RESEARCH PLAN

1. Specific Aims. Stroke mortality is increased by as much as 40% in the southeastern United States relative to the rest of the country, and this "Stroke Belt" has been recognized for more than 60 years. There is no consensus concerning the cause(s) of the Stroke Belt. Hypertension is the risk factor with the largest population attributable risk for stroke and shows potential clustering in the southeastern US based upon state-wide maps. Conclusive evidence for clustering of hypertension in this region has been unavailable, because the lack of directly measured blood pressure assessments in a nationwide cohort with substantial geographic heterogeneity has prevented the use of spatial statistics to properly address such a question.

The REasons for Geographic and Racial Differences in Stroke (REGARDS) study recruited over 30,000 participants with directly measured hypertension data, which has often been limited to self-report in other nation-wide studies. The lack of directly measured hypertension data has limited the use of spatial methods to: (1) formally test for clustering of hypertension using individual-level data; or (2) create a statistically robust high-resolution map of hypertension risk. By leveraging the geographic heterogeneity of REGARDS participants' residential addresses and their coordinates, we can achieve both of these goals. We are hindered because REGARDS oversampled the Stroke Belt, which has plausibly higher rates of hypertension (called 'preferential sampling'). Preferential sampling can bias estimates of disease rates for the whole population because of the disproportionately large number of cases in the sample compared to using simple random sampling. The overall goal of this project is to overcome this complication and formally evaluate and map hypertension risk for the continental US. We will pursue this overall goal with the following specific aims:

Aim 1. Develop novel extensions for tests of disease clustering that perform well in the presence of preferential sampling. *Hypothesis: Preferential sampling will cause tests for disease clustering to have incorrect power levels, and thus the tests will need modification.*
Aim 2. Create a high-resolution map of hypertension risk for the continental US. *Hypothesis: A Bayesian hierarchical model will be effective for creating the map while handling the large dataset, the large area over which we are mapping hypertension risk, the heterogeneous data density, and the preferential sampling.*

The expected outcomes from this project are: (1) improved methods for testing for disease clustering in the presence of preferential sampling; and (2) a statistically robust map of hypertension risk for the continental US. New understanding of the nationwide patterns of hypertension risk could result in more efficient allocation of public health efforts in the fight against hypertension, potentially leading to the prevention of future strokes.

2. Background and Significance

2.1. Current knowledge about the Stroke Belt/Buckle. Stroke mortality is increased by as much as 40% in the Southeastern United States (Howard et al. 1995; Lanska 1993), and this "Stroke Belt" has been recognized for more than 60 years. (Borhani 1965). The Stroke Belt traditionally consists of North Carolina, South Carolina, Georgia, Tennessee, Alabama, Mississippi, Louisiana, and Arkansas. More recently a "Stroke Buckle" has also been identified, which appears to have higher rates of stroke mortality than even the rest of the Stroke Belt (Howard et al. 1997). The Stroke Buckle contains the coastal plains of North Carolina, South Carolina, and Georgia. There is no consensus concerning the cause(s) of the Stroke Belt or

Buckle, although several hypotheses have been proposed and evaluated (Howard 1999). Traditional risk factors have explained only a minimal amount of the elevated stroke mortality in the Stroke Buckle and Belt, compared with the rest of the US (Howard 1999).

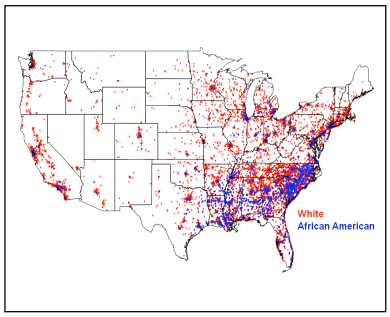

Figure 1. Map of REGARDS participant locations.

One aim of the REGARDS study is to find the reasons for this geographic disparity in stroke mortality. To accomplish this goal, the study recruited participants from all over the continental US, but oversampled the Stroke Belt and Stroke Buckle regions in order to avoid having unstable stroke rate estimates in these areas of generally low population density [see Figure 1, which is reproduced from (Howard et al. 2011)]. Specifically, REGARDS recruited 35% of the participants from the Stroke Belt, 21% from the Stroke Buckle, and 44% from the rest of the continental US.

An initial thought was that stroke incidence might be higher in the Stroke Belt and Buckle, thus creating higher rates of stroke mortality. Recently, people living in the Stroke Belt and Buckle have been found to have more strokes compared to people living in the rest of the US, but this increased stroke incidence is less pronounced than the increased stroke mortality (Howard et al. 2011). The reason(s) for people dying of strokes more often in southeastern US is still largely unknown.

2.1.1. Current understanding of the geographic distribution of hypertension.

Hypertension is the risk factor with the largest population attributable risk for stroke (Ohira et al. 2006), and it shows potential clustering in the Southeastern US, based upon maps using data aggregated at the state level, as shown in Figure 2, called a chloropleth map (CDC 2011; Hall et al. 1997; Jones et al. 1999). Whether there is true geographic variation in hypertension risk over the entire US has not been formally investigated using the information available in point-level spatially referenced data. The REGARDS study has obtained point-level data on participants' residential locations (i.e., latitude and longitude), creating an opportunity to analyze the geographic distribution of hypertension at a higher resolution than any previous study. By using the sampled locations of residential addresses, we can create a smoothed map that shows the geographic variation in hypertension at a fine scale, similar to the temperature map shown in Figure 3.

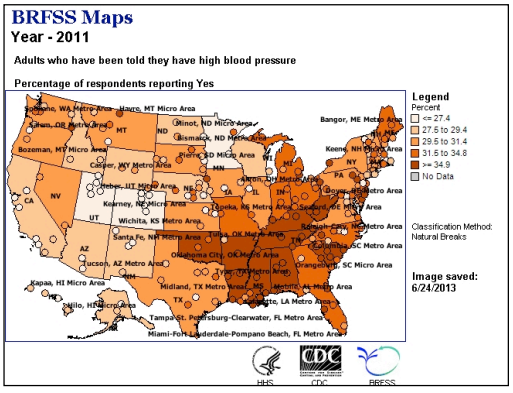

Figure 2. State chloropleth map of hypertension awareness (CDC 2011)

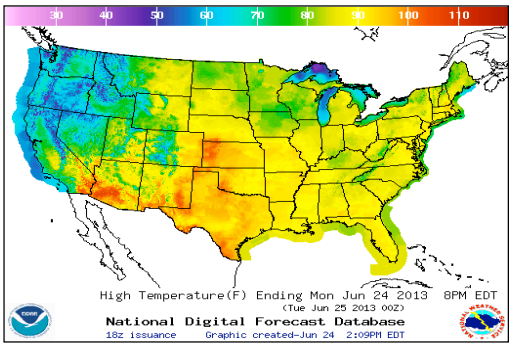

Figure 3. High-resolution map of temperature (NOAA)

2.1.2. Hypertension data in the REGARDS cohort.

REGARDS recruited 30,239 participants from 2003 to 2007, and continue to be followed every 6 months for suspected stroke events. During the baseline in-home visit, a trained health professional measured both systolic and diastolic blood pressure (each an average of two measurements). If systolic blood pressure was greater than or equal to 140 mmHg, diastolic blood

pressure was greater than or equal to 90 mmHg, or the participant reported current use of blood pressure medication, the participant was labeled “hypertensive.”

2.2. Objectives of spatial analyses. When studying the geographic distribution of diseases, there are two main goals: (1) testing whether risk for the disease varies across the region of interest; and (2) mapping the disease risk over the region by predicting risk at locations that were not sampled. The first goal is referred to as testing for disease *clustering*. In our case, tests for clustering would indicate whether risk for hypertension changes over geographic space. Formal statistical tests for clustering are useful because without them a simple map could be misleading. For example, readers of a map may not know whether displayed differences in risk in different regions were statistically different. Tests for disease clustering, using cases and controls, have been developed and are commonly implemented in public health research. These methods often assume random sampling of participant locations, and have both parametric and nonparametric implementations (Cuzick and Edwards 1990; Diggle and Chetwynd 1991; Moore and Carpenter 1999; Waller and Gotway 2004).

Mapping disease risk by predicting risk at locations that were not sampled is also a well-established procedure. It can be done by constructing a hierarchical model and using robust prediction methods in a framework called ‘model-based geostatistics’ (Diggle et al. 1998). These methods assume that no preferential sampling has occurred.

2.3. The research problem. The problem of interest here is two-fold: (1) developing tests for disease clustering that perform well in the presence of preferential sampling; and (2) producing a high-resolution map of hypertension risk for the continental US.

2.4. Gaps project will fill. A high-resolution map of hypertension risk for the continental US is currently unavailable to the stroke research community. We are also unsure whether preferential sampling of subregions in an epidemiologic study affects tests of disease clustering. This project will fill these two gaps.

2.5. Importance and relevance of this project to stroke. The statistically robust map of hypertension risk will provide evidence for or against hypertension as a contributor to the presence of the Stroke Belt, as well as a mechanism for creating future maps for comparison. If hypertension does not appear to be significantly elevated in the Stroke Belt, it can be excluded as a potential cause of the Stroke Belt. If hypertension does appear to be elevated in the Stroke Belt, then mapping its risk will be the first step in providing evidence for hypertension as a cause of the Stroke Belt and/or Buckle. Our approach will also reveal geographic patterns in hypertension risk at a level of detail that is undetectable in state chloropleth maps. Such information could change how efficiently public health efforts are allocated in the fight against hypertension. Additionally, future studies investigating the geographic distribution of stroke or its risk factors might have to oversample certain areas to obtain stable estimates. The novel methodology developed in this research project could lead to more robust inference in these cases of oversampling compared to traditional methods, meaning policy-makers can make funding decisions that more reliably lead to the prevention of future strokes or treatment of its risk factors.

3. Preliminary Study. To explore the feasibility of the proposed research, we conducted a pilot simulation study on the performance of the difference in *K* functions test of clustering. The

methodology for the simulation reflects the overall methodology outlined in the analysis plan of section 4.1.1.

3.1. Background of *K* function and statement of research problem. One measure of the variability in how many disease cases appear per unit area is the *K* function (Ripley 1976). The general form of the *K* function is the *expected* number of points within a given distance *h* of a *randomly chosen* point, divided by the overall expected density of points in the entire region of interest (also called the "intensity"). The *K* function was later used to develop a test for overall disease clustering, by comparing the difference in estimated *K* functions between a set of cases and a set of appropriate controls, at some specific distance *h* (hereafter referred to as the *K* test) (Diggle and Chetwynd 1991). One advantage of this type of test is that spikes in the number of cases in a region that are due solely to spikes in population density do not register as disease clustering. Only increases in the number of cases above and beyond increases in the number of controls will register as disease clustering, which makes sense given that areas with more people at risk will generally result in more cases, regardless of whether there is disease clustering at work. The *K* test will also indicate whether there is an unexpectedly small number of cases in some areas. A key assumption for the *K* test is that a stationary process generated the intensity of the number of points in the study region (Waller and Gotway 2004). If there is preferential sampling (oversampling of areas with expected higher numbers of disease cases), then this assumption is violated. It is unknown how the *K* test performs under preferential sampling.

3.2. Simulation methodology. To perform a preliminary investigation of the performance of the *K* test under preferential sampling, and to demonstrate the feasibility of Aim 1, we used Monte Carlo methods to compare the rejection rate of the *K* test in four different conditions: (1) constant risk across the study region (null hypothesis), with random sampling; (2) constant risk across the study region (null hypothesis), with preferential sampling (or simply oversampling); (3) varying risk across the study region (alternative hypothesis), with random sampling; and (4) varying risk across the study region (alternative hypothesis), with preferential sampling. Evaluation of the *K* test under constant risk provided the empirical type 1 error rate, and evaluating it in the presence of varying risk provided the empirical power of the test. The assigned nominal type 1 error rate was 0.05, and the distance at which clustering was tested was half the length of one side of the square region of interest. Five hundred Monte Carlo simulations were performed, with simulated datasets of about 1,000 cases and about 1,000 controls for each iteration of the 500 Monte Carlo simulations. Preferential sampling was done by sampling locations from a multinomial distribution with probabilities of being in three equally sized subregions of 20%, 30%, and 50%, respectively. When varying the risk of being a case across the region, the same subregions were used and probabilities of being a case were 10%, 30%, and 60%, respectively. Therefore, in the preferential sampling case (condition 4), approximately 50% of the data points were sampled from the subregion with the highest risk of being a case (60%).

Table 1. Results of preliminary simulation

	No clustering (Null)	**Clustering (Alternative)**
Random Sampling	0.054	0.566
Preferential Sampling	0.04	0.83

3.3. Results. Table 1 shows the empirical rejection rates of the *K* test in all four situations.
As seen in the table, preferential sampling did not appear to affect the type 1 error rate of the *K* test, as the empirical rejection rates were near the nominal type 1 error rate of 0.05. Differences between preferential sampling and random sampling are

evident in the power for the rejecting the null hypothesis. Preferential sampling appears to provide much greater power to detect overall clustering when it exists compared to designs with random sampling.

3.4. Discussion. The preliminary simulation study indicates: (1) preferential sampling affects the performance of the *K* test; and (2) we have the ability to complete the proposed research described in Aim 1. While in this specific case preferential sampling provided greater empirical power to the test, we do not know whether this increase correctly represented an increase in true power. There is still much to explore regarding the distance at which clustering should be tested and the intensity of the clustering. The low power of the test under random sampling was not worrisome because the test might not have been conducted at the scale at which the clustering truly existed. The main purposes of this preliminary simulation were to show plausibility of the hypothesis that preferential sampling can affect tests of disease clustering and that we can feasibly carry out the proposed work.

We have no preliminary results addressing Aim 2 due to the increased computational burden involved. (The proposed models are fit using computationally intensive Markov Chain Monte Carlo algorithms and will require considerable computing time to complete.) I have completed didactic coursework in Bayesian methods and Bayesian disease mapping, which will assist in completion of Aim 2. Additionally, the consultant for this project, Dr. Lance Waller, has extensive experience in Bayesian hierarchical modeling of spatial data, which, in conjunction with the experience of other committee members, will provide me with the necessary assistance to successfully complete Aim 2.

4. Research Design and Methods

4.1. Develop novel extensions for tests of disease clustering that perform well in the presence of preferential sampling. This aim will have two components: (1) investigating the performance of three popular tests of disease clustering in the presence of preferential sampling; and (2) modifying the tests appropriately so that they achieve appropriate type 1 error rates and power levels.

4.1.1. Investigating the performance of three popular tests of disease clustering in the presence of preferential sampling. Testing for disease clustering is a key goal of any spatial analysis of disease. John Snow informally assessed clustering when investigating the 1854 cholera epidemic in London (Snow 1855). In our study of hypertension, we are concerned with determining whether the risk for the disease varies over the region of interest, or clusters. While there are many different tests for disease clustering, we will concern ourselves with three popular ones: (1) a test which compares a summary statistic of the intensity function of the cases with a summary statistic of the intensity function of the controls (Wheeler 2007); (2) the *K* test (Diggle and Chetwynd 1991); and (3) the Cuzick-Edwards nearest neighbor test (Cuzick and Edwards 1990). These three tests have recently been compared in an empirical study of clustering of childhood leukemia rates in Ohio (Wheeler 2007). These three tests often assume random sampling of diseased and non-diseased participants, or at least that no preferential sampling occurred. Little is understood about how these three methods would perform when preferential sampling has occurred, such as in the REGARDS data set. Although the REGARDS study oversampled regions of specifically high stroke mortality, it is reasonable to expect from our knowledge of the relationship between hypertension and stroke that the Stroke Belt and Stroke Buckle have high rates of hypertension as well (Ohira et al.

2006). The goal of this aim will be to evaluate the statistical performance of 3 tests of disease clustering in the presence of preferential sampling.

Analysis plan. We will use Monte Carlo simulations to investigate the performances of the test of intensity functions, the *K* test, and the Cuzick-Edwards nearest neighbor test. The empirical rejection error rate will be calculated under the null hypothesis (constant risk or no clustering, which provides the empirical type 1 error rate) and the alternative hypothesis (nonconstant risk or clustering, which provides empirical power), in the presence of random sampling and preferential sampling. (Technically, preferential sampling under the null hypothesis would simply be oversampling.) Table 2 shows this analysis plan in tabular format.

Table 2. Conditions to simulate.

	No clustering (Null)	**Clustering (Alternative)**
Random Sampling	Type 1 error rate	Power
Preferential Sampling	Type 1 error rate	Power

Our simulated region will be roughly the size of the continental US, with the regions of preferential sampling to be located roughly at the Stroke Belt, Stroke Buckle, and the rest of the US, with sampling probabilities for the approximately 30,000 observations that reflect the REGARDS dataset: 20% in the Stroke Buckle, 35% in the Stroke Belt, and 45% from the rest of the US. This simulation will be implemented in the R programming language, with the added “sp” and “splancs” packages (Bivand et al. 2008; Pebesma and Bivand 2005; R Core Team 2013; Rowlingson and Diggle 1993).

4.1.2. Modifying the tests appropriately so that they achieve appropriate type 1 error rates and power levels. Our goal will be to develop tractable and interpretable extensions of existing tests that are appropriate to use when preferential sampling has occurred. As was shown in the preliminary analysis (Section 3), preferential sampling can affect the estimated power of the *K* test. More exploration of how preferential sampling affects the theoretical underpinnings of these tests for disease clustering is needed to understand whether our estimated rejection rates reflected the true type 1 error or power. The general method of conducting all three of these tests is to randomly re-label all of the locations sampled as cases or controls, and then recalculate the test statistic (Waller and Gotway 2004). Under the null hypothesis of no disease clustering, the random labeling should not affect the value of the test statistic. We hypothesize that the random labeling process will have to be modified in order to accommodate the oversampling. Other options for modifications of the tests are weighting of the observations or introducing Bayesian methods into the test.

4.2. Create a high-resolution map of hypertension risk for the continental US. There are many ways to create maps of a variable when that variable is continuous and is observed at point-level locations. An example of this situation is shown in Figure 3, where temperature is observed at each latitude and longitude location of the weather stations within the US. When the variable of interest is binary, such as in the case of disease status, then mapping becomes more complicated. The complication is that the surface we are interested in is not a surface of 1’s and 0’s indicating disease status, but a surface of the probability of having hypertension at each location. In other words, the variable we are interested in is not the variable we directly observe. In these situations, the surface in which we are actually interested, but do not directly observe, is called a latent process (Diggle et al. 1998). Hierarchical models are often natural tools to use when investigating a latent process, which in our case is risk of hypertension. We will treat the individual observations (hypertensive or not) as independent of one another, but will accommodate a spatial correlation structure for the risk of hypertension (i.e., at the latent

process level). After the model is constructed, prediction of hypertension risk at new locations that were not observed is relatively straightforward. The goal will be to produce an overall map of hypertension risk (and associated prediction errors), as well as maps specifically for men, women, blacks, and whites. This aim will also have two steps: (1) construction of the hierarchical statistical model; and (2) prediction of hypertension risk for all locations on a pre-specified fine grid that covers the continental US.

4.2.2. Construction of the hierarchical model. The underlying hypothesis is that risk of hypertension, after accounting for systematic factors such as age, weight, sex, obesity, and race, varies smoothly with respect to geographical location. Such smoothness reflects effects of unobserved factors that vary from location to location (e.g., lifestyle). This type of smooth process can be modeled using spatial covariance functions that assume, *a priori*, that the risk of hypertension between two individuals decreases monotonically as geographic distance increases. We will model the risk of hypertension, $p(s_i)$, for the *i*-th individual at location s_i (here s_i is a vector of latitude and longitude coordinates), where $i = 1,\ldots,n$, with the following generalized linear mixed model:

$$\log\left(\frac{p(s_i)}{1-p(s_i)}\right) = X_i\beta + f(s_i)\mu + S(s_i)$$

where X_i is a vector of relevant covariates (likely to be gender and race) for the *i*-th observation, β is a vector of fixed effect parameters, $f(s_i)$ is a function of the *i*-th location coordinates, μ is a vector of fixed effect parameters for the function $f(s_i)$, and $S(s_i)$ is a random effect of the spatial process at the *i*-th location. We usually assume that: (1) $S(\cdot)$ is a Gaussian process with mean 0 and variance-covariance matrix $\Sigma(\theta)$, where Σ is an *n* x *n* positive definite covariance matrix formed by evaluating a valid spatial covariance function (e.g., Matérn) at pairs of points in space, and θ is a set of parameters indexing the covariance function; and (2) that the process that generated the locations (s_i) is stochastically independent of $S(\cdot)$ [i.e., no preferential sampling (Diggle et al. 1998)].

We have decided that a Bayesian framework will be appropriate for fitting the above model. We will determine appropriate prior distributions for all parameters in the model, which will then be used to fit the model using Markov Chain Monte Carlo methods. The final fitted model will provide posterior distributions of the model parameters.

Challenges. The large number of observations in this dataset can create a computational burden. Solving for the likelihood of this model requires n^3 computations, where *n* is the number of observations (approximately 30,000 in the REGARDS study) (Cressie and Johannesson 2008; Eidsvik et al. 2010; Fuentes 2007). Therefore, we will use specialized Markov Chain Monte Carlo (MCMC) methods. Another challenge is the mean risk for hypertension probably changes over the region, given the large size of the region of interest (the continental US). A spatial process with a constant mean, often called a stationary process, is a conventional assumption in model-based geostatistics. Inclusion of the $f(s_i)$ term will help to remove the mean effects at different locations and help satisfy the stationarity assumption. Third, the heterogeneous data density of REGARDS across the US can create prediction intervals for hypertension risk that are much wider in regions with fewer observations (e.g., the western part of the US compared to the eastern part of the US). Therefore, we shall use

Bayesian methodology to “borrow” information from neighboring areas to improve local estimation. While it may seem strange to borrow information from a densely sampled region like North Carolina to inform estimates in a sparsely sampled region like Montana, the small sample size in Montana could lead to unreasonably variable/unstable risks. Bayesian methods are one way to alleviate the impact of the small sample sizes in these regions in a statistically robust manner. Finally, the preferential sampling will require joint modeling of the process that produced the sampled locations and the risk of hypertension, following previous suggestions for using model-based geostatistics in the presence of preferential sampling (Diggle et al. 2010).

4.2.3. Prediction of hypertension risk for all locations on a pre-specified fine grid that covers the continental US. Prediction of hypertension risk will be done by sampling from the posterior distribution of hypertension risk at locations on a pre-specified fine grid that covers the continental US, in order to provide a high-resolution map. The scale of the prediction grid will be no smaller than the shortest distance between two observed locations in the REGARDS data set.

4.3. Limitations. No research is without limitations, and the proposed research has some that are worth noting. The proposed model requires several assumptions, including the choice of the functional form for fixed effects as well as the choice of co-variance function for the random effect. Incorrect specification of any of these elements can lead to model misspecification. We will address this challenge by evaluating alternative specifications for the model and choosing a particular one using an appropriate model comparison criteria (e.g., AIC, BIC). A minor limitation of the proposed research is computational complexity, which requires extra time and resources to overcome. However, we believe that the benefits of using the methods we have proposed (e.g., Bayesian methods) outweigh the additional computational complexity they contribute to the project.

4.4. Restriction of the study sample to blacks and whites. The REGARDS study was begun with two specific purposes: (1) to investigate potential causes of the geographic disparity in stroke mortality between the Stroke Belt, the Stroke Buckle, and the rest of the US; and (2) to investigate potential causes of the *racial* disparity in stroke mortality between blacks and whites (Howard et al. 2005; Tassone et al. 2009). We acknowledge that there are potentially other racial/ethnic disparities regarding stroke mortality, but they were not the subject of the research proposed by REGARDS. Thus, our study sample for this project is limited to blacks and whites of both genders.

5. Ethical Aspects of the Proposed Research. As with any scientific research on human subjects, there are concerns about ethics. All of the Institutional Review Boards of the participating institutions in the REGARDS study approved the original research. Access to the REGARDS data is restricted to those with IRB approval to use the data. I have achieved and maintained both HIPAA and IRB training through the University of Alabama at Birmingham, and have obtained approval from the REGARDS executive committee to conduct this project. The scope of this project is within the realm of the consent form for the REGARDS study. This project also has specific ethical concerns. Location of participant residential addresses is potentially identifiable health data, and thus must be treated with care. When displaying the locations of participants in the REGARDS dataset, all locations will be randomly “perturbed” in order to conceal the real locations of the participants (Armstrong et al. 1999).

TRAINING/CAREER GOALS

Having lived in the Stroke Belt my whole life, I am committed to finding the source of our geographic disparity in stroke mortality. **My overall goal for the fellowship period is** to become an expert in the spatial distribution of hypertension in the continental US.

This fellowship award will provide me with opportunities that would be unavailable otherwise, such as collaboration with an expert in spatial statistics (Dr. Lance Waller), training in the use of GIS, monetary support to present work at conferences devoted specifically to spatial statistics and conferences devoted specifically to cardiovascular disease, and support for time spent teaching workshops and/or courses on spatial statistics to biostatistics students and clinical/public health professionals.

The award will help me gain the following **skills:**

- learn the pathophysiology and deleterious effects of stroke;
- identify, use, and evaluate appropriate analysis tools for spatially referenced datasets that are large (on the order of tens of thousands), have heterogeneous observation density across the region of interest, and violate traditional assumptions of geostatistical methods (e.g., constant mean across the region of interest);
- develop tractable and *interpretable* statistical methodology in the context of stroke; and
- manage and display spatially referenced data using geographical information systems (GIS).

These skills describe a pipeline for conducting and disseminating independent research, and they will prepare me to continue on to my overall goals.

My overall career goals are: (1) to understand the causes of stroke by collaborating with stroke experts to apply spatial statistics; and (2) to develop new spatial statistics methodology that helps answer complex research questions about cardiovascular diseases. Use of spatial statistics is not limited to the geographic scale, but can also be applied to a wide array of spatial scales, such as imaging. Moreover, the analytical skills gained during this fellowship have immediate applications to other areas of biomedical research where risk of disease differs according to geography.

After graduation, I will pursue a postdoctoral fellowship with an expert in spatial statistics. While my first overall career goal is to perform applied research, a strong background in the theory of spatial statistics will prepare me to adapt to new, exciting, and complex spatial health problems. This postdoctoral position will eventually lead to an academic appointment, where I will continue to investigate causes of cardiovascular diseases in general, and stroke in particular.