Journal of Preventative Medicine*. 14(3 Suppl):53-59.

Salah, A., Y. Kamarianakis, S. Chlif, N. Alaya, and P. Prastacos. 2007. Zoonotic cutaneous leishmaniasis in central Tunisia: spatio-temporal dynamics. *International Journal of Epidemiology*. 36(5):991-1000.

SAS®. Statistical Analysis Software. Copyright © 2006 - 2008 by SAS Institute Inc., Cary, NC, USA.

Steere, A., E. Taylor, G. McHugh, and E. Logigian. 1993. The overdiagnosis of Lyme disease. *The Journal of the American Medical Association*. 269(14):1812-1816.

Svenungsson, B., and G. Lindh. 1997. Lyme borreliosis--an overdiagnosed disease? *Infection*. 25(3):140-143.

Tennessee Department of Health (TDH CEDS). 2010. Tennessee Department of Health, Communicable and Environmental Disease Services: Reportable Diseases and Events. http://health.state.tn.us/ceds/notifiable.htm accessed October 5, 2010.

Tennessee Department of Health (TDH WebAim). 2010. Tennessee Department of Health, Communicable and Environmental Disease Services: The Communicable Disease Interactive Data Site: Available at: http://health.state.tn.us/Ceds/WebAim/WEBAim_criteria.aspx

Tennessee Wildlife Resources Agency. Tennessee Land Use/Land Cover Landsat TM imagery 1997. TN SDS. Tennessee Spatial Data Service metadata files. http://www.tngis.org/frequently_accessed_data.html accessed October 5, 2010.

US Census Bureau. 2009. 2005-2009 American Community Survey. Available at: http://factfinder.census.gov/servlet/ADPGeoSearchByListServlet?ds_name=ACS_2009_5YR_G00_&_lang=en&_ts=313948522194

US Department of Agriculture (USDA). 2000. Summary Report: 1997 National Resources Inventory (revised December 2000), Natural Resources Conservation Service, Washington, DC, and Statistical Laboratory, Iowa State University, Ames, Iowa, 89p.

Warring, W., and J. Ruffin. 1946. A tick-borne epidemic of tularemia. *The New England Journal of Medicine*. 234:137-140.

Wilfert, C., J. MacCormack, K. Kleeman, R. Philip, E. Austin, V. Dickinson, and L. Turner. 1984. Epidemiology of Rocky Mountain spotted fever as determined by active surveillance. The *Journal of Infectious Diseases*. 150(4):469-79.

Wimberly, M., A. Baer, and M. Yabsley. 2008. Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *International Journal of Health Geographics*. 7:15.

Winters, A., R. Eisen, S. Lozano-Fuentes, C. Moore, W. Pape, and L. Eisen. 2008. Predictive spatial models for risk of West Nile virus exposure in eastern and western Colorado. *American Journal of Tropical Medicine and Hygiene*. 79(4):581-590.

Yang, K., X. Zhou, X. Wu, P. Steinmann, X. Wang, G. Yang, J. Utzinger, and H. Li. 2009. Landscape pattern analysis and Bayesian modeling for predicting *Oncomelania hupensis* distribution in Eryuan County, People's Republic of China. *American Journal of Tropical Medicine and Hygiene*. 81(3):416-23.

Yiannakoulias, L. and L. Svenson. 2009. Differences between notifiable and administrative health information in the spatial–temporal surveillance of enteric infections. *International Journal of Medical Informatics*. 78(6):417-424.

Yiannakoulias, N., D. Schopflocher, and L. Svenson. 2009. Using administrative data to understand the geography of case ascertainment. *Chronic Diseases in Canada*. 30(1):20-28.

Young, J. 1998. Underreporting of Lyme disease. *The New England Journal of Medicine*. 338(22):1629.

Table 2-1: Summary statistics and results from general linear mixed models for yearly per 100k disease incidence comparisons between MCO and TDH data

	MCO		TDH		
	Yearly mean	SD	Yearly mean	SD	F-value
Lyme Disease	3.76	0.80	0.49	0.13	835.44*
Babesiosis†	0.01	0.02	0.00	0.00	-
Rocky Mtn. spotted fever	2.75	0.46	2.32	1.17	14.45*
Human Monocytic Ehrlichiosis†	0.06	0.06	0.59	0.34	-
Tularemia†	0.19	0.21	0.05	0.04	-
La Crosse Viral Encephalitis†	0.04	0.05	0.10	0.13	-
West Nile Virus†	0.08	0.09	0.08	0.12	-

* Significant at *P* < 0.05

† Mixed-models did not converge

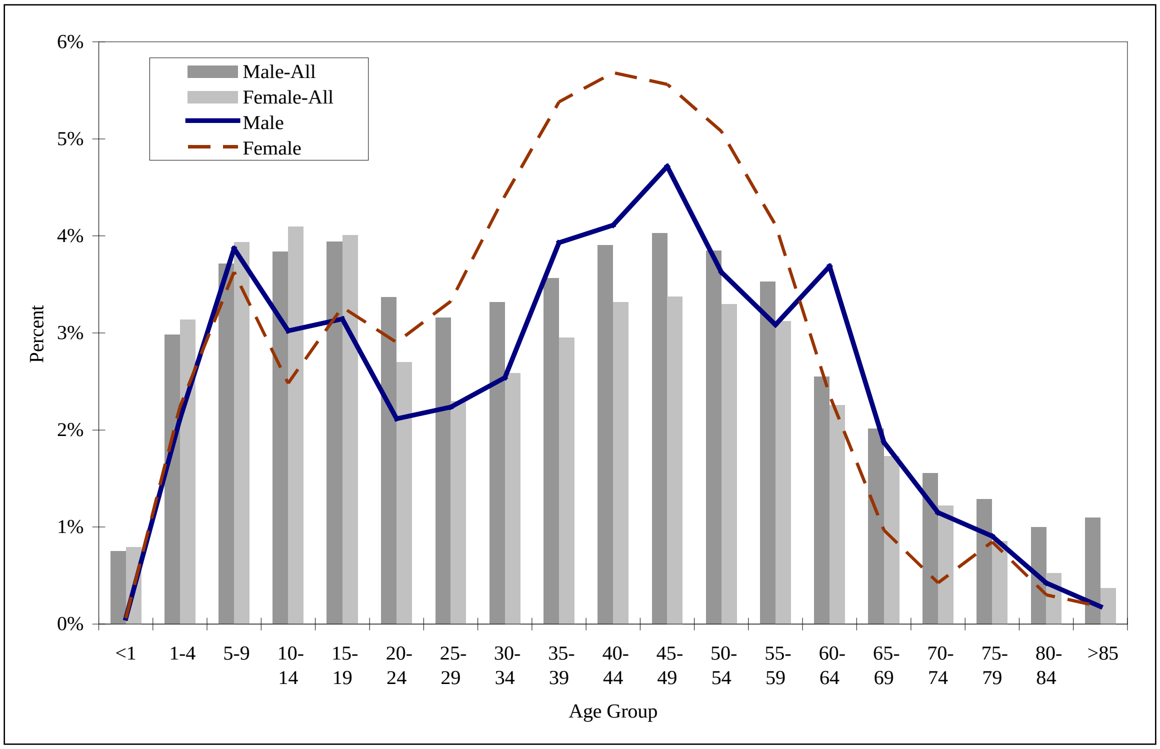


Figure 2-1: Distribution of gender across age groups for patients with one of the described diseases (lines) compared to the entire MCO population (columns)

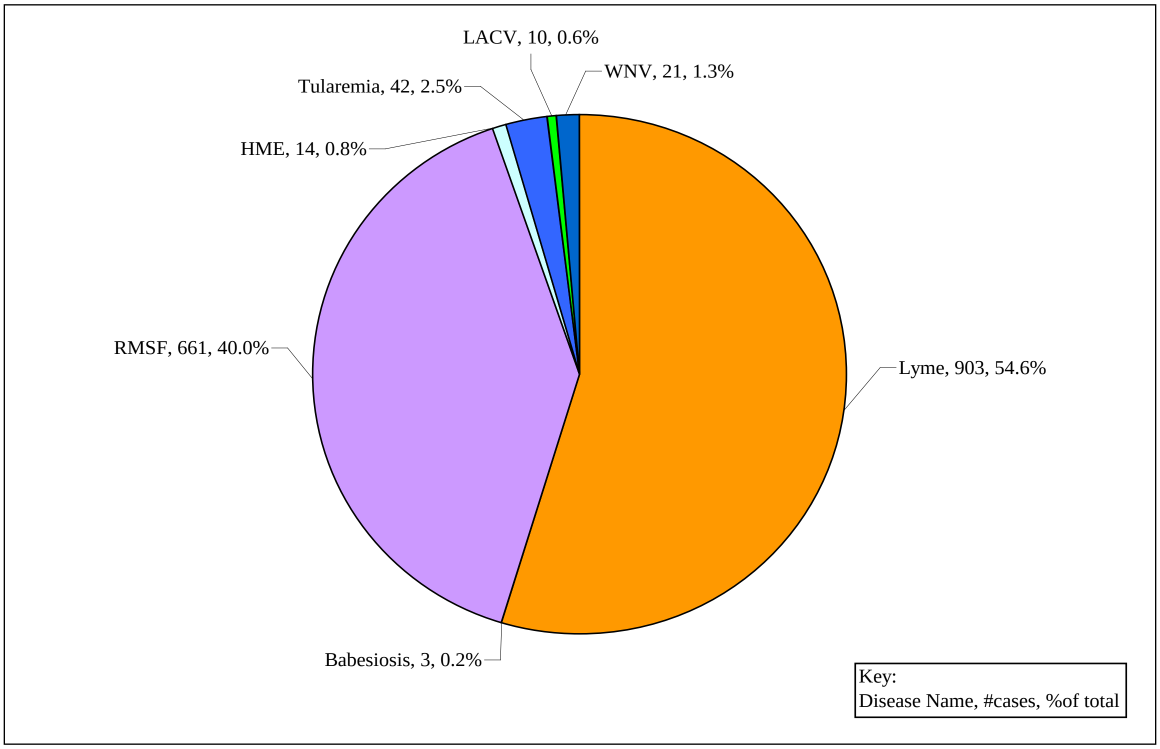


Figure 2-2: Percent distribution of medically diagnosed zoonotic diseases in Tennessee for the 2000-09 study period

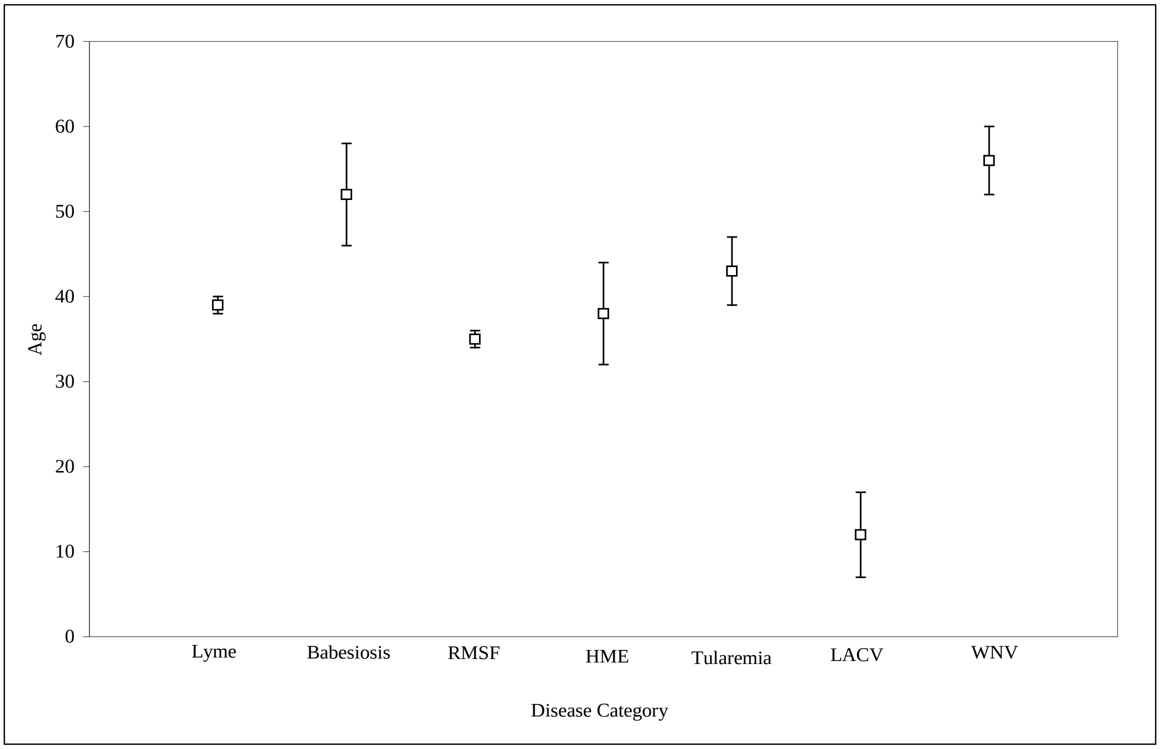


Figure 2-3: Mean (white square) and standard errors (bars) of patient age across disease type

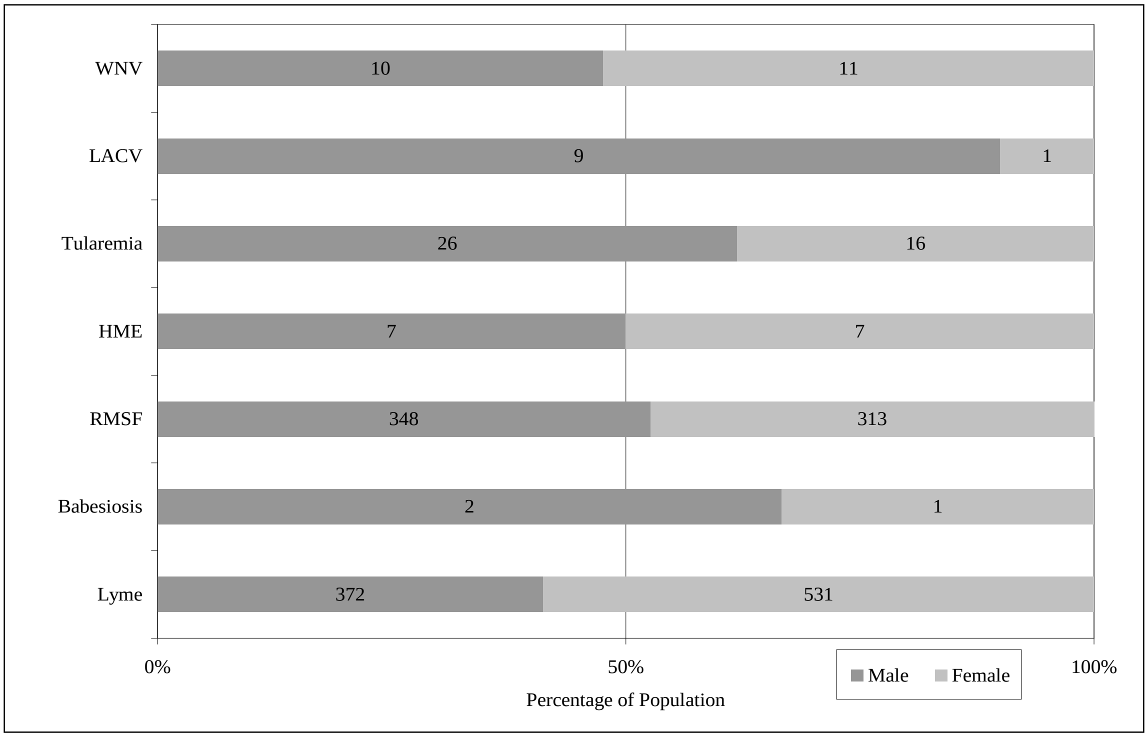
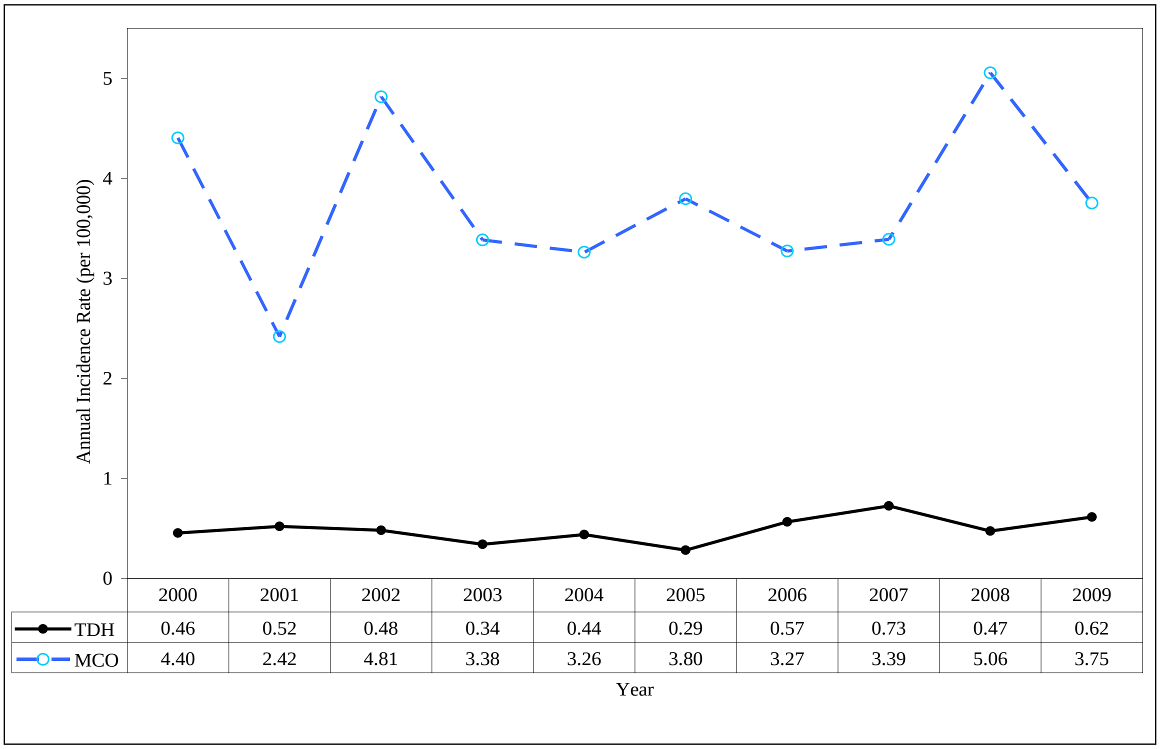


Figure 2-4: Distribution of gender across disease type



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
—●— TDH	0.46	0.52	0.48	0.34	0.44	0.29	0.57	0.73	0.47	0.62
--○-- MCO	4.40	2.42	4.81	3.38	3.26	3.80	3.27	3.39	5.06	3.75

Figure 2-5: Temporal comparison of Lyme disease incidence rates using MCO medical claims data and Tennessee State Health Department (TDH) reported data

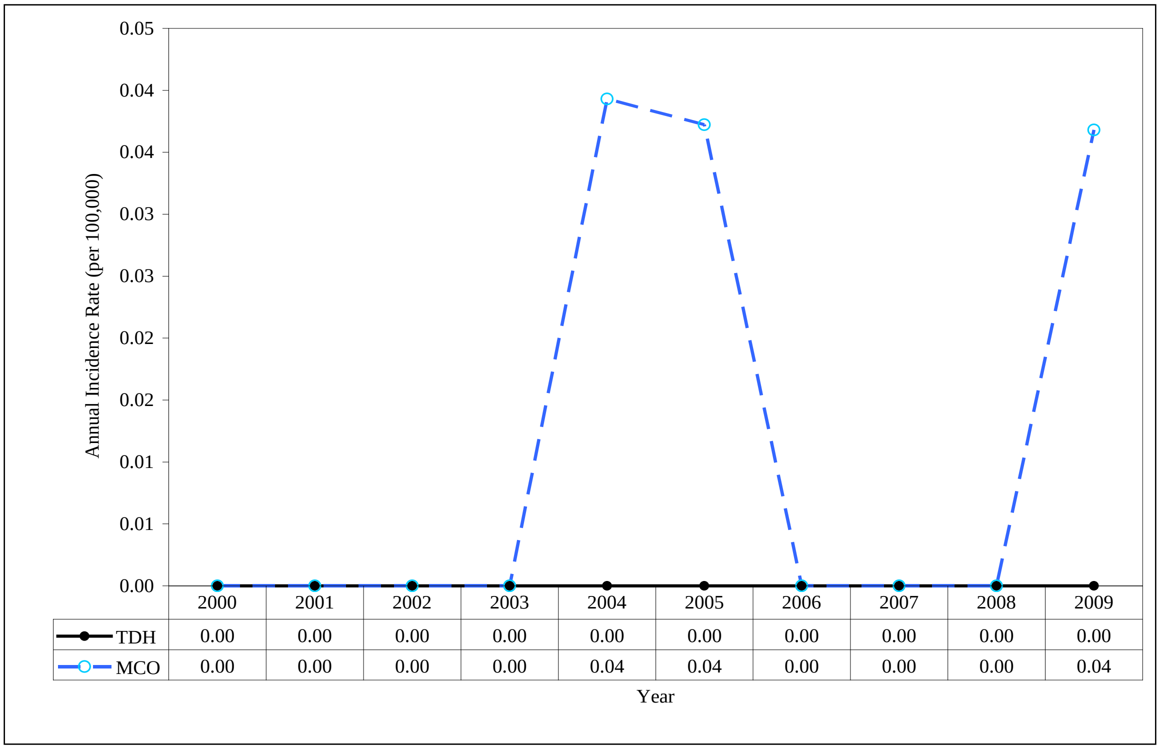
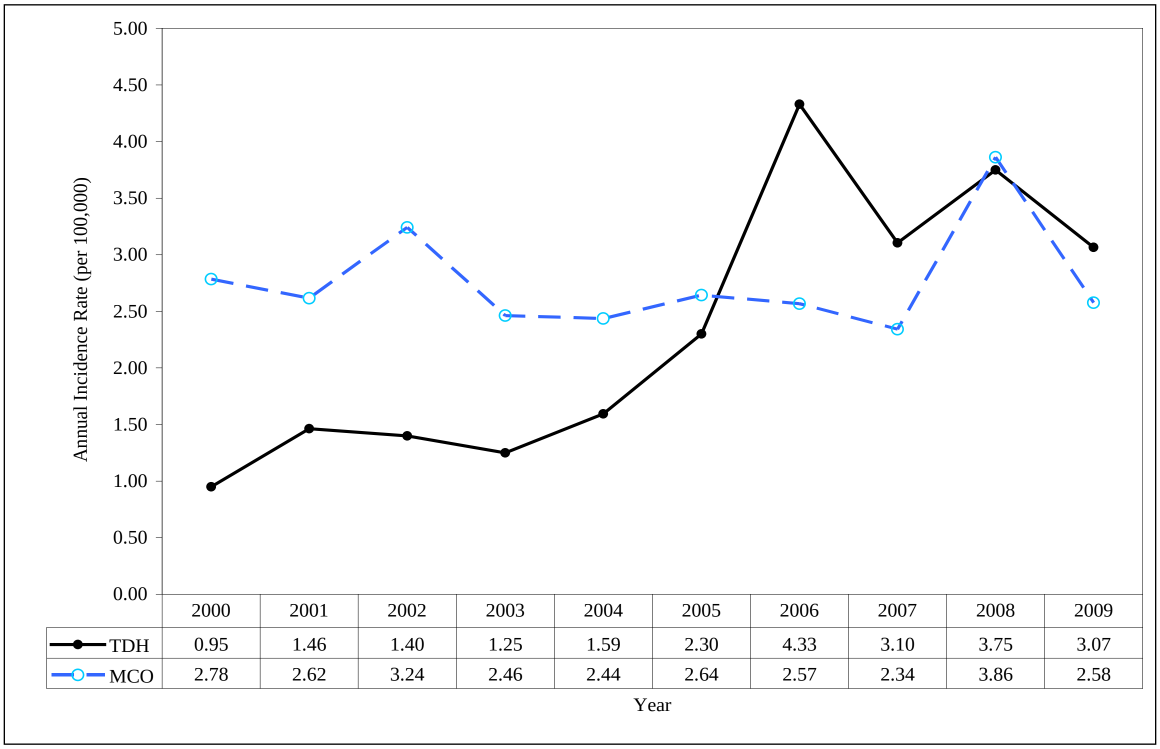
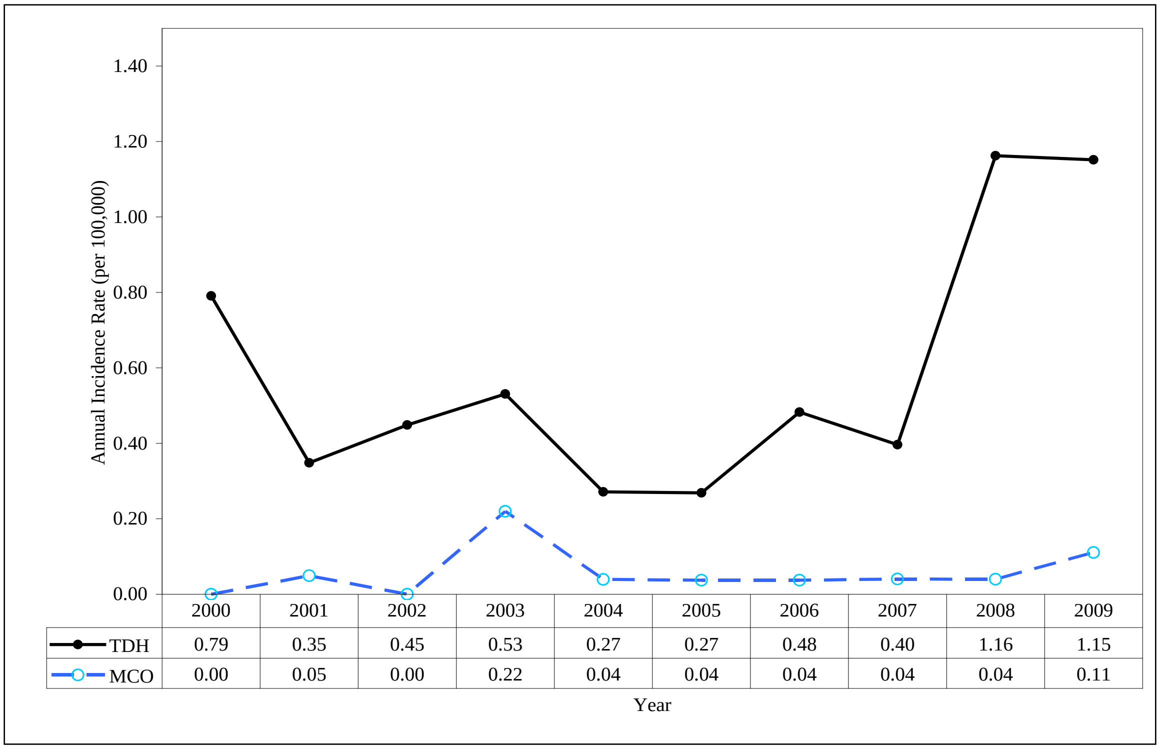


Figure 2-6: Temporal comparison of babesiosis disease incidence rates using MCO medical claims data and Tennessee State Health Department (TDH) reported data



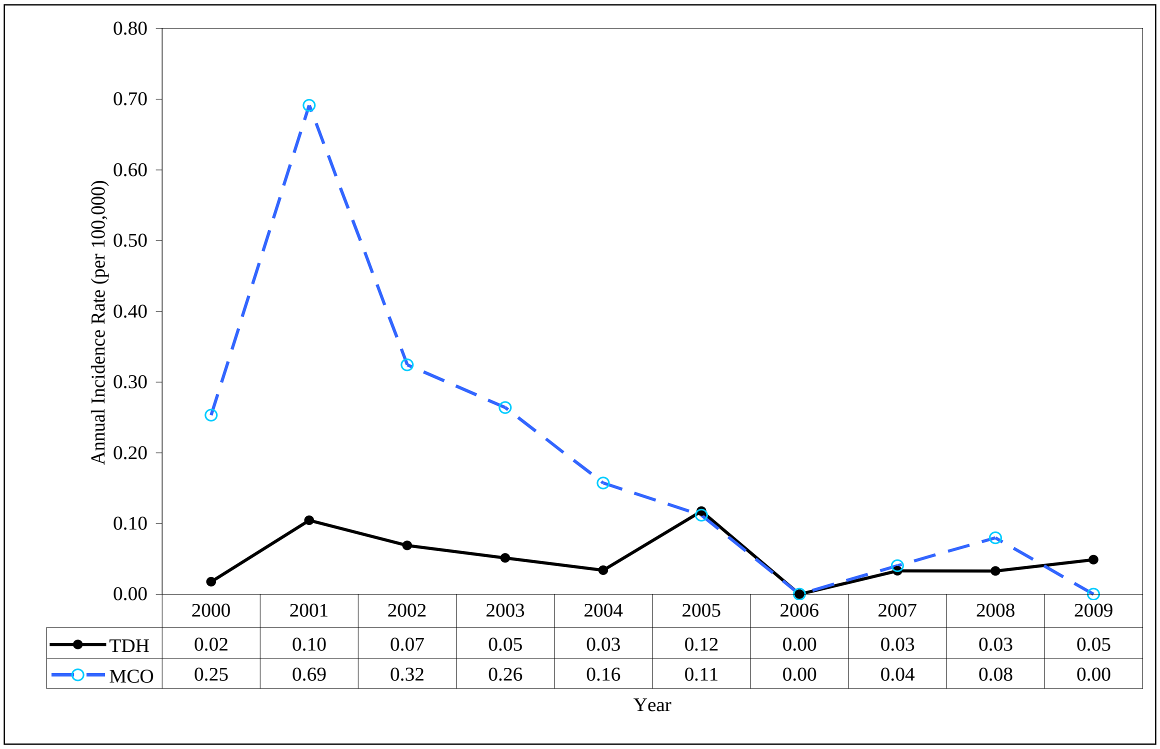
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
TDH	0.95	1.46	1.40	1.25	1.59	2.30	4.33	3.10	3.75	3.07
MCO	2.78	2.62	3.24	2.46	2.44	2.64	2.57	2.34	3.86	2.58

Figure 2-7: Temporal comparison of RMSF disease incidence rates using MCO medical claims data and Tennessee State Health Department (TDH) reported data



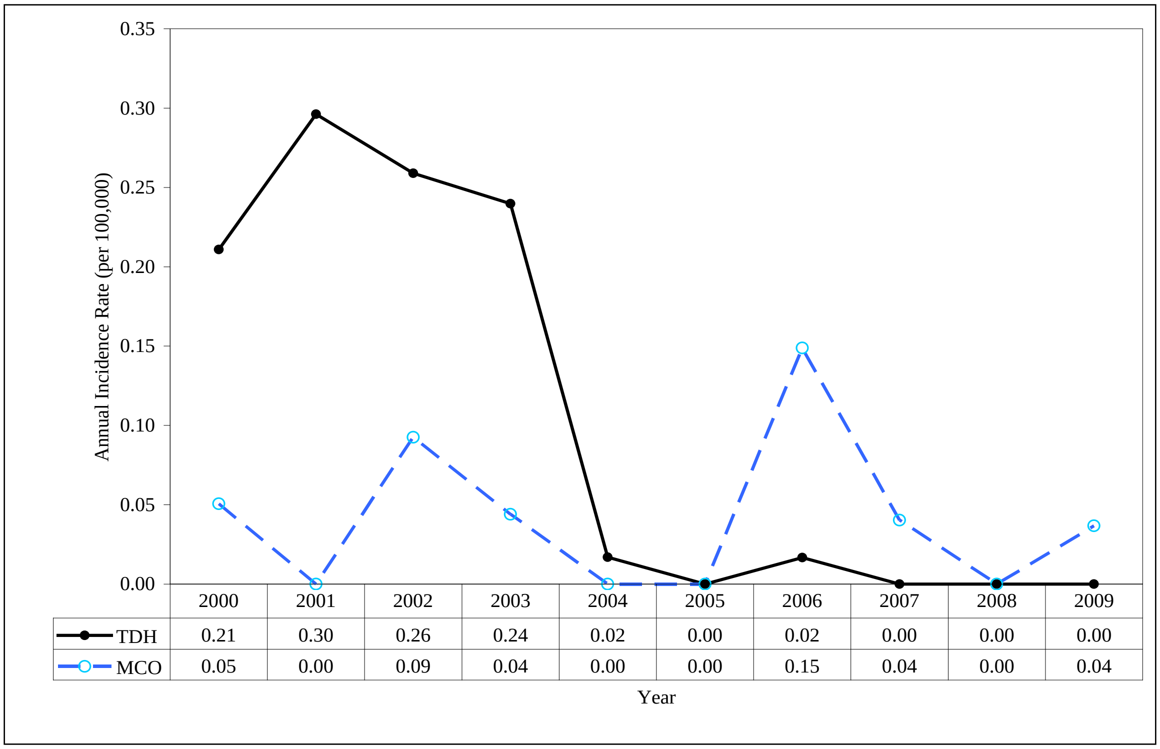
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
TDH	0.79	0.35	0.45	0.53	0.27	0.27	0.48	0.40	1.16	1.15
MCO	0.00	0.05	0.00	0.22	0.04	0.04	0.04	0.04	0.04	0.11

Figure 2-8: Temporal comparison of HME disease incidence rates using MCO medical claims data and Tennessee State Health Department (TDH) reported data



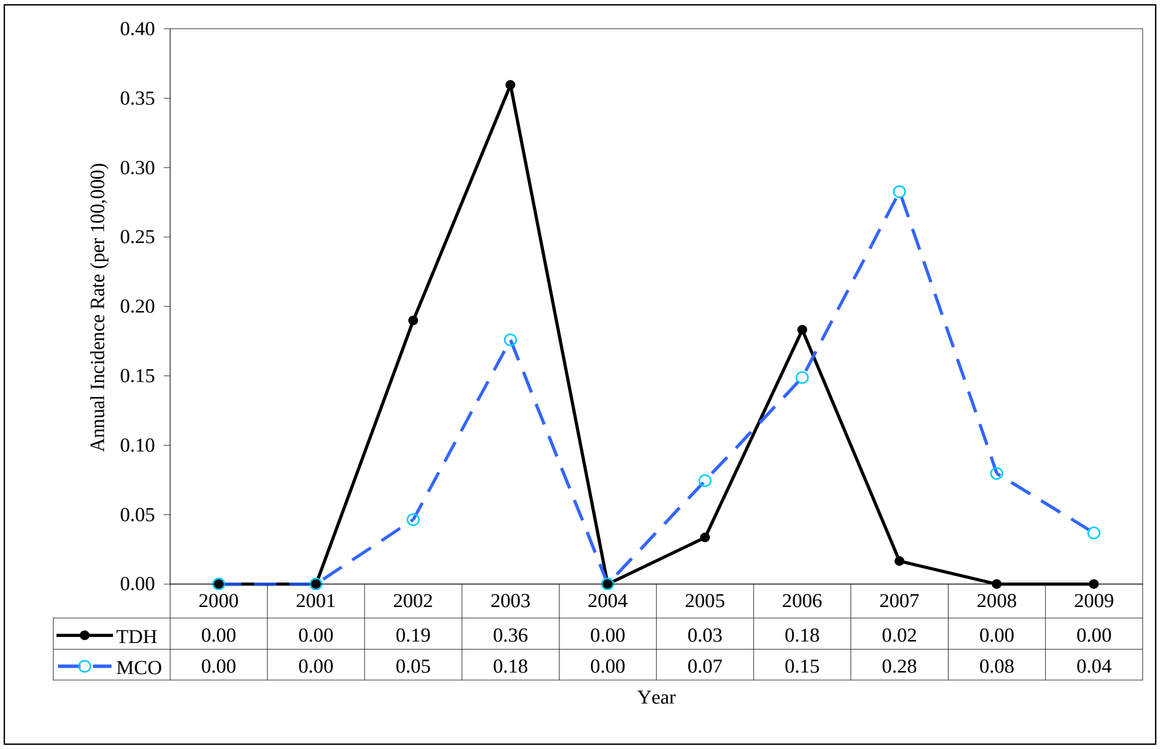
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
● TDH	0.02	0.10	0.07	0.05	0.03	0.12	0.00	0.03	0.03	0.05
○ MCO	0.25	0.69	0.32	0.26	0.16	0.11	0.00	0.04	0.08	0.00

Figure 2-9: Temporal comparison of tularemia disease incidence rates using MCO medical claims data and Tennessee State Health Department (TDH) reported data



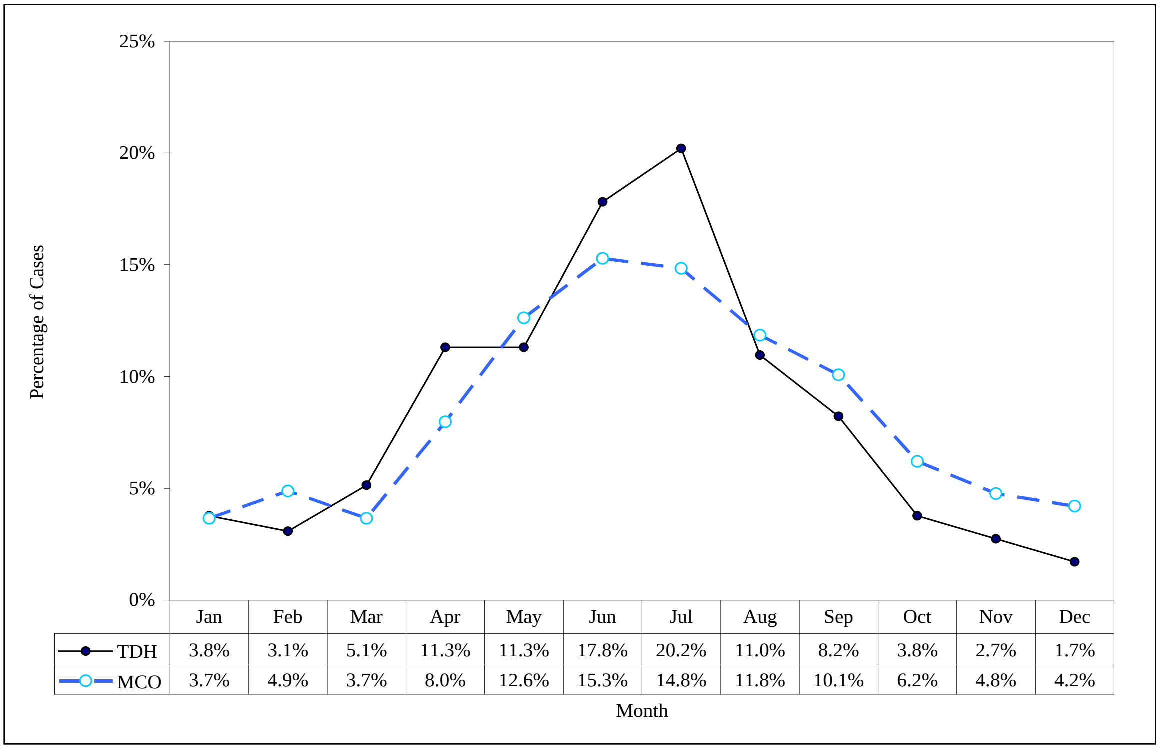
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
● TDH	0.21	0.30	0.26	0.24	0.02	0.00	0.02	0.00	0.00	0.00
○ MCO	0.05	0.00	0.09	0.04	0.00	0.00	0.15	0.04	0.00	0.04

Figure 2-10: Temporal comparison of La Crosse viral encephalitis (LACV) disease incidence rates using MCO medical claims data and Tennessee State Health Department (TDH) reported data



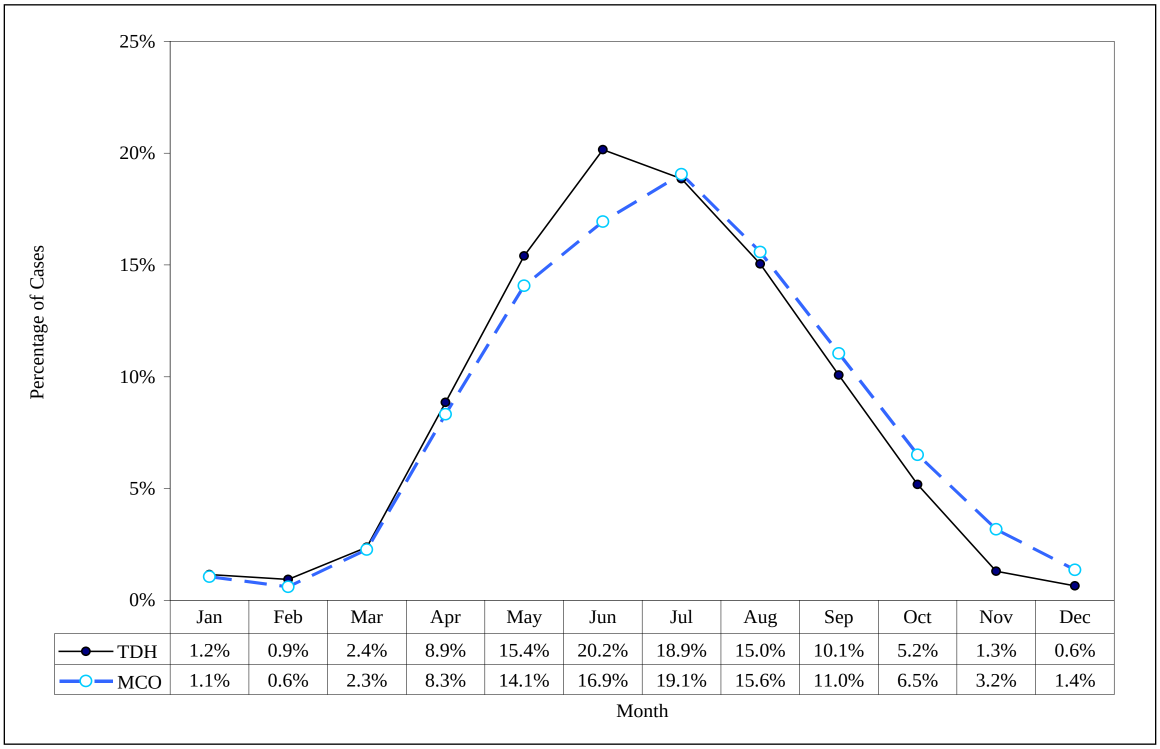
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
TDH	0.00	0.00	0.19	0.36	0.00	0.03	0.18	0.02	0.00	0.00
MCO	0.00	0.00	0.05	0.18	0.00	0.07	0.15	0.28	0.08	0.04

Figure 2-11: Temporal comparison of West Nile virus (WNV) disease incidence rates using MCO medical claims data and Tennessee State Health Department (TDH) reported data



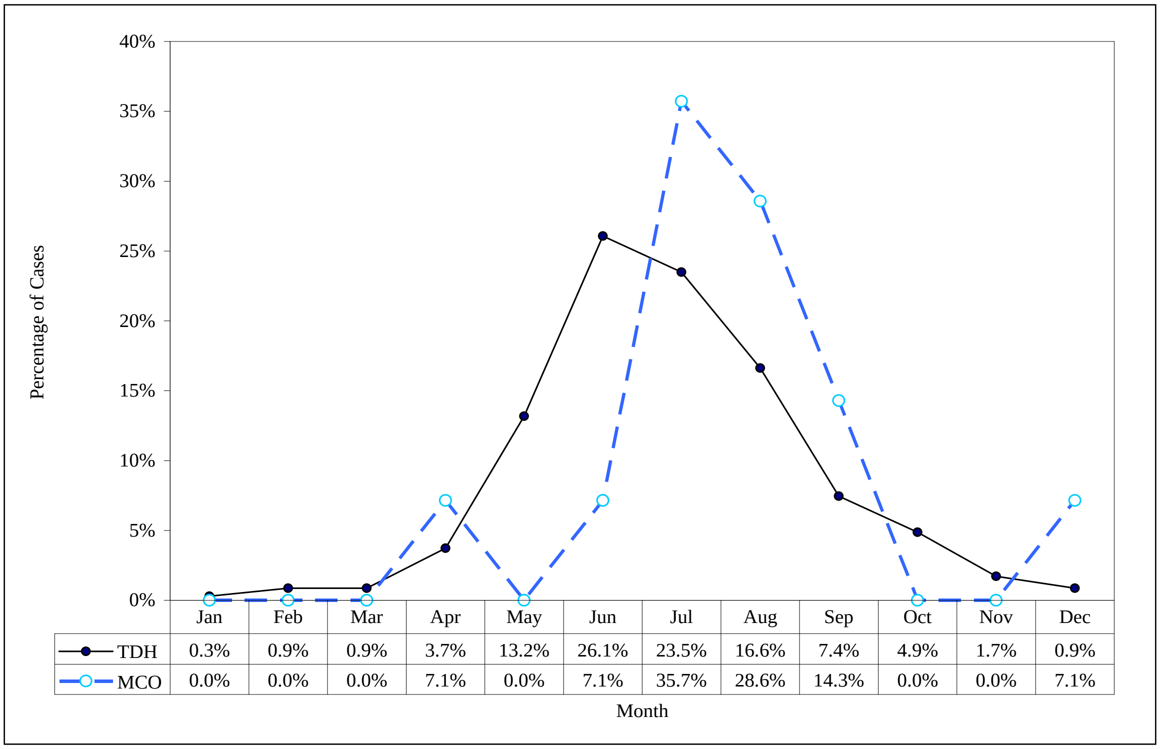
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TDH	3.8%	3.1%	5.1%	11.3%	11.3%	17.8%	20.2%	11.0%	8.2%	3.8%	2.7%	1.7%
MCO	3.7%	4.9%	3.7%	8.0%	12.6%	15.3%	14.8%	11.8%	10.1%	6.2%	4.8%	4.2%

Figure 2-12: Comparison of the seasonal distribution of Lyme disease cases for MCO data (diagnosis date according to medical claims) versus TDH data (estimated date of exposure according to state records)



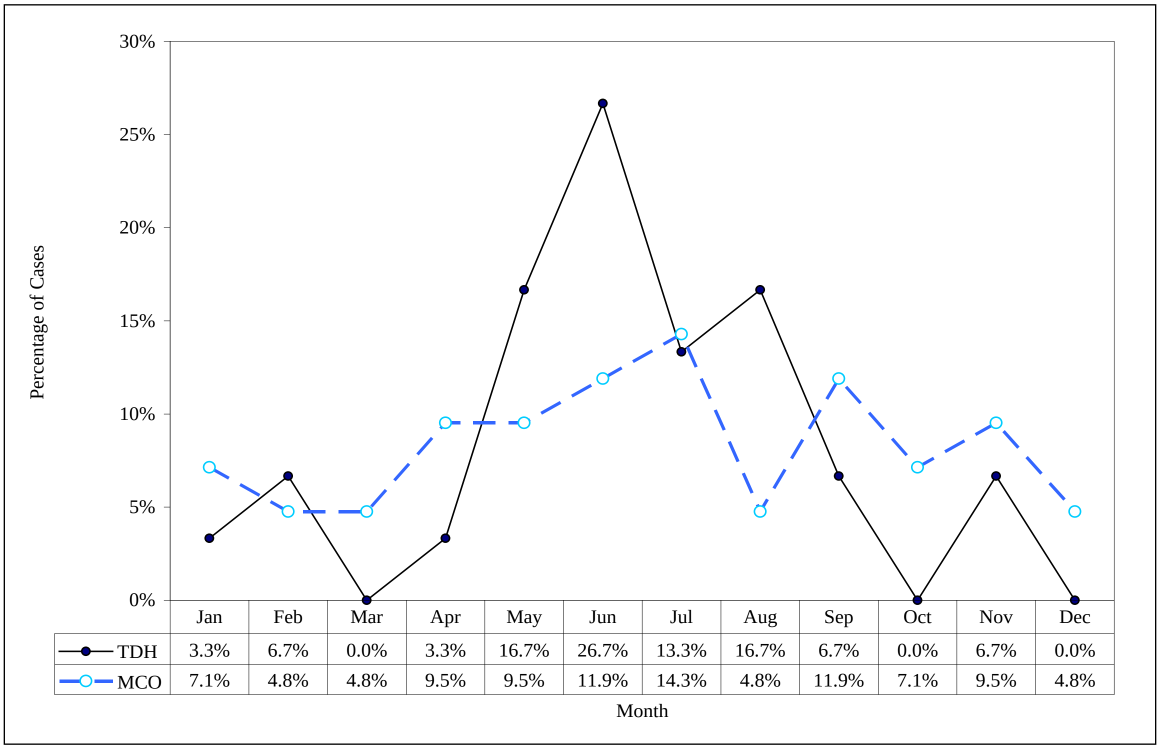
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
─●─ TDH	1.2%	0.9%	2.4%	8.9%	15.4%	20.2%	18.9%	15.0%	10.1%	5.2%	1.3%	0.6%
─○─ MCO	1.1%	0.6%	2.3%	8.3%	14.1%	16.9%	19.1%	15.6%	11.0%	6.5%	3.2%	1.4%

Figure 2-13: Comparison of the seasonal distribution of Rocky Mountain spotted fever cases for MCO data (diagnosis date according to medical claims) versus TDH data (estimated date of exposure according to state records)



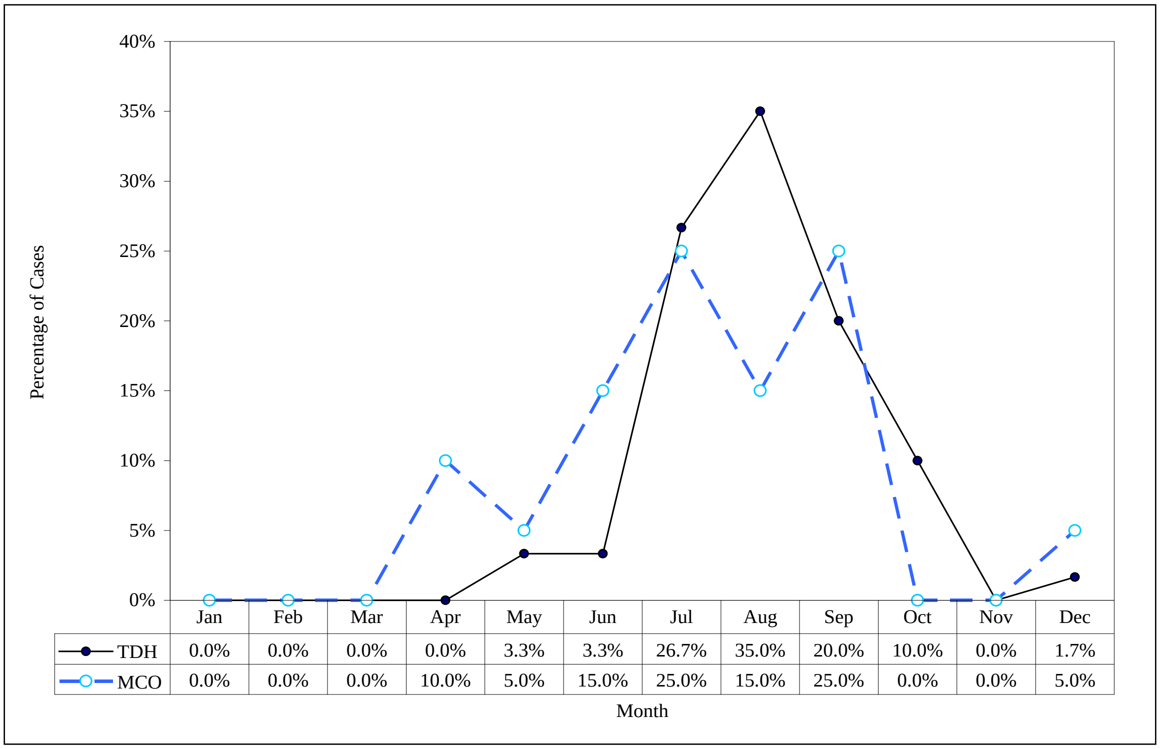
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TDH	0.3%	0.9%	0.9%	3.7%	13.2%	26.1%	23.5%	16.6%	7.4%	4.9%	1.7%	0.9%
MCO	0.0%	0.0%	0.0%	7.1%	0.0%	7.1%	35.7%	28.6%	14.3%	0.0%	0.0%	7.1%

Figure 2-14: Comparison of the seasonal distribution of human monocytic ehrlichiosis cases for MCO data (diagnosis date according to medical claims) versus TDH data (estimated date of exposure according to state records)



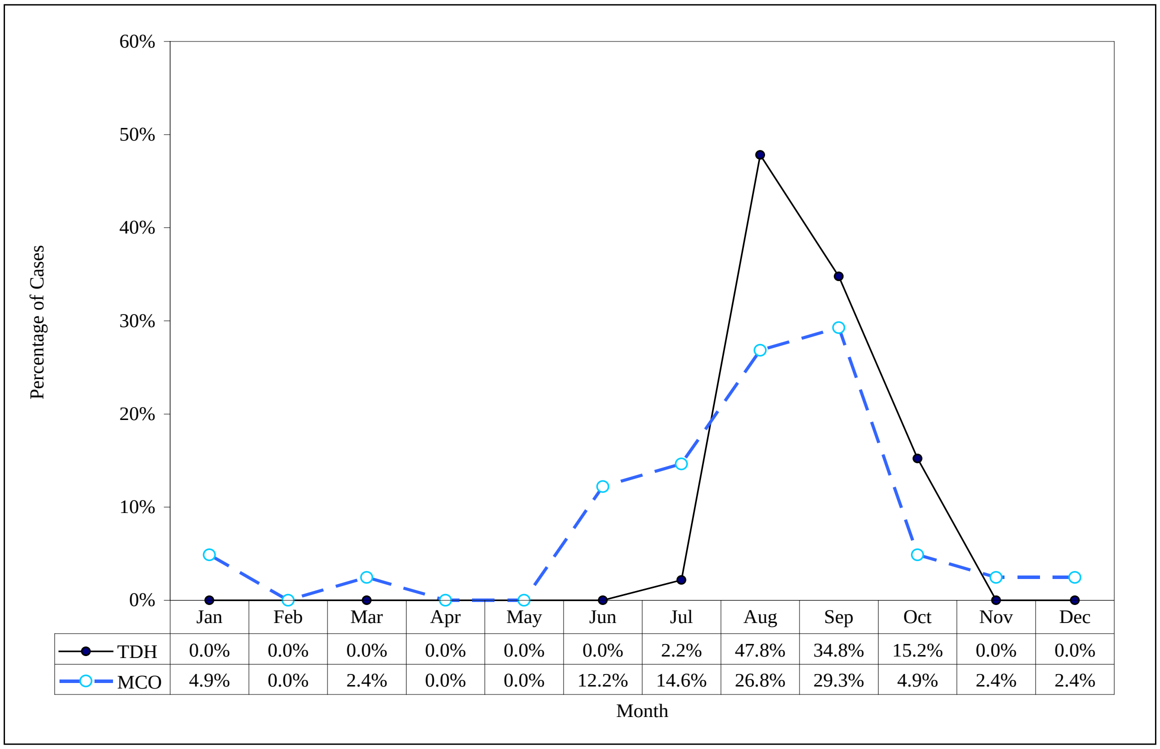
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
● TDH	3.3%	6.7%	0.0%	3.3%	16.7%	26.7%	13.3%	16.7%	6.7%	0.0%	6.7%	0.0%
○ MCO	7.1%	4.8%	4.8%	9.5%	9.5%	11.9%	14.3%	4.8%	11.9%	7.1%	9.5%	4.8%

Figure 2-15: Comparison of the seasonal distribution of tularemia cases for MCO data (diagnosis date according to medical claims) versus TDH data (estimated date of exposure according to state records)



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
TDH	0.0%	0.0%	0.0%	0.0%	3.3%	3.3%	26.7%	35.0%	20.0%	10.0%	0.0%	1.7%
MCO	0.0%	0.0%	0.0%	10.0%	5.0%	15.0%	25.0%	15.0%	25.0%	0.0%	0.0%	5.0%

Figure 2-16: Comparison of the seasonal distribution of La Crosse viral encephalitis cases for MCO data (diagnosis date according to medical claims) versus TDH data (estimated date of exposure according to state records)



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
● TDH	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.2%	47.8%	34.8%	15.2%	0.0%	0.0%
○ MCO	4.9%	0.0%	2.4%	0.0%	0.0%	12.2%	14.6%	26.8%	29.3%	4.9%	2.4%	2.4%

Figure 2-17: Comparison of the seasonal distribution of West Nile virus cases for MCO data (diagnosis date according to medical claims) versus TDH data (estimated date of exposure according to state records)

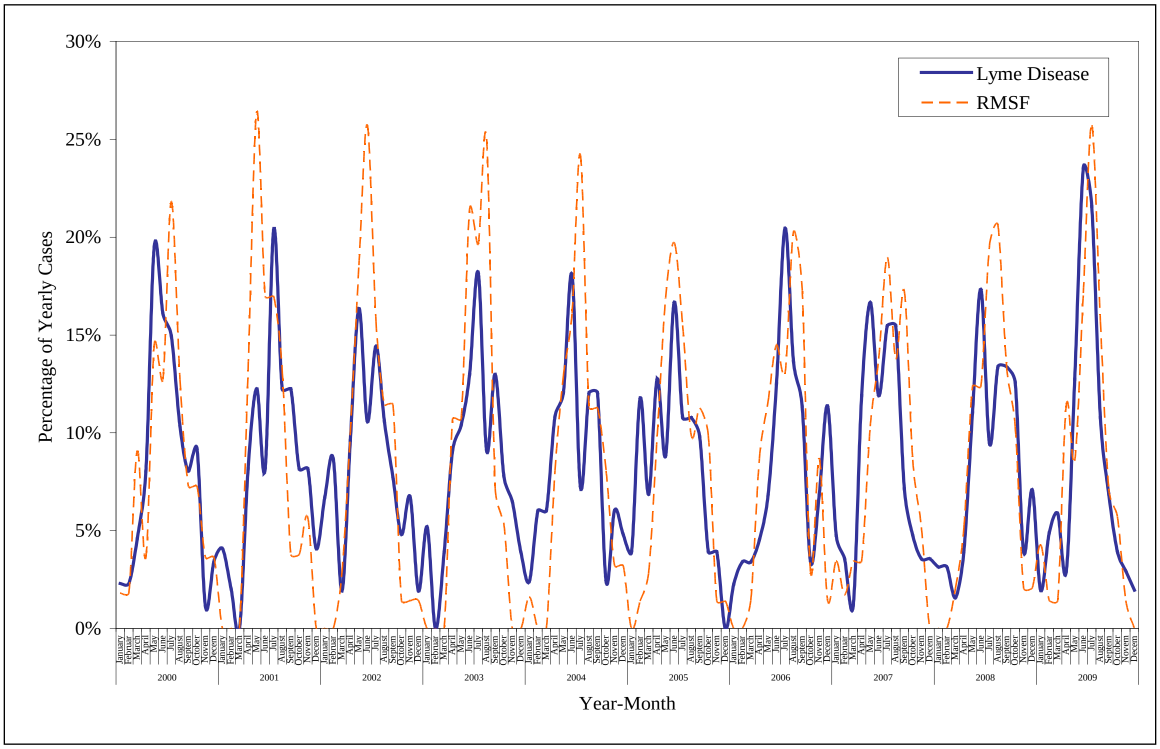


Figure 2-18: Seasonality of Lyme disease (solid line) compared to Rocky Mountain spotted fever (dashed line) over the entire 2000-09 study period using MCO administrative medical claims data

CHAPTER 3

SPATIO-TEMPORAL DIFFERENCES OF ARTHROPOD-BORNE INFECTIONS USING ADMINISTRATIVE MEDICAL CLAIMS DATA AND STATE REPORTED SURVEILLANCE DATA

ABSTRACT

When considered separately, notifiable disease registry systems and administrative medical claims data have positive and negatives attributes within disease surveillance efforts. Combined however, these data sources could provide a more complete source of information. Using a spatio-temporal scan statistic, zoonotic case information derived from a state registry system (TDH) was compared with administrative medical claims information derived from a managed care organization (MCO) to statistically validate when and where these data sources differ. Study observations included case information for four tick-borne (Lyme disease [LD], ehrlichiosis, Rocky Mountain spotted fever [RMSF], tularemia) and 2 mosquito-borne diseases (West Nile virus [WNV], La Crosse viral encephalitis [LACV]) known to occur in Tennessee during 2000-09. A total of 103 clusters were detected indicating when/where case volume was greater than expectation. Of these, 9 were statistically significant (*P*<0.05) with 7 from TDH data. Considering only the significant clusters, there was no spatial or temporal overlapping between data sources. Findings suggest MCO data and TDH registry data each add unique important disease information. This study

further supports the need to integrate administrative and clinical registry data sources in order to provide a more comprehensive set of case information.

INTRODUCTION

Syndromic surveillance is the identification of disease indicators based on cases yet to be confirmed through, for example, laboratory results. Case information is usually defined through administrative medical data sources such as emergency room reports and hospital inpatient data. The intent of surveillance is to detect disease outbreaks quickly in order to increase response time (Mandl et al. 2003; Kuldorff et al. 2005). Until recently, little emphasis has been given to the importance in using administrative medical claims data for research in and surveillance of communicable diseases (Yiannakoulias and Svenson 2009; Chapter 2). However, because of the need for rapid outbreak detection, administrative medical data as a supplemental resource for disease surveillance is gaining more attention (Buckeridge 2005; Yiannakoulias and Svenson 2009).

Notifiable diseases are infectious diseases for which regular, frequent, and timely reporting of individual diagnosed cases aids in prevention and control. The National Notifiable Disease Surveillance System (NNDSS), with oversight from the Centers for Disease Control and Prevention (CDC), serves as the nation’s comprehensive source of data on reportable notifiable diseases (CDC NNDSS 2010). When considering using notifiable disease registry systems and administrative medical claims data in surveillance efforts, both data sources have positive and negatives attributes. Disease registry systems provide case information for health care organizations and providers, public health officials, government and regulatory agencies, and others concerned with information about potentially preventable diseases. However, significant underreporting of notifiable diseases exists even though state regulations or contractual obligations may require it (Marier 1977; Meek et al. 1996; Young 1998; Koo and Caldwell 1999; Bailey et al. 2005; Rosen 2009; Yiannakoulias and Svenson 2009; Chapter 2). Further, registry data is often presented at a comparatively more granular spatial scale (*e.g.*, county) compared to administrative medical data. Administrative medical claims data may be more comprehensive than disease registry data but at the expense of potential over-reporting due to misdiagnosis or premature diagnosing without confirmed laboratory tests, coding errors, and variability in provider practice patterns. Together though, these two data sources could provide a valuable combination of information for spatio-temporal surveillance (Yiannakoulias and Svenson 2009; Chapter 2).

Other than the work presented in Chapter 2, only one other study examining the differences between administrative data obtained from a health insurer and notifiable registry data for zoonotic case information is known (Yiannakoulias and Svenson 2009). Both studies indicate significant underreporting of cases, as well as spatial and temporal variation of information between the two data sources. Chapter 2 highlights that significant spatio-temporal differences exist at the aggregated population level. Yiannakoulias and

Svenson (2009) go on to test for spatio-temporal clustering differences to specifically highlight where and when statistical separation occurs for *Escherichia coli* O157:H7 infections derived from the two data sources.

This study compares information derived from state reported cases of zoonotic infections with administrative medical claims information derived from a large southeastern managed care organization (MCO). The analysis presented in Chapter 2 and is expanded here to compare the spatio-temporal clustering information generated from these two data sources. The intent is to statistically validate when and where these data sources differ by examining case information on four tick-borne (Lyme disease [LD], ehrlichiosis, Rocky Mountain spotted fever [RMSF], tularemia) and 2 mosquito-borne diseases (West Nile virus [WNV], La Crosse viral encephalitis [LACV]) known to occur in Tennessee.

METHODS

Study Area

The study area for this project was described in Chapter 2, but briefly, Tennessee is considered a southeastern state and is approximately bounded within the southernmost west coordinate (-90.309200, 34.995800) to the northern most east coordinate (-81.646900, 36.611900). The spatial sampling unit consists of the 95 counties within Tennessee.

Disease Case Data

Case data for the four diseases was extracted from two separate data sources and compared for spatial and temporal differences. The first data source was medically diagnosed cases extracted from a MCO claims data warehouse. Described earlier in Chapter 2, all medical claims having a primary or secondary arthropod-borne disease diagnosis code of interest (see below) were extracted for January 1, 2000-December 31, 2009. Although 3 records of babesiosis were observed in the MCO database (Chapter 2), no observations existed in TDH data and this disease was therefore excluded from analyses. Medical claims having one of the following diagnosis codes were retained for study:

Tick-Borne Diseases:

- Borreliosis - Lyme disease (LD) (ICD-9 code: 088.81)
- Ehrlichiosis - human monocytic ehrlichiosis (HME) (ICD-9 code: 082.41)
- Rickettsiosis - Rocky Mountain spotted fever (RMSF) (ICD-9 Diagnosis Code: 082.0)
- Tularemia (ICD-9 code: 021)

Mosquito-Borne Diseases:

- La Crosse viral encephalitis (LACV) (ICD-9 code: 062.5)
- West Nile virus (WNV) (ICD-9 code: 066.4)

Any patient receiving medical services for one of the selected diseases prior to the start of the study period (January 1, 2000) or after the study period (December 31, 2009) was removed from the analysis. For MCO medical claims data, space and time are represented as the county of residence for the patient at the time medical services were rendered, respectively.

The second data source was an extract provided by the Tennessee Department of Health (TDH), Center for Environmental and Communicable Diseases (CEDS) detailing all notifiable diseases reported to the state of Tennessee during the study period (TDH WebAim 2010). Because TDH serves as the compiler of all data sources to the state level, this data represents a theoretically complete set of reported cases for the state. For TDH data, space and time are represented as the resident county for the infected person and when the exposure likely occurred, respectively.

Statistical Analyses

For each data source and disease, a retrospective space-time permutation analysis was conducted to determine if significant space-time disease clusters were similar between data sources. The space-time scan statistic methodology is described in detail in Kulldorff et al. (2005). Briefly, a scan statistic is created by moving a cylindrical window over each county centroid, where the circular base represents the size of the search radius space around the centroid and the cylinder height represents a pre-defined time duration. Significant cluster detection is determined using this scan statistic by creating a relatively infinite number of overlapping cylinders to define the scanning window, each being a possible candidate for a disease cluster. Within each cylinder, the actual and expected number of disease cases, along with a Poisson generalized likelihood ratio (GLR) is calculated. Under the Poisson assumption, the generalized likelihood ratio (GLR) for any given scan window is calculated as:

$$\left(\frac{c}{E[c]}\right)^{c}\left(\frac{T-c}{T-E[c]}\right)^{T-a} I \qquad \text{Equation 2: Poisson GLR}$$

where

T = total number of cases

c = actual number of cases within the scan window

E[c] = expected number of cases within the window under the null hypothesis

I = indicator function which is equal to 1 if c > E [c] or 0 otherwise (Kulldorff 1997)

To detect clusters with high rates, *I* was set to 1 (*i.e.*, observed value should be higher than the expected value). Using Monte Carlo simulation (Dwass 1957), the actual GLR is compared to simulated GLRs within the cylinder. Relative risk (RR) for a significant cluster is calculated as the observed number of cases divided by the expected number of cases. For clusters where RR > 1, this indicates the observed number of diseases cases is greater than expectation. Statistical significance is defined in terms of a p-value, and is computed as $p=R/(S+1)$, where R is the rank of the GLR for the actual observation and S is the number of simulated cases. For example, if you simulate 999 cases, you thus obtain 999 GLR values. You then rank order these 999 GLRs from highest to lowest, where the highest GLR indicates the highest probability a cluster exists at that site. You then insert the actual GLR into this rank ordered list, and if the actual GLR is higher than the 50th highest simulated GLR, then the cluster is statistically significant at an alpha of 0.05 (*i.e.*, 50 / 999+1). Irrespective of the actual *P* value itself (*i.e.*, does not have to be below 0.05), the cluster with the highest *P* value is considered the primary cluster and all subsequent clusters in P value rank order are considered secondary. This analysis adjusts for any potential purely spatial and/or temporal variation, does not require a control comparison, and is most appropriate when information about the population-at-risk is unavailable or irrelevant (Kulldorff et al. 2005). SaTScan™ software v9.0.1 (Kulldorff 2010) was used for all cluster detection analysis. Specific software settings for these analyses included a retrospective space-time permutation probability model scanning for areas of high disease incidence, time aggregation of 1 month, a maximum spatial cluster size equal to 25% of the at-risk population, maximum temporal cluster size equal to 25% of the study period, a maximum of 999 Monte Carlo replications, and secondary clusters could not entirely overlap other previously reported clusters. Maps of significant clusters were generated using Maptitude™ v5.0 GIS software (Caliper Corporation 2008).

RESULTS

Case volume results for the study period are presented in detail elsewhere (Chapter 2). Briefly, within the MCO claims data there were 1,651 diagnosed cases distributed across the six diseases compared to 2,166 TDH registered cases. Raw count values are less important in this study because the underlying populations are different, and therefore counts are expected to vary. More importantly is the distribution of cases within a data source and the space-time clustering of these cases.

LD cases contributed to the majority of disease cases for MCO data (54.7%), while RMSF cases accounted for the majority of TDH cases (64.1%) (Figure 3-1). Across all six diseases, 103 clusters had a RR score greater than 1, indicating the observed number of diseases cases was greater than expectation. Relative risk varied significantly across the diseases, but the largest RR values are associated with smaller case volume. Of the 103 clusters, 9 were statistically significant (*P*<0.05) with 7 from TDH data and 2 from MCO data. Considering each disease separately and only examining statistically significant clusters, there was no spatial or temporal overlapping between data sources (Table 3-1). Spatial and/or temporal overlapping occurred between data sources if significance was not considered (data not presented). Both data sources produced statistically significant LD and RMSF clusters, while only TDH data produced a significant WNV cluster. No significant clusters were found for HME, tularemia, or LACV. A cluster ID was assigned to all space-time clusters in order to cross-reference information in Table 3-1 to Figure 3-2 – Figure 3-4 for temporal and spatial cluster maps. The cluster ID is a concatenation of disease, data source and cluster order. The first letter in the ID denotes the first letter of the disease (L=Lyme disease; R=Rocky Mountain spotted fever; W=West Nile virus), the second letter denotes the data source (M=MCO; T=TDH) and the number represents the cluster order based on relative risk values (1=first; 2=second, etc.). The top portion of Figure 3-2 – Figure 3-4 illustrates the spatial overlay of the clusters, while the bottom portion indicates the temporal overlay of the clusters.

DISCUSSION

This study compared the spatio-temporal information of arthropod-borne zoonotic disease cases derived from two data sources, administrative medical claims data and a state notifiable disease registry system. No attempt was made to make case-level comparisons between the data sources. Rather, analyses were aggregated to the data source scale in order to make generalizations about spatio-temporal clustering similarities and differences between these two systems. Unlike the sampling universe in Yiannakoulias and Svenson (2009) in which nearly all patients are insured by public insurance, our observed values (*i.e.*, disease cases) for TDH and MCO are not necessarily independent, nor are they necessarily non-independent. MCO cases are not necessarily a pure subset of the state numbers and vice versa. For example, a patient could have LD and the diagnosing clinician reports the case to the state. That patient may or may not be a member of the MCO. If the patient was a member of the MCO, this would be both a state case and a MCO case. If they are a non-MCO member, this is a state case, but not an MCO case. Conversely, a patient could have LD and be a MCO member but the diagnosing clinician does not report the case. This would indicate a MCO case that is not a state case.

MCO cluster LM1 and TDH clusters LT1 and LT2 were statistically significant LD clusters but did not overlap in space or time. Cluster LT2 had no radius and was therefore centered on the county centroid. The outermost northwest portion of the LM1 cluster was approximately 12 km from cluster LT2, and the clusters were separated by about 1.5 years in time. MCO cluster LM1 temporally correlates with earlier MCO findings (Chapter 2) suggesting sharp peaks in LD rates during the 2002 time period. LM1 and LT1 outer cluster limits are separated by less than 45 km. Both data sources thus provide valuable disease case information not captured in one another. No LD clusters occurred in the western portion of the state. This mostly agrees with the findings of others reporting detection of *Borrelia burgdorferi*, the causative bacterium of LD, in the middle to eastern portions of the state (Haynes et al. 2005; Shariat et al. 2007; Jordan et al. 2009). LD clusters were not necessarily confined to summer months, unlike the seasonality trends where the majority of cases occurred in the summer months (Chapter 2). We interpret this finding as the ability of the scan statistic to adjust for purely temporal abnormalities by testing thousands or even millions of overlapping space-time clusters (Kulldorff et al. 2005). That is, disease surveillance is only effective if you can detect in a timely manner when case volume is abnormal. If you were to simply compare case volume on a month by month basis, for example, your results would suggest LD outbreaks occur in May or June compared to previous months simply because this is when the disease is most prevalent (Chapter 2). Conclusions could be flawed because the temporal look-back period is not long enough, nor it is variable as it is in the scan statistic (Kulldorff et al. 2005). Thus, after adjusting for temporal case volume, significant clusters appear throughout a year from both data sources with no clear pattern. All of this suggests the eastern portion of Tennessee may be a high-risk area for LD monitoring.

RMSF is the most commonly reported tick-borne disease in Tennessee (Moncayo et al. 2010). In 2009 there were 1,393 cases reported nationwide, with 184 (13%) occurring in Tennessee (CDC MMWR 2010) making it the 3rd highest case count in the US. In a study of RMSF disease severity, Tennessee ranked 2nd only to North Carolina in the percentage of fatal RMSF cases (Adjemian et al. 2009). A significantly large RMSF cluster (RM1) was detected from MCO data in the western portion of the state, centered in Haywood County and extending out 87 km, touching 28 Tennessee counties and completely inscribing 10 counties.

The center of RM1 was located 105 km from the cluster center of 6 fatal RMSF cases reported by Adjemian et al. (2009), and was completely inscribed within its 250 km radius. Temporally for RM1, this 6-month cluster coincides with the most prevalent months (April – October) of infected cases (Chapter 2), with nearly 92% of all MCO cases occurring during this time of year. TDH cluster RT1, RT3, and RT4 are all located in the middle portion of Tennessee, suggesting this rather large area should be monitored more closely for RMSF outbreaks. TDH cluster RT2 in Monroe County is further east than the other TDH clusters, and is temporally long beginning in May 2003 and lasting 52 months (August 2005). This long duration temporally agrees with earlier findings (Chapter 2) where TDH rates began to rise dramatically in 2003 and peaked in 2006. Our findings in the west agree with the known increased risk of RMSF in western Tennessee (Adjemian et al. 2009; Moncayo et al. 2010), but go further to suggest the middle and eastern portions of the state should also be monitored for heightened RMSF infections. As with LD, statistically significant RMSF clusters for MCO and TDH did not overlap in space or time, thus providing further evidence to support data integration.

A significantly high volume of WNV cases occurred in Fall 2003. One year earlier in August 2002, a significant cluster in Dyer County was detected in TDH data (WT1), though it went undetected in MCO claims data. Our findings generally agree both temporally and spatially with others. The largest WNV epidemic ever recorded in US history occurred in 2002, with 4,156 human cases and 284 deaths. Shelby County, Tennessee is located 3 counties south of Dyer County. It has consistently reported the highest number of human WNV cases, and from 2002 through 2006, 136 human WNV cases were reported to the state of which 66% occurred in Shelby County (Ozdenerol et al. 2008). Further, high volumes of laboratory confirmed WNV infections were detected in the Tennessee Valley area for the July – September 2002 time period in nearby Paris, TN, less than 35 km from the Dyer County cluster edge (Cupp et al. 2007).

Overall, results are mixed when comparing spatial and temporal clustering between data sources. This agrees with the preponderance of evidence suggesting the need to integrate electronic administrative data with clinical registry data (*e.g.*, NAHDO; Doebbeling et al. 1999. Virnig and McBean 2001) in order to provide more comprehensive information than either single source. Disease surveillance and retrospective health care studies require monitoring of incidence rates across space and over time. Therefore, sample sizes can be limited within the space, time, or space-time dimension. For example, less than 2% of all peer-reviewed publications in the journal Ecology were from studies lasting more than 5 years (Tilman 1989). Because of limited data resources, public health officials and researchers should make full use of existing data sources, both administrative and clinical registries. Limited sample sizes can be inherit in studies covering large geographic areas simply due to logistical and cost constraints, and therefore the alternative is to reduce the study area and scope (e.g. Letcher et al. 2002). However, a small scale spatial study can introduce unwanted bias in the results because biological organisms may exhibit differential responses at different spatial scales (Wiens and Milne 1989; Zimmerman et al. 2007). Medical claims data are recorded within the healthcare system every time a patient visits their doctor or hospital for a medical service, fills a prescription medicine, or seeks consultation from a physician. Of particular interest is the amount of available data from health plans, as well as the temporal and spatial granularity of captured data elements from each medical encounter. Medical claims data contain, among other things, the patient's ZIP code at the time of service, date of medical service, and medical diagnosis codes which describe the reason why the patient is seeking medical care. The geographic element of a patient's residence location combined with the date of diagnosis provides both a spatial and temporal "stamp" of what the patient was exposed to, and potentially when and where the exposure may have occurred. Health plans may provide a centralized warehouse of rich data spanning many years, supporting more large-scale longitudinal disease studies (Roos et al. 1987; Schull et al. 2006) and surveillance activities.

Our study is not without limitations for some of the reasons outlined in Chapter 2. Statistical significance was not reached in all clusters and insignificance could simply be determined based on our parameter settings within the SaTScan™ tool (Sugumaran et al. 2009), the chosen spatial scale (Winters et al. 2008; Lloyd 2010), or limited sample size for certain diseases.

CONCLUSIONS

Findings suggest administrative claims data offer disease case information not captured in clinical registry systems and vice-versa, thus supporting the need for integrating data to provide a more comprehensive data source. Less than one-third of all US states placed contractual obligations on Medicaid contracts for MCOs to report communicable diseases (Mauery et al. 2003). Therefore, health plans themselves could engage in direct reporting of notifiable diseases because they process the medical claims containing the diagnoses information. Supplemental reporting of communicable diseases by health plans could centralize the reporting to health departments or the CDC, thereby expediting the ability to identify potential disease clusters (Mauery et al. 2003). Medical claims data may aid in the study and tracking zoonoses as it could be used to improve both the temporal and spatial scale of study through the use of long-term longitudinal data covering a large geographic expansion. Additionally, claims data could supplement the current reporting of notifiable diseases to the CDC. This effort may help bridge the disease incidence gap created by health care providers' underreporting and thus allow for more effective tracking and monitoring of infectious zoonotic diseases across time and space.

REFERENCES

Adjemian, J., J. Krebs, E. Mandel, and J. McQuiston. 2009. Spatial clustering by disease severity among reported Rocky Mountain spotted fever cases in the United States, 2001–2005. *American Journal of Tropical Medicine and Hygiene*. 80(1):72-77.

Bailey, T., M. Carvalho, T. Lapa, W. Souza, M. Brewer. 2005. Modeling of under-detection of cases in disease surveillance. Annals of Epidemiology. 15(5):335-343.

Buckeridge, D. 2005. A method for evaluating outbreak detection in public health surveillance systems that use administrative data, Ph.D. Dissertation, 2005. Available at: http://bmir.stanford.edu/file_asset/index.php/1002/BMIR-2005-1082.pdf

Caliper® Corporation. 2008. Maptitude™ GIS software v5.0 build 370.

Centers for Disease Control and Prevention (CDC MMWR). 2010. Morbidity Mortality Weekly Report (MMWR): Provisional cases of selected notifiable diseases, United States, week ending January 2, 2010. http://www.cdc.gov/mmwr/index.html accessed October 4, 2010

Centers for Disease Control and Prevention (CDC NNDSS). 2010. National Notifiable Diseases Surveillance System. Available at: (http://www.cdc.gov/ncphi/disss/nndss/nndsshis.htm

Cupp, E., H. Hassan, X. Yue, W. Oldland, B. Lilley, and T. Unnasch. 2007. West Nile Virus Infection in Mosquitoes in the Mid-South USA, 2002–2005. *Journal of Medical Entomology*. 44(1):117-125.

Doebbeling, B., D. Wyant, K. McCoy, S. Riggs, R. Woolson, D. Wagner, R. Wilson, C. Lynch. 1999. Linked insurance-tumor registry database for health services research. *Medical Care*. 37(11):1105-1115.

Dwass, M. 1957. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*. 28:181–187.

Haynes, J., P. Lee, R. L. Seipelt, and S. Wright. 2005. Detection of *Borrelia burgdorferi* sequences in a biopsy from a Tennessee Patient. *Journal of the Tennessee Academy of Science*. 80(3-4): 57-59.

Jordan, B., K. Onks, S. Hamilton, S. Hayslette, and S. Wright. 2009. Detection of *Borrelia burgdorferi* and *Borrelia lonestari* in Birds in Tennessee. *Journal of Medical Entomology*. 46(1): 131-138.

Koo, D., and B. Caldwell. 1999. The role of providers and health plans in infectious disease surveillance. *Effective Clinical Practice*. 2(5):247-252.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics-Theory and Methods*. 26(6):1481-1496.

Kulldorff, M., R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari. 2005. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*. 2(3):e59.

Kulldorff, M. 2010. Information Management Services. SaTScan™: Software for the spatial and space–time scan statistics, version 9.0.1 [computer program]. Available at: http://www.satscan.org. Accessed 14 September 2010.

Letcher, B., G. Gries, and F. Juanes. 2002. Survival of stream-dwelling Atlantic salmon: Effects of life history variation, season, and age. *Transactions of the American Fisheries Society*. 131:838–854.

Lloyd, C. 2010. Exploring population spatial concentrations in Northern Ireland by community background and other characteristics: an application of geographically weighted spatial statistics more options. *International Journal of Geographical Information Science*. 24(8):1193-1221.

Mandl, K., J. Overhage, M. Wagner, W. Lober, P. Sebastiani, F. Mostashari, J. Pavlin, P. Gesteland, T. Treadwell, E. Koski, L. Hutwagner, D. Buckeridge, R. Aller, and S. Grannis. 2003. Implementing syndromic surveillance: A practical guide informed by the early experience. *Journal of the American Medical Informatics Association*. 11(2):141-50.

Marier, R. 1977. The reporting of communicable diseases. *American Journal of Epidemiology*. 105(6):587-590.

Mauery D., B. Kamoie, S. Blake, and J. Levi. 2003. Public health communicable disease reporting laws: Managed care organizations' laboratory contracting practices and their implications for state surveillance and reporting statutes Centers for Disease Control and Prevention. Available at:

Meek, J., C. Roberts, E. Smith Jr, M. Cartter. 1996. Underreporting of Lyme disease by Connecticut physicians. *Journal of Public Health Management and Practice*. 2(4):61-65.

Moncayo, A., S. Cohen, C. Fritzen, E. Huang, M. Yabsley, J. Freye, B. Dunlap, J. Huang, D. Mead, T. Jones, J. Dunn. 2010. Absence of *Rickettsia rickettsii* and occurrence of other spotted fever group rickettsiae in ticks from Tennessee. *American Journal of Tropical Medicine and Hygiene*. 83(3):653-657.

National Association of Health Data Organizations (NAHDO). Administrative Data and Disease Surveillance: An Integration Toolkit. Available at: http://www.nahdo.org/LinkClick.aspx?fileticket=GDfdqyYO2VQ%3d&tabid=129

Ozdenerol, E., E. Bialkowska-Jelinska, G. Taff. 2008. Locating suitable habitats for West Nile Virus-infected mosquitoes through association of environmental characteristics with infected mosquito locations: a case study in Shelby County, Tennessee. *International Journal of Health* Geographics. 7(12)

Roos L. Jr, J. Nicol, and S. Cageorge. 1987. Using administrative data for longitudinal research: comparisons with primary data collection. *Journal of Chronic Diseases*. 40(1):41-49.

Rosen, M. 2009. Investigating the maintenance of the Lyme disease pathogen, *Borrelia burgdorferi*, and its Vector, *Ixodes scapularis*, in Tennessee. Master's Thesis, University of Tennessee. Available at: http://trace.tennessee.edu/utk_gradthes/554

Schull, M., T. Stukel, M. Vermeulen, and D. Alter. 2006. Study design to determine the effects of widespread restrictions on hospital utilization to control an outbreak of SARS in Toronto, Canada. *Expert Review of Pharmacoeconmics & Outcomes Research*. 6(3):285-292.

Shariat, B., J. Freimund, S. Wright, C. Murphree, and J. Thomas. 2007. Borrelia infection rates in winter ticks (*Dermacentor albipictus*) removed from white-tailed deer (*Odocoileus virginianus*) in Cheatham County, Tennessee. *Journal of the Tennessee Academy of Science*. 82(3-4): 57-61.

Sugumaran, R., S. Larson, and J. Degroote. 2009. Spatio-temporal cluster analysis of county-based human West Nile virus incidence in the continental United States. *International Journal of Health Geographics*. 8:43.

Tennessee Department of Health (TDH WebAim). 2010. Tennessee Department of Health, Communicable and Environmental Disease Services: The Communicable Disease Interactive Data Site: Available at: http://health.state.tn.us/Ceds/WebAim/WEBAim_criteria.aspx

Tilman, D. 1989. Ecological experimentation: strengths and conceptual problems. In: Lichens, G.E. (ed). Long-term studies in ecology. Springer-Verlag, New York, New York. 136-157p.

Virnig, B. and M. McBean. 2001. Administrative data for public health surveillance and planning. *Annual Review of Public Health*. 22:213–30.

Wiens, J. and B. Milne. 1989. Scaling of ‘landscapes’ in landscape ecology, or, landscape ecology from a beetle’s perspective. *Landscape Ecology*. 3:87–96.

Winters, A., R. Eisen, S. Lozano-Fuentes, C. Moore, W. Pape, and L. Eisen. 2008. Predictive spatial models for risk of West Nile virus exposure in eastern and western Colorado. *American Journal of Tropical Medicine and Hygiene*. 79(4):581-590.

Yiannakoulias, L. and L. Svenson. 2009. Differences between notifiable and administrative health information in the spatial–temporal surveillance of enteric infections. International Journal of Medical Informatics. 78(6):417-424.

Young, J. 1998. Underreporting of Lyme disease. *The New England Journal of Medicine*. 338(22):1629.

Zimmerman, G., R. Gutiérrez, and W. Lahaye. 2007. Finite study areas and vital rates: sampling effects on estimates of spotted owl survival and population trends. *Journal of Applied Ecology*. 44:963–971.

Table 3-1: Spatio-temporal county-level cluster summary of statistically significant clusters per data source

Disease	Data Source	Cluster ID*	County[†]	Cluster Radius[‡]	Time Period of Cluster	*P* value	Num. of Cases	Relative Risk (RR)**
Lyme disease	MCO	LM1	Monroe	76.0	1/02-3/04	0.006	68	1.95
	TDH	LT1	Sullivan	92.0	1/09-7/09	0.004	13	5.02
		LT2	Cumberland	0.0	7/00-8/00	0.034	4	23.84
Rocky Mountain spotted fever	MCO	RM1	Haywood	86.7	4/09-10/09	0.002	25	3.15
	TDH	RT1	Marshall	0.0	10/00-11/00	0.001	3	347.25
		RT2	Monroe	67.5	5/03-8/05	0.001	65	2.08
		RT3	Smith	74.9	5/07-8/08	0.001	119	1.66
		RT4	Hickman	52.8	12/05-11/06	0.006	57	2.04
Human Monocytic Ehrlichiosis	no significant clusters							
Tularemia	no significant clusters							
La Crosse viral enceph.	no significant clusters							
West Nile virus	TDH	WT1	Dyer	67.9	8/02-8/02	0.007	5	6.57

* Refer to Figure 3-2 - Figure 3-4 for cluster ID location and time comparison

[†] County location for center of cluster

[‡] Radius = 0 indicates cluster is centered on county centroid; Displayed as a dot on maps.

** Calculated as the number of observed cases divided by the number of expected cases

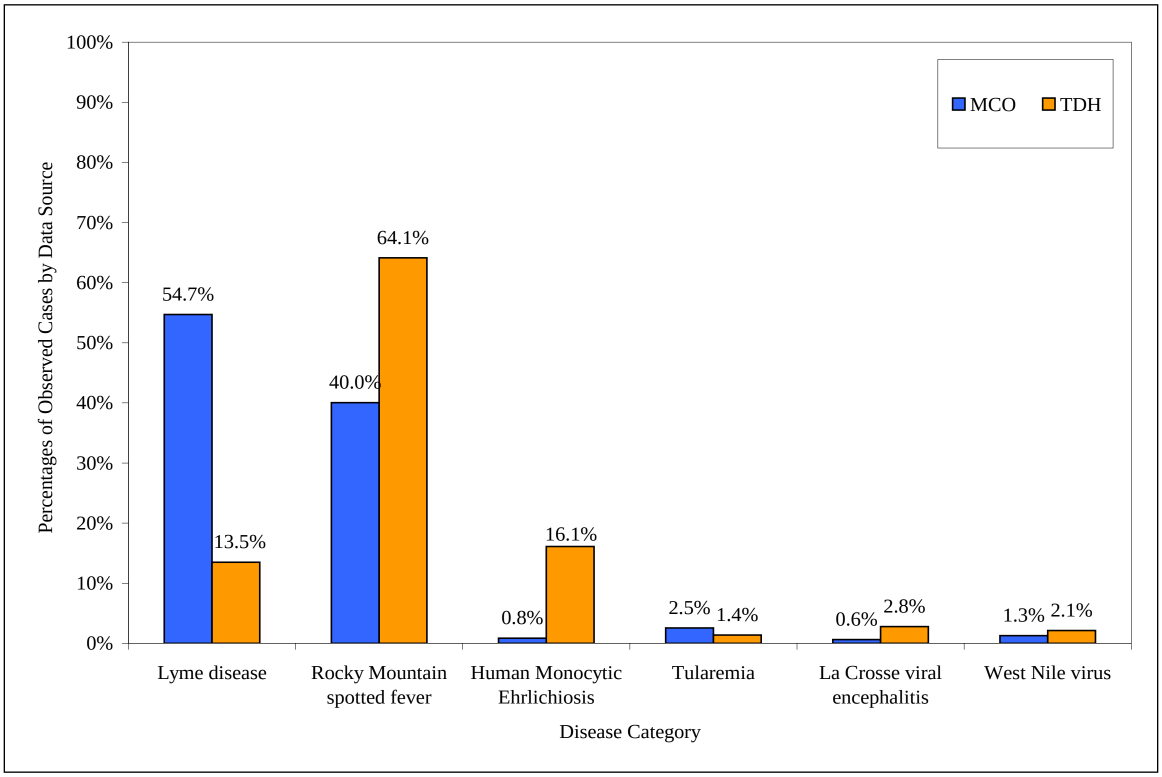


Figure 3-1: Summary of disease distribution by data sources: managed care organization (MCO) adminsitrative claims data and the Tennessee Deaprtment of Health (TDH) notifiable disease registry

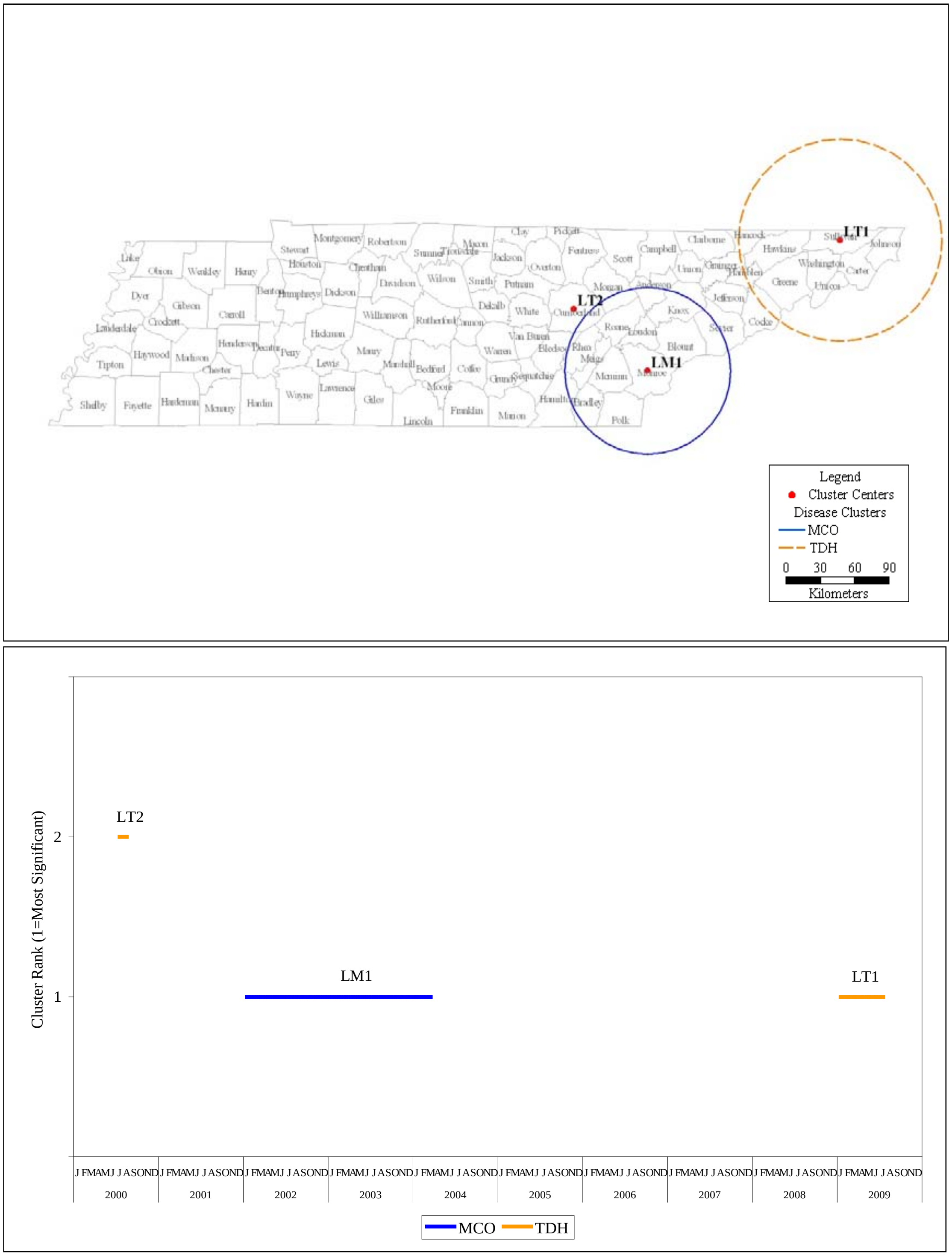


Figure 3-2: Spatial (top) and temporal (bottom) comparison of significant Lyme disease clusters created using two data sources: managed care organization (MCO) administrative claims data and the Tennessee Department of Health (TDH) notifiable disease registry

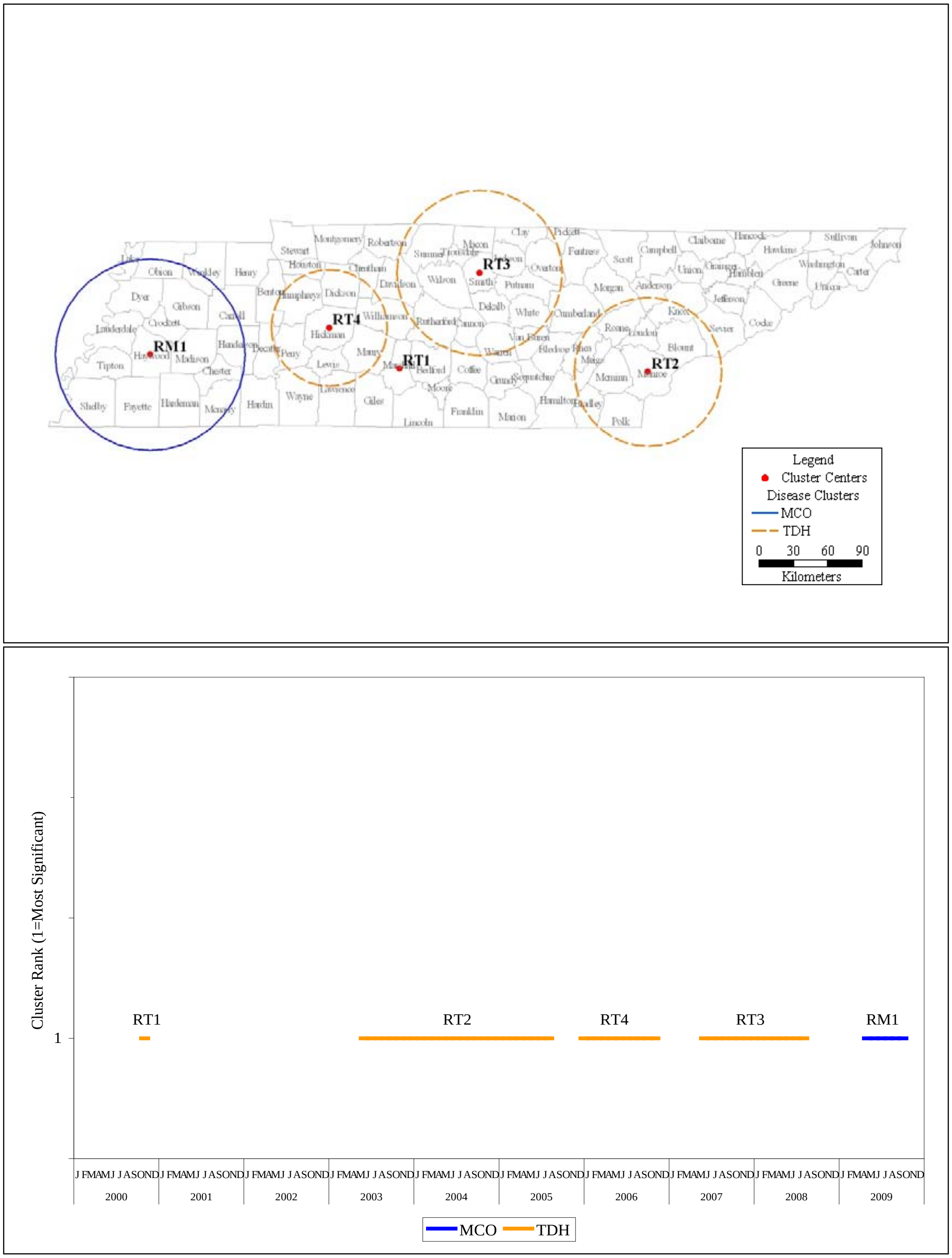


Figure 3-3: Spatial (top) and temporal (bottom) comparison of significant RMSF clusters created using two data sources: managed care organization (MCO) administrative claims data and the Tennessee Department of Health (TDH) notifiable disease registry

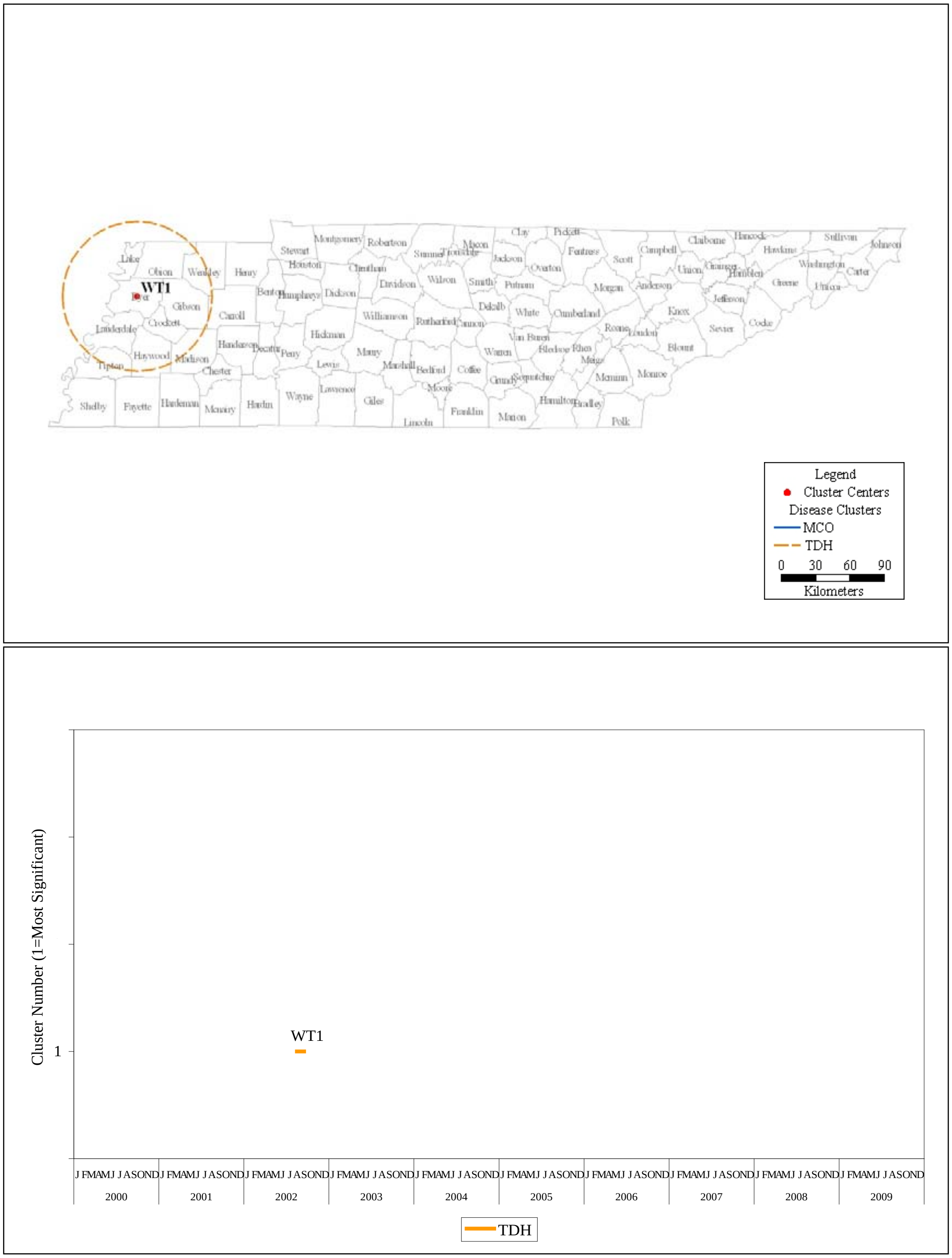


Figure 3-4: Spatial (top) and temporal (bottom) comparison of significant WNV clusters created using two data sources: managed care organization (MCO) administrative claims data and the Tennessee Department of Health (TDH) notifiable disease registry

CHAPTER 4

SPATIALLY EXPLICIT MULTI-SCALE MODELS FOR EXPLAINING THE OCCURRENCE OF INFECTIOUS ZOONOTIC DISEASES

ABSTRACT

Zoonotic diseases can be transmitted via an arthropod vector, and it is often of interest to create disease incidence risk maps based on underlying associative factors within the surrounding landscape of known occurrences. A major limitation however is the ability to track disease incidence at a meaningful geographic scale. It has been shown that administrative medical claims data is useful in the tracking of zoonotic diseases and provides disease case information at the ZIP code level. Four separate modeling techniques were compared (stepwise logistic regression, classification and regression tree, gradient boosted tree [GBT], neural network [NNET]) to describe the occurrence of 2 tick-borne diseases known to occur in Tennessee (Lyme disease [LD], Rocky Mountain spotted fever [RMSF]) as they relate to socio-demographic, geographic, and habitat characteristics. Areas higher in disease prevalence were not necessarily the same areas having high predicted risk of disease infection. Of 615 ZIP codes modeled, LD occurred in 49.9% and RMSF in 46.8%. GBT best explained LD occurrence (misclassification rate: 0.232; average squared error: 0.187; ROC: 0.789). RMSF incidence was best explained with a NNET algorithm (misclassification rate: 0.288; average square error: 0.232; ROC: 0.696). Covariates most useful in explaining LD and RMSF were similar and included co-

occurrences of RMSF and LD, respectively, amount of forested and non-forested wetlands, pasture/grasslands, and urbanized/developed lands, population counts, and median income levels of the underlying census population.

INTRODUCTION

Because zoonotic diseases are transmitted via an arthropod vector, it is often of interest to understand vector habitat in the epidemiologic study of diseases. It is common in spatial epidemiology to describe vector habitat and then create causal inference risk maps of potentially high-risk areas based on habitat preferences (Wimberly et al. 2008; Winters et al. 2008). These geospatial mapping exercises outline areas having high probabilities of vector prevalence, and then infer disease risk based on probable presence or absence. For example, abundance of the tick genus *Ixodes*, one of which is the vector primarily responsible for the transmission of Lyme disease (LD), is associated with temperature, landscape slope (Lane and Stubbs 1990), forested areas with sandy soils (Kitron et al. 1992), and increasing residential development (Aronoff 1989). Tularemia incidence is positively associated with dry forested habitat areas (Eisen et al. 2008). Populations of people living within forested areas and on specific soils are at higher risk of contracting LD (Glass et al. 1995; Killilea et al. 2008). Human monocytic ehrlichiosis (HME or Ehrlichia chaffeensis) is more associated with wooded habitats compared to neighboring grassy areas (Gaff and Schaefer 2010).

A major limitation in the study of such diseases however is the ability to comprehensively track disease incidence at a meaningful geographic scale (Killilea et al. 2008). Data aggregations and disease incidence rates are most often presented at the county level (Wimberly et al. 2008; Eisen and Eisen 2007;

Sugumaran et al. 2009). Unfortunately, county level assessments compared to ZIP code level analyses may mask smaller isolated high risk areas as well as obscure within county variability (Mostashari et al. 2003; Eisen et al. 2006). In 2007, the Centers for Disease Control and Prevention (CDC) called for a means to improve data collection methods to determine probable pathogen exposure sites based specifically on patient activity spatial patterns (Eisen and Eisen 2007). This suggests geocoding the residential location (street address or ZIP code) of the infected patient and conducting a radial search around that point to examine the underlying landscape (Wieczorek et al. 2006). However, data describing possible pathogen exposure sites are limited (Glass et al. 1995; Eisen and Eisen 2007), and means to collect this information can be very costly. Therefore, studies within the wildlife and ecological sciences are often limited in predictive power due to the inability to generate large sample sizes, either because of costs, data availability, or both (Bissonette 1999).

Administrative medical claims data contain, among other things, a patient's ZIP code at the time of service, date of medical service, and medical diagnoses describing the reason(s) why the patient is seeking medical care. The use of administrative claims data in the study of zoonotic diseases was previously discussed (Chapters 2-3). Use of this data is relatively easy and inexpensive to work with, and could represent a volume rich source of persons diagnosed with zoonotic diseases. The geographic element of a patient's residence location combined with the diagnosis provides spatially explicit information regarding what the patient was exposed to, and potentially where the exposure may have occurred. Spatially explicit disease case models created using data from managed care organizations (MCO) do not exist. It is the purpose of this study to determine if meaningful exploratory spatial models can be constructed at the ZIP code level to help describe the occurrence of 2 tick-borne zoonotic diseases known to occur in Tennessee (LD and Rocky Mountain spotted fever [RMSF]). The general research hypothesis is certain landscape and socio-demographic factors are useful in explaining zoonotic disease presence.

METHODS

Study Area

The study area for this project was described in Chapter 2, but briefly, Tennessee is considered a southeastern state and is approximately bounded within the southernmost west coordinate (-90.309200, 34.995800) to the northern most east coordinate (-81.646900, 36.611900). Estimated land cover percentages for the state are as follows: open water (2.7%), forested wetland (3.0%), non-forested wetland (0.4%), grassland/pasture (37.2%), cropland (5.8%), upland deciduous forest (40.6%), upland mixed forest (4.4%), upland coniferous forest (3.6%), urban/developed (1.9%), and non-vegetated (0.2%) (Tennessee Wildlife Resources Agency 1997).

Disease Case Data

Medically diagnosed cases of LD and RMSF from January 1, 2000-December 31, 2009 were collected from the electronic data warehouse system of a large MCO located in Tennessee. These diseases were selected because they occurred in at least 20% of the sample units (*i.e.*, ZIP codes), and therefore would not be potentially plagued by issues related to rare event modeling. The process of data collection was described in detail in Chapter 2, but briefly, zoonotic disease cases within the study area of Tennessee were extracted from MCO claims data warehouse if they had any of following diagnosis codes for LD (ICD-9 code: 088.81) and RMSF (ICD-9 Diagnosis Code: 082.0). Disease cases without at least 3 separate line items in the claims system were removed. Any patient receiving medical services for one of the selected diseases prior to the start of the study period or after the study period was removed from the analysis.

Spatial Sample Unit

This study uses two types of spatial data: 1) disease occurrence data at the ZIP code level extracted from medical claims and 2) underlying spatial data to describe the socio-demographic, geographic, and habitat characteristics surrounding the ZIP code centroid. ZIP codes can have either a geographic centroid or population-weighted centroid. A geographic centroid is defined by the US Census Bureau as the center of the tabulation area as it relates to the geographic extremes of the physical boundaries of the polygon. A population-weighted centroid is the center of the tabulation area as determined by where the majority of the population is located within the polygon. For this study, the geographic centroid was converted to a population weighted ZIP code centroid to create the spatial sample units. This weighted-average transformation was accomplished using the underlying inscribed census block population counts within the enclosing ZIP code to calculate an adjusted longitude (x_z) and latitude (y_z), following this formula:

$$x_z = \frac{\sum_{i=1}^{z} p_i x_i}{\sum_{i=1}^{z} p_i}, \quad y_z = \frac{\sum_{i=1}^{z} p_i y_i}{\sum_{i=1}^{z} p_i}$$ Equation 3: Population-weighted centroid conversion

where

x_z = transformed population-weighted x-coordinate for ZIP code z

p_i = the population of the i^{th} census block within ZIP code z

x_i = the x-coordinate value of the i^{th} census block.

Repeat for the y-coordinate.

Dependent (Response)Variable

For the purposes of this study, spatial models are considered to be exploratory models at the ZIP code level, and separate models were built for each of the 2 studied diseases. Two separate modeling exercises were conducted across the 2 diseases using different dichotomous (*i.e.*, binary) response variables. The first approach assigned a value of 1 to all ZIP codes if the disease in question was present at any time during the study period, otherwise the ZIP is assigned a value of 0. ZIP codes with a value of 1 are hereafter considered ‘case’ sites.

The second modeling approach assigned a 1 to only those ZIP codes with a z-score greater than zero. This was done to explain characteristics of ZIP codes having an observed disease case volume above expectation relative to all other ZIP codes. The observed number of cases in a ZIP code was the per 100k rate averaged over the study period, and the expected number of cases within a ZIP code was derived from the statewide incidence rate averaged over the entire study period. Thus ZIP code rates were proportional to the member population within that ZIP code. A z-score was calculated for each ZIP code using the standard formula:

$$z_i = \frac{y_i - \overline{y_j}}{\sigma_j}$$

Equation 4: Standard z-score calculation

where

z_i = z-score for ZIP code *i*

y_i = observed per 100k rate of cases in ZIP code *i* averaged over the entire study period

$\overline{y_j}$ = mean of the disease rate cases averaged across the set of *j* ZIP codes

σ_j = standard deviation of the disease rate cases across the set of *j* ZIP codes

Independent Variables

Underlying socio-demographic, geographic, and habitat characteristics of the landscape surrounding the population-weighted ZIP code centroid served as explanatory variables. Clinical variables representing the per100k rate of other zoonotic diseases (LD, RMSF, human monocytic ehrlichiosis, babesiosis, tularemia, La Crosse viral encephalitis, and West Nile virus) within the ZIP code were also included. Independent variables in the model are considered multi-level because data aggregations were done at 2 spatial scales, 1.6 km and 8 km. Socio-demographic factors included total population count and median income from the 2000 US Census Bureau estimates within 1.6 km and 8 km of the ZIP centroid. Geographic factors included continuous distance (km) to the nearest river/stream and the number of river kilometers within the 2 radial aggregation bands. Habitat characteristics included the amount (km^2) of land use type and wetland type (described below) within the 2 radial aggregation bands.

Land use data was downloaded from the Tennessee Spatial Data Server (TSDS) and is a generalized version of the detailed vegetation map that was prepared in compliance with the National Gap Analysis Program effort. The 10 land cover types were derived from classification techniques performed on Landsat Thematic Mapper imagery and included open water, forested wetland, non-forested wetland, pasture/grassland, cropland, upland deciduous forest, upland mixed forest, upland coniferous forest, urban/developed, and non-vegetated (barren land & strip mines/rock quarries/gravel pits). The strip mines/rock quarries/gravel pits class were taken from ancillary data sets and added to the classification file. The forest classes were extracted from satellite imagery and reclassified. Forest communities were interpreted from aerial videography acquired in April 1995 and correlated to the satellite imagery (Tennessee Wildlife Resources Agency 1997).

Digital wetland areal data was downloaded from the TSDS and is sourced from the National Wetlands Inventory (NWI) data base. The US Fish and Wildlife Service (USFWS) and the US Geological Survey (USGS) are the Federal agencies primarily responsible for providing geospatial information relative to the Nation's wetlands. This data layer represents the extent, approximate location and type of wetlands and deepwater habitats in the conterminous United States. These data delineate the areal extent of wetlands and surface waters as defined by Cowardin et al. (1979). Certain wetland habitats are excluded from the National mapping program because of the limitations of aerial imagery as the primary data source used to detect wetlands. This data layer was digitized from USGS topographic base maps. Alpha-numeric codes describing the type of wetland are attributed to each digitized polygon and correspond to the wetland and deepwater classifications. For example, “L1UB1Hx” indicates the delineated area as:

- L: Lacustrine (System)
- 1: Limnetic (Subsystem)
- UB: Unconsolidated Bottom (Class)
- 1: Cobble-Gravel (Subclass)
- H: Permanently Flooded (Water Regime modifier)
- X: Excavated (Modifier)

There were a total of 567 different described wetland types in the Tennessee NWI wetlands data layer. To reduce the amount of potential explanatory variables, the top 11 wetland types by area were selected (Table 4-1). This reduced set of wetland areas account for approximately 90% of the entire landscape, so little information was lost and provided a refined basis for predictive modeling.

All continuous independent variables (*i.e.*, covariates) were transformed using a quantitative binning procedure. This was done to improve model performance so as to not restrict the relationships between covariates and response to only linear interpretations. For each covariate, 4 bins were created using quantiles to generate groups by splitting the data into bins having approximately the same frequency of observations. For example, the covariate "median income" could be separated into 4 bins, where INCOME_BIN_1 has all observations with an income less than $29,000, INCOME_BIN_2 ($29-$33,000), INCOME_BIN_3 ($33-$39,000) and INCOME_BIN_4 (>$39.000). These transformed variables are then treated as ordinal dummy variables in the modeling procedures. When modeling a particular disease, geographic co-occurrence of all other diseases was included as a binary indicator (0,1 where 1 indicates another disease was also recorded in the ZIP code).

Patient level characteristics (*e.g.*, age, gender, comorbidities) were excluded from analyses because the intent of this study was to determine what geographically based risk factors could explain disease occurrence. Additionally, we aimed to produce risk factors that could be replicated in other environments without requiring known case/patient level information.

Analytical Modeling Techniques

Four separate modeling techniques were compared (stepwise logistic regression, classification decision tree, gradient boosted tree, neural network) to determine which model type performs best (*i.e.*, champion model). The modeling dataset consisted of 615 ZIP code records with 2 different binary response variables (evidence of disease, above average incidence according to z-score) and all aforementioned explanatory variables. The dataset was partitioned into two mutually exclusive data sets, a training data set, and a validation data set. The training data set was used for preliminary model fitting, and then once the model was built, the validation data set was used to fine-tune (to help prevent over-fitting) and assess the final adequacy of the model. The data partitions were created using stratified sampling (stratified by the binary response variable), and the training data set included approximately 80% (n=490) of the observations, and the validation set contained the remaining 20% (n=125).

Stepwise logistic regression (SLR) is a variable selection algorithm that begins with no candidate variables in the model, and then systematically adds effects that are significantly associated with the response variable (Efroymson 1960). Effects can be subsequently removed if it is not significantly associated with the response once another variable enters the model. This selection process continues until either 1) no other effect in the model meets the 'stay significance level' or 2) the user defined number of iterations criterion is met. The entry significance level value was set to 0.5 to ensure effects with potential were considered, while stay significance was set to a more conservative 0.05 to guard against Type I errors (concluding that a factor was influential when in fact it was not).

A classification and regression (CART) decision tree (Breiman et al. 1984) is a commonly used algorithm in data mining and machine learning techniques. Classifications are used with nominal targets, while regression trees are used with continuous targets. A tree is created by applying a series of simple interpretable rules to the data in a recursive partitioning factor using a splitting criterion. These rules are then used to classify new observations into a series of tree nodes. One of the major benefits of a decision tree is its ability to use missing data which can often be as informative as known data, unlike regression techniques which cannot process this information directly. A classification tree was created using the Pearson Chi-square p-value statistic as a splitting criterion. Maximum threshold p-values for variable consideration in the splitting criterion were set to 0.2 with a Bonferroni adjustment (to account for multiple comparisons), and the minimum number of acceptable observations for a categorical value was set at 15.

Gradient boosting within classification and regression trees (GBT) is an emerging technique in data mining algorithms that has been shown to outperform traditional decision tree approaches (De Ville 2006; Elith et al. 2008). Boosting is an adaptive method designed to improve predictive performance by combining multiple simple models into one overall “ensemble” model (Friedman 2001; Friedman 2002). Boosting is described in detail elsewhere (Friedman 2001), but briefly, this approach recursively resamples the data to generate results that form a weighted average of the resampled data set. The successive samples are adjusted to accommodate previously computed inaccuracies. This continues until a user-defined limit is reached, then each tree within the series is combined to form a single final algorithm explaining the response variable.

A neural network (NNET) is a type of model that is designed to mimic the neurophysiology of the human brain, in that it attempts to “learn” as it moves along the data and examines it. These types of models are referred to as feedforward backpropagation networks (Lapedes and Farber 1987). As with the gradient boosting technique, they are typically used when understanding the effects of the model are less important compared to model performance. That is, the output of the model cannot be readily interpreted as the aforementioned SLR and CART techniques can. In a neural network, there are three kinds of units in the modeling procedure:

1. Input units obtain the values of covariates and standardize those values;
2. Hidden units perform internal computations, providing the nonlinearity that makes neural networks powerful; and
3. Output units compute predicted values and compare those predicted values with the values of the response variable

Each unit produces a single computed value and this computed value is passed along the connections to other hidden or output units. Output units (*i.e.*, predicted values) are compared with the response variable value to compute the error function in an attempt to minimize the error. For this project, the multilayer perceptron (MLP) method which is the most common network technique was utilized. The MLP was leveraged because they are best used when prior knowledge of the relationship between inputs and targets is unknown.

Model Comparisons

All models were built using SAS® Enterprise Miner™ (SAS® 2009). A model champion was chosen using the overall misclassification rate applied to the validation dataset, which represents the percentage of all incorrectly predicted observations. In addition, the following model fit statistics were examined: receiver operator characteristic (ROC) curves, averaged squared error, sensitivity, specificity, and positive predictive values (PPV). ROC curves plot sensitivity (true positive) on the y-axis and 1-specificity (false positive) on the x-axis, which can be used to visually interpret how well models perform relative to one another. Models with a steep initial rise then level off are comparatively better than models with curves that follow the 45 degree diagonal. To provide interpretation for GBT and NNET models, the original complete data set (n=615) was scored with the predictive algorithms produced by the final GBT and NNET models. This scoring calculated a predictive probability ranging from 0-1 for each observation (*i.e.*, ZIP code), detailing the likelihood that the disease in question would be present in the ZIP code. We then applied an explanatory CART model to the data to determine which independent variables were most associated with predicted probabilities greater than 0.5 (Wall and Cunningham 2000).

RESULTS

Of the 615 ZIP codes modeled, LD occurred in 49.9% (n=307), RMSF occurred in 46.8% (n=288), and LD or RMSF occurred in 97% (n=595) of the ZIP codes. Approximately 33% (n=204) of the ZIP codes had at least one case of LD and one case of RMSF. Of the 307 ZIP codes with LD, 51 had above average incidence rates of LD (*i.e.*, z-score >0). Of the 288 ZIP codes with RMSF, 48 had above average incidence rates (*i.e.*, z-score >0). Lastly, 2% (n=12) of all ZIP codes had above average incidence rates for both RMSF and LD.

The average LD rate across all ZIP codes and the entire study period was 4.56 per 100k (SD: 9.46). The highest average LD rate (81.3 per 100k; n = 2) occurred in ZIP code 38564 within the Knoxville region of Jackson County. The highest raw count of LD cases (n = 29) occurred in ZIP code 37830 of Anderson County (Knoxville region). The average RMSF rate across all counties and the entire study period was 4.05 per 100k (SD: 9.32). The highest average RMSF rate (98.1 per 100k; n = 1) occurred in ZIP code 37140 within the Nashville region of Hickman County. The highest raw count of RMSF cases (n = 28) occurred in ZIP code 38401 of Maury County (Nashville region). Approximately 38% of the LD cases occurred in the Nashville regional area (middle of state), and only 5% occurred in the Johnson City area (northeast potion of state). Similarly, 45% of the RMSF cases occurred in the Nashville regional area and only 3% occurred in the Johnson City area (Table 4-2; Figure 4-1; Figure 4-2).

Exploratory models examining ZIP codes having at least one occurrence of LD or RMSF successfully converged across all 4 modeling procedures. For the LD models, the GBT outperformed all others with a misclassification rate of 0.232, average squared error of 0.187 and ROC value of 0.789 (Table 4-3; Figure 4-3) using misclassification rate as the champion model selection criterion. Covariates most useful in explaining LD occurrence within the GBT model were co-occurrences of RMSF, amount of forested and non-forested wetlands, upland deciduous forests and urbanized/developed lands, population counts, median income, and wetland type PUBHh (Palustrine Unconsolidated Bottom

Permanently Flooded Dike/Impounded). Occurrence of RMSF was best explained using a neural network algorithm (misclassification rate=0.288; average square error=0.232; ROC=0.696) (Table 4-3; Figure 4-4). Similar to the LD model, covariates most useful in explaining RMSF occurrence within the NNET model were co-occurrences of LD, amount of forested and non-forested wetlands, pasture/grasslands, and urbanized/developed lands, and population counts.

The algorithms from the champion models were used to score the validation data set (n=125). Areas higher in disease prevalence were not necessarily the same areas having high predicted risk of disease infection (Figure 4-5; Figure 4-6). Table 4-3 provides a comprehensive assessment of all modeling outcomes for LD and RMSF and details covariates useful in explaining the variability in disease occurrence. A ZIP code was predicted to be a "case" site if the posterior probability was greater than or equal to 0.50, and therefore all model fit statistics are based on this predicted probability threshold. The symbols denote the general direction of the data, where a "+" indicates a positive relationship between the covariate and the response (*i.e.*, as the covariate increases, the likelihood of a disease case occurring also increases), a "-" indicates a negative relationship between the covariate and the response (*i.e.*, as the covariate increases, the likelihood of a disease case occurring decreases), and a "+ / -" indicates a non-linear relationship (*i.e.*, in some ranges of the covariate, the likelihood of a disease case occurring decreases while in other ranges, likelihood of disease increases). Note that the interpretations of the signs are only generalizations for two reasons: first, not all modeling procedures can be directly interpreted, and second, raw data were transformed using the binning procedure to segment each variable into groups thus allowing for non-liner interpretations. Additionally, p-values for covariates are not reported because only the SLR procedure produces this type of interpretable statistic.

Model fit was adequate for both LD and RMSF. Figure 4-7 displays the performance of each model against the posterior probability predictions. The dotted 45^{o} line represents a perfect model fit based on the predictions from the algorithm. For example, within the posterior probability range of 0.50 – 0.60 you would expect from a perfect model that approximately 50-60% of the ZIP codes actually had a disease case. Additionally, this chart can be used to determine the optimal posterior probability that should be used as a threshold to assign a predicted classification of “case” to the ZIP code. Moving the threshold value of the prediction can thus alter the model fit statistics because model evaluation is based in part on the ability to predict a “case”.

Exploratory models using the z-score to define ZIP codes with above average incidence rates were unsuccessful across all modeling types. The models did not pick any successful covariates to explain the above average incidence rates, and therefore each algorithm simply predicted all observations to have below average incidence rates. No other results are reported for these models.

DISCUSSION

Results from this study suggest LD and RMSF incidence rates are associated with varying landscape characteristics. Disease incidence was explained reasonably well within the spatially explicit models at the ZIP code level using administrative medical claims data as a source for diagnosed cases. It is believed this is the first study that has attempted to use claims data for modeling the spatial characteristics of zoonotic diseases. This study also supports the need to collect and study disease incidence at the ZIP code level as opposed to a more coarse county level.

Three out of the four models suggested that LD incidence increased with increasing urbanization. Two different covariates reflect urbanization in this study: urbanization as a land use type and population counts. Both covariates indicated a consistently positive relationship with disease risk across the 4 models. Assuming urbanization is indicative of residential habitation, others have also suggested that residential factors were associated with increased risk of LD (Steere et al. 1977; Maupin et al. 1991). Others found LD risk to be reduced in highly developed areas (Glass et al. 1995). It is likely that land use types between studies are different and therefore produce different findings. Glass et al. (1995) specifically described highly developed areas as multiunit residential neighborhoods, and found these areas to be negatively associated with risk of LD. The urbanization variable used in our study is defined in terms of land use type, not actual physical representations of housing structures. Further, Glass et al. (1995) report an adjusted odd ratio upper confidence limit equal to 1 for this urbanization variable, which denotes the possibility that no significant association exists (*i.e.*, in statistics, an odds ratio of 1 indicates the independent variable does not have any statistical influence on the outcome varaible).

LD incidence was significantly associated with both forested and non-forested wetland areas. In a comprehensive review of literature related to LD risk, Killilea et al. (2008) found that LD was consistently associated with forested areas. A probable explanation is these land use types provide valuable habitat for host abundance (McLean et al. 1993; Ginsberg et al. 2005; Ogden et al. 2008). A crude analysis between disease incidence and forested wetland area suggest a positive correlation when forested wetlands account for up to 2.5% of the surrounding sample area. However, disease incidence declines when the amount of forested wetlands is above this amount. Similarly, a positive correlation exists between disease incidence and upland deciduous forests when this land use type accounts for up to 24% of the surrounding sample area. Above this amount and the relationship becomes negative. Glass et al. (1995) reported persons living in forested areas had elevated risk (OR: 3.7; 95% CI: 1.2 – 11.8) of LD exposure. This non-linear relationship between disease incidence within deciduous forests and non-forested wetlands may result from the complex vector-host interaction. For example, an area that is 100% forested may not be inhabited by humans and, therefore, reduces the possibility of disease transmission from vector to host. Consequently, an area that is 100% urbanized may eliminate vector habitat, thus removing all chances of a vector-host interaction. *Borrelia burgdorferi*, the causative agent of LD, may occur in urban and suburban development areas as well as in isolated park/forest preserves where deer, rodents, and birds can thrive (Magnarelli et al. 1995). Kitron et al. (1992) reported that *I. scapularis* were most abundant on sandy soils with deciduous forests.

The positive association between LD occurrence and median incomes may be more an artifact of the data source rather than an actual correlation. The data source is from persons with health insurance, both commercially insured and government subsidized programs for those who cannot afford coverage (*i.e.*, Medicaid). Relatively wealthier persons have more access to care and tend to disproportionately utilize medical services compared to lower income persons (Wilkinson and Pickett 2006; Lusardi et al. 2010).

Covariates explaining RMSF incidence were mostly similar to LD and thus similar interpretation of results are assumed. However, one notable difference was RMSF was significantly associated with the amount of pasture/grassland within all 4 models. The American Dog tick (*Dermacentor variabilis*) is the most commonly identified species responsible for transmitting the *Rickettsia rickettsii* bacterial organism that causes RMSF in humans. *D. variabilis* is considered an ixodid tick (hard-shell tick) and these are commonly found in grassland areas including pastures, old fields, clearings around homes, and brushy habitats (Liu et al. 1995; Parola and Raoult 2001).

When evaluating either LD or RMSF, the co-occurrence of the each other was significant throughout all 8 models. There are two, though possibly more, likely explanations for this relationship. As previously mentioned, significant explanatory covariates were similar for each disease. Therefore, it is plausible that suitable habitat features are overlapping for the tick vectors (Parola and Raoult 2001). Another possible reason for this interaction is both diseases have similar clinical presentations, thus cases may be misdiagnosed between the two diseases (Masters et al. 2003). In highly endemic areas within the US where awareness of RMSF is high, many patients receive an alternate diagnosis when initially seeking medical attention. Cases not laboratory confirmed are frequently not RMSF and laboratory confirmation using weak diagnostic criteria may lead to false-positives (Helmick et al. 1984). Because of the possibility of misdiagnoses, it is recommended that clinicians receive confirmatory laboratory results prior to making a definitive clinical diagnosis.

Areas higher in disease prevalence were not necessarily the same areas having high predicted risk of disease infection. This supports our original project intent to illustrate the need to build spatially explicit models. Traditional risk maps can highlight temporally static areas where case volumes are high relative to other spatial units. This approach benefits from its simplicity, however it lacks statistical validation and does not account for other influencing factors and is influenced by population.

Limitations in study include the inability to definitively confirm a diagnosed case of LD and/or RMSF as such. Land use and wetlands data do not necessarily reflect the same temporal period as the diagnosed disease case. The champion models for LD and RMSF were the GBT and NNET, respectively. Although they performed well, these modeling procedures do not produce directly interpretable results. Therefore, the ability to describe the quantitative impact of the covariates without deriving them from the SLR or CART results is limited.

CONCLUSIONS

Findings from this study suggest that administrative medical claims data is a viable source to study and map disease risk for LD and RMSF. Spatial models predicting disease risk are favorable to defining risk by mapping areas of high incidence. Spatial factors associated with medically diagnosed cases of zoonoses agree with other literature using actual CDC reported cases. Little work exists using more advanced non-linear modeling techniques like those used in this study and it is recommended to explore these options as they may provide better results than traditional regression-based approaches. Administrative medical claims data is relatively easy to access given the appropriate permissions, relatively no cost once access is granted and provides the researcher with a volume rich dataset from which to study.

REFERENCES

Aronoff, S. 1989. Geographic Information Systems: A management perspective. Ottawa, Canada: WDL Publications. 294 p.

Bissonette, J. 1999. Small sample size problems in wildlife ecology: a contingent analytical approach. *Wildlife Biology*. 5:65-71.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA. 358pp.

Cowardin, L., V. Carter, E. Golet, and E. LaRoe 1979. Classification of wetlands and deepwater habitats of the United States. US Fish and Wildlife Service FWS/OBS 79/31. 103 pp.

De Ville, B. 2006. Decision trees for business intelligence and data mining: using SAS Enterprise Miner. SAS® Publishing.

Efroymson, M. 1960. Multiple Regression Analysis, in Mathematical Methods for Digital Computers, eds. A. Ralston and H. S. Wilf, New York: Wiley, chap. 17.

Eisen, L., and R. Eisen. 2007. Need for improved methods to collect and present spatial epidemiologic data for vectorborne diseases. *Emerging Infectious Diseases*. 13(12):1816-1820.

Eisen, R., R. Lane, C. Fritz, and L. Eisen. 2006. Spatial patterns of Lyme disease risk in California based on disease incidence data and modeling of vector-tick exposure. *American Journal of Tropical Medicine and Hygiene*. 75(4):669–676.

Eisen, R., P. Mead, A. Meyer, L. Pfaff, K. Bradley, and L. Eisen. 2008. Ecoepidemiology of tularemia in the southcentral United States. *American Journal of Tropical Medicine and Hygiene*. 78(4):586–594.

Elith, J., J. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*. 77(4):802–813.

Friedman, J. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. 29(5):1189-1232.

Friedman, J. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis - Nonlinear methods and data mining*. 38(4):367-378.

Gaff, H., and E. Schaefer. 2010. Metapopulation models in tick-borne disease transmission modelling. *Advances in Experimental Medicine and Biology*. 673:51-65.

Ginsberg, H., P. Buckley, M. Balmforth, E. Zhioua, S. Mitra and F. Buckley. 2005. Reservoir competence of native north American birds for the Lyme disease spirochete, *Borrelia burgdorferi. Journal of Medical Entomology*. 42(3):445-449.

Glass, G., B. Schwartz, J. Morgan III, D. Johnson, P. Noy, and E. Israel. 1995. Environmental risk factors for Lyme disease identified with geographic information system. *American Journal of Public Health*. 85(7):944-948.

Helmick, C., K. Bernard, and L. D'Angelo. Rocky Mountain spotted fever: clinical, laboratory, and epidemiological features of 262 cases. 1984. *The Journal of Infectious Diseases*. 150(4):480–488.

Killilea, M., A. Swei, R. Lane, C. Briggs, and R. Ostfeld. 2008. Spatial dynamics of Lyme disease: A Review. *EcoHealth*. 5(2):167–195.

Kitron, U., C. Jones, J. Bouseman, J. Nelson, and D. Baumgartner. 1992. Spatial analysis of the distribution of *Ixodes dammini* (Acari*: Ixodidae*) on white-tailed deer in Ogle County, Illinois. *Journal of Medical Entomology*. 29(2):259-266.

Lane, R., and H. Stubbs. 1990. Host-seeking behavior of adult *Ixodes pacificus* (Acari: Ixodidae) as determined by flagging vegetation. *Journal of Medical Entomology*. 27(3):282-287.

Lapedes, A., and R. Farber. 1987. Nonlinear signal processing using neural networks: Prediction and system modeling. Technical report LA-UR87-2662, Los Alamos National Laboratory.

Liu, Q., G. Chen, Y. Jin, M. Te, L. Niu, S. Dong, and D. Walker. 1995. Evidence for a high prevalence of spotted fever group rickettsial infections in diverse ecologic zones of inner Mongolia. *Epidemiology and Infection*. 115(1):177-183.

Lusardi, A., D. Schneider, and P. Tufano. 2010. The Economic Crisis and Medical Care Usage. Harvard Business School. Unpublished working paper 10-079.

Magnarelli, L., A. Denicola, K. C. Stafford and J. F. Anderson. 1995. Borrelia-Burgdorferi in an urban-environment - white-tailed deer with infected ticks and antibodies. *Journal of Clinical Microbiology*. 33(3):541-544.

Masters, E., G. Olson, S. Weiner, and C. Paddock. 2003. Rocky Mountain spotted fever: a clinician's dilemma. *Archives of Internal Medicine*. 163(7):769–774.

Maupin, G., D. Fish, J. Zultowsky, E. Campos, and J. Piesman. 1991. Landscape ecology of Lyme disease in a residential area of Westchester County, New York. *American Journal of Epidemiology*. 133(11):1105-1113.

McLean, R., S. Ubico, C. Hughes, S. Engstrom and R. Johnson. 1993. Isolation and characterization of *Borrelia burgdorferi* from blood of a bird captured in the Saint-Croix river valley. *Journal of Clinical Microbiology*. 31(8):2038-2043.

Mostashari, F., M. Kulldorff, J. Hartman, J. Miller, and V. Kulasekera. 2003. Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases*. 9(6):641-646.

Ogden, N., R. Lindsay, K. Hanincova, I. Barker, M. Bigras-Poulin, D. Charron, A. Heagy, C. Francis, C. O'Callaghan, I. Schwartz, and R. Thompson. 2008. Role of migratory birds in introduction and range expansion of I*xodes scapularis* ticks and of *Borrelia burgdorferi* and *Anaplasma phagocytophilum* in Canada. *Applied and Environmental Microbiology*. 74(12):3919-3919.

Parola, P. and D. Raoult. 2001. Ticks and Tickborne Bacterial Diseases in Humans: An Emerging Infectious Threat. *Clinical Infectious Diseases*. 32(6):897-928.

SAS Institute Inc. 2009. SAS® Enterprise Miner 6.1: Single-User Installation Guide. Cary, NC: SAS Institute Inc.

Steere, A., S. Malawista, and D. Snydman. 1977. Lyme arthritis: an epidemic of oligoarticular arthritis in children and adults in three Connecticut communities. Arthritis and Rheumatism. 20(1):7-17.

Sugumaran, R., S. Larson, and J. Degroote. 2009. Spatio-temporal cluster analysis of county-based human West Nile virus incidence in the continental United States. *International Journal of Health Geographics*. 8:43.

Tennessee Wildlife Resources Agency. 1997. Tennessee Land Use/Land Cover Landsat TM imagery. Tennessee Spatial Data Service metadata files. http://www.tngis.org/frequently_accessed_data.html accessed October 5, 2010.

Wall, R., and P. Cunningham. 2000. Exploring the potential for rule extraction from ensembles of neural networks. In: Griffith J, O'Riordan C, editors. Proceedings of the 11th Irish Conference on Artificial Intelligence and Cognitive Science. Trinity College, Dublin, Computer Science Technical Report TCD-CS-2000-24. 52-68p.

Wieczorek J, Q. Guo, and R. Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*. 18:745–67.

Wilkinson, R., and K. Pickett. 2006. Income inequality and health: A review and explanation of the evidence. *Social Science & Medicine*. 62:1768–1784.

Wimberly, M., A. Baer, and M. Yabsley. 2008. Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *International Journal of Health Geographics*. 7:15.

Winters, A., R. Eisen, S. Lozano-Fuentes, C. Moore, W. Pape, and L. Eisen. 2008. Predictive spatial models for risk of West Nile virus exposure in eastern and western Colorado. *American Journal of Tropical Medicine and Hygiene*. 79(4):581-590.

Table 4-1: Top 11 wetland types by area in Tennessee and selected for study

Wetland Type	Description	Area (sq km)	Percent of Total Area	Cumulative %
L1UBHh	Lacustrine Limnetic Unconsolidated Bottom Permanently Flooded Dike/Impounded	2726.4	30.9%	30.9%
PFO1A	Palustrine Forested Broad-Leaved Deciduous Temporary Flooded	1812.1	20.5%	72.0%
R2UBH	Riverine Lower Perennial Unconsolidated Bottom Permanently Flooded	1347.9	15.3%	82.0%
PFO1C	Palustrine Forested Broad-Leaved Deciduous Permanently Flooded	1061.7	12.0%	90.8%
PUBHh	Palustrine Unconsolidated Bottom Permanently Flooded Dike/Impounded	254.2	2.9%	84.6%
PFO6F	Palustrine Forested Deciduous Semipermanently Flooded	205.3	2.3%	86.3%
PFO1F	Palustrine Forested Broad-Leaved Deciduous Semipermanently Flooded	109.6	1.2%	86.5%
PUBHx	Palustrine Unconsolidated Bottom Permanently Flooded Excavated	96.7	1.1%	87.4%
R2UB3H	Riverine Lower Perennial Unconsolidated Bottom Mud Permanently Flooded	84.9	1.0%	88.3%
PEM1A	Palustrine Emergent Persistent Temporary Flooded	79.5	0.9%	89.1%
PEM1C	Palustrine Emergent Persistent Seasonally Flooded	68.5	0.8%	89.8%

Table 4-2: Regional summary of disease distribution (Lyme disease and Rocky Mountain spotted fever) for the 2000-09 study period within Tennessee according to medically diagnosed claims data

	Lyme disease *N* (%)		Rocky Mountain spotted fever *N* (%)		Total (%)	
Nashville	343	(38%)	296	(45%)	639	(41%)
Knoxville	271	(30%)	149	(23%)	420	(27%)
Chattanooga	96	(11%)	87	(13%)	183	(12%)
Jackson	80	(9%)	80	(12%)	160	(10%)
Memphis	69	(8%)	26	(4%)	95	(6%)
Johnson City	44	(5%)	23	(3%)	67	(4%)
Totals	903		661		1,564	

Table 4-3: Model summary statistics for spatially explicit models describing the occurrence of medically diagnosed cases of Lyme disease and Rocky Mountain spotted fever for the 2000-09 study period within Tennessee

	Lyme Disease (LD)				Rocky Mountain Spotted Fever (RMSF)			
Model Type	GBT	SLR	NNET	CART	GBT	SLR	NNET	CART
Model Performance								
Misclassification Rate	0.232*	0.272	0.288	0.296	0.304	0.312	0.288*	0.296
Average Square Error	0.187	0.182	0.253	0.206	0.230	0.210	0.232	0.213
ROC	0.789	0.812	0.688	0.674	0.702	0.727	0.696	0.712
PPV	83.7%	75.0%	77.1%	85.7%	69.8%	68.5%	72.5%	70.4%
Sensitivity	66.1%	67.7%	59.7%	48.4%	62.7%	62.7%	62.7%	64.4%
Specificity	87.3%	77.8%	82.5%	92.1%	75.8%	74.2%	78.8%	75.8%
Input Variables**								
Land cover								
Forested Wetland	+ / -		-				+	+ / -
Non-Forested Wetland	+ / -						-	
Pasture/Grassland					+	+ / -	+ / -	+
Upland Deciduous Forest	+ / -	+ / -			-			
Urban/Developed	+	+	+		+		+ / -	
Wetland Type								
PUBHh	-							
Geographic								
Distance to River			+ / -					
Demographic								
Population Counts	+	+	+	+	+	+	+	+ / -
Median Income	+	+	+ / -	+				
Clinical								
Lyme Dis. Co-occurrence					+	+	+	+
RMSF Co-occurrence	+	+	+	+				

* Best model chosen using lowest misclassification rate on validation dataset

** Denotes aggregations were made at 1.6 and 8km where applicable

NOTE: Variables missing from this table indicate non-significance across all models and plus and minus signs indicate direction of relationship

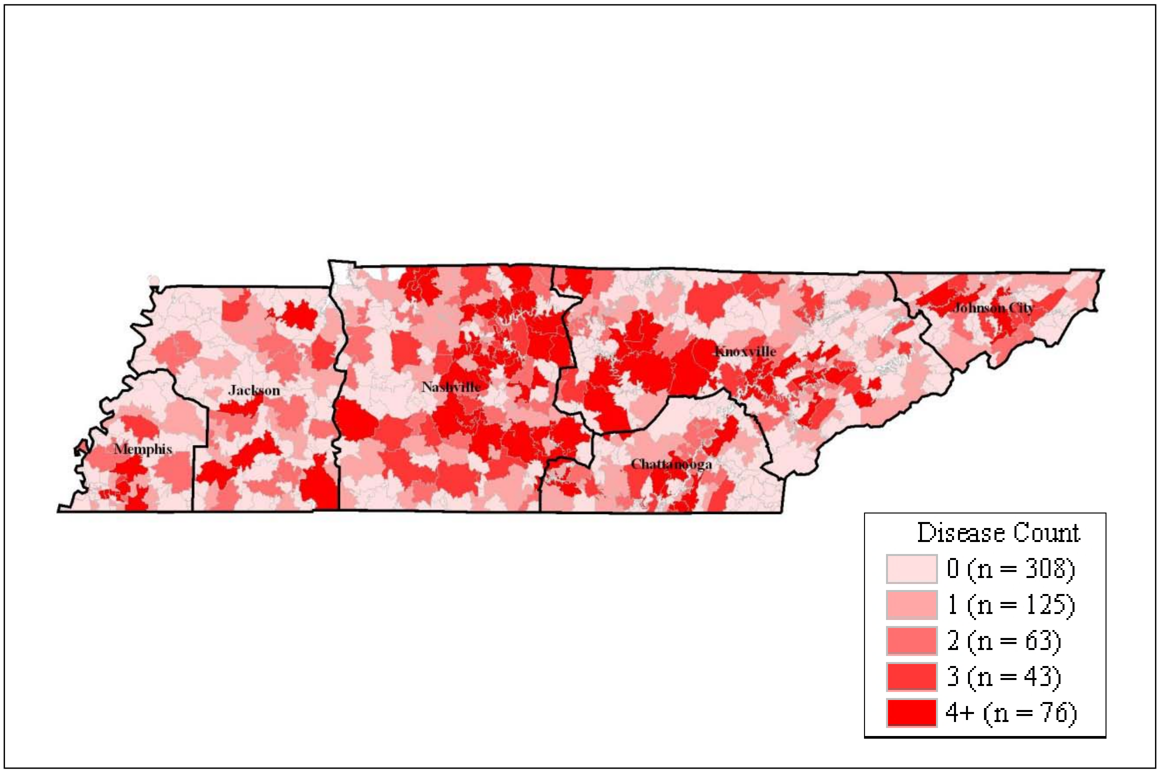


Figure 4-1: Spatial distribution of medically diagnosed Lyme disease cases (raw count) within Tennessee ZIP codes during the 2000-09 study period: Dark black outlines define regional areas

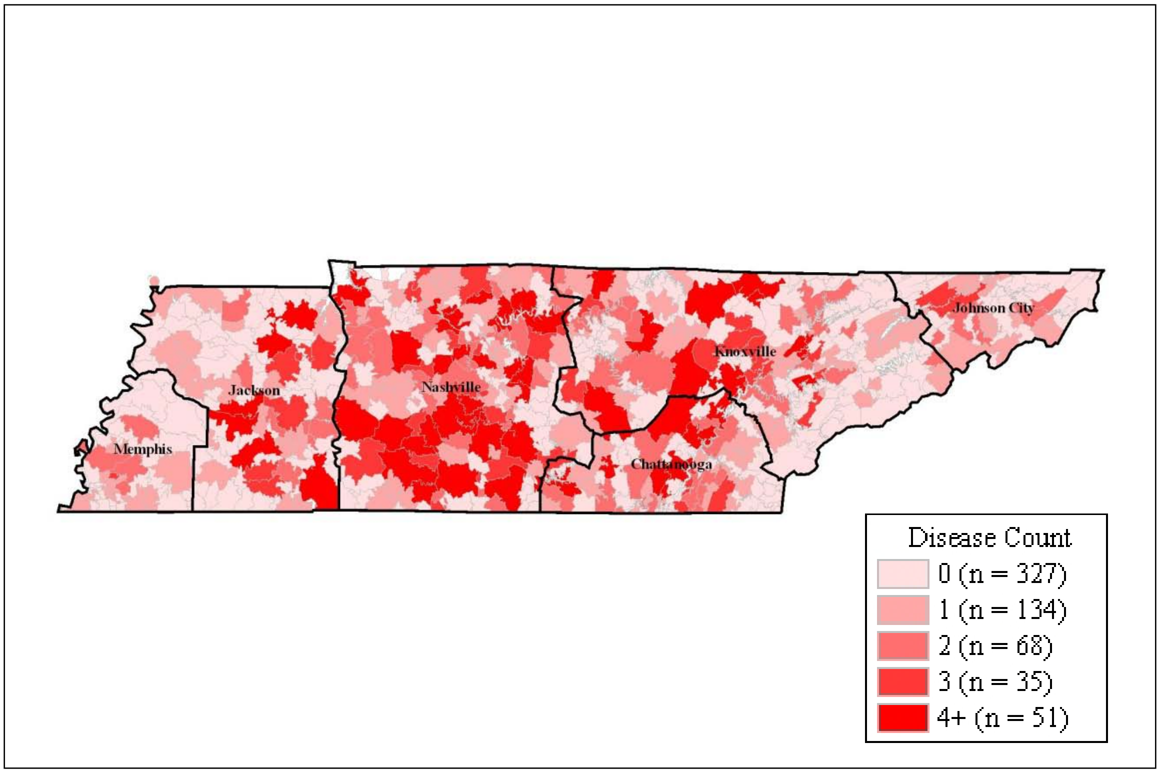


Figure 4-2: Spatial distribution of medically diagnosed Rocky Mountain spotted fever cases (raw count) within Tennessee ZIP codes during the 2000-09 study period: Dark black outlines define regional areas

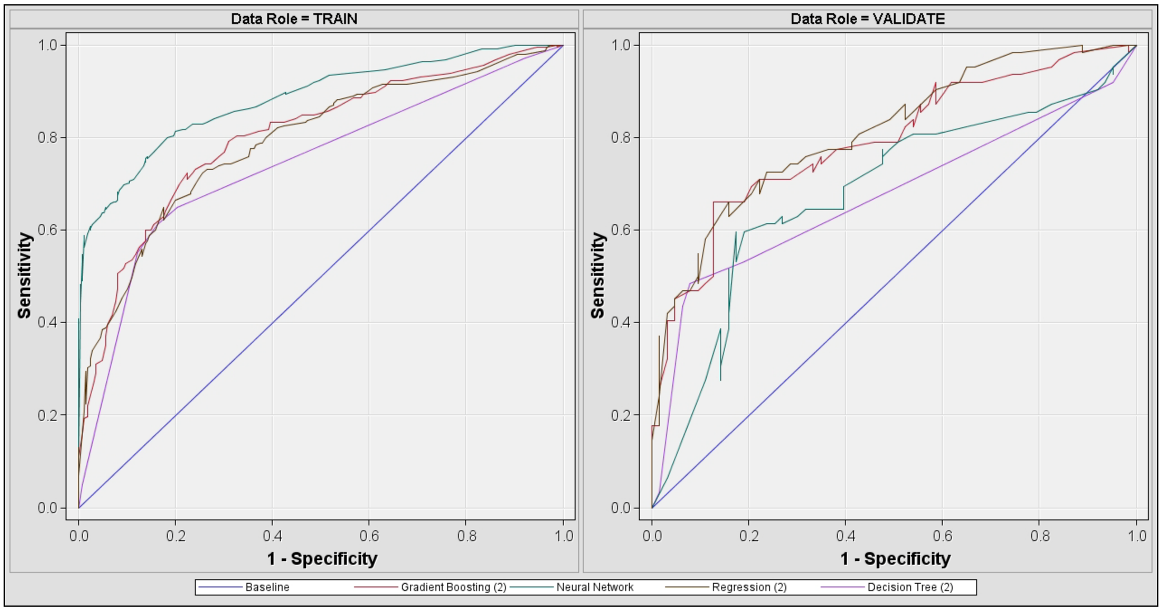


Figure 4-3: Receiver operator characteristic (ROC) curves for spatial models explaining occurrence of medically diagnosed cases of Lyme disease for the 2000-09 study period within Tennessee

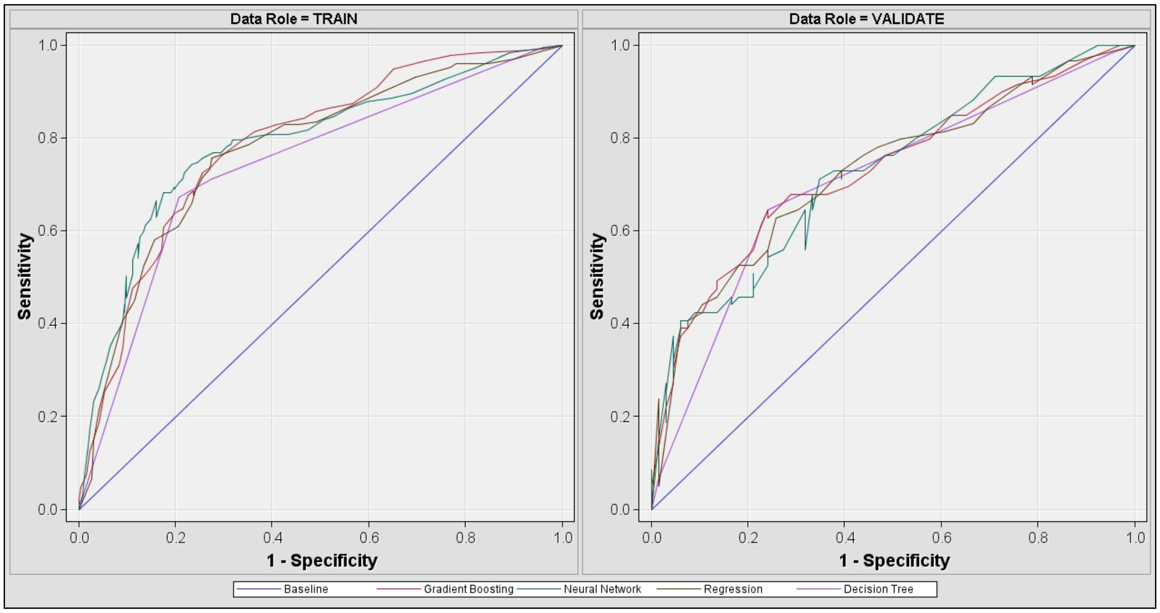


Figure 4-4: Receiver operator characteristic (ROC) curves for spatial models explaining occurrence of medically diagnosed cases of Rocky Mountain spotted fever for the 2000-09 study period within Tennessee

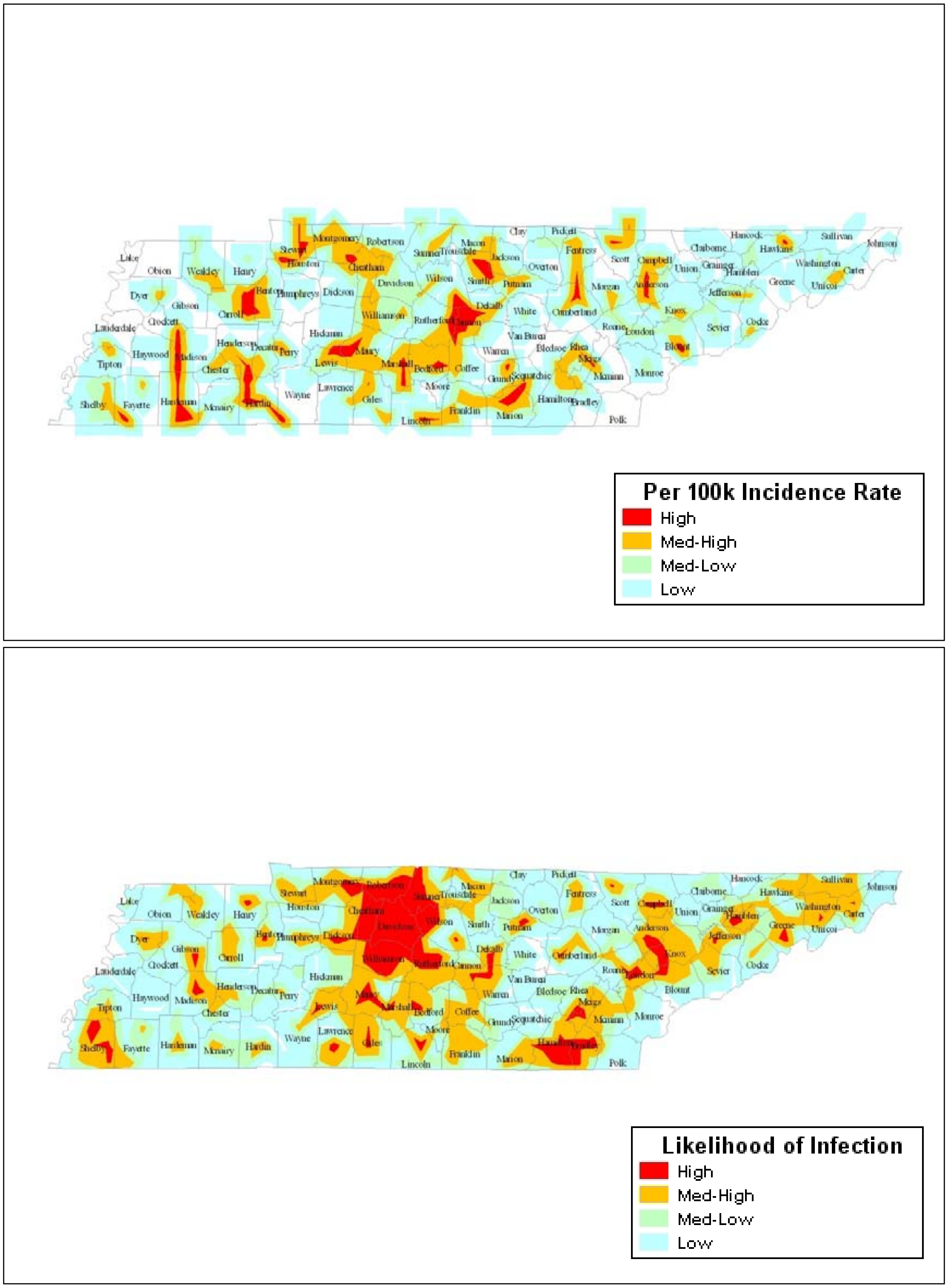


Figure 4-5: Delineated risk areas for Lyme disease according to raw disease incidence per 100k rates (top) and predicted probabilities from spatial predictive models (bottom)

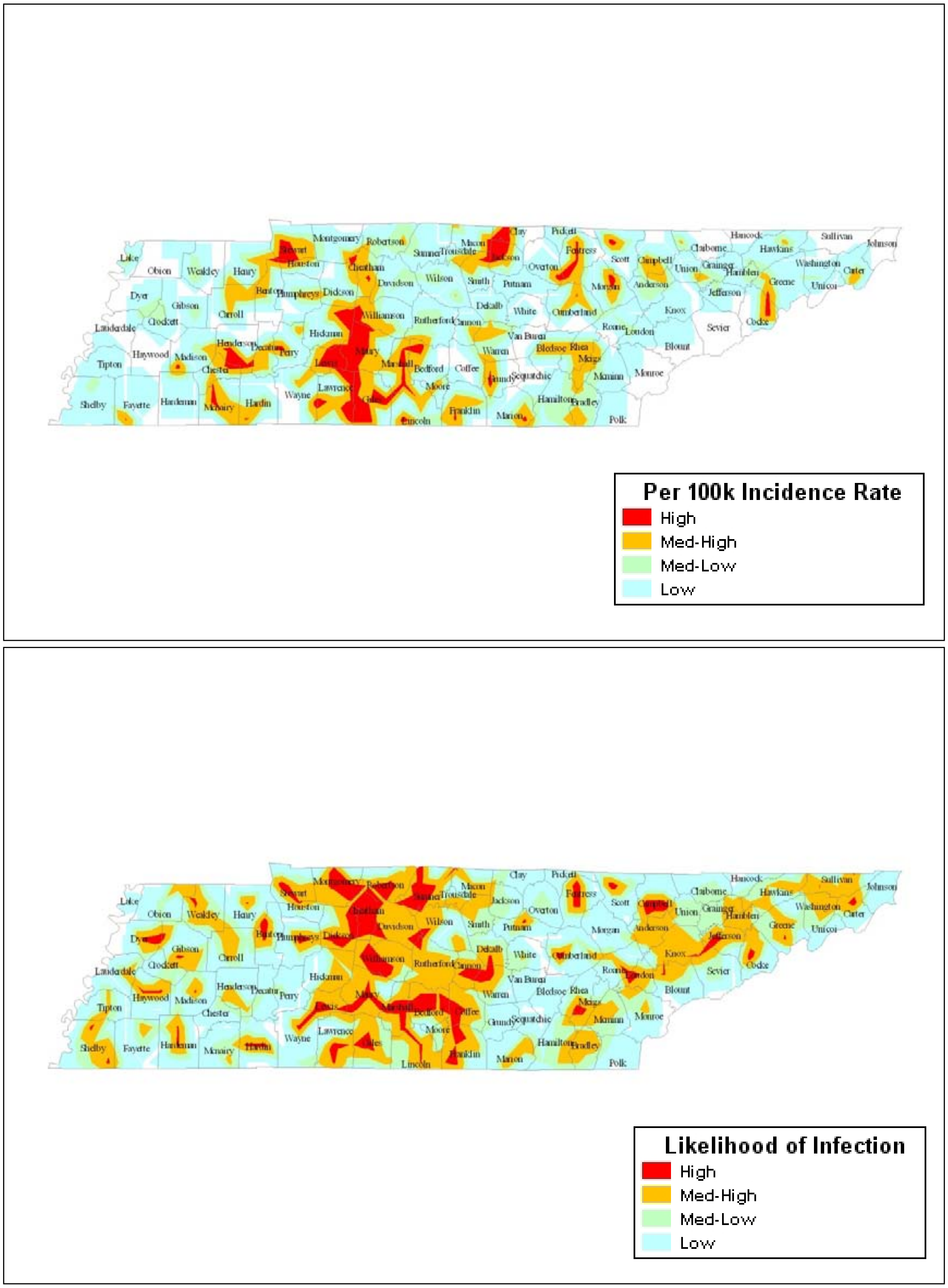


Figure 4-6: Delineated risk areas for RMSF according to disease incidence per 100k rates (top) and predicted probabilities from spatial predictive models (bottom)

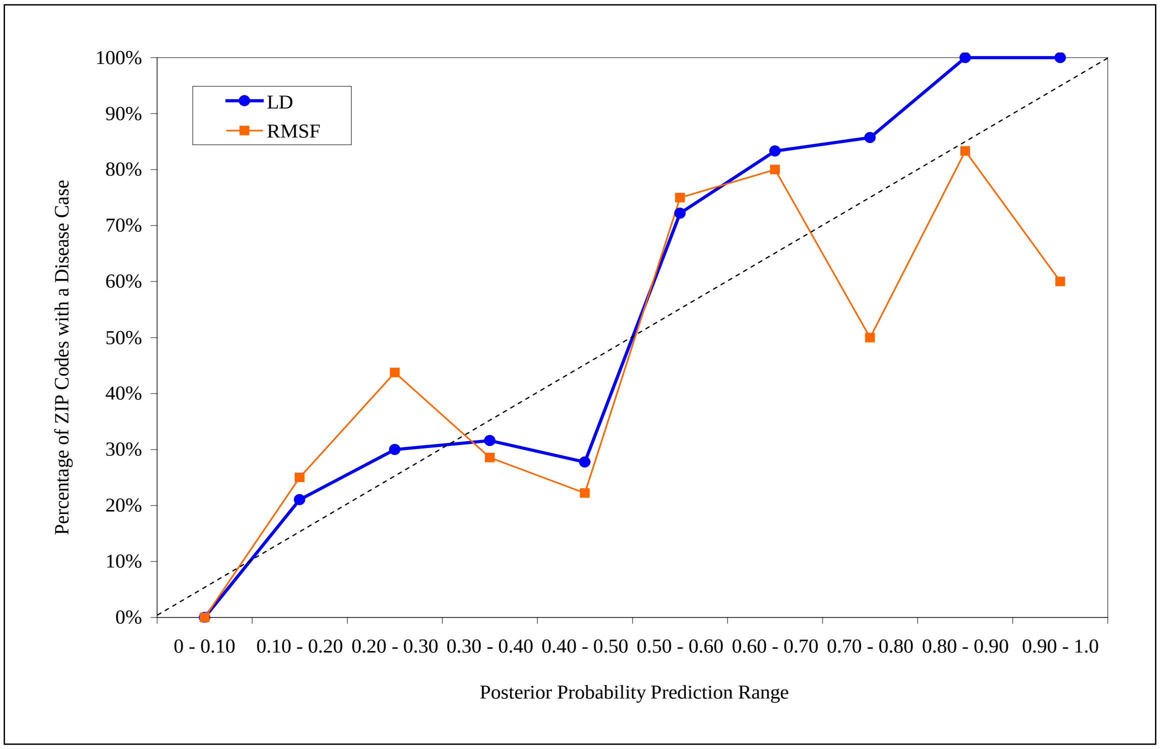


Figure 4-7: Performance of champion models as a function of the posterior probability predictions on the validation datasets

CHAPTER 5

USING A RETROSPECTIVE SPACE-TIME PERMUTATION SCAN STATISTIC FOR DETECTING CLUSTERS OF ARTHROPOD-BORNE ZOONOTIC DISEASES

ABSTRACT

Determining when and where disease prevention efforts should be targeted is a major focus in the study of zoonotic diseases. Space-time scan statistics were developed to detect statistically significant clusters of disease incidence where the observed amount is above expectation. This provides a means to study disease distribution over space and time, as well as the underlying factors influencing disease presence. The objective of this study was to determine if any significant spatial and/or temporal clusters existed for five tick-borne (Lyme disease [LD], babesiosis, ehrlichiosis, Rocky Mountain spotted fever [RMSF], tularemia) and 2 mosquito-borne diseases (West Nile virus [WNV], La Crosse viral encephalitis) known to occur in Tennessee. A cross-sectional sampling was performed for 10 consecutive years (2000-2009) across 615 population-weighted ZIP code centroids in Tennessee. Disease incidence data were extracted from administrative medical claims data from a large southeastern managed care organization. SaTScan™ software was used to detect significant clusters using a retrospective space-time permutation analysis. Overall, 1,654 unique cases were distributed across the 7 studied diseases and 3 statistically significant clusters were detected. A significant LD cluster (*P* = 0.006,

RR = 2.22, radius = 43.4 km) was detected in northeast Tennessee around ZIP code 37710. A significant RMSF cluster (*P* = 0.018, RR = 3.26, radius = 87.4 km) was detected in west Tennessee around ZIP code 38006. A significant WNV cluster was located near the RMSF cluster. Findings suggest these significant cluster areas have underlying geographic/habitat features explaining their existence, and ZIP code scale analyses may provide enhanced information compared to county-level assessments. Focused disease/vector prevention efforts in non-endemic areas are warranted.

INTRODUCTION

In ecological studies, sample units may represent an observation taken at some location (space) and/or at some temporal event (time). Determining when and where disease prevention efforts should be targeted is a major focus in the study of zoonotic diseases (*i.e.*, diseases that can be transferred from/through animals to humans). Combining epidemiologic methods to identify disease risk with geospatial analytics provide an opportunity to study disease distribution over space and time, as well as the underlying factors influencing disease presence (Glass et al. 1995; Sugumaran et al. 2009; Chapters 3 and 4). Additionally, it is important to identify significant disease clusters in order to implement appropriate public health precautions (Steere et al. 1977; Rogers and Randolph 2003; Iyengar 2005; Sugumaran et al. 2009).

Identification of significant clusters is confounded by the statistical property that any geographic region under study will always contain some high-rate area by chance alone (Kuldorff et al. 1998). To address this issue, space-time scan statistics were developed to detect non-randomly occurring clusters while accounting for multiple statistical testing (Kulldorff 1997; Iyengar 2005; Kulldorff 2010). Scan statistics, like those incorporated into the SaTScan™ software package (Kulldorff 2010), have been widely implemented in various fields of study including, but not limited to, forestry (Coulston and Riitters 2003), wildlife biology (Miller et al. 2002; Porcasi et al. 2006; Spindler et al. 2009), and infectious diseases (Chaput et al. 2002; Mostashari et al. 2003; Brooker et al. 2004).

Of the approximate 1,415 species of infectious organisms known to be pathogenic to humans, 868 (61%) are zoonotic. Of all recently emerging pathogens, 75% are zoonotic and are twice as likely to be associated with emerging diseases compared to non–zoonotic pathogens. This recent emergence of zoonoses in the US has been attributed to climate change, reforestation, increases in reservoir and vector populations, residential preferences, and increased outdoor recreational activities (Taylor et al. 2001). Currently in Tennessee, there are approximately 70 communicable diseases required to be reported to the Tennessee State Health Department for tracking purposes. However, problems exist with disease surveillance because multiple systems are implemented through multiple agencies with little cross-coordination, thus creating unnecessary duplication of efforts and inefficient use of resources. Further, wildlife diseases are rarely covered in surveillance efforts and no definition exists for what triggers a response for action (Dunn 2005).

In Tennessee, disease surveillance relies on a system where significant underreporting of diseases is known to exist and publicly reported data from the State Health Department is available only at the county level (Chapters 1 – 3). Administrative medical claims data extracted from a managed care health plan are effective in measuring zoonotic disease incidence across time and space (Chapters 2 – 4). Routinely collected administrative data is an inexpensive comprehensive source of disease information well-suited for retrospective study and disease surveillance. However, only one known study leverages claims data as a source for studying spatio-temporal zoonotic disease clustering (Yiannakoulias and Svenson 2009). The objective of this study was to determine if any significant spatial and temporal clusters existed for five tick-borne (Lyme disease [LD], babesiosis, ehrlichiosis, Rocky Mountain spotted fever [RMSF], tularemia) and 2 mosquito-borne diseases (West Nile virus, La Crosse viral encephalitis) known to occur in Tennessee. If using medical claims data is a viable approach, surveillance tracking of infectious zoonotic diseases across time and space could improve by utilizing this resource.

METHODS

Study Area

The study area for this project was described in Chapter 2, but briefly Tennessee is considered a southeastern state and is approximately bounded within the southernmost west coordinate (-90.309200, 34.995800) to the northern most east coordinate (-81.646900, 36.611900). The spatial sampling unit consisted of the 615 population-weighted ZIP code centroids within Tennessee (Chapter 4).

Disease Case Data

The collection of disease incidence data from the managed care organization (MCO) data warehouse was described earlier. Briefly, all medical claims having a primary or secondary arthropod-borne disease diagnosis code of interest (see below) were extracted for the study period January 1, 2000-December 31, 2009. Medical claims having one of the following diagnosis codes were retained for study:

Tick-Borne Diseases:

- Babesiosis (ICD-9 code: 088.82)
- Borreliosis - Lyme disease (LD) (ICD-9 code: 088.81)
- Ehrlichiosis - human monocytic ehrlichiosis (HME) (ICD-9 code: 082.41)
- Rickettsiosis - Rocky Mountain spotted fever (RMSF) (ICD-9 Diagnosis Code: 082.0)
- Tularemia (ICD-9 code: 021)

Mosquito-Borne Diseases:

- La Crosse viral encephalitis (LACV) (ICD-9 code: 062.5)
- West Nile virus (WNV) (ICD-9 code: 066.4)

Any patient receiving medical services for one of the selected diseases prior to the start of the study period or after the study period was removed from the analysis. Disease cases were aggregated to the ZIP code on the medical claim, which represents the ZIP code of residence for the patient at the time medical services were rendered. Population-weighted ZIP code centroids were geocoded and used in the cluster detection analysis as the spatial sample unit. The day, month, and year of the diagnosis were also extracted from the claims data.

Retrospective Space-Time Permutation Analysis

Previous work has shown LD and RMSF incidence rates vary geographically (Chapters 2 – 4). Traditional risk maps can highlight temporally static areas where case volumes are high relative to other spatial units (*e.g.*, Figure 5-1). Spatial kriging, a geospatial interpolation process, can smooth out these risk maps so that risk is not clearly defined by ZIP code boundaries (*e.g.*, Figure 5-2). These approaches benefit from their simplicity, however, they lack the statistical rigor (Chapter 4) and capability to simultaneously vary across time and space. To overcome this issue, a retrospective space-time permutation analysis was conducted for each selected disease to determine if any significant space-time clusters exist within Tennessee and throughout the study period. This methodology is described in detail in Kulldorff et al. (2005). Briefly, a scan statistic is created by moving a cylindrical window over each ZIP code centroid, where the circular base represents the size of the search radius space around the centroid and the cylinder height represents a pre-defined time duration. Significant cluster detection is determined using this scan statistic by creating a relatively infinite number of overlapping cylinders to define the scanning window, each being a possible candidate for a disease cluster. Within each cylinder, the actual and expected number of disease cases, along with a Poisson generalized likelihood ratio (GLR) is calculated. Under the Poisson assumption, the GLR for any given scan window is calculated as:

$$\left(\frac{c}{E[c]}\right)^{c}\left(\frac{T-c}{T-E[c]}\right)^{T-a} I$$

where T is the total number of cases, c is the actual number of cases within the scan window, E[c] is the expected number of cases within the window under the null hypothesis, and *I* is the indicator function which is equal to 1 if c > E [c] or 0 otherwise (Kulldorff 1997). To detect clusters with high rates, *I* was set to 1 (*i.e.*, observed value should be higher than the expected value). Using Monte Carlo simulation (Dwass 1957), the actual GLR is compared to simulated GLRs within the cylinder. Relative risk (RR) for a significant cluster is calculated as the observed number of cases divided by the expected number of cases. Statistical significance is defined in terms of a p-value, and is computed as $p=R/(S+1)$, where *R* is the rank of the GLR for the actual observation and *S* is the number of simulated cases. For example, if you simulate 999 cases, you thus obtain 999 GLR values. You then rank order these 999 GLRs from highest to lowest, where the highest GLR indicates the highest probability a cluster exists at that site. You then insert the actual GLR into this rank ordered list, and if the actual GLR is higher than the 50th highest simulated GLR, then the cluster is statistically significant at an alpha of 0.05 (*i.e.*, 50 / 999+1). This analysis adjusts for any potential purely spatial and/or temporal variation, does not require a control comparison, and is most appropriate when information about the population-at-risk is unavailable or irrelevant (Kulldorff et al. 2005). SaTScan™ software v9.0.1 (Kulldorff 2010) was used for all cluster detection analysis. Specific software settings for these analyses included a retrospective space-time permutation probability model scanning for areas of high disease incidence, time aggregation of 1 month, a maximum spatial cluster size equal to 25% of the at-risk population, maximum temporal cluster size equal to 25% of the study period and a maximum of 999 Monte Carlo replications. Maps of significant clusters were generated using Maptitude™ v5.0 GIS software (Caliper Corporation 2008).

RESULTS

Overall disease case results are presented earlier in Chapter 2, but briefly, 1,654 unique cases were distributed across the 7 studied diseases and used in the cluster detection analyses. The majority of disease cases were LD (n = 903; 55%), followed by RMSF (n = 661; 40%). The remaining 5 diseases made up the residual 5% of disease cases. Davidson County accounted for 9.7% (n=88) of all LD cases and 21.4% (n = 3) of HME cases. Maury County had the highest number of RMSF cases (n = 47; 7.1%) while Shelby County had the highest number of WNV (n = 8; 38.1%) cases.

A significant LD cluster (53 cases, *P* = 0.006, RR = 2.22) was detected centering in the northeastern area of Tennessee (36.164667 N, 84.236972 W;

radius = 43.4 km) around ZIP code 37710 (Anderson County), approximately 35 km northwest of Knoxville (

Figure **5-3**). This cluster encapsulated 51 ZIP code areas and was specifically associated with the time period beginning October 2001 and ending August 2003. A secondary cluster of cases, though non-significant (18 cases, *P* = 0.189, RR = 3.69), was centered around ZIP code 38483 of Lawrence County (35.430774 N, 87.323665 W; radius = 66.7 km), approximately 95 km southwest of Nashville. This cluster was specifically associated with the time period beginning August 2003 and ending May 2004. There were 25 other non-significant clusters detected throughout the state with p-values greater than 0.5 (50% of Monte Carlo replications). To reduce the amount of information, we only present the significant cluster information (Table 5-1) but graphically show all secondary clusters in the map figures for spatial reference (

Figure **5-3** - Figure 5-5).

A significant cluster of RMSF (24 cases, *P* = 0.018, RR = 3.26) was detected centering in the western side of Tennessee (35.707710 N, 89.084874 W; radius = 87.4 km) around ZIP code 38006 (Crockett County), approximately 105 km northeast of Memphis (Figure 5-4). This geographically large cluster encapsulated 108 ZIP code areas and was specifically associated with the time period beginning April 2009 and ending October 2009. A secondary cluster of cases, though non-significant at alpha = 0.05 (4 cases, *P* = 0.10, RR = 36.72), was centered around ZIP code 37082 of Cheatham County (36.088518 N, 87.122007 W; radius = 9.6 km), approximately 31 km west of Nashville. This cluster was specifically associated with the time period beginning May 2006 and ending June 2006. There were 23 other non-significant clusters detected throughout the state with p-values greater than 0.5 (50% of Monte Carlo replications) (Table 5-1; Figure 5-4).

A significant cluster of WNV (4 cases, *P* = 0.044, RR = 5.25) was detected centering in the western side of Tennessee (35.752529 N, 89.538019 W; radius = 64.5 km) around ZIP code 38063 (Lauderdale County), approximately 83 km northeast of Memphis (Figure 5-5). This cluster encapsulated 40 ZIP code areas and was specifically associated with the time period beginning July 2006 and ending September 2006 (Table 5-1).

The most likely cluster of HME was not statistically significant (2 cases, *P* = 0.57, RR = 7.00), and was centered around ZIP code 37174 of Maury County (35.728158 N, 86.910828 W; radius = 17.8 km) approximately 50 km south of Nashville. The most likely cluster of tularemia was not statistically significant (3 cases, *P* = 0.37, RR = 10.5), and was centered around ZIP code 38341 of Benton County (35.874121 N, 88.087316 W; radius = 34.5 km) approximately 190 km northeast of Memphis and 120 km southwest of Nashville. The most likely cluster of LACV was not statistically significant (2 cases, *P* = 0.930, RR = 5.00), and was centered around ZIP code 37705 of Anderson County (36.216522 N, 84.014630 W; radius = 14.1 km) approximately 20 km east of the significant

LD cluster center. Overall, 3 cases of babesiosis, were found, however no clusters were detected (Table 5-1).

DISCUSSION

Arthropod-borne zoonotic diseases are known to vary geographically and occur in significant clusters (*e.g.*, Eisen et al. 2008; Adjemian et al. 2009). Often, spatio-temporal modeling of these diseases is conducted at the county level or higher spatial scale (*e.g.*, Mostashari et al. 2003; Wimberly et al. 2008; Adjemian et al. 2009). While this scale may be appropriate for multi-state initiatives, it can mask smaller isolated high risk areas as well as obscure within county variability (Mostashari et al. 2003; Eisen et al. 2006). Analyses at a finer spatial scale like ZIP codes could improve disease surveillance activities while simultaneously protecting the identity of infected patients. Further, significant underreporting of zoonotic diseases by diagnosing clinicians exists (Marier 1977; Meek et al. 1996; Young 1998; Koo and Caldwell 1999; Figueiras et al. 2004) and could be improved using administrative medical claims data (Chapters 2 and 3). Thus the importance of the current study was our ability to successfully demonstrate spatio-temporal modeling at the ZIP code scale using medical claims data from a health plan.

LD is the most frequently reported vector-borne disease in the US. In 2009 there were 29,780 cases reported nationwide, with 32 occurring in Tennessee (CDC MMWR 2010). LD is caused by the bacterium *Borrelia* *burgdorferi,* which is transmitted to humans via the Blacklegged or deer tick (*Ixodes scapularis*), the same tick responsible for transmitting babesiosis and certain forms of ehrlichiosis. Spatial clustering of LD is common in endemic areas in the northeastern US (Steere et al. 2004; Doll 2008 *unpublished*), but is not considered endemic in Tennessee. Infected tick vectors are considered rare (ALDF 2010) and in a sample of nearly 900 blacklegged ticks, no evidence of *Borrelia burgdorferi* was found within the state of Tennessee (Rosen 2009). Previous work disputes these findings, suggesting LD incidence may be 7 times higher (3.7 vs. 0.49 per 100k) than state reported values (Chapter 2). Further, the current study suggests LD varies geographically within Tennessee and indicates the presence of a significant cluster in the northeast part of the state. There is reasonable evidence to suggest the infection occurred within/near the patient's residence (Maupin et al. 1991; Glass et al. 1995; Cromley et al. 1998; Eisen et al. 2006) and not while traveling to an endemic area. With LD on the rise nationwide (CDC NCEZID 2010), there is a need for more active surveillance in non-endemic states and improved reporting to address underreporting.

The significant LD cluster northwest of Knoxville, TN encompassed 51 ZIP codes. The Knoxville cluster center was located approximately 80km from the county-level based significant LD cluster from previous work (LM1 in Chapter 3), and was nearly one-half its size. The comparatively smaller size was expected, given that the data is at a smaller spatial scale and one would expect better granularity. However, the spatial displacement of cluster centers (80 km) was unexpected. Additionally, the time periods of the clusters were similar, but the Knoxville ZIP cluster started 2 months before and ended 7 months before the county-level cluster. This suggests that ZIP code data could not only provide enhanced spatial scale, but may produce fundamentally different results compared to larger spatial scales. We performed a post-hoc analysis to compare the ZIP codes within the Knoxville cluster to all other Tennessee ZIP codes outside the cluster. Findings from this analysis support earlier work (Chapter 4) that LD is more prevalent in urbanized areas of greater populations, as well as forested areas. Compared to non-cluster areas, ZIP codes in the Knoxville cluster had over 5 times the median amount of urbanized area within an 8 km band surrounding the centroid and median population counts were nearly 3 times higher. The median amount of upland coniferous forested area was approximately 2.5 times greater within the cluster compared to outside the cluster. Findings go on to suggest the occurrence of LACV and tularemia was 4.7 and 2.2 times higher, respectively, in the Knoxville cluster compared to non-cluster ZIP codes.

RMSF is the most severe tick-borne rickettsial illness in the US and is caused by the *Rickettsia rickettsii* bacterial organism (CDC NCEZID 2010). Infections occur most commonly in the southeastern and south central US and are typically transmitted from the bite of an infected American Dog tick (*Dermacentor variabilis*). Symptoms include the development of a rash within 2 to 4 days after the onset of fever, and can be non-descript or mimic other illnesses with headache, muscle pain, nausea, and lack of appetite. In 2009 there were 1,393 cases reported nationwide, with 184 (13%) occurring in Tennessee (CDC MMWR 2010) making it the 3rd highest case count in the US. In a study of RMSF disease severity, Tennessee ranked 2nd only to North Carolina in the percentage of fatal RMSF cases (Adjemian et al. 2009).

The significant RMSF cluster detected in western Tennessee (Crockett County) was spatially large (87.4 km radius) encompassing 108 ZIP codes but temporally small (7 months). The Crockett cluster was nearly identical to a previously detected significant RMSF cluster using county-level data (Cluster RM1 in Chapter 3). The ZIP and county-level clusters had centers located approximately 20 km apart, both were approximately equal in size and covered the same time period. The Crockett cluster center was located only 90 Euclidean kilometers from the cluster center of six fatal RMSF cases reported by Adjemian et al. (2009), and was completely inscribed within its 250 km radius. The Adjemian et al. cluster represented 26% of all fatal RMSF cases reported during their 5 year study. The eastern most edge of the Crockett cluster was only 1.5 km away from overlapping the Adjemian et al. cluster center.

Closer examination of the 108 inscribed ZIP codes support earlier findings that RMSF incidence is associated with the presence of forested wetlands (Chapter 4). ZIP codes within the Crockett cluster had over 50 times the median amount of surrounding forested wetland habitat, 5 times the amount of cropland (unfounded in earlier results), number of WNV cases were 1.8 times higher and tularemia cases 4.5 times higher than the remaining areas of the state. Contrary to earlier findings, the cluster area was less populated with lower LD rates compared to non-cluster ZIP codes. The southwest edge of the Crockett cluster narrowly misses including the Memphis population and comprises the more rural parts of Tennessee. This supports others that very complex interactions are at work and no single attribute can drive high incidence rates (Holman et al. 2001; Goddard 2008; Adjemian et al. 2009). Because our data were aggregated to the ZIP code scale rather than the county, this may better delineate the focus area Adjemian and colleagues suggest is needed for studying RMSF infections.

The West Nile virus (WNV) was first detected in the US in 1999 and became notifiable in 2002. WNV is spread to humans through the bite of an infected mosquito, typically thought to be the *Culex pipiens* mosquito, which become infected after feeding on infected birds. Though the virus quickly spread across the US from 1999 through 2001, neuroinvasive disease incidence remained low until 2002 when large outbreaks in the Midwest and Great Plains occurred. Approximately 80 percent of people infected with WNV are asymptomatic. Less than 1% of people infected will have severe life-threatening symptoms, such as high fever, neck stiffness, stupor, disorientation, coma, tremors, convulsions, muscle weakness, vision loss, numbness, and paralysis (CDC NCEZID 2010). There were 329 reported cases of non-neuroinvasive West Nile virus in 2009, 4 of which occurred in Tennessee. Additionally, there were 361 reported cases of neuroinvasive West Nile virus in 2009, 4 of which occurred in Tennessee (CDC MMWR 2010).

Similar in geographic locale to RMSF, a significant WNV cluster was also detected in west Tennessee (Lauderdale County) and was over 85% inscribed within the RMSF Crockett cluster. This was an interesting find because a nearly identical (spatial and temporal) WNV cluster was detected using county-level MCO data, but it was not statistically significant (*P* = 0.702) (Chapter 3, not reported). ZIP codes within the WNV cluster had similar attributes to the Knoxville LD and Crockett RMSF clusters. The median amount of croplands and forested and non-forested wetlands within the WNV cluster were 10 to 20 times higher inside the cluster compared to ZIP codes outside of the cluster. Additionally, urbanization was higher within the cluster. Several wetland types were more prevalent inside the cluster compared to outside, including emergent and semi-permanently flooded deciduous forested wetlands. It is well established that mosquitoes thrive in wetland habitat areas and degraded wetlands can provide ideal habitat for WNV carrying mosquitoes. Mosquito larvae feed on algal blooms created by microbial growth in nutrient rich contaminated waters. Filling or draining wetlands may not provide the necessary habitat for mosquito predators and thereby increases mosquito outbreaks. Restoring damaged or degraded wetlands could help control the spread of WNV, as healthy wetlands can sustain numerous species of mosquito-eating fish, amphibians, insects and birds (US EPA 2004). Early warning systems designed to detect an uprising in WNV can be effective and efficient means to preventing WNV (Mostashari et al. 2003; Gosselin et al. 2005).

Limitations of this study include the possibility that identified clusters are the result of some unmeasured variable that could vary geographically, such as climate, demographics, or clinician diagnostic abilities/patterns. Using administrative claims data, we cannot definitively know if the clinician diagnosed cases meet the CDC criteria for confirmed or probable. However, results support earlier findings relating disease occurrence to favorable habitat conditions. The permutation scan statistic is susceptible to changes in the underlying population over long periods of time, and significance may be biased by this population change rather than an actual disease incidence change. However, the significant clusters detected in this study were not localized to the latter part of the study period, thus it can be assumed the overall changing population was not an issue.

CONCLUSIONS

This study successfully demonstrated spatio-temporal modeling at the ZIP code scale using medical claims data from a MCO is possible, and may provide enhanced information compared to county-level assessments. Significant clusters of LD, RMSF, and WNV were detected in Tennessee during the 2000-09 study period. These significant cluster areas have underlying geographic/habitat features that help explain their existence. Further work investigating clusters while adjusting for potential confounding effects such as demographic and geographic factors is warranted. Additionally, findings suggest that focused disease/vector prevention efforts in non-endemic areas are warranted.

REFERENCES

Adjemian, J., J. Krebs, E. Mandel, and J. McQuiston. 2009. Spatial clustering by disease severity among reported Rocky Mountain spotted fever cases in the United States, 2001–2005. *American Journal of Tropical Medicine and Hygiene*. 80(1):72-77.

American Lyme disease Foundation (ALDF). 2006. Ehrlichiosis fact sheet. http://www.aldf.com/Ehrlichiosis.shtml; Accessed November 25, 2010

Brooker, S., S. Clarke, J. K. Njagi, S. Polack, B. Mugo, B. Estambale, E. Muchiri, P. Magnussen, and J. Cox. 2004. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. *Tropical Medicine and International Health*. 9(7):757-766.

Caliper® Corporation. 2008. Maptitude™ GIS software v5.0 build 370.

Centers for Disease Control and Prevention (CDC NCEZID). 2010. National Center for Emerging and Zoonotic Infectious Diseases; http://www.cdc.gov/ncezid/ accessed September 21, 2010

Centers for Disease Control and Prevention (CDC MMWR). 2010. Morbidity Mortality Weekly Report (MMWR): Provisional cases of selected notifiable diseases, United States, week ending January 2, 2010. http://www.cdc.gov/mmwr/index.html accessed October 4, 2010

Chaput, E., J. Meek, and R. Heimer. 2002. Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut. *Emerging Infectious Diseases*. 8(9):943-948.

Coulston, J. and K. Riitters. 2003. Geographic analysis of forest health indicators using spatial scan statistics. *Environmental Management*. 31(6):764-773

Cromley, E., M. Cartter, R. Mrozinski, and S. Ertel. 1998. Residential setting as a risk factor for Lyme disease in a hyperendemic region. *American Journal of Epidemiology.* 147(5):472-477.

Doll, M. 2008. Spatial analysis of Lyme disease in Howard County, Maryland. PHASE student final presentation, Maryland Department of Health and Mental Hygiene 5/16/2008. Retrieved 12/21/2010 from http://cha.maryland.gov/phase/ppt/2007/MargaretDoll.pdf.

Dunn, J. 2005. Zoonotic Diseases. Unpublished. Regional Epidemiology Meeting. November 17, 2005. Tennessee Department of Health, Centers for Disease Control and Prevention.

Dwass, M. 1957. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*. 28:181–187.

Eisen, R., R. Lane, C. Fritz, and L. Eisen. 2006. Spatial patterns of Lyme disease risk in California based on disease incidence data and modeling of vector-tick exposure. *American Journal of Tropical Medicine and Hygiene*. 75(4):669–676.

Eisen, R., P. Mead, A. Meyer, L. Pfaff, K. Bradley, and L. Eisen. 2008. Ecoepidemiology of tularemia in the southcentral United States. *American Journal of Tropical Medicine and Hygiene*. 78(4):586–594.

Figueiras, A., E. Lado, S. Fernández, and X. Hervada. 2004. Influence of physicians' attitude on under-notifying infectious diseases: a longitudinal study. *Public Health*. 118(7):521-526.

Glass, G., B. Schwartz, J. Morgan III, D. Johnson, P. Noy, and E. Israel. 1995. Environmental risk factors for Lyme disease identified with geographic information system. *American Journal of Public Health*. 85(7):944-948.

Goddard, J. 2008. Infectious Diseases and Arthropods. Humana Press; 2nd edition. 251p.

Gosselin, P., G. Lebel, S. Rivest, M. Fradet. 2005. The Integrated System for Public Health Monitoring of West Nile virus (ISPHM-WNV): a real-time GIS for surveillance and decision-making. *International Journal of Health Geographics*. 4:21.

Holman, R., C. Paddock, A. Curns, J. Krebs, J. McQuiston, and J. Childs. 2001. Analysis of risk factors for fatal Rocky Mountain spotted fever: Evidence for superiority of tetracyclines for therapy. *The Journal of Infectious Diseases*. 184(11):1437–1444.

Iyengar, V. 2005. Space-time clusters with flexible shapes. CDC Morbidity Mortality Weekly Report (MMWR). Supplement 54:71-76.

Koo, D., and B. Caldwell. 1999. The role of providers and health plans in infectious disease surveillance. *Effective Clinical Practice*. 2(5):247-252.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics-Theory and Methods*. 26(6):1481-1496.

Kulldorff, M., W. Athas, E. Feuer, B. Miller, and C. Key. 1998. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health*. 88(9):1377-1380.

Kulldorff, M., R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari. 2005. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*. 2(3):e59.

Kulldorff, M. 2010. Information Management Services. SaTScan™: Software for the spatial and space–time scan statistics, version 9.0.1 [computer program]. Available at: http://www.satscan.org. Accessed 14 September 2010.

Marier, R. 1977. The reporting of communicable diseases. *American Journal of Epidemiology*. 105(6):587-590.

Maupin, G., D. Fish, J. Zultowsky, E. Campos, and J. Piesman. 1991. Landscape ecology of Lyme disease in a residential area of Westchester County, New York. *American Journal of Epidemiology*. 133(11):1105-1113.

Meek, J., C. Roberts, E. Smith Jr, M. Cartter. 1996. Underreporting of Lyme disease by Connecticut physicians. *Journal of Public Health Management and Practice*. 2(4):61-65.

Miller, M., I. Gardner, C. Kreuder, D. Paradies, K. Worcester, D. Jessup, E. Dodd, M. Harris, J. Ames, A. Packham, and P. Conrad. 2002. Coastal freshwater runoff is a risk factor for *Toxoplasma gondii* infection of southern sea otters (*Enhydra lutris nereis*). *International Journal for Parasitology*. 32(8):997-1006.

Mostashari, F., M. Kulldorff, J. Hartman, J. Miller, and V. Kulasekera. 2003. Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases*. 9(6):641-646.

Porcasi X, Catalá SS, Hrellac H, Scavuzzo MC, Gorla DE. 2006. Infestation of rural houses by *Triatoma infestans* (Hemiptera: *Reduviidae*) in southern area of Gran Chaco in Argentina. *Journal of Medical Entomology*. 43(5):1060-1067.

Rogers, D., and S. Randolph. 2003. Studying the global distribution of infectious diseases using GIS and RS. *Nature Reviews Microbiology*. 1(3):231-237.

Rosen, M. 2009. Investigating the maintenance of the Lyme disease pathogen, *Borrelia burgdorferi*, and its Vector, *Ixodes scapularis*, in Tennessee. Master's Thesis, University of Tennessee. Available at: http://trace.tennessee.edu/utk_gradthes/554

Spindler, B., S. Chipps, R. Klumb, and M. Wimberly. 2009. Spatial analysis of pallid sturgeon *Scaphirhynchus albus* distribution in the Missouri River, South Dakota. *Journal of Applied Ichthyology*. 25:8-13.

Steere, A., S. Malawista, and D. Snydman. 1977. Lyme arthritis: an epidemic of oligoarticular arthritis in children and adults in three Connecticut communities. Arthritis and Rheumatism. 20(1):7-17.

Steere, A., J. Coburn, and L. Glickstein. 2004. The emergence of Lyme disease. *The Journal of Clinical Investigation*.113(8):1093-1101.

Sugumaran, R., S. Larson, and J. Degroote. 2009. Spatio-temporal cluster analysis of county-based human West Nile virus incidence in the continental United States. *International Journal of Health Geographics*. 8:43.

Taylor, L., S. Latham, and M. Woolhouse. 2001. Risk factors for human disease emergence*. Philosophical Transactions of the Royal Society of London*. 356(1411):983-989.

US Environmental Protection Agency (US EPA). 2004. Wetlands and West Nile virus: Fact sheet. EPA 843-F-04-010 Office of Water August 2004 edition. Available at http://www.epa.gov/owow/wetlands/pdf/WestNile.pdf

Wimberly, M., A. Baer, and M. Yabsley. 2008. Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *International Journal of Health Geographics*. 7:15.

Yiannakoulias, L. and L. Svenson. 2009. Differences between notifiable and administrative health information in the spatial–temporal surveillance of enteric infections. *International Journal of Medical Informatics*. 78(6):417-424.

Young, J. 1998. Underreporting of Lyme disease. *The New England Journal of Medicine*. 338(22):1629.

Table 5-1: Spatio-temporal cluster analyses output statistics for statistically significant clusters of arthropod-borne zoonotic diseases

	ZIP Code Center	Latitude	Longitude	Radius (km)	Time Period of Cluster	*P* value*	Num. of Cases	Relative Risk (RR)†
Lyme disease	37710	36.1647	-84.2370	43.4	10/01-8/03	0.006	53	2.22
Rocky Mountain spotted fever	38006	35.7077	-89.0849	87.4	4/09-10/09	0.018	24	3.26
Human monocytic ehrlichiosis				no significant clusters				
Tularemia				no significant clusters				
La Crosse viral encephalitis				no significant clusters				
West Nile virus	38063	35.7525	-89.5380	64.5	7/06-9/06	0.044	4	5.25

* *P* value derived from 999 Monte Carlo simulations

† Relative risk (RR) calculated as the number of observed cases divided by the number of expected cases. RR>1 indicates case rates above expectation

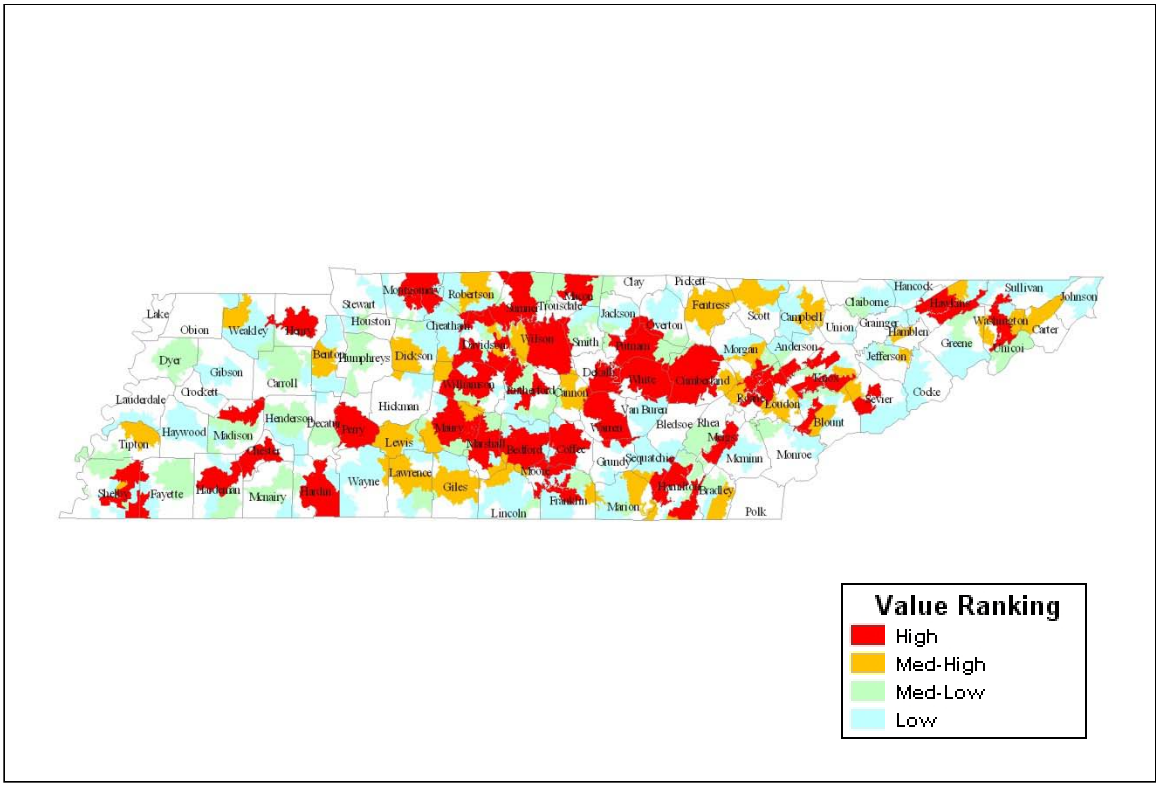


Figure 5-1: Lyme disease risk map created by simple aggregation of raw counts to ZIP codes for the 2000-09 study period within Tennessee

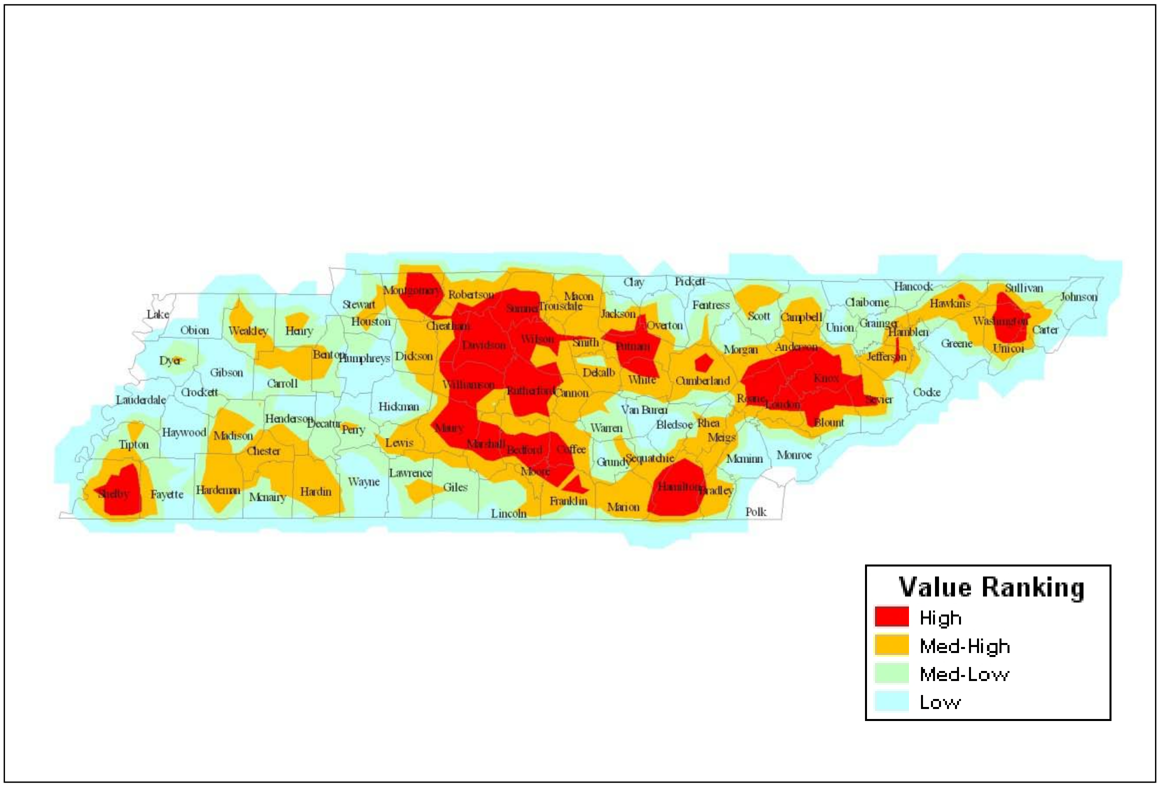


Figure 5-2: Lyme disease risk map created by spatial kriging (geospatial interpolation method) of raw counts to ZIP codes for the 2000-09 study period within Tennessee

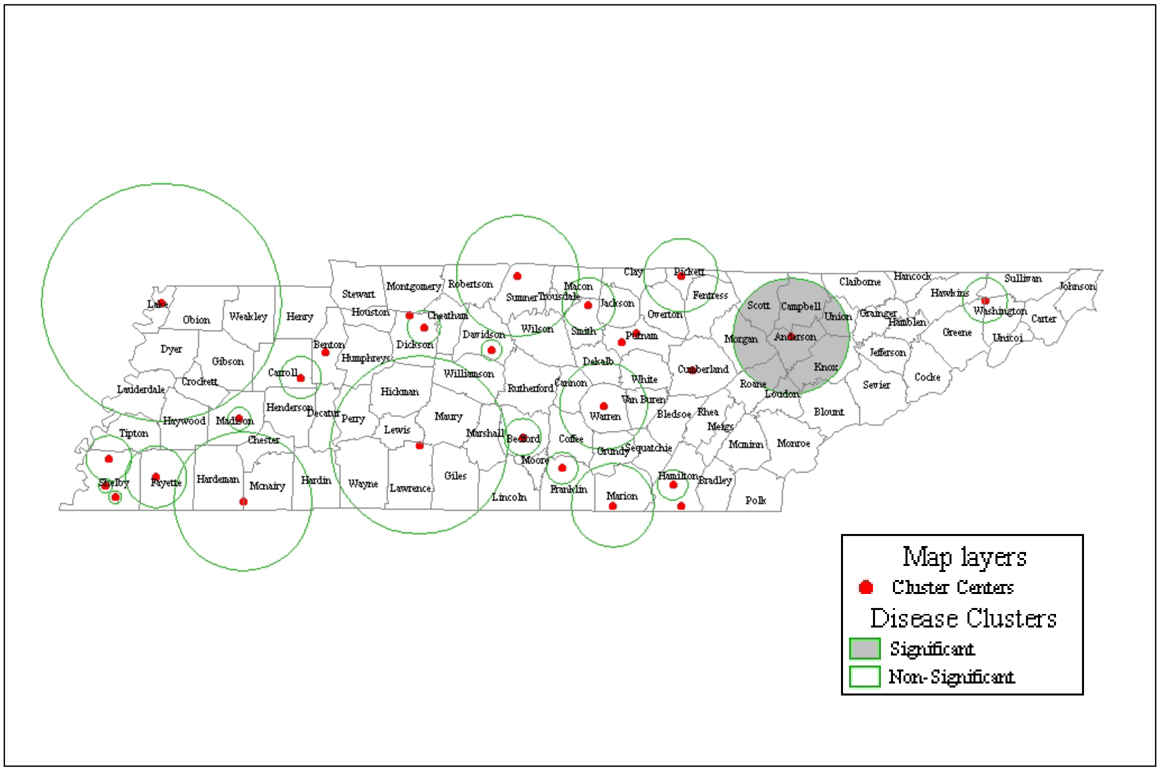


Figure 5-3: Location and radius of clusters of increased rates of medically diagnosed Lyme disease cases identified in Tennessee for the 2000-09 study period.
NOTE: Statistically significant clusters are shaded in grey.

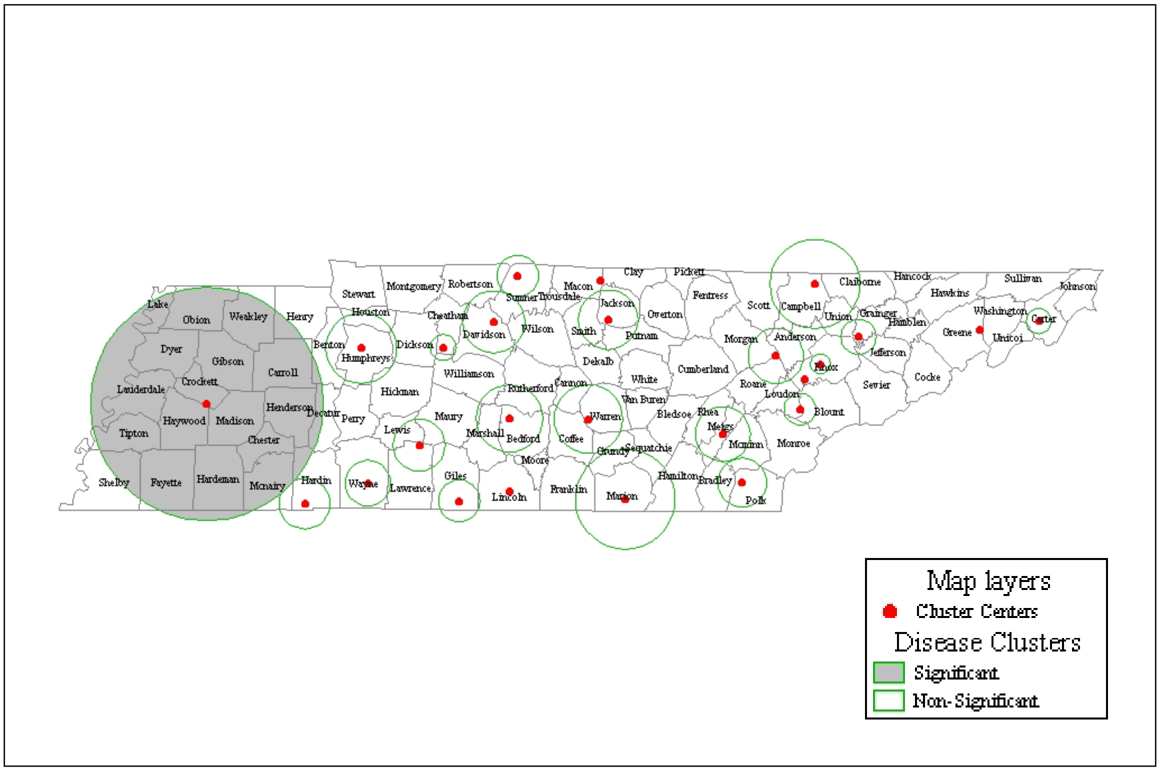


Figure 5-4: Location and radius of clusters of increased rates of medically diagnosed Rocky Mountain spotted fever (RMSF) cases identified in Tennessee for the 2000-09 study period.

NOTE: Statistically significant clusters are shaded in grey.

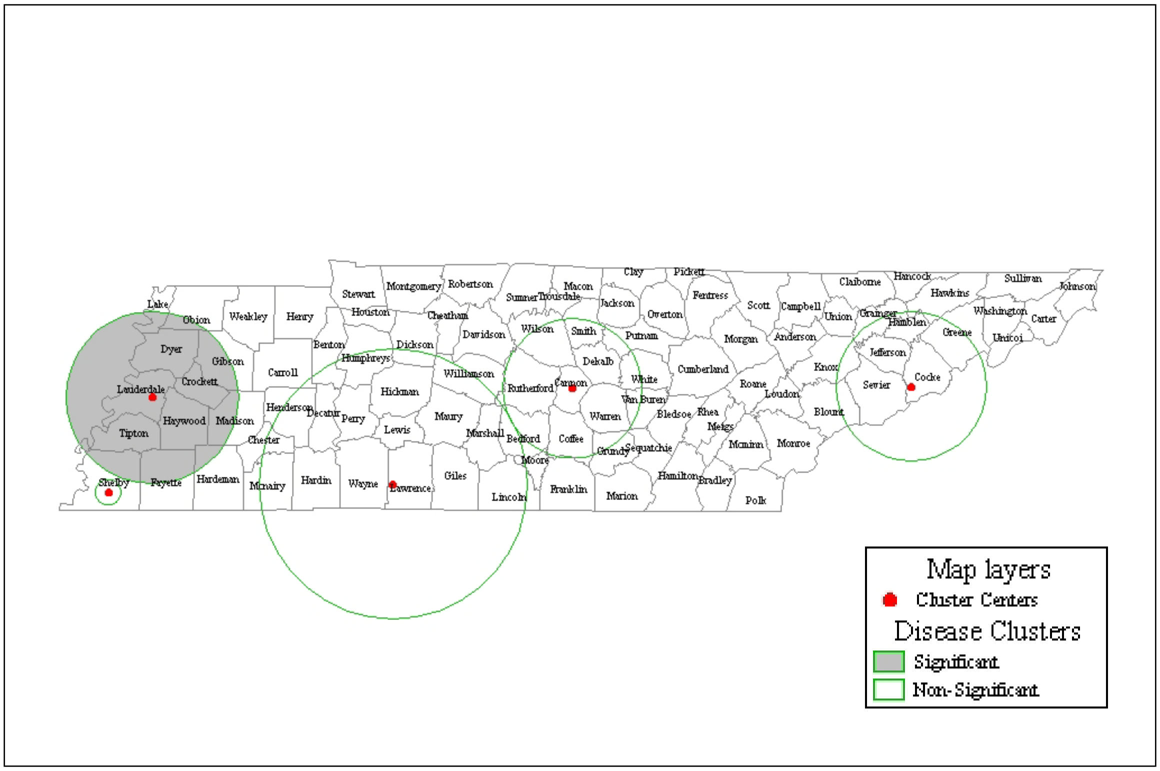


Figure 5-5: Location and radius of clusters of increased rates of medically diagnosed West Nile virus (WNV) cases identified in Tennessee for the 2000-09 study period.

NOTE: Statistically significant clusters are shaded in grey