2014

A Spatiotemporal Model for Repeated Imaging Data

Brandon George
University of Alabama at Birmingham

A SPATIOTEMPORAL MODEL FOR REPEATED IMAGING DATA

by
BRANDON J. GEORGE

INMACULADA B. ABAN, COMMITTEE CHAIR
LESLIE A. MCCLURE
HEMANT K. TIWARI
LOUIS J. DELL'ITALIA
THOMAS S. DENNEY
HIMANSHU GUPTA

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2014

A SPATIOTEMPORAL MODEL FOR REPEATED IMAGING DATA

BRANDON J. GEORGE

BIOSTATISTICS

ABSTRACT

Longitudinal imaging studies have increased in popularity as clinical researchers seek to investigate how phenomena within the body change over time. Analysis of data from these studies is complicated by correlation between repeated measures over time and different locations in the body.

To address this problem we propose the use of a linear model with a separable parametric correlation structure. This model considers spatial and temporal correlation independently and incorporates the correlation using parametric functions that have the potential to be much more efficient than an unstructured approach. Our model also has the ability to control for time- and space-varying covariates, which previously used summary methods cannot do.

Results from a simulation study that investigates the effects of correlation structure selection on statistical inference about a treatment-by-time interaction are reported. Using the true correlation structure conserves the Type I error rate and maximizes power versus other structures, while misspecified structures may inflate the type I error rate or reduce power. If the misspecified structure can closely approximate the true correlation function then the Type I error rate is conserved and the loss of power from the true structure is negligible. For the considered conditions, information criteria are highly accurate at choosing a working correlation structure that conserved the Type I error rate.

Our model is compared to summary methods through a simulation study that considers inference on a treatment-by-time effect. Our model more reliably conserves the Type I error rate and has greater statistical power than summary methods in space and time. The practice of analyzing spatial regions separately is found to have poor statistical properties. The presence of missing data does not change the qualitative results.

Finally, we apply our model to the UAB SCCOR study, which considered MRI-derived outcomes from a longitudinal clinical trial in mitral regurgitation patients assigned to medical therapy or placebo. This study provided the motivation for this dissertation and inspired the scenarios used in the simulation studies. Here we discuss practical considerations of applying our model to real data such as how to choose a working correlation structure and how to handle missing data.

Keywords: spatiotemporal, correlation, spatial, longitudinal, cardiology, imaging

DEDICATION

I dedicate my dissertation research to my wife, Stephanie Brosius, for all her love and support throughout these long years, and to my mother and late father, Kathy and Thomas George, for the love of learning and desire for self-improvement they instilled in me. Without them this work would not have been possible.

ACKNOWLEDGEMENTS

I would like to sincerely thank my committee chair, Dr. Inmaculada Aban. Her guidance and encouragement were indispensible to this work. I could not have asked for a better advisor, and through her I was able to recognize my potential as a researcher. I feel blessed to have had a mentor who not only pushed me to excel but also celebrated my accomplishments along the way.

I would like to thank all of my committee members for the role they played in my work: to Drs. Hemant Tiwari and Leslie McClure, for their assistance in becoming a better statistician and presenter; to Drs. Louis Dell'Italia, Himanshu Gupta, and Thomas Denney, for not only the data that formed the core of my methodological and applied work but for their encouragement to be a better collaborative statistician and to focus on what is ultimately important in a model.

I would like to thank Dr. Chun Schiros for the lovely cMR images she provided. I would also like to thank Dr. Sean Simpson at Wake Forest School of Medicine for his wisdom and insight into separable covariance structures, and his generous support while investigating the same statistical problem.

I would also like to thank the Biostatistics department, Dr. Hemant Tiwari, and the National Heart, Lung, and Blood Institute (via T32HL079888) for their financial support in my graduate training.

I would like to finally thank my classmates Matthew Loop, Hwasoon Kim, and Guoquio Wang, for the support and advice they have given throughout graduate school.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table *Page*

SELECTING A SEPARABLE PARAMETRIC SPATIOTEMPORAL COVARIANCE STRUCTURE FOR LONGITUDINAL IMAGING DATA

COMPARING SUMMARY METHODS AND A SPATIOTEMPORAL MODEL IN THE ANALYSIS OF LONGITUDINAL IMAGING DATA

APPLYING A SPATIOTEMPORAL CORRELATION MODEL FOR LONGITUDINAL IMAGING DATA

LIST OF FIGURES

Figures *Page*

LITERATURE REVIEW

SELECTING A SEPARABLE PARAMETRIC SPATIOTEMPORAL COVARIANCE STRUCTURE FOR LONGITUDINAL IMAGING DATA

COMPARING SUMMARY METHODS AND A SPATIOTEMPORAL MODEL IN THE ANALYSIS OF LONGITUDINAL IMAGING DATA

APPLYING A SPATIOTEMPORAL CORRELATION MODEL FOR LONGITUDINAL IMAGING DATA

LIST OF ABBREVIATIONS

AR-1	autoregressive-1 correlation function
AUC	area under the curve
CS	compound symmetric correlation function
EXP	exponential correlation function
IC	information criteria
MAT	Matérn correlation function
LEAR	linear exponent autoregressive correlation structure
LRT	likelihood ratio test
LVEF	left ventricular ejection fraction
LVESD	left ventricular end-systolic dimension
MR	mitral regurgitation
MRI	magnetic resonance imaging
R/T	radius of curvature-to-wall thickness
REML	restricted maximum likelihood
SCCOR	Specialized Centers of Clinically Oriented Research
SPH	spherical correlation function
TOEP	Toeplitz correlation function
UN	unstructured correlation model

INTRODUCTION

Introduction and Motivation

Since the early days of Fisher and Gosset, most statistical methods have arisen from a desire to solve a particular problem. Methodological research grows from an investigator's encounter with a set of data that simply cannot be appropriately analyzed by any known methodologies.

In this sense, our work is much like many that have come before it. Our motivation comes from the data collected in the SCCOR (Specialized Centers of Clinically Oriented Research) study at University of Alabama at Birmingham (UAB). It is composed of three-dimensional magnetic resonance imaging (MRI) scans of the subject's heart for patients with multiple possible cardiovascular diseases such as myocardial infarction, mitral valve regurgitation, and left ventricular hypertrophy. In general, each patient's data consists of five of these scans, collected six months apart. For this research, we focused on the part of the study that looked at patients with mitral regurgitation (MR) enrolled in a clinical trial testing a medical therapy.

In this body of work we propose a method that models both spatial and temporal correlation with separable parametric structures, and investigate ways to choose between multiple structures. We also investigate how such a model compares to summary measures in space and time that have been previously used to study such datasets. Lastly, we discuss how our model can be implemented in practice using the UAB SCCOR data as an example.

The difficulty in this application is that the raw data are collected at different points, referred to as voxels, within the three-dimensional MRI image. These values cannot be assumed to be independent, as there is most likely spatial correlation for values within the

same image. Similarly, when looking at the values at the same location in the same patient taken at different time points, one would expect to see some level of temporal correlation. Therefore, the most appropriate method to analyze this dataset will account for both spatial and temporal correlation. Furthermore, not every patient had every observation and rarely adhered to the six month schedule, so an ideal method needs to be able to handle unevenly spaced repeated measures with missing data.

One particularly interesting application of such a spatiotemporal model is that it can be used to draw inferences about remodeling in the left ventricle over time. Indeed, one reason clinical researchers are excited about imaging is that it can be utilized to observe more sensitive clinical outcomes. Rather than relying on a less specific and more variable outcome such as mortality, investigators can get a direct look at how an intervention can change the course of the disease. Since imaging can be expensive, especially for MRI, it is highly desirable to get as much information as possible from the collected images. This method is also extendable to other body parts and imaging modalities, such as images of the brain or PET scans.

A specific motivating factor is that current evaluation of mitral regurgitation patients is done by looking at global parameters for left ventricular geometry and function. We feel that this represents a possible improvement in clinical interpretation and practice, and that relating ventricular geometry and function at a segment-level basis to disease progression could increase sensitivity in analysis and eventually diagnosis. In particular, we wish to examine the radius of curvature-to-wall thickness ratio in patients with mild and severe mitral regurgitation as a measure of disease progression to determine the effectiveness of medical therapy. These topics will be explained more fully in the next section.

Background and Literature Review

Clinical

Mitral Regurgitation and Left Ventricular Remodeling. Although we wish for our model to be generalizable to many different applications of spatiotemporal medical data, it is important to consider the immediate problem we hope the method can solve. In this case, we need to have a solid understanding of what is known to be the natural history of mitral regurgitation so that the appropriate MRI parameters can be used to identify those changes. In particular, we wish to know how the left ventricle remodels in patients with mitral regurgitation in the absence of medical or surgical intervention. For this application we are interested in properties that vary over the left ventricle; global descriptors such as ejection fraction do not have concerns with spatial correlation and have already been handled in SCCOR using basic mixed models.

Mitral regurgitation is defined as backflow from the left ventricle through a faulty mitral valve into the low-pressure left atrium. The greater the amount of backflow, the 'worse' the MR. The body then attempts to compensate for this loss of cardiac output by pumping harder resulting in left ventricular hypertrophy. Evidence suggests that one avenue of compensation is an increase in the activation of the systemic sympathetic nervous system[26]. *In vitro* experiments have shown that consistent exposure to norepinephrine has severely deleterious effects to cardiac myocytes[23]. In patients with MR we see that after the initial compensation and hypertrophy the ventricular wall begins to bulge and suffer harmful structural changes; this secondary change is called *decompensation*. This decompensation has been seen to happen along with increased markers of oxidative stress and pathology indicating myofibrillar degradation[2]. The primary effects of the secondary remodeling seem to be an increase in stroke volume and decrease in ejection fraction. In particular, in dogs it has been found that MR leads to:

- reduced cardiac output[29],

- an increase in LV end-diastolic (LVED) volume-to-LV mass ratio,
- a decrease in the LVED circumferential curvature at the midwall,
- an increase in the LVED radius/wall thickness ratio[14].

All together, there seems to be a 'ballooning' of the left ventricle, where the chamber gets bigger but the wall does not get correspondingly thicker. In more technical terms, there seems to be an increase in the *sphericity* of the left ventricle such that it becomes more spherical rather than the normal 'long-ellipsoid' shape. To use a sports analogy, it looks more like a basketball than a football. Because the wall does not increase its thickness along with this increase in radius, the wall stresses must increase according to the laws of Laplace. It has been seen in the SCCOR study that both the maximum shortening (E_{min}) and rotation across the left ventricle differ between patients with MR and healthy individuals[31]. Previous work has shown that in healthy mammalian hearts, the radius-to-thickness (R/T) ratio is approximately constant from the base to the apex of the left ventricle[4]. Previous results from the SCCOR dataset indicate that the R/T ratio is elevated across the whole LV in untreated MR patients, and that this increase is significantly more pronounced in patients with more advanced MR[30].

There have been mixed results with medical therapies, where beta-blockers have been the main focus of research. This approach of beta-blockers is done due to evidence of an elevated adrenergic response in MR patients[27] and promising results in canines[36], but SCCOR's results are not as strongly positive[1]. With a lack of effective medical intervention, the standard intervention is surgical repair of the mitral valve. Although patients can be identified or diagnosed with MR from murmurs or low ejection fraction, surgery is typically not done unless the patient has an abnormally large left ventricular end systolic dimension (LVESD). A common rule of thumb is that the patient should only receive surgery when the LVESD is greater than 40mm. There have been concerns among cardiologists that this cutoff results in surgical intervention coming too late to prevent irreversible

damage to the myocardium, a view that has been supported by previous analyses of the SCCOR data[2, 30].

Currently, clinical practice focuses on global MRI parameters in mitral regurgitation patients. The assessment of left ventricular geometry is done by looking at the LVESD, the diameter of the left ventricle at the base near the mitral and aortic valves. The assessment of left ventricular function is done by looking at the left ventricular ejection fraction (LVEF), the proportion of blood in the left ventricle in diastole that exits into the aorta in systole. However, these measures are less than ideal; the aforementioned sphericity and damage to the myocardium have been observed in MR patients with 'good' LVEF[2] and with LVESD smaller than 40mm[30]. Furthermore, it has been observed that LVEF decreases post-surgery which has thus far been unexplained but may also be indicative of irreversible myocardial damage[30].

It is important to understand the connection between left ventricular geometry and function. Through Laplace's Law and hemodynamics, the two are inextricably linked. However, since the heart can compensate for MR and maintain overall function (as in LVEF[2]), clinicians rely on left ventricular geometry to identify how far along in the disease course a patient is. Since the increase in LVESD is not very sensitive and usually comes later in the disease (potentially too late to prevent permanent damage), we propose that the R/T ratio be used as an index of the sphericity of the left ventricle. It would be of great interest to show that the R/T ratio (representing geometry) is associated with the abnormal strain and rotation (representing function), as it could possibly move us closer to establishing a more sensitive measure of MR progress that can result in better patient outcomes.

Presentation and Interpretation of Cardiac Imaging Data. Cardiac magnetic resonance (CMR) imaging is the clinical gold standard for describing the structural and functional

characteristics of cardiac motion. When constructing a plan for statistical analysis, it is important to understand where this CMR data comes from and what it means.

The first thing to consider is how the heart is described in CMR imaging, so that we can all be discussing the same part of the heart. There are three main axes that are used for imaging the ventricles, shown in Figure 1: short axis, horizontal long axis, and vertical long axis[8].

It is useful to divide the left ventricle into segments so that particular regions can be clearly identified and discussed. The most common way to do this is with the AHA's 17 segment model, presented in Figure 2. The view is from the short axis, looking down into the left ventricle from the base towards the apex. The concentric rings represent the bullet shape of the LV; it can also be considered a cylinder cut into three pieces with a hemispherical cap. It is common to assign one value to a segment as a summary of all of the voxels included in that area; it is also common to summarize the segments themselves into the basal, mid, and apical rings. The 17th segment, the apex, is usually omitted from these types of analyses. One can also examine the different segments grouped by the coronary artery that feeds them, shown in Figure 3[8].

A variety of imaging processing tools have been developed to extract important clinical values from CMR images. Ventricular wall thickness is easy to measure, as one can measure the length of lines drawn radially outward across the myocardium[7]. However, we often wish to know something about the motion of the ventricular walls, which can be more difficult. A common way to measure motion is through the use of tagging.

Tagging is the process where the magnetic field produced by the MRI machine is modulated in such a way that the myocardium is temporarily magnetized in a certain pattern; this pattern comes out in the final image as a grid of dark lines against the light-colored tissue[32]. Because the tissue itself is holding the grid, movement of the wall can be seen as deformation of the grid which allows us to track the movement of individual

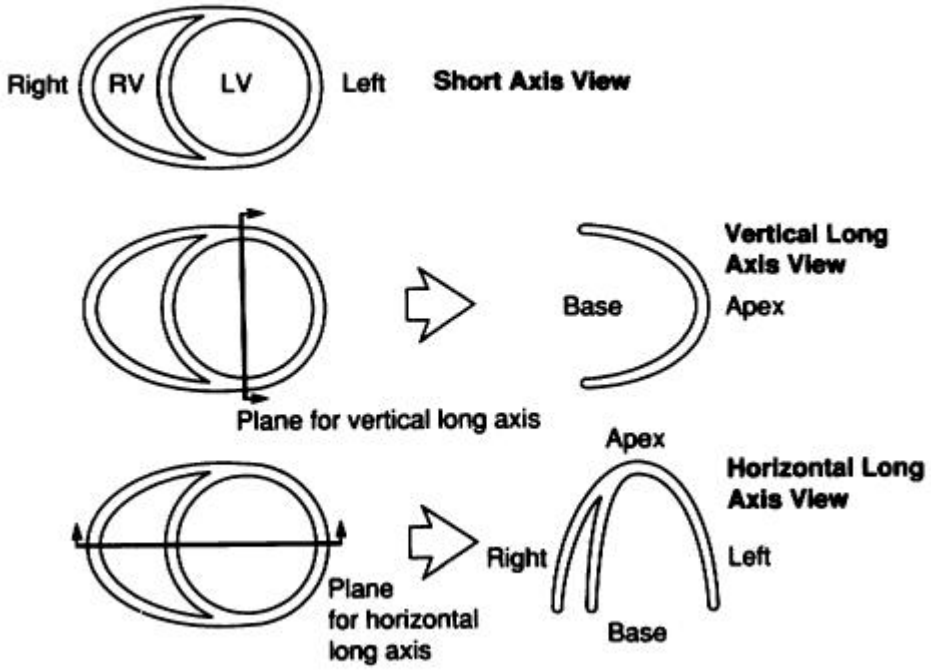

Figure 1. Three orientations for viewing the left ventricle[8].

Note: From "Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart" by M. D. Cerqueira, et. al., 2002, *Circulation*, 105, p. 540. Copyright 2002 by the American Heart Association. Reprinted with permission: license number 3191020494022.

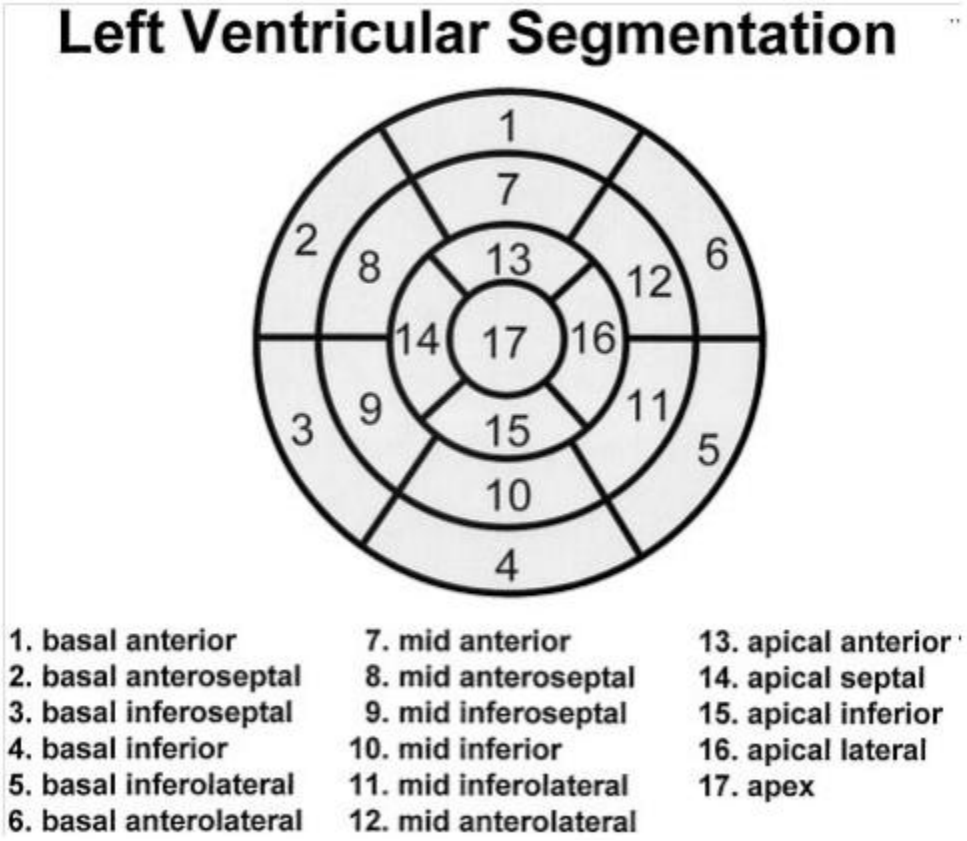

Figure 2: The AHA's 17-segment model of the left ventricle[8].

Note: From "Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart" by M. D. Cerqueira, et. al., 2002, *Circulation*, 105, p. 542. Copyright 2002 by the American Heart Association. Reprinted with permission: license number 3191020494022.

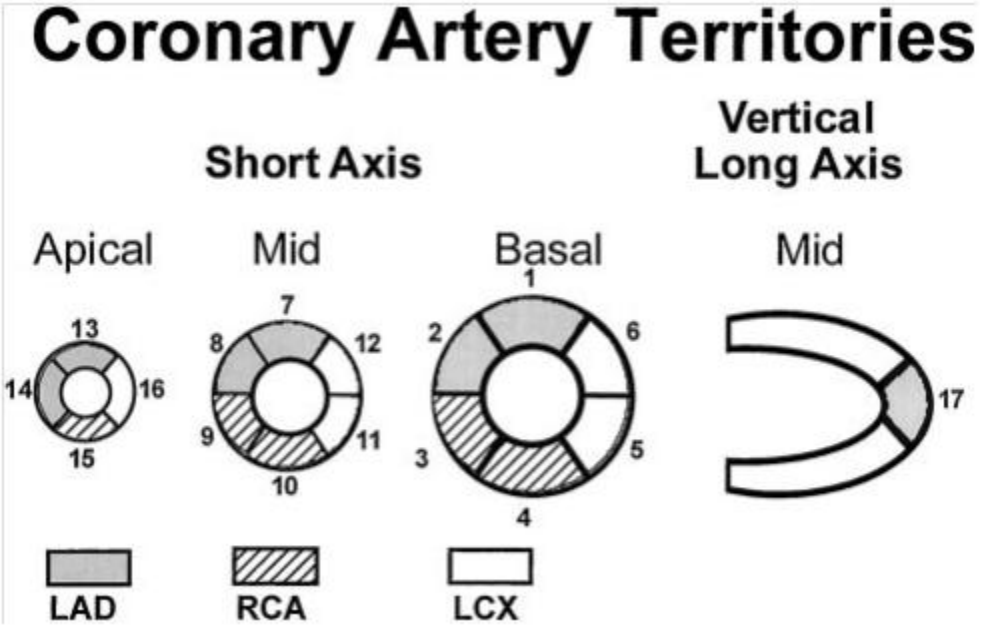

Figure 3: The 17 segment model with the coronary arteries that feed them (left anterior descending, right coronary, and left circumflex)[8].

Note: From "Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart" by M. D. Cerqueira, et. al., 2002, *Circulation*, 105, p. 542. Copyright 2002 by the American Heart Association. Reprinted with permission: license number 3191020494022.

points. Three-dimensional movement can be observed as well by imposing and imaging grids in orthogonal planes[38]. The computation needed to tie a series of images together to truly capture the motion of a cardiac cycle is non-trivial. Fortunately, harmonic phase (HARP) analysis can do these calculations quickly by treating the image as a signal and moving it to the frequency domain with a Fourier transform. This speed is highly desirable, as it lessens the patient burden by reducing the length of time the subject must hold their breath[7]. This procedure of measuring cardiac wall motion through tagging and HARP was used in the collection of the SCCOR data[1].

Once the motion of the heart can be fully observed, it is possible to generate a strain map of the ventricle. *Strain* is a unitless quantity describing the amount of deformation (stretching or compression) and is defined as

$$\frac{L - L_0}{L_0}$$

where L is the current length and L_0 is the original length. Strain in the ventricle is usually examined as being in the radial, circumferential, and longitudinal directions, although one may also look at the principle strains which are in the direction of the greatest elongation and greatest compression. Having a three-dimensional map of strain in the left ventricle over the cardiac cycle has immense clinical implications. The systolic strain maps allow one to assess the contractility of the myocardium and also to detect abnormalities in the pattern of contraction. This asynchrony can be indicative of arrhythmias[25]. Diastolic strains lets one quantify the elasticity of the myocardium as it relaxes as blood fills the ventricle[7].

Since we are studying abnormalities in the LV, it is crucial to know what the strain map looks like for a healthy subject. In healthy patients, circumferential strain increases from the base to the apex and from the epicardium to the endocardium, and tends to be higher in the anterior and lateral regions than the inferior region. Longitudinal strain also

increases from base to apex and has the same transmural pattern of increasing from the epicardium to the endocardium (through the ventricular wall from the inside to outside of the heart). The transmural pattern holds for radial strain as well, but there is not a consensus about how radial strain changes between the base and apex. It has been observed that older patients tend to have slower relaxation in diastole for both circumferential and longitudinal strains[32]. Slow relaxation of circumferential strain has also been seen in patients with hypertrophy[25].

In addition to strain, torsion is an important parameter to measure. Looking up from the apex, in healthy individuals the base of the LV rotates clockwise while the apex rotates counter-clockwise. This wringing motion is thought to improve the efficiency of the heart, in part by creating suction during the 'un-wringing' in diastole that helps fill the LV. Patients with hypertrophy tend to experience greater torsion during the cardiac cycle, while older patients experience slower relaxation of torsion in diastole[32].

Spatial Statistics

History and Background. The use of the spatial component in analysis is not a new development. For instance, consider the work by John Snow on the 1854 cholera epidemic in London which many consider to be the start of modern epidemiology. He recorded the residences of all the cholera victims in the Soho neighborhood, and noted the locations on a map of the area that included the locations of public water pumps. By doing so, he was able to see that although there was spatial variability in the case locations, there was a distinct clustering around the Broad Street pump.

This story is well-known among biostatisticians and epidemiologists. There are many reasons for this, such as the successful civic response to close the Broad Street pump, to the deductive nature of his study that has the feel of exciting detective work with a clear culprit. In a way, it is an example of spatial statistics at not only its best, but also at its

easiest. The outcome was binary (cholera or no cholera) and clusters of cases are much easier to see in acute diseases, especially infectious diseases. Unfortunately, in practice the spatial resolution may be much larger than a single neighborhood, and the disease could be of a more chronic nature with a far more complex etiology.

To handle these more complicated situations, formal statistical methods are necessary. Some of the greatest development has followed the example of John Snow and has focused on the clustering of events like disease cases over a geographic region. In their book *Applied Spatial Statistics for Public Health Data*, Waller and Gotway[37] (2004) detail the many different methods that can be used to address this issue of clustering as well as the complexities involved in defining the problem. When looking at clusters, one must decide whether the research hypothesis pertains to detecting what are distinct clusters or to finding if there is a general pattern of clustering; many methods have been produced which answer both of these types of questions. Another wrinkle in this kind of cluster analysis is the nature of the location data: some datasets have the exact location for each event, while others only have counts aggregated over whole subregions. Although ultimately one assumes the events are the result of a stochastic Poisson point process, the way this assumption is implemented varies between study types. Unfortunately, regional count data introduces the modifiable area unit problem, which is that the study results could change merely if the subregion boundaries were redrawn.

Spatial Exposure Data. In the SCCOR data, we are primarily concerned with continuous values distributed over space rather than discrete binary events. These measures of continuous spatially-distributed random variables are called *exposure data* by Waller and Gotway[37]. For example, an agricultural study may look at rainfall at different farms across the state as the exposure of interest. Note that these kinds of variables can be the dependent variable instead of a covariate. In this type of dataset, values $Z(s_1), Z(s_2), ..., Z(s_N)$ for some variable Z are collected at N different locations; typically the variable is continu-

ous and the locations coincide with event locations. We assume that the values observed are realizations of a *random field*, a random or stochastic process defined as $[Z(s) : s \in D]$ for a study region D. Note that for a fixed s, $Z(s)$ is a random variable. The random field is considered *stationary* if $E[Z(s)] = \mu$ for all s, second-order stationary if $\text{Cov}[Z(s_i), Z(s_j)] = C(s_i - s_j)$ for all $s_i, s_j \in D$. In other words, stationarity means that the mean (and possibly covariance) of a random field does not depend on the location within the field. A random field is *isotropic* if the covariance function $C(\bullet)$ depends only on distance and not direction.

Ultimately, we want to be able to describe this spatial autocorrelation within the random field. One approach is based on *semivariograms*, defined as

$$\text{Var}(Z(s_i) - Z(s_j)) = 2\gamma(s_i - s_j), s_i, s_j \in D \tag{1}$$

where the function $\gamma(\bullet)$ is the semivariogram. The properties of a semivariogram are:

1. $\gamma(s - u) = \gamma(u - s)$. Direction doesn't matter for a given two points. This means that a corresponding covariance structure is symmetric.

2. $\gamma(s - s) = \gamma(0) = 0$.

3. With spatial lag h between points s and u, $\gamma(h)/h^2 \to 0$ as $h \to \infty$. As points go apart, the semivariogram goes to zero.

4. The semivariogram is *conditionally negative definite*, meaning

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \gamma(s_i - s_j) \leq 0 \tag{2}$$

   such that $a_1, \ldots, a_m$ are real numbers whose sum is zero.

5. A semivariogram is isotropic if it depends only on the length of the spatial lag, not the direction.

A common method for interpreting semivariograms is to plot them against the spatial lag distance h. An increasing semivariogram implies that closer values are more related than distant ones. A typical function will increase but level off after a certain distance; this plateau is called the *sill*.

If the random field is second order stationary, then

$$\gamma(h) = C(0) - C(h). \tag{3}$$

If $C(h) \to 0$ as $h \to \infty$ then C(0) is the sill of the semivariogram. We can get the correlation as a function of the spatial lag, called the *correlogram*, as

$$\rho(h) = C(h)/C(0).$$

The semivariogram is desirable to use as its estimation is slightly easier than the estimation of the covariance function since it does not rely on estimating the mean of the random field.

If the random field is assumed to be isotropic, the estimation of the semivariogram begins by choosing a parametric model. Common models include spherical, exponential, power, Matérn, and Cardinal-Sine functions. Much like choosing an error distribution in linear models, these functions have a general shape and multiple parameters that affect the exact shape. Most models have parameters relating to the sill, the intercept (called a *nugget effect*), the slope, and the distance to reaching the sill (called the *range*). The choice of parametric semivariogram function is based on both dimensionality of the study region and the expected underlying covariance structure. For example, of the models listed above only the Cardinal-Sine semivariogram can model negative correlations.

Estimation of a parametric semivariogram is done by computing estimates of its

parameters. Considering that $E[Z(s_i)] = \mu$, we can rearrange Equation 1 so that

$$2\gamma(h) = \text{Var}(Z(s+h) - Z(s)) = E[(Z(s+h) - Z(s))^2] - E[(Z(s+h) - Z(s))]^2$$

where the second term goes to zero. To estimate the remaining part, we just square and sum pairs of observations that have the same spatial lag in both distance and direction. The method of moments estimator is

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum N(h)[Z(s_i) - Z(s_j)]^2, h \in \mathbb{R}^2 \tag{4}$$

where $N(h)$ is the set of distinct pairs defined by $s_i - s_j = h$, and $|N(h)|$ is the number of distinct pairs. Note that this approach only gives point estimates at observed vectors of h. Assuming isotropy can remove the directional requirement and increase the number of eligible pairs. For irregularly spaced data, the problem of having enough pairs at a given lag can be helped by specifying a *tolerance region* where the pairs are grouped ahead of time into distance regions, and the estimate for a region is assigned to the region's average $\|h\|$.

Note that the empirical semivariogram $\hat{\gamma}$ may not be conditionally nonnegative definite, and certainly will not have values for all possible h. Thus, we want to fit some of those parametric models to our empirical points. One approach to model fitting is based on least squares, where we try to find the values of the parameters $\boldsymbol{\theta}$ that minimize $\sum_{j=1}^{K} [\hat{\gamma}(h_j) - \gamma(h_j|\boldsymbol{\theta})]^2$. Unfortunately, the values of $\hat{\gamma}$ are not independent and have heterogeneous variance (at least due to differing number of lags going into each lag region). A generalized least squares (GLS) approach is then needed. The pure GLS approach requires a covariance matrix and ends up needing the fourth-order moments of the random field, so the simplified weighted least squares method may be better. This approach works by

choosing $\boldsymbol{\theta}$ to minimize the weighted residual sum of squares (WRSS)

$$WRSS(\boldsymbol{\theta}) = \frac{1}{2}\sum_{j=1}^{K} \frac{N(h_j)}{[\gamma(h_j|\boldsymbol{\theta})]^2}\left[\hat{\gamma}(h_j) - \gamma(h_j|\boldsymbol{\theta})\right]^2. \tag{5}$$

Alternatively, the maximum likelihood (ML) approach can be used to estimate semivariogram parameters. If we assume that $Z(s_1), \ldots, Z(s_M)$ are a vector drawn from a multivariate Normal distribution $N_M(\boldsymbol{\mu}, \Sigma(\boldsymbol{\theta}))$, then the MLEs are the values that minimize

$$l(\boldsymbol{\theta}) = \log(|\Sigma(\boldsymbol{\theta})|) + (\mathbf{Z} - \mathbf{1}\mu)^T \Sigma(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{1}\mu) + N\log(2\pi). \tag{6}$$

A usable approximation of $\Sigma(\boldsymbol{\theta})$ is the inverse of the Fisher information matrix. One benefit of using the MLEs is that one can use a likelihood ratio test to compare nested models. Of course, Akaike's information criterion (AIC) can also be used to choose a model for the semivariogram.

Recall that the AIC works by finding the model that gives the highest likelihood, penalized by the number of parameters in the model to prevent overfitting. Mathematically,

$$\text{AIC} = 2\ln(k) - 2\ln(L)$$

where k is the number of parameters and L is the maximum value of the likelihood for the given model. In this case, the 'best' model is the one with the smallest (most negative) value for the AIC. Another option is the Bayesian information criterion (BIC), defined as

$$\text{BIC} = k\ln(n) - 2\ln(L)$$

where n is the sample size. Again, the smaller the value of the BIC, the 'better' the model.

There are methods to estimate semivariograms for anisotropic random fields as well.

Anisotropy comes in two forms: geometric and zonal. In *geometric anisotropy*, the different directional semivariograms have the same shape and sill, but the range changes with direction. We also assume that the maximum range of all directions a_{max} occurs at direction ϕ, and the minimum ranges occur at $\phi \pm 90^o$, making an ellipse. Therefore, we can calculate the *reduced distance* between two observations by taking a transformation that effectively turns the ellipse into a unit circle.

We consider *zonal anisotropy* to be where the range is constant but the sill changes over directions. This model is useful when one direction (such as elevation) is fundamentally different from the others. One method for dealing with zonal anisotropy is to model the semivariogram as the sum of two semivariograms, one isotropic with sill c_{min} and range a and the other with constant sill $c_{max} - c_{min}$ and an anisotropic range. This range is a in the direction of ϕ and overwhelmingly large perpendicular to that, which makes that component effectively nil in the final model.

Often times it is of interest to interpolate values of $Z(s)$ that were not observed based off of locations that did have their value of Z recorded. The most common method for accomplishing this is *Kriging*. In essence, Kriging makes a prediction based off of a weighted average of the observed values $\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i)$ where the weights λ_i are a function of the semivariogram. Many variations of Kriging exist to deal with issues such as surface smoothness, probability maps, and regional exposure data.

Spatial Regression Models. In the previous section, we discussed how to define the spatial correlation in a single spatially distributed variable using semivariogram models. However, it is common for data to contain multiple variables distributed across space with the research question of how they relate to each other. According to Waller and Gotway[37], the most common way to deal with this situation is with spatial regression models.

The simplest model we can use for spatial data is independent multivariate regres-

sion, given by the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{7}$$

where Y are the N outcomes at locations $s_1, ..., s_N$, X is an N x p covariate matrix, β are the regression parameters, and $\boldsymbol{\epsilon} \sim Normal_N(\mathbf{0}, \boldsymbol{\Sigma})$. Note that covariate matrix X contains the possible predictors measured at the same locations as the outcomes. The variance-covariance matrix $\boldsymbol{\Sigma}$ can be defined as

$$\boldsymbol{\Sigma} = \text{Var}(\mathbf{Y}) = \text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I} \tag{8}$$

which means that each of the observations are independent. For this model to be valid, all of the spatial variation in Y needs to be explained by the covariates in X. This is a likely invalid assumption, so further modifications will be necessary.

By assuming a multivariate normal distribution for the outcome variables, $\mathbf{Y}$, the β parameters and σ^2 can be estimated with maximum likelihood methods. Specifically, by setting the derivative of the log-likelihood to zero we can derive the *score equations*. For independent observations, the results of $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\hat{\sigma}^2 = [(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})]/[N-p]$ are identical to ordinary least squares (OLS) estimation. To make inferences with the model, note that the variance of the parameter estimates can be found through the *information matrix* to be $\widehat{\text{Var}(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ which can be used for t-tests or confidence intervals on the $\hat{\boldsymbol{\beta}}$s. Also, note that often times it is preferable to use the *restricted maximum likelihood* (REML) instead, as it gives less biased estimates than traditional ML in these cases of correlated observations.

Although REML was popularized for use in repeated measures studies by Harville (1977), it can be used in spatial data as well since the underlying difficulty in the data (correlation) is the same. In these cases REML works by taking the likelihood of a set of contrasts of the outcome rather than the outcomes themselves. The contrasts, defined

as the matrix $\mathbf{u}$, are such that $E[\mathbf{uY}] = \mathbf{0}$; this reduces the number of effective observed outcomes which corrects for the loss of degrees of freedom due to estimating the $\boldsymbol{\beta}$s.

In most cases, the assumption of independent observations conditional on the spatial covariates is unrealistic. There is almost always some residual spatial autocorrelation among the outcomes that is not accounted for by the covariates. Instead, we consider where

$$\text{Var}(\boldsymbol{\epsilon})_{ij} = \Sigma_{ij} = \text{Cov}\left(Y(s_i), Y(s_j)\right) \geq 0 \tag{9}$$

represents the residual spatial correlation between the outcomes at s_i and s_j. We stick with the overall regression model where $\mathbf{Y} \sim Normal_N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ but now we consider $\mathbf{X}\boldsymbol{\beta}$ to be the *large-scale variation* or *spatial trend* in the outcome and the $\boldsymbol{\epsilon}$ to be the *small-scale variation*.

In order to compute estimates for β, the covariance structure needs to be known. A simple model assumes that the overall structure is known up to a constant, with $\boldsymbol{\Sigma} = \sigma^2\mathbf{V}$ where $\mathbf{V}$ is known. This modification to the least squares approach results in the estimates being $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ and $\hat{\sigma}^2 = [(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})]/[N - p]$.

More commonly, we don't know what V is either. Even if we make the simplifying assumption of constant variance ($\Sigma_{ii} = \sigma^2$), there are still N(N-1)/2 terms to estimate in the matrix. Since we only have N observations, additional structure needs to be assumed for $\boldsymbol{\Sigma}$. A common approach to this problem is to choose and fit a parametric covariance function. Recall equation 2.3, such that $\text{Cov}(s_i, s_j) = C(s_i - s_j) = C(0) - \gamma(s_i - s_j)$; thus, by choosing a semivariogram model we can drastically reduce the number of covariance components, denoted $\boldsymbol{\theta}$, we need to estimate.

Estimation of these semivariograms is more difficult than before, though, since the presence of the $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ trend violates the assumption of stationarity. $\mathbf{Y}$ is no longer stationary since its mean is no longer constant over the study region, but rather the mean is

a function of the values of $\mathbf{X}$ at a certain point. Equation 2.1 now becomes

$$E[(Y(s_i) - Y(s_j))]^2 = 2\gamma(s_i - s_j) + \left[\sum_{k=0}^{p} \beta_k[x_k(s_i) - x_k(s_j)]\right]^2$$

so we can see that the addition of a trend introduces a bias. This bias could be corrected for if $\boldsymbol{\beta}$ was known, but we need $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ to estimate $\boldsymbol{\beta}$. Therefore, the two sets of parameters need to be estimated simultaneously. This can be done using either iteratively reweighted generalized least squares (IRWGLS) or maximum likelihood.

IRWGLS works much like the EM algorithm and has the following steps:

1. Guess the values of $\hat{\boldsymbol{\beta}}$.
2. Calculate the residuals as $\mathbf{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.
3. Use the residuals to estimate the parameters of the chosen semivariogram model and get an estimate of the covariance matrix, $\widehat{\boldsymbol{\Sigma}(\boldsymbol{\theta})}$.
4. Re-estimate $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\widehat{\boldsymbol{\Sigma}(\boldsymbol{\theta})}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}(\boldsymbol{\theta})}^{-1}\mathbf{Y}$.
5. Repeat steps 2-4 until convergence criteria are met (the estimates stop changing).

Unfortunately, the semivariogram estimation from residuals in step 3 contains bias. Hopefully this bias would be small, but maximum likelihood may be preferable.

Maximum likelihood estimation is more difficult than in the simple ordinary least squares model, since there is typically not a nice closed form for the score equations for the derivative with respect to θ. Therefore, iterative algorithms such as Newton-Raphson are needed to find the values of $\hat{\boldsymbol{\theta}}$ that maximize the log-likelihood. A common trick to reduce the number of parameters to fiddle with at any given iteration is to maximize the *concentrated log-likelihood*, which is found by plugging $(\mathbf{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Y}$ in for $\boldsymbol{\beta}$ in the likelihood function. By doing so one needs to only maximize over the different θs, since by the invariance property of MLEs one can get the MLE for $\boldsymbol{\beta}$ by plugging $\hat{\boldsymbol{\theta}}$

into the previous formula. Furthermore, the variances of the MLEs can be found from the information matrix. In addition to these nice properties, one can also use a likelihood ratio test to compare nested parametric models.

Regardless of how the parameters are estimated, care needs to be taken in interpreting the results. Due to how the spatial variation in Y is divided up based on covariates and a predicted correlation structure, one can actually get a similar fit for two models with different covariates and covariance matrices. It should also be noted that model fitting may be problematic for noisy outcomes with a set of predictors that do not well explain the outcome variability.

If interpolation of outcome values is desired, the regression model can be used via the *universal Kriging predictor*. This form of Kriging is similar to before, but the predicted value must also contain a spatially weighted average of the $\mathbf{X}\hat{\boldsymbol{\beta}}$. This method is preferable to ordinary Kriging methods that ignore covariates, or even models that subtract the covariate effect, as spatial covariates may greatly reduce or simplify the covariance in the outcome.

We must also consider the assumptions of the linear model with spatial data. The traditional general linear model assumes independence, but the observations are no longer independent due to the spatial correlation. The revised assumptions of this model are:

1. Homogeneity of variance among the outcomes. $\text{Var}(\mathbf{Y}) = \Sigma_{ii} = \sigma^2$ for all i.
2. The outcome has the specified covariance structure. Note the lack of independence, and that an estimated structure is something the statistician is assuming.
3. Linearity between predictors and the outcome, as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$.
4. Existence of the error variance, and that the mean of $\boldsymbol{\epsilon}$ is zero.
5. The outcomes have a multivariate normal distribution.

As we can see, it is still quite similar to the set of assumptions of the general linear model. Of course, we still need to examine the residuals after fitting the model to check these model assumptions.

One limitation of semivariogram-based estimation of the covariance matrix is that it relies on having the exact distances between observations. When one has summary data for entire regions in space, the concept of distance is less useful since inter-centroid distances are not always representative of how things are truly related in space. One approach is the use of *autoregressive* models, where the outcome for one region is regressed on nearby outcome values.

By regressing the residuals in the model on one another simultaneously, we get the formula

$$\epsilon(s_i) = \nu(s_i) + \sum_{j=1}^{N} b_{ij}\epsilon(s_j), \quad b_{ii} = 0 \tag{10}$$

where b_{ij} are the spatial dependence parameters and the residuals of the residuals $\nu(s_i)$ are independently distributed with mean zero and variance σ_i^2. Combining this result with equation (2.7) gives us the new autoregressive model

$$Y(s_i) = \mathbf{X}(s_i)'\boldsymbol{\beta} + \sum_{j=1}^{N} b_{ij}[Y(s_j) - \mathbf{X}(s_j)'\boldsymbol{\beta}] + \nu(s_i). \tag{11}$$

Thus, we now have where the spatial dependence not explained by the covariates is encapsulated in the weighted average of these spatial deviations. The non-spatial variability is taken up by the ν term. The parameters now affecting the variance/covariance of Y are the $\{\sigma_i^2\}$ and the $\{b_{ij}\}$, such that the covariance matrix of Y can be expressed as

$$\boldsymbol{\Sigma} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Sigma}_\nu(\mathbf{I} - \mathbf{B}')^{-1} \tag{12}$$

where $\mathbf{B}$ is the matrix of b_{ij} and $\boldsymbol{\Sigma}_\nu = diag\{\sigma_1^2, ..., \sigma_N^2\}$.

As before, we would like to impose a structure on $\mathbf{B}$ so there will be a reasonable number of parameters to estimate. We can do so by assuming that a general relationship structure scaled by a constant is known, such as $\mathbf{B} = \rho\mathbf{W}$. Another way to simplify the spatial dependence parameters is to restrict them to only have non-zero values when in a certain neighborhood set describing nearby regions. The resulting model can be described by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} - \rho\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{Y} + \boldsymbol{\nu}. \tag{13}$$

We can see that in this model the residual spatial variability is explained due to spatially lagged components based on the covariates and other outcome values in neighboring regions. Other common simplifying assumptions are that $\mathbf{Y}$ is multivariate normal and $\boldsymbol{\Sigma}_\nu = \sigma^2\mathbf{I}$.

Estimation of the parameters within $\mathbf{B}$ can be done by maximizing the concentrated log likelihood as described above. Similarly, the variances of these parameters can be found through the information matrix. It should be noted that assumptions about the structure of $\mathbf{B}$ that reduce the number of parameters can greatly reduce the difficulty in finding valid estimates.

It is also important to test whether this autoregression is even necessary. After all, correlation between observations reduces your power since each observation contains less unique information. The simplest way to check for spatial autocorrelation is by assuming $\mathbf{B} = \rho\mathbf{W}$ and testing the null hypothesis that $\rho = 0$. This can be done in one of three ways. The first involves using Moran's I to check for spatial dependence in an OLS model. The second is to take the variance estimate for $\hat{\rho}$ from the information matrix and do a Wald's test. Lastly, a likelihood ratio test can be performed comparing the autoregressive model with the OLS model since (under the simplifying assumption mentioned above) the two are nested.

In the models before, we are specifying the joint probability of all of the observa-

tions in $\mathbf{Y}$. It may be useful to look at $Y(s_i)$ as being conditional on $\mathbf{Y}_{-i}$. We can then discuss the conditional mean

$$E\left[Y(s_i)|\mathbf{Y}_{-i}\right] = \mathbf{x}(s_i)'\boldsymbol{\beta} + \sum_{j=1}^{N} c_{ij}[Y(s_j) - \mathbf{x}(s_j)'\boldsymbol{\beta}] \tag{14}$$

and variance

$$\text{Var}[Y(s_i)|\mathbf{Y}_{-i}] = \sigma_i^2 \tag{15}$$

where c_{ij} are the spatial dependence parameters with $c_{ii} = 0$ and $c_{ij} = 0$ when s_j and s_j are not in the same neighborhood set. In other words, if two points are too distant from one another (not in the same neighborhood) they have no dependence on one another. The joint distribution of all outcomes can be derived from the individual conditional probabilities, provided they are Gibbs random fields (Hammersley-Clifford theorem). The theorem will hold true under the assumption of Gaussian conditional distributions and a multivariate normal joint distribution. The variance of this multivariate distribution is

$$\boldsymbol{\Sigma} = (\mathbf{I} - \mathbf{C})^{-1}\boldsymbol{\Sigma}_C \tag{16}$$

where $\boldsymbol{\Sigma}_C = diag(\sigma_1^2, ..., \sigma_N^2)$. There is the restraint that $\sigma_j^2 c_{ij} = \sigma_i^2 c_{ji}$ to ensure a valid symmetric covariance matrix. We can also note that when variance σ^2 is constant among Y's, we can relate the conditional model to the simultaneous one by $\mathbf{C} = \mathbf{B} + \mathbf{B}' - \mathbf{B}\mathbf{B}'$.

Because of the close relationship between simultaneous and conditional autoregressive models, the variance components of the conditional autoregressive models can be estimated in the same manner. One benefit to the conditional model is that the least squares estimators are now consistent, although you can get the same plus more from maximum likelihood with some additional computation.

In summary, spatial autoregressive models can be quite useful for describing the relationships between spatial variables. Great care should be taken in deciding what model

of autocorrelation to use, as the choice greatly affects the results. It is wise to try several different models and compare them to decide which best represents the data itself. Later on, we shall see a modern example of how an analysis of cardiac imaging data approaches this problem of multiple candidate correlation structures.

Longitudinal Data Analysis

Longitudinal data, also referred to as repeated measures, deals with analyzing data collected at multiple time points from the same subject. Specific methods are needed to account for the correlation within a subject over time as the measures are no longer independent. In this way, longitudinal data presents similar problems as spatial data; the main difference is that there is only one dimension and that the relationships are truly just going forward in time. Quite often, it is the temporal relationship itself that a study is interested in quantifying. For example, a drug study may be interested in the trajectory of a patient's cholesterol while on a certain statin. As each patient's measure will have a unique time response, typically the overall goal of longitudinal methods is to summarize the response in some generalizable way.

There are two common complications to longitudinal data. The first is when the observations are unevenly spaced, such that the length of time between observations is not constant over the study period. This can make estimation of covariances between observations difficult, as two adjacent pairs of observations will probably not have the same correlation if the difference in time is different. This is compounded when different subjects are observed at different time points. For example, subject A has their second visit after five weeks while subject B has theirs after only four weeks. The other complication is from missed observations, which leave a gap in the timeline for a patient or a truncated response curve. Both types of complication are quite common in real longitudinal datasets and should always be addressed at the risk of losing information and power or misrepresenting the data itself.

Here, we will describe four general types of methods for dealing with continuous longitudinal data as well as discuss their strengths and weaknesses. The four types are univariate reduction, MANOVA with correlated errors, mixed models, and growth curve modeling.

Note that we are not considering the approach of general estimating equations (GEEs). Although most commonly seen in discrete data analysis, this method is capable of analyzing correlated data with continuous outcomes as well. GEEs work by considering a 'quasi-likelihood,' where a distribution is considered for the marginal mean of the outcome rather than the outcome itself[3]. Previous work has shown that the estimators from GEEs are fairly efficient, but not as efficient as pure likelihood-based methods[13]. Since in our assumed model we will have the full likelihood expression, it is unnecessary to use the quasi-likelihood and thus we will avoid using GEEs in favor of the more efficient maximum likelihood approach.

Univariate Reduction and Area Under the Curve. The simplest approach to longitudinal data is where all of the repeated measures are reduced to a single variable[13]. Although this method makes reporting and interpreting a result extremely simple, it comes at the cost of losing a great deal of detail and information about the actual shape of the temporal relationship. Specifically, two subjects with different temporal trajectories can end up with the same summary value as their outcome. Furthermore, one generally collects information about covariates at each time point in addition to the outcome and it is unclear how these time-varying covariates can be translated into a single appropriate value for analysis.

Matthews *et al.* (1990) consider there to be two common types of ways to summarize longitudinal data: slopes and areas under the curve (AUC)[24]. For the first, one calculates the slope of a straight line fit to the response variable versus time for each subject, and uses that as the subject's outcome variable. This approach is limited to cases where the time response is approximately linear, which may not always be the case. It should be

noted that the widely used pre-post t-test could be considered a special case of this method for only two time points.

The AUC is more generalizable than the slope, as it does not assume a linear trend over time. Indeed, it doesn't assume any overarching shape for the temporal trajectory: it just calculates the area under a subject's response curve for the course of the study. When calculating the AUC one generally estimates the response curve by connecting observations with a straight line and using the trapezoidal rule. If the curve is assumed to have an exponential shape the logarithmic trapezoidal method can be used instead. In cases where the length of observation is different, such as loss to follow-up, it is acceptable to divide the subject's AUC by the length of the subject's follow-up. There are different definitions for AUC in other settings, such as receiver-operator characteristic (ROC) curves or pharmacodynamics, but for simple longitudinal data with continuous outcomes the trapezoidal rule is almost always used. Matthews *et al.* suggest choosing a summary method that relates to your research question; a study on growth may prefer using the slope, while a study that looks at overall outcome may prefer the AUC.

A benefit of AUC is that in the presence of missing observations one can just have the curve connect across the gap; this can have a negative impact on the variance. Although the lack of an assumed shape is a strength of the method, the lack of information regarding the shape is a definite weakness. Furthermore, there could be an infinite number of possible curves that give the same AUC value, making the measure much less useful when the shape of the time response is heterogeneous among subjects. Unequal length of follow-up should also be considered when calculating the AUC.

ANOVA/MANOVA with Correlated Errors. Another class of methods are based on the Analysis of Variance (ANOVA) design, described in *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods* (2008)[13]. Indeed, repeated measures ANOVA is one of the earliest methods used to handle longitudinal data[39].

Let us first define Y_{ij} as the outcome variable for subject i at their j^{th} observation, $\boldsymbol{\beta}$ as a vector of fixed parameters, and $\mathbf{X}_{ij}$ as the ANOVA design matrix. The model itself is

$$Y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + b_i + e_{ij} \tag{17}$$

where the random effect for individual i is $b_i \sim N(0, \sigma_b^2)$ and the error term $e_{ij} \sim N(0, \sigma_e^2)$. Recall, ANOVA partitions the variance into a part explained by the model and a part unexplained by the model. Repeated measures ANOVA seeks to further partition that unexplained variance into variance within a subject and left-over error variance. A common approach is to impose compound symmetry on the error structure within an individual, such that $Var(Y_{ij}) = \sigma_b^2 + \sigma_e^2$ and $Cov(Y_{ij}, Y_{ik}) = \sigma_b^2$. This assumption may not be the valid as compound symmetry is only applicable when the effect is randomly allocated, which is not the case in repeated measures where we presume that the temporal correlation decreases between more distant observations. The assumption of constant error variance over time and subjects may also be unrealistic. Furthermore, the model is not flexible to missing values and does not account for uneven spacing, so as a whole this model is limited.

Regular multivariate ANOVA (MANOVA) lets us analyze multivariate outcomes, which can be extended to this problem since longitudinal data is essentially multivariate with temporal correlation. This extension is referred to as repeated measures MANOVA. A notable variant of this approach was developed by Box in 1950 and is based on the transforming the outcome to multiple polynomial contrasts of the repeated-measures. This variant, called *profile analysis*, is flexible to different shaped time responses as it does not care what the time-response truly is: the behavior over time is totally encompassed in the construction of the linear contrasts. It also does not require the restrictive compound-symmetry assumption on the covariance; the only assumption is that the covariance is the same for each subject. Unfortunately, to achieve all of these nice properties it requires that the data must be balanced and not have missing data[22].

A similar approach exists based on a linear model, described by Albert (1999)[3]. Let us consider a px1 vector of covariates as $\mathbf{X}_{ij}$ with a normally distributed outcome Y_{ij}. A potential linear model is thus

$$Y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij} \tag{18}$$

where ϵ_{ij} is a time-series error structure. Quite often, one of the variables in $\mathbf{X}$ is time which allows for us to sync the discrete j^{th} observation with the continuous time at which it occurred. The trick then comes in the definition of the error term. One potential solution is an autoregressive moving-average (ARMA) model, defined as

$$\epsilon_t = \theta \epsilon_{t-1} + \gamma a_{t-1} + a_t$$

where $a_t \sim N(0, \sigma_\epsilon^2)$, θ is the first-order autoregressive effect, and γ is the moving-average effect. This allows for the repeated measures to be linked together through time, since the random term is a function of the previous errors. In essence, the moving average represents how the dependent variable moves over time as a random walk. Unfortunately, this method does not do well with missing and irregularly spaced observations either. Inclusion of time as a covariate also implies that one is explicitly defining the time-outcome relationship, and the assumed relationship may not be valid across all subjects.

Mixed Models. One common approach is to use random-effects models, where the parameters of a linear model are random variables with their own distribution. For example, an outcome may have a linear relationship with time, but the slope of that relationship varies from person-to-person. Albert (1999) describes the difference between fixed and random effects anecdotally; fixed effects are covariates that are explicitly fixed (like treatment group assignment or baseline values) while random effects are things that can produce varying results between subjects. Random effects can also capture individuals' deviations from their group averages[3].

The random-effects model works by taking a first stage where the linear model is conditioned on the random effects and then taking a second stage where the random effects are given a distribution (usually multivariate normal). This results in a multivariate normal outcome with a particular covariance matrix. Some benefits include: no requirement of balanced data, can explicitly model within- and between-subject variability, and the ability to control for covariates. The main difficulty is in determining the covariance structure and calculating its variance components.

The use of a mixed model for longitudinal data was first formally defined by David Harville in 1977[18]. In 1982, Nan Laird and James Ware developed a more general version, which is the focus of the discussion below[20].

Let us define the outcome for the i^{th} individual as $\mathbf{y}_i$ for $i = 1, ..., m$, where $\mathbf{y}_i$ is composed of n_i observations and has a $N_{n_i}(\boldsymbol{\mu}_i,\ \boldsymbol{\Sigma})$ distribution. We define the stage one model as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{19}$$

where $\mathbf{X}$ is the design matrix for the p population parameters $\boldsymbol{\alpha}$, $\mathbf{Z}$ is the design matrix for the k individual parameters in $\mathbf{b}_i$, and error terms $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{R}_i)$ with $n_i \times n_i$ covariance matrix $\mathbf{R}_i$. Note that the error terms are independent of the parameter vectors. The second stage gives $\mathbf{b}_i$ the distribution $N_k(\mathbf{0}, \mathbf{D})$. The marginal distribution of the y's thus has mean $\mathbf{X}_i\boldsymbol{\alpha}$ and covariance matrix $\mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$. A simplifying assumption would be to set $\mathbf{R}_i = \sigma^2\mathbf{I}$, making the observations independent. This is called the *conditional-independence model*.

We define the set $\boldsymbol{\theta}$ to be the variance components of $\mathbf{R}_i$ and $\mathbf{D}$. Therefore, to fully specify the model from the data we need to be able to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. Note that the structure of $\mathbf{D}$ is user-defined and thus has a variable number of parameters to estimate. One structure proposed by Harville is that $\mathbf{D} = diag[\theta_1\mathbf{I}, ..., \theta_c\mathbf{I}]$, where the random effects are divided into c independent groups he called 'levels.'

If the variance components are known, estimation of $\boldsymbol{\alpha}$ and $\mathbf{b}_i$ is simple and is done with the following formulas based on maximum likelihood:

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^{m} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i, \tag{20}$$

and

$$\hat{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\alpha}}). \tag{21}$$

Laird and Ware derived formulas for the variance of the two estimators as well. If we know $\mathbf{V}_i$, the computation is straightforward. Of course, it is a far more common case where the variance components must be estimated as well. This results in a more difficult situation, where $\boldsymbol{\alpha}$ must be known to calculate $\hat{\boldsymbol{\theta}}$, but $\boldsymbol{\theta}$ is needed to work out $\hat{\boldsymbol{\alpha}}$. Thus, the two sets of parameters must be estimated simultaneously. This is typically done by maximizing the maximum likelihood (ML) or restricted maximum likelihood (REML) using numeric methods.

Assuming the conditional-independence model, and that $\mathbf{D}$ is a nonnegative-definite matrix, Laird and Ware also found that from maximum likelihood

$$\hat{\sigma}^2 = \sum_{i=1}^{m} \mathbf{e}_i^T \mathbf{e}_i \Big/ \sum_{i=1}^{m} n_i = t_1 \Big/ \sum_{i=1}^{m} n_i \tag{22}$$

and

$$\hat{\mathbf{D}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{b}_i \mathbf{b}_i^T = \mathbf{t}_2 / m, \tag{23}$$

so the sufficient statistics for $\boldsymbol{\theta}$ are t_1 and the $k(k+1)/2$ unique components of $\mathbf{t}_2$. Taking the expectation of these statistics, we get

$$\hat{t}_1 = E \left\{ \sum_{i=1}^{m} \mathbf{e}_i^T \mathbf{e}_i | \mathbf{y}, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}} \right\} = \sum_{i=1}^{m} \left[\hat{\mathbf{e}}_i(\hat{\boldsymbol{\theta}})^T \hat{\mathbf{e}}_i(\hat{\boldsymbol{\theta}}) + tr\ var\{\mathbf{e}_i | \mathbf{y}, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}\} \right] \tag{24}$$

and

$$\hat{\mathbf{t}}_2 = E\left\{\sum_{i=1}^{m} \mathbf{b}_i \mathbf{b}_i^T | \mathbf{y}, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}\right\} = \sum_{i=1}^{m} \left[\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})^T + var\{\mathbf{b}_i | \mathbf{y}, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}\}\right] \tag{25}$$

where $\hat{\mathbf{e}}_i(\hat{\boldsymbol{\theta}}) = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}) - \mathbf{Z}_i\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})$.

Therefore an EM algorithm can be constructed that follows these steps:

1. Start with an initial value for $\hat{\boldsymbol{\theta}}$.
2. E-step: Calculate $\hat{\mathbf{t}}$ from the current value of $\hat{\boldsymbol{\theta}}$ by equations (24) and (25).
3. M-step: Using the values of $\hat{\mathbf{t}}$ from the E-step, calculate $\hat{\boldsymbol{\theta}}$ from equations (22) and (23).
4. Cycle through E- and M-steps until convergence.

Like most implementations of an EM algorithm, this approach is very powerful. Unfortunately, in practice it can be slow to converge, especially when the parameter is near the edge of the parameter space. This can be difficult with the constraint that $\sigma^2 > 0$ or $\theta_i > 0$, since for distant observations the estimate covariance may be quite small and actually near zero.

It should be noted that maximization of the likelihood can also be done using the Newton-Raphson algorithm, as discussed by Harville. This carries the usual difficulties of the Newton-Raphson algorithm, such as ensuring the global maximum is found instead of a local one, and the imposition of the parameter space. It also requires the calculation of second derivatives, which may result in a the method requiring great deal more work to implement than the EM algorithm.

Restricted maximum likelihood is based off of the reduction of the likelihood function to adjust for the loss of degrees of freedom in small datasets and large number of parameters. Laird and Ware described a scenario where the REML function is the marginal

distribution of $\mathbf{y}_i$ and $\boldsymbol{\theta}$, and the parameter estimates can be found using the above EM algorithm with the E-step slightly changed to reflect how things are no longer conditional on $\hat{\boldsymbol{\alpha}}$. Conversely, Harville defines the REML as being composed of N-p contrasts. This results in a much more complicated likelihood function, and may require the use of approximations for $\hat{\boldsymbol{\alpha}}$ and $\hat{\mathbf{b}}_i$ if the new function is too complex. Estimation of Harville's REML-based parameters follows the same procedure as above.

Laird and Ware's version of the mixed model is generally better to use, as it allows for a greater flexibility in the definition of the covariance structure compared to Harville's independent 'levels.' The design matrices are similarly less restricted under the Laird-Ware model. In the case of small sample sizes, the REML version of estimation is suggested to correct for possible bias.

Growth Mixture Modeling. As defined in *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, the goal of *growth mixture models* is to capture the unobserved heterogeneity between subjects during the time course of the study[13]. The models utilize both random effects and finite mixtures. For a simple model to start, consider for subject i at time point j

$$Y_{ij} = \eta_{0i} + \eta_{1i} a_{ij} + \kappa_i w_{ij} + \epsilon_{ij} \tag{26}$$

where the outcome Y is continuous, w is a time-varying covariate, a_{ij} is the index (1, 2,...,T) of the time of the observation, and η_{0i} and η_{1i} are the random intercept and slope of the growth process, respectively. We may also consider a time-invariant predictor (like gender) as X. κ is a random slope for the time-varying covariate effect. $\epsilon \sim N$ are the residual errors.

The random slopes and intercepts are defined as

$$\eta_{0i} = \alpha_0 + \gamma_0 x_i + \zeta_{0i} \tag{27}$$

$$\eta_{1i} = \alpha_1 + \gamma_1 x_i + \zeta_{1i} \tag{28}$$

$$\kappa_i = \alpha_2 + \gamma_2 w_{ij} + \zeta_{2i} \tag{29}$$

where α and γ are model parameters and $\zeta \sim N$ are residuals. Note that this model can be considered a special case of the traditional mixed model; let us define

$$\Lambda_i = \begin{pmatrix} 1 & a_{i1} \\ 1 & a_{i2} \\ \vdots & \vdots \\ 1 & a_{iT} \end{pmatrix}.$$

Thus, in the traditional model we consider

$\mathbf{X}_i = (\Lambda_i \mathbf{w}_i \Lambda_i x_i \mathbf{w}_i x_i)$	Fixed effect covariates: time-invariant and time-variant, as well as time indices
$\boldsymbol{\beta} = (\alpha_0,\ \alpha_1,\ \alpha_2,\ \gamma_0,\ \gamma_1,\ \gamma_2)^T$	Fixed eff. parameters: parameters from the random slopes and intercept
$\mathbf{Z}_i = (\Lambda_i \mathbf{w}_i)$	Random effect covariates: time-varying covariates and their time index
$\mathbf{b}_i = (\zeta_{0i},\ \zeta_{1i},\ \zeta_{2i})^T$	Random effect parameters: random intercept and slopes residuals
$\mathbf{e}_i = \boldsymbol{\epsilon}_i$	

for the two models to be equivalent.

This model can be further broken into four components, two fixed effects and two random effects. It also relates to factor analysis when $\Lambda_i = \Lambda$. Also consider when $a_{it} = a_t$, which is done to make the time index into a parameter; this is typically done to assess deviation from the linear growth model. Also note that a non-parametric version of the growth mixture model can be fit for non-normal outcomes.

The maximum likelihood estimates of the model parameters can be calculated using the EM algorithm. Muthén and Asparouhov (*Longitudinal Data Analysis*) suggest having the estimation step find the posterior distributions of the $\boldsymbol{\eta}$'s, but a non-Bayesian approach should be functional as well.

An extension of the growth mixture model is the multilevel mixture model. It functions by grouping the subjects into clusters, where subjects within a cluster share fixed and random effects. As such, this results in additional parameters to estimate so the improvement in model fit is offset by a greater loss of degrees of freedom and greater computational cost. Also, it is not entirely clear what the most appropriate procedure is for grouping subjects into clusters and if done by 'eyeballing' the growth curves a great deal of bias could be introduced.

As a whole, growth mixture modeling seeks to add greater complexity to the typical mixed model so that the shape of the growth curve and the effects of covariates can have a finer effect on the outcome in the model. The growth mixture model focuses its efforts on describing the actual curve of the outcome versus time, which is quite different from the temporal correlation structure the mixed model is built around. It seems that the choice between the two should be decided by what the application at hand requires.

Comparison of Longitudinal Methods: Summary Statistics Versus Mixed Models. Despite the prevalence of longitudinal data, there has been very little work done to compare the different choices of method. Although the more complex methods will give a better fit to the data, these models may be more difficult to interpret and certainly spend more of the data on estimating nuisance parameters such as variance components. It is of great interest to know how much power a statistician gives up by using a simpler model such as summary statistics to analyze longitudinal data.

A recent paper by Zucker *et al.* (2012) reported on a simulation study that ex-

amined the power of mixed models and summary statistics[40]. The simulation was set up in the manner of a two-arm repeated measure trial with J observations over time and equal group sizes. The assumption of no missing data was also made. They defined the outcome as being continuous and denoted Y_{ij} for subject i at time t_j with $j = 1, ..., J$. Two different summary measures were used, the unweighted $\frac{1}{J}\sum_{j=1}^{J} Y_{ij}$ and the weighted $\frac{1}{\sum_j t_j}\sum_{j=1}^{J} t_j Y_{ij}$. Two forms of the traditional mixed model were used, with and without the group-specific effect. In other words, with and without the implicit assumption of the baseline values being equal between arms. The unstructured covariance structure was also used. Three scenarios were used, in which the groups always started at the same place but the number of observations varied from 4 to 6 and different variance structures were tried. The scenarios were inspired by three different studies, and were simulated from a mixed model.

When testing for a treatment by time interaction, they found that across all scenarios the mixed model without the baseline treatment difference was the most powerful, with the weighted summary measure close behind. It should be noted that this is not surprising, since the mixed model was used to generate the data in the first place. The mixed model with the treatment-intercept performed worse, far worse in the case with a small number of observations. Zucker *et al.* rationalized this by fitting two parameters for a line is very inefficient when the number of observations is as small as 4, but changing that to just one parameter ameliorates that problem substantially. This result may be of questionable use, however, since their simulation was done on data where there truly was no baseline difference between groups. It is unlikely that these trends would hold when there is an actual difference at baseline. On average randomization should eliminate baseline differences, but for small sample sizes there may be a small difference that could throw off the model fit if unaccounted for.

They found that when testing for a treatment effect on the intercept of the model

(here there were actual baseline differences), the mixed model was generally most efficient. However, in cases where the number of observations is small the summary measures performed better.

As a whole, there seems to be evidence that the summary measures can actually compete with mixed models on longitudinal data, especially when the number of observations is small. It seems that the weighted summary measure is more powerful than the unweighted; this is relevant to the model choices presented before as the AUC is a kind of time-weighted summary measure. The authors noted that work needs to be done in this area that extends these findings to unbalanced data, which we hope to do in terms of missing observations. Their work was not conclusive in deciding between mixed models and summary measures as the most powerful, which suggests that there exists the equipoise needed for us to compare different longitudinal methods in this proposed research.

Spatiotemporal Modeling

It is possible to extend our previous knowledge of spatial processes to spatiotemporal models by simply considering them a special case of spatial data with one additional dimension. For the imaging data, instead of three-dimensions the data is considered to be four-dimensional.

Unfortunately, this approach has several intrinsic limitations. First, time goes forward, so although a later time can be correlated with an earlier one, it does not make sense to say that the previous observation is affected by a later one. Also, defining the lags between observations in time can be more complicated than defining spatial lags. Lastly, although the spatial observations may be effectively continuous over time, temporal observations are generally treated as discrete observations.

Following the theory described in *Handbook of Spatial Statistics* (2010)[15], let us first consider a stationary spatial process with covariance function C_S, whose realizations

change over time with velocity vector $\mathbf{V}$. The resulting process is considered to be stationary with covariance structure $C(h,u) = E[C_S(\mathbf{h} - \mathbf{V}u)]$ with distance vector $\mathbf{h}$ and time u. The *frozen field* model is a special case where velocity is constant over time. Spatial dispersion models are a common implementation of this idea, with most applications in meteorology.

For a normally distributed outcome, let us attempt to decompose spatial Gaussian processes based on their moments by defining the process Y at point s and time t as

$$Y(s,t) = \mu(s,t) + \eta(s,t) + \epsilon(s,t) \tag{30}$$

where μ is a trend function, η is stationary process with mean zero and 'continuous sample paths,' and ϵ is the error term with mean zero. We assume the covariance structure of ϵ to be separable such that

$$Cov[\epsilon(s+\mathbf{h}, t+u),\ \epsilon(\mathbf{s},t)] = aI(\mathbf{h},t) = (\mathbf{0},0) + bI(\mathbf{h}=\mathbf{0}) + cI(u=0) \tag{31}$$

where I is the indicator function and $b, c > 0$. The 'b' term is purely spatial, while the 'c' term is purely temporal.

The simplest form of the trend function is where it decomposes into independent spatial and temporal components ($\mu(\mathbf{s},t) = f(\mathbf{s}) + g(t)$). Here, we can use the standard spatial approaches for the spatial component and assign an appropriate function for the temporal component. Although this allows for flexibility in defining the temporal trend, it also requires the statistician to assume a certain function whose validity needs to be scrutinized.

We define a second-order stationary process to be where the covariance does not

depend on location, such that

$$Cov[\eta(s+\mathbf{h}, t+u),\ \eta(\mathbf{s}, t)] = C(\mathbf{h}, u)$$

where C is the space-time covariance function. The pure spatial and temporal covariances can be found by setting u or $\mathbf{h}$ to 0, respectively. This second-order stationarity is related and equivalent to *strict stationarity*, which is translation invariant meaning the relationship does not change with the location of the points. In the event that the data is not stationary, a transformation or different aggregation may serve to produce approximately stationary results.

Now that space-time covariance structures have been introduced, their properties need to be discussed. A space-time covariance function is considered to be *symmetric* if the direction of a temporal or spatial does not matter. In other words, if

$$C(\mathbf{h}, u) = C(\mathbf{h}, -u) = C(-\mathbf{h}, u) = C(-\mathbf{h}, -u).$$

Inherent directionality in the system, such as a unidirectional current, prevents the system from being symmetric.

A popular and powerful simplifying assumption is one of separability. We consider a space-time covariance function to be *separable* if it can be written as the product of a spatial covariance function and a temporal covariance function. In other words, if C is the product of a spatial function and a temporal function. A benefit of a separable model is that it makes the inversion of the correlation matrix less intensive as it can be expressed as the Kronecker product of spatial and temporal correlation matrices; this is highly desirable when the number of observations per subject grows large and the cost of inverting the matrix becomes high. This creates a simpler model with fewer parameters, but prevents the consideration of a space-time interaction[16]. The implications of violating the sepa-

rability assumption have not been well studied, but work has been done to validate tests of separability[35]. Because of the simplicity of a separable space-time covariance structure, we consider it a good starting point for our model.

Often times, we cannot assume that the space-time covariance function is separable. This occurs when there is a certain interaction between space and time that needs to be addressed in the model. Thus, nonseparable covariance functions have been developed[9]. It should be noted that these typically also assume symmetry for the ease of computation.

In cases of a flow in the outcome variable over space, it is reasonable to think of correlation between two points in time as being able to be represented by the correlation between two locations for a given velocity and identical time lag. This is considered to be *Taylor's hypothesis* which is satisfied if there is a velocity vector $\mathbf{v}$ in space such that $C(0, u) = C(\mathbf{v}u, 0)$. The frozen field model is an example of this. It is uncertain how valid such an assumption is for geometric imaging parameters, as the myocardium does not move to different parts of the left ventricle.

The *product-sum* method breaks up the space-time covariance function into

$$C(\mathbf{h}, u) = a_0 C_S^0(\mathbf{h}) C_T^0(u) + a_1 C_S^1(\mathbf{h}) + a_2 C_T^2(u) \tag{32}$$

where C_S are spatial functions and C_T are temporal. This method allows for specific changes in the covariance function due to space or time[10].

A different option is

$$C(\mathbf{h}, u) = \phi\left(\sqrt{a_1 ||\mathbf{h}||^2 + a_2 u^2}\right) \tag{33}$$

where ϕ is a continuous function such that $\phi(0) = 1$, which creates spatial anisotropy and satisfies Taylor's hypothesis. The limitation to this method is that it assumes that space and time are interchangeable, so that a certain change in space can be replicated

with a proportional change in time. The use of a single 'spatiotemporal distance' greatly simplifies the covariance structure, but prevents the situation where space and time have unique effects on covariance.

Previous Methods for Spatiotemporal Analysis of Longitudinal Imaging Data

In recent years there has been some research to develop statistical methods that can model the spatiotemporal correlation inherent in longitudinal imaging data. Two major works stand out, however: Bowman and Waller's work from 2004 and the work by Simpson *et al.* from 2014.

Cardiac Imaging Analysis by Bowman and Waller. Bowman and Waller (2004) were concerned with the unique analytical problems that are associated with using biomedical imaging for research, specifically the use of single photon emission computed tomography (SPECT) to image perfusion of the heart[5]. This imaging modality gives a three-dimensional image of the hearts perfusion, which was used by the researchers to examine the effects of a myocardial infarction on perfusion. The study looked at perfusion of the left ventricle at rest and during stress at two time points, one two days after the infarct and the other one year post-MI. The overall goal was to characterize both the changes in perfusion under different physiological conditions after an infarct and how remodeling changes perfusion throughout the left ventricle over time. This study is highly relevant to this research, as the goal of making a spatiotemporal model that can deal with cardiac MRI data is essentially the same as what we want to do. However, our data has more than two time points, and examines multiple different physical characteristics of the left ventricular myocardium other than perfusion.

Statistical challenges arise since this data involves all of the 600 voxels in a three-dimensional image. Also, each subject and condition typically has roughly two-dozen

images taken to reduce error. Although one could conceivably have a cardiologist make a categorical interpretation of the hearts image as an outcome, this approach is less powerful than a quantitative assessment and is subject to error with how the image is interpreted. The traditional approach has been to run *t*-tests at each voxel and create a map of the *t*-statistics. The problem with this approach is twofold. First, an image has a large number of voxels so the issue of multiple testing inflating the Type I error becomes a real concern. Second, the values at one voxel are known to be related to the values of the surrounding voxels, so this spatial correlation in the data makes the implicit assumption of independent voxels in the *t*-test approach invalid. Furthermore, the repeated examination of a patients heart raises the concern of temporal correlation within a patient from one physiological condition to another and from the first year to the second.

Their analysis looked at 1 patient under four conditions pertaining to stress/rest crossed with 2 days/ 1 year post-MI. The images were comprised of 600 voxels, which were condensed to 20 sectors in three cylindrical rings with a hemispherical cap. These four regions related to the regions from the base to the apex of the left ventricle. They could also be divided up into thirds based on what major coronary artery fed that part of the left ventricular wall (left anterior descending, left circumflex, right coronary artery). Note that at each condition, the image was taken multiple times to make a more stable estimate for each sector; a typical sector had 20-22 samples.

The model proposed by Bowman and Waller is a mixed effects model. The model for the perfusion $\mathbf{Y}_k$ for subject *k* is

$$\mathbf{Y}_k = \mathbf{X}_k\boldsymbol{\beta} + \mathbf{Z}_k\mathbf{d}_k + \mathbf{e}_k \tag{34}$$

where $\boldsymbol{\beta}$ is the fixed effect parameter vector, $\mathbf{X}_k$ are the fixed effect covariates for subject k, $\mathbf{Z}_k$ are the random effect covariates for subject k, $\mathbf{d}_k$ is the random effects parameter vector, and $\mathbf{e}_k$ are the error terms. Furthermore, it is assumed that $\mathbf{d}_k \sim N(\mathbf{0}, \boldsymbol{\Delta})$ which is

independent of $\mathbf{e}_k \sim N(\mathbf{0}, \sigma^2\mathbf{V}_k)$. Thus, the overall covariance matrix for $\mathbf{Y}_k$ is

$$\mathbf{\Sigma}_k = \mathbf{Z}_k\mathbf{\Delta}\mathbf{Z}_k' + \sigma^2\mathbf{V}_k \tag{35}$$

Furthermore, the estimator for the standard fixed effects $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k=1}^{K}\mathbf{X}_k'\hat{\mathbf{\Sigma}}_k^{-1}\mathbf{X}_k\right)^{-1}\left(\sum_{k=1}^{K}\mathbf{X}_k'\hat{\mathbf{\Sigma}}_k^{-1}\mathbf{Y}_k\right) \tag{36}$$

and was estimated using iterative algorithms. It should be noted that there we no real covariates in the model, so the βs were just the average perfusion of a certain sector under a certain condition across all of the samples.

The spatial and between-condition correlation was handled through different structures of $\mathbf{V}_k$. One approach dubbed *between sector correlation* worked by defining $\mathbf{V}_k = diag(\mathbf{V}_{k1}, \mathbf{V}_{k2}, ..., \mathbf{V}_{kC})$ treating the C different conditions as independent. $\mathbf{V}_{kc}$ was the unstructured correlation matrix between the sectors and samples in subject k at condition c. A second approach was *within sector correlation* and defined $\mathbf{V}_k = diag(\mathbf{V}_{k\bullet 1}, \mathbf{V}_{k\bullet 2}, ..., \mathbf{V}_{k\bullet 20})$ where $\mathbf{V}_{k\bullet s}$ was the unstructured correlation between the different conditions for a given subject k and sector s. This approach treats the different sectors as being independent. So, is essence the between sector correlation looks at the spatial correlation and ignores the temporal aspect, while the within sector correlation models the temporal relationship while ignoring the spatial aspect. However, it is not clear that the within sector correlation is really representative of the correlation between conditions since the autoregressive structure they used does not make sense in a 2x2 design and seems to instead reflect repeated samples within a given condition.

A third approach utilizes both the random effects and the correlated errors. The initial term $\mathbf{Z}_k\mathbf{\Delta}\mathbf{Z}_k'$ refers to the between sector covariance while the second term, $\sigma^2\mathbf{V}_k$, corresponds to the within sector covariance. The value of $\mathbf{\Delta}$ depends on what form is

chosen for $\mathbf{d}_k$; different choices correspond to different assumptions about the relationship between regions, or the desired number of regions to examine. For instance, one could look at the covariance between the four sides (inferior/anterior, septal/lateral) of the LV instead of the 20 sectors by changing $\mathbf{Z}_k$ to be a matrix of 0's and 1's denoting whether an observation came from a particular side. We typically assume that the covariance of $\mathbf{d}_k$ is the same for all conditions; in other words, that the random effect for each sector is independent of temporal or physiological condition. This can be denoted by $\mathbf{\Delta} = diag(\Delta_c, ..., \Delta_c)$. In this approach, the random effect is needed to account for the covariance between sectors.

The advantage of the mixed model with correlated errors is that you may get a more accurate model, depending on the covariance structure chosen. The main disadvantage is that you need to estimate the variance components of $\mathbf{\Delta}$, which adds computational burden and reduces degrees of freedom. Interpretation of the results may also become more complex since we are adding parameters to the model.

In their example, Bowman and Waller attempted to fit six different covariance structures to their model. Four of them only looked at between sector correlation. One of these used an unstructured covariance structure between all sectors, another assumed compound symmetry ($\sigma_{ij} = \sigma_1^2 + \sigma^2 I[i = j]$), and two let the sectors within a given region (four rings or three artery regions) be unstructured while assuming regions were independent. A fifth model used only within sector covariance with an assumed autoregressive structure, and the sixth utilized both compound symmetric between and autoregressive within sector covariance. The authors proposed that AIC be used to determine which covariance structure was most appropriate.

Interestingly, the 'best' of the six structures was found to be the unstructured covariance structure. This model found that sectors within a region tended to be positively correlated, while sectors in the base were negatively correlated with those at the apex. If this is truly the case, it is not surprising that independence between regions or compound

symmetry failed to fit the data well. Also, the authors noted that they were unable to fit an unstructured between sector model with a within sector autoregressive model due to computational limitations; this finding has implication on this project, such that we will not be able to simply let everything have unstructured covariance.

The work by Bowman and Waller provides a strong base to build a better spatiotemporal model from. Although they attempted to relate observations at two different times and two different conditions, it is not clear how their model handles correlation at successive observations. Only one model was tried with a spatiotemporal covariance structure, and the temporal component appeared to refer to multiple samples in one condition and did not utilize a proper distance-based covariance structure. Unfortunately, our preliminary work suggested that the unstructured-by-autoregressive model they used was not viable for our dataset (it failed to converge) so the Bowman-Waller model is not directly applicable to the SCCOR dataset.

Brain Imaging Analysis by Simpson et al.. More recently, the research group of Simpson, Edwards, Muller, and Styner (2014) have considered a linear model with a separable parametric correlation structure[34]. Their application looked at caudate morphology measured via MRI scans taken at multiple points in time. Much like the 17-segment model we will use for the left ventricle, Simpson *et al.* mapped the MRI data defining the morphology to a 21-segment model of the caudate. The application driving their research was very similar to ours as they were interested in whether medical therapy affected the caudate shape over time in patients with schizophrenia.

The model they proposed is designed for data collected on subjects with S_i spatial and T_i temporal observations for $i = 1, ..., N$. They assume that the observations for subject i have a multivariate normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance $\sigma^2 [\boldsymbol{\Gamma}_i \otimes \boldsymbol{\Omega}_i]$ where $\boldsymbol{\Gamma}_i$ is a $T_i \times T_i$ temporal correlation matrix and $\boldsymbol{\Omega}_i$ is a $S_i \times S_i$ spatial correlation matrix. The values of the correlation matrices are determined by parametric correlation

functions. In their application they used maximum likelihood estimation with the justification of a large sample size ($N = 296$, $S_iT_i \in (63, 147)$). With the exception of using ML estimation instead of REML, their described model is identical to the one we propose. It should be noted that that our respective groups worked independently; their paper was submitted June 2013 and published in February 2014, while this dissertation was proposed in August 2013. Happily, once the coincident research paths were brought to light at ENAR 2014 there has been nothing but positive discussion between the groups.

Although both works describe the same model framework, the paper by Simpson *et al.* focuses on the application of a particular parametric function crossed with itself. The linear exponent autoregressive (LEAR) structure was previously proposed by the group and found to have excellent flexibility for having only two parameters[33]. The LEAR$\otimes$LEAR structure was compared with other crossed structures when applied to the caudate study data in terms of AIC and inference about predictors. The paper did not provide simulation results, with the only conclusion being that for their particular dataset the LEAR model provided superior fit to the other parametric functions considered. We feel that there is still a significant gap into defining the practical considerations of applying such a model that our research seeks to address.

Research Goals

In this section we will discuss the goals of the three papers presented in this dissertation and how they fit together into a body of work.

Paper 1: Selecting a Separable Parametric Spatiotemporal Covariance Structure for Longitudinal Imaging Data

In the first paper, we lay out the theoretical framework for our spatiotemporal model. In particular, we describe the multivariate normal structure of the outcomes and

how it relates to the likelihood function and REML estimation of the parameters. We then outline the main assumptions of the covariance structure such as separability and the use of parametric spatial and temporal correlation functions.

The paper then describes the results from a simulation study that investigated practical considerations for applying such a model. The abilities of several information criteria to select the true correlation structure was quantified. However, in practice one never knows the true structure and it could be argued that there is not a true parametric structure for real data. Therefore, we looked at the effects on Type I error rate and statistical power when making inferences about predictors when a correlation structure that did not necessarily match the one generating the data was fitted. Specifically, inferences about a treatment-by-time interaction were examined as it was most relevant to the goals of the motivating clinical trial. Those results give us an idea of how badly inference can go if an inappropriate correlation structure is used, as well as a sense of what structures provide a "good enough" fit for a certain true correlation structure. To support this effort, we simulated data from several combinations of spatial and temporal correlation structures as well as different degrees of correlation. Sample size was also varied between simulation conditions to assess whether correlation structure misspecification could be powered out of. Furthermore, we considered the Type I error rate when an information criteria-chosen structure was used, giving a quantification for how safe a given criterion is in terms of not falsely rejecting the null hypothesis.

Finally, the first paper considered an application of our model to the UAB SCCOR data. We demonstrate how information criteria can be used to select a working separable correlation structure and how the choice can be checked by comparing multiple information criteria or plotting observed versus predicted correlations. We also examine the inferences made under different working correlation structures to reinforce the assertions made by the simulation study results. As REML was used, we considered how the choice of predictors

can affect the estimation and selection of parametric correlation functions.

Overall, the first paper presents and validates a linear model with a separable parametric correlation structure for use in practice.

Paper 2: Comparing Summary Methods and a Spatiotemporal Model in the Analysis of Longitudinal Imaging Data

The first paper sought to demonstrate our model's reliability and how it could be used in practice, while our second paper looks to quantify the benefit of using our model instead of the previously used summary methods. Although the prospect of using all of the observed imaging data together is highly appealing to clinical investigators, before this stimulation study it was unclear whether or not the complexity of such a model is justified by improved reliability and efficiency. Summary methods were appealing not just because of how they can eliminate the need to model correlation between observations but with how they can sometimes make interpretation easier.

The body of the second paper reports the results of a simulation study that looked at how well our model compared to summary measures in terms of statistical inference about a treatment-by-time effect. Combination of spatial (regional averages analyzed together or separately, and a global average) and temporal (slope, endpoint, and area-under-the-curve analysis) summary measures were looked at for data generated under a linear time course. Like in the previous paper, different generating correlation structures, degree of correlation, and sample size were considered. The results from using a Wald and a Kenward-Roger corrected F-test are compared. Although that paper showed that the Type I error rate of our model was reliably conserved when the true correlation structure is used, it was of keen interest to quantify the gain in power compared to the more easily implemented summary methods. We also show that the common practice of analyzing regional averages separately has extremely poor statistical properties due to the multiple correlated tests performed.

This paper also considered the effects of missing data on the different models. A benefit of some summary measures is to sum or average over missing data, resulting in no loss in the number of observations in the analysis. However, the theory behind them suggests that such averaging can still result in unequal variances which could cause adverse behavior in the standard errors of estimators. We hoped to determine how the summary methods themselves handled missing data when compared to our proposed model.

Paper 3: Applying a Spatiotemporal Correlation Model for Longitudinal Imaging Data

The goal of the first two papers were to convince statisticians that our proposed model has merit and should be used in lieu of the easier summary methods. Conversely, the goal of the third paper is to convince clinical investigators to go along with an analysis plan based on our proposed method. To do so, we explain the function and assumptions of our model in more easily accessible language, and give an example where we apply our model to an outcome from the UAB SCCOR study. A different MRI parameter than the first paper was used, and observations from all subjects were used (compared to only subjects with complete sets as in the first paper).

This paper walks through a full analysis of longitudinal imaging data using our model and points out the steps where clinical and statistical investigators need to work together to produce the best results. This includes selection of predictors, selection of a working correlation structure, and the drawing of inferences. Care is taken to explain the benefit of using time- and space-varying predictors that would be excluded by summary methods. We also examine the results of inference between our model and summary methods when applied to the UAB SCCOR data, and use degrees of freedom to explain how much information is truly lost in the summation. Finally, we dispense advice for investigators when using our model.

SELECTING A SEPARABLE PARAMETRIC SPATIOTEMPORAL COVARIANCE STRUCTURE FOR LONGITUDINAL IMAGING DATA

BRANDON GEORGE, INMACULADA ABAN

Published online by *Statistics in Medicine*
DOI: 10.1002/sim.6324

Format adapted and errata corrected for dissertation

ABSTRACT

Longitudinal imaging studies allow great insight into how the structure and function of a subject's internal anatomy changes over time. Unfortunately, the analysis of longitudinal imaging data is complicated by inherent spatial and temporal correlation: the temporal from the repeated measures, and the spatial from the outcomes of interest being observed at multiple points in a patients body. We propose the use of a linear model with a separable parametric spatiotemporal error structure for the analysis of repeated imaging data. The model makes use of spatial (exponential, spherical, and Matérn) and temporal (compound symmetric, autoregressive-1, Toeplitz, and unstructured) parametric correlation functions.

A simulation study, inspired by a longitudinal cardiac imaging study on mitral regurgitation patients, compared different information criteria for selecting a particular separable parametric spatiotemporal correlation structure as well as the effects on Type I and II error rates for inference on fixed effects when the specified model is incorrect. Information criteria were found to be highly accurate at choosing between separable parametric spatiotemporal correlation structures. Misspecification of the covariance structure was found to have the ability to inflate the Type I error or have an overly conservative test size, which corresponded to decreased power.

An example with clinical data is given illustrating how the covariance structure procedure can be done in practice, as well as how covariance structure choice can change inferences about fixed effects.

INTRODUCTION

One of the most common approaches to analyzing correlated data is the use of linear regression models with correlated errors. Through this method the investigator can specify the covariance structure between observations that can be assumed to be related. For example, in longitudinal studies the successive observations on a subject are considered to have temporal correlation, with observations closer in time being more highly correlated than those further apart. Similarly, in imaging studies the observations from different parts of the same individual are considered to have spatial correlation that follows the same pattern of decreasing correlation with distance. Therefore longitudinal imaging studies need to be analyzed using methods that account for the spatiotemporal correlation in the data. Note that here we refer to longitudinal imaging studies as those that take a handful of successive images days, months, or even years apart; we do not refer to functional imaging studies that take hundreds of images seconds or fractions of a second apart.

In spatial statistics, this relationship between distance and correlation is typically defined by a parametric covariance structure. Commonly used spatial structures include the exponential, spherical, and Matérn structures. For temporal correlation, typically the autoregressive-1 (AR-1) and Toeplitz structures are used to define decreasing correlation with time. The compound symmetric structure can also be used for distance-independent correlation, while the unstructured correlation model allows for the most precise fit to the data. Note that when separability between space and time is assumed, these temporal and spatial structures can be combined in a spatiotemporal model[1]. Even if the correlation between observations is a nuisance and only fixed effect parameters are of interest, use of the proper covariance matrix is essential to prevent the estimates of the standard errors of the parameters from being erroneous[2].

Selecting a Covariance Structure

In practice, however, the investigator will not know *a priori* which covariance function will best represent the data. A three-step approach, proposed by Diggle[3] and refined by Wolfinger[4], starts by (1) choosing fixed effects for the model, then (2) fitting different covariance structures, and (3) finally choosing between covariance structures using either formal testing or examination of the variogram. The variogram is a plot of covariance between observations as a function of distance and is the spatial statistician's preferred option for spatial covariance structure selection[5]. Unfortunately, in spatiotemporal models it can be difficult to 'eyeball' which of several semivariograms would be appropriate since space and time would be on different axes or different graphs. Furthermore, this procedure does not take the complexity of the model into consideration. Most spatial covariance functions will have five or fewer parameters so overfitting is not as much of a concern in that field, but when the unstructured covariance matrix is brought into consideration one must be acutely aware of model parsimony.

Early on, formal testing for choosing covariance structures had been done using likelihood ratio tests (LRT). For example, Schaalje *et al.*[6] looked at an example of how one could choose between common longitudinal correlation structures using the LRT, and Grady and Helms[7] used the LRT to decide between multiple covariance structures and random effects. However, as Grady and Helms noted the LRT is only a valid test for nested models which works decently well in longitudinal structures where everything is nested within the unstructured model and compound symmetry and AR-1 are nested within Toeplitz. Unfortunately, compound symmetry and AR-1 are not nested within one another, making the LRT useless for choosing directly between the two. The problem of non-nested covariance structures is greater in spatial statistics, where beyond exponential within Matérn there is very little nesting of models. Thus, the likelihood ratio test is not a valid choice for choosing between spatiotemporal covariance structures.

A more helpful tool for choosing a spatiotemporal structure is the use of information criteria (IC). As noted by Wolfinger[4], information criteria can be used instead of the LRT in the third step of Diggle's algorithm[3]. The most commonly used information criteria are AIC, BIC, HQIC, CAIC, and AICC, whose forms and preference for parsimony will be discussed later. In general, these information criteria provide a metric that quantifies how well the model fits the data with a penalty that increases with model complexity. They can be used to choose between covariance structures[2] or between different parameterizations involving random effects[8] with no requirement that the covariance matrices be nested.

Several other methods have also been previously studied for their ability to choose between covariance structures in longitudinal data analysis. Keselman *et al.*[9] investigated a whole battery of statistical tests for model selection, while Wang and Schaalje[10] examined the use of predictive criteria such as the R^2 statistic and predictive error sum of squares (PRESS). Ibrahim *et al.*[11] looked into smooth clipped absolute deviation (SCAD) and the adaptive least absolute shrinkage and selection operator (ALASSO) as model selection tools.

A large amount of work has been done to assess the accuracy of information criteria in the selection of the true longitudinal covariance structure[8, 12, 13, 14, 15, 16, 17]. In addition to comparing the different information criteria against one another, many of these simulation studies have examined how study parameters and peculiarities in the data affect accuracy. Some of these parameters and conditions include sample size, number of follow-up visits, balance between treatment groups, skewness of the outcome, degree of correlation, and what sorts of covariance structures generated the simulated data and which were candidates for selection. Although much has been done with choosing temporal correlation structures for longitudinal studies, there has been little work in choosing spatiotemporal covariance structures like those seen in longitudinal imaging studies. Simpson *et al.*[18] considered information criteria for the selection of correlation functions with

an end goal of application to longitudinal imaging data, but their simulation studies on IC accuracy were limited to one dimension of correlation. Therefore we sought to investigate which information criterion is best suited for choosing between spatiotemporal covariance structures and how accurate we can expect it to be.

Effects of Covariance Structure on Fixed Effect Inference

Since selection methods are not perfect, it is necessary to assess how statistical inference changes when the covariance structure is misspecified. There has been a large amount of research into how different factors and covariance structures affect the Type I error rate and power of tests of fixed effects as well as the bias and standard error of the corresponding parameter estimates[9, 12, 13, 16, 18, 19, 20]. Much like the studies of accuracy, these studies examined the effects of sample size, treatment group balance, skewness, and temporal covariance structure misspecification on inference. Many of these studies overlapped with investigations of information criteria, and reported the Type I error rate and bias observed when either AIC or BIC was used to choose a covariance structure. Again, these studies have simulated purely longitudinal data and so there remains a need to examine Type I and II error rates in a spatiotemporal model under similar conditions.

Motivation

Our work here has been motivated by a multi-aim longitudinal cardiac imaging study that looked at the effects of medical therapy on patients with mitral regurgitation (MR)[21]. These patients randomized to treatment or placebo at baseline and were observed every six months for up to two years. At baseline and every follow-up visit, the patients underwent three-dimensional cardiac magnetic resonance (CMR) imaging to observe the geometry and function of the left ventricle. The data from these 3-D images was condensed into 16 data values corresponding to the segments defined by the American

Heart Association's model for the left ventricle[22]. Due to the spatial relationship between these segments and the repeated measures over time, a spatiotemporal model is needed to properly model the entirety of the data. Also, since this arm of the study has a small sample size characteristic of imaging studies a separable parametric spatiotemporal model is preferred in lieu of an unstructured covariance matrix. The goal of this arm of the study was to determine if two different treatment strategies affected the remodeling of the left ventricle post-surgery; in other words, we wish to know if the time course post-surgery is different for the different treatment strategies. In terms of a statistical model, we are interested in making inferences about the treatment-by-time fixed effect term. Previously, longitudinal studies of CMR imaging data have focused on global outcomes such as ejection fraction, which removes the concern with spatial correlation, or have taken summary values from the segments and analyzed them separately[23]. In our example we wish to investigate an outcome (end-systolic wall thickness) at all of the segments of the left ventricle simultaneously, since we feel that the time course of the outcome may be different at different locations in the heart. The outcome has also been observed to differ between the base and apex of the left ventricle, so we wish to utilize a linear regression framework to control for and possibly make inferences on the regions of the heart as fixed effects.

Previous methodological work into cardiac imaging has fallen short of a true spatiotemporal model. Bowman and Waller[24] constructed a mixed model that incorporated spatial correlation as well as covariance between related conditions, but the spatial aspect did not use parametric spatial covariance structures and the conditions did not follow that typical longitudinal form of several follow-up visits. Simpson *et al.*[25] applied a separable parametric spatiotemporal model to neuroimaging data, but their work looked at only a limited set of parametric correlation functions and did not report simulation results as to how that choice affects statistical inference. Seals[27] built a spatial model for the left ventricle based on the 16-segment model and cardiac imaging data. That study also examined how accurate AIC and BIC are at choosing the correct covariance structure as well as how

the choice of covariance structure influences the inference of fixed effects. Ultimately, the spatial model by Seals provides the foundation for this spatiotemporal model.

The goal of this paper is to propose and examine a spatiotemporal model appropriate for analyzing longitudinal imaging data. We will present a linear model and describe how a separable, parametric covariance structure can be used to capture the inherent spatial and temporal correlation within such studies. The results of simulations will also be discussed, where the accuracy of various information criteria used to select a spatiotemporal covariance structure were evaluated. The simulations also looked at the effects on Type I and Type II error rates when the fitted covariance structure was different than the one used to generate the data. Finally, we will present an example with real longitudinal imaging data to illustrate how the covariance structure selection can be done and how fitting different covariance structures can affect inferences on fixed effects.

PROPOSED STATISTICAL MODEL, COVARIANCE STRUCTURES, AND INFORMATION CRITERIA

Linear Model with Correlated Errors

We propose the use of a linear model with a spatiotemporal error structure. First, consider a dataset with no missing values where N subjects were imaged at J points in time and each image contains K outcome values. Let us define Y_{ijk} as the observed outcome of subject i at time j and location k such that $i = 1, ..., N$, $j = 1, ..., J$, and $k = 1, ..., K$. Let us also define $\mathbf{Y}_i$ as the $JK \times 1$ vector of observations from subject i and $\mathbf{Y}_{ij\bullet}$ as subject i's K observed outcomes at time j, so that

$$\mathbf{Y}_i = \mathbf{Y}_{i\bullet\bullet} = \{\mathbf{Y}_{i1\bullet}, \mathbf{Y}_{i2\bullet}, ..., \mathbf{Y}_{iJ\bullet}\}, \quad \mathbf{Y}_{ij\bullet} = \{Y_{ij1}, Y_{ij2}, ..., Y_{ijK}\}.$$

Therefore, the dataset contains NJK observations divided up into correlated blocks of size JK. Using the framework described by Jennrich and Schluchter[28], the linear model has

the form

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{1}$$

where $\mathbf{X}_i = \mathbf{X}_{i\bullet\bullet}$ is a $JK \times p$ matrix containing the values of the fixed effects variables for subject i, $\boldsymbol{\beta}$ are the corresponding p regression parameters to be estimated, and $\boldsymbol{\epsilon}_i \sim MVN_{JK}(0, \sigma^2\boldsymbol{\Sigma})$ and is independent and identically distributed for all N subjects, given no missing data. The σ^2 is the variance of the errors and Σ is their correlation matrix, which assumes homogeneity of variance across repeated observations. If there are missing observations, one could define the errors to have a correlation matrix $\boldsymbol{\Sigma}_i$ which is a subset of $\boldsymbol{\Sigma}$. In this application, we wish for the correlation matrix $\boldsymbol{\Sigma}$ to follow a separable parametric spatiotemporal structure; the parameters of this matrix are contained in the vector $\boldsymbol{\theta}$ so that we can refer to the correlation matrix as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

When $\boldsymbol{\theta}$ and thus $\boldsymbol{\Sigma}$ is known, the fixed effect parameters $\boldsymbol{\beta}$ are easily estimated by solving the normal equations where $\mathbf{V}$ is a block diagonal matrix comprised of N blocks of $\sigma^2\boldsymbol{\Sigma}$:

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \tag{2}$$

with the property that the estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$[29]. This estimator comes from maximizing the log-likelihood ℓ of the form[14]

$$\ell(\boldsymbol{\beta}|\mathbf{Y}) = -\frac{NJK}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\sigma^2\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{Y}_{i\bullet\bullet} - \mathbf{X}_{i\bullet\bullet}\boldsymbol{\beta})'(\sigma^2\boldsymbol{\Sigma})^{-1}(\mathbf{Y}_{i\bullet\bullet} - \mathbf{X}_{i\bullet\bullet}\boldsymbol{\beta}). \tag{3}$$

In the case where $\boldsymbol{\Sigma}$ is known, inferences can be made on $\boldsymbol{\beta}$ using either a Wald's test or, preferably, a likelihood ratio test.

Unfortunately, in practice we do not know the values of $\boldsymbol{\theta}$ and must estimate them from the data. Although maximum likelihood estimation is possible, the estimator for $\boldsymbol{\theta}$ is biased as it fails to account for the degrees of freedom lost in estimating $\boldsymbol{\beta}$[30]. There are also concerns with the likelihood function being multi-modal in maximum likelihood

estimation[31]. Therefore, it is more common to see $\{\sigma^2, \boldsymbol{\theta}\}$ estimated with restricted maximum likelihood (REML) which greatly reduces the bias. It does so by projecting the data into a space where all of the fixed effects of $\mathbf{X}\boldsymbol{\beta}$ are removed from the likelihood function. The $\hat{\boldsymbol{\theta}}$ are then found by maximizing the restricted log-likelihood

$$\begin{aligned} REML_1(\sigma^2, \boldsymbol{\theta}|\mathbf{Y}) = &- \frac{NJK - p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{N}\mathbf{X}'_{i\bullet\bullet}\mathbf{X}_{i\bullet\bullet}\right| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{N}\mathbf{X}'_{i\bullet\bullet}[\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1}\mathbf{X}_{i\bullet\bullet}\right| - \frac{1}{2}\sum_{i=1}^{N}\log|\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})| \\ &- \frac{1}{2}\sum_{i=1}^{N}(\mathbf{Y}_{i\bullet\bullet} - \mathbf{X}_{i\bullet\bullet}\hat{\boldsymbol{\beta}})'[\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1}(\mathbf{Y}_{i\bullet\bullet} - \mathbf{X}_{i\bullet\bullet}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{4}$$

where the constant term has been modified and the second and third terms have been added when compared to the full likelihood function[14]. Note that in many software implementations the second term is dropped when reporting the maximum restricted likelihood as it is not a function of θ; this quantity is referred to as $REML_2$ and is defined as

$$REML_2(\sigma^2, \boldsymbol{\theta}|\mathbf{Y}) = REML_1(\sigma^2, \boldsymbol{\theta}|\mathbf{Y}) - \frac{1}{2}\log\left|\sum_{i=1}^{N}\mathbf{X}'_{i\bullet\bullet}\mathbf{X}_{i\bullet\bullet}\right|. \tag{5}$$

If inferences about $\boldsymbol{\theta}$ are of interest, it was noted by Harville[29] that $\hat{\boldsymbol{\theta}}$ follows an asymptotically normal distribution.

The actual REML estimation is more complicated than it may seem since the form of $REML$ contains $\hat{\boldsymbol{\beta}}$, but $\hat{\boldsymbol{\beta}}$ needs $\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})$ to be estimated before it can be calculated. Thus, iterative algorithms are needed to estimate all of the parameters in our model. One possibility is the EM algorithm, most famously described by Laird and Ware[30], which has a low computational burden per iteration but may require many iterations. Another common choice is the Newton-Raphson algorithm, described by Jennrich and Schluchter[28] and refined by Lindstrom and Bates[32], which has a greater computational burden per step in calculating matrix derivatives but ultimately requires far fewer steps than the EM

algorithm. A third option is the Fisher scoring algorithm, also discussed by Jennrich and Schluchter[28], which provides the base of the model estimation software we utilized[33]. These algorithms address the cyclic nature of estimating $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ by switching between the two estimations, using the most recent estimate of one to estimate the other and repeating until convergence. In this approach, the typical initial step is to estimate $\hat{\boldsymbol{\beta}}$ using least-squares estimation then using that to begin the REML estimation of $\hat{\boldsymbol{\theta}}$.

One limitation of REML estimation is that its estimates are specific to a certain structure of $\mathbf{X}$. If one wished to compare a full model to one with a reduced number of fixed effects, the REML procedure will project the data into two different spaces and the likelihoods cannot be compared. Thus, traditional likelihood ratio tests cannot be used to compare nested fixed effects models when REML is used. In practice inferences about fixed effects under REML estimation have been restricted to a hybrid of ML and REML estimation[4] or, more commonly, Wald's tests. Previous research has suggested that when a Wald's test is used it is highly advisable to use some form of degrees of freedom correction such as Kenward-Roger's adjustment[4, 12, 13, 20]. There have been efforts made by Welham and Thompson[34] to use LRTs in REML estimation by altering the full model's projection to go to the same space as the reduced model, but to our knowledge their method has not been implemented in common statistical software.

Separable Parametric Covariance Structures

Once an initial set of fixed effects has been chosen, the parametric structure of the correlation matrix must be determined. Ultimately, the only restriction on $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is that it must be positive definite, though it is practically always symmetric as well. The main idea behind parametric covariance functions is that 'close' observations are more highly related than 'far' observations. There is also the consideration that two pairs of observations with identical 'distances' apart should have similar if not identical correlations. We will discuss below which parametric covariance models we considered.

When choosing a spatiotemporal model, one must think about how space and time relate in the problem at hand. A simplifying assumption is that the spatial and temporal correlations are independent; in this case one can construct a separable model where the correlation between two observations is the product of the spatial and temporal correlation functions. Using matrix algebra, one can create a full correlation matrix from two independent sources of correlation by taking the direct/Kronecker product (denoted $\otimes$) of the two correlation matrices[35, 36]. Separable covariance structures for repeated measures imaging data have been used previously by Simpson *et al.*[25], who have also investigated a test for separability[26].

Nonseparable spatiotemporal models exist, and frequently rely on Taylor's hypothesis that spatial and temporal correlation use the same function and are connected through a velocity[1]. This kind of spatiotemporal structure is most common in meteorology, where the difference in weather between points A and B is the same as the difference at point A now and point A at a later time once the front over point B has moved to A. While this form may be useful in medical imaging that examines signal transduction, for imaging studies such as ours that look at anatomical structure the Taylor hypothesis is entirely inappropriate. Thus, we will focus on investigating a separable spatiotemporal model which is reasonable based on the nature of the problem as well as successful previous use of the assumption[25].

We investigated four temporal and three spatial correlation functions, chosen for how they are frequently fitted in practice and appear in most common statistical software. The three spatial structures are exponential (EXP), spherical (SPH), and Matérn (MAT). The temporal structures are compound symmetry (CS), autoregressive-1 (AR-1), Toeplitz (TOEP), and the unstructured model (UN). The details of these correlation functions are given in Table 1. Note that we have assumed the spatial structures are isotropic: the correlation between two points depends only on the magnitude of the distance between them

and not the direction/orientation.

Information Criteria

In analyzing longitudinal imaging data with our model, one would need to choose a particular combination of spatial and temporal correlation structures when fitting the data. Studies on longitudinal methods had seen promising results in the use of information criteria to select temporal correlation structures, and so we wish to extend their use to spatiotemporal models. The most common information criteria (AIC[37], AICC[38], CAIC[39], BIC[40], HQIC[41]) all have the form of the log-likelihood function, ℓ, penalized by the number of parameters. The information criteria that came after the AIC also weight the penalty according to the sample size. However, the choice of likelihood and sample size is not immediately clear when REML is part of the model estimation. Should the maximum likelihood be used for ℓ, or should $REML_1$ or $REML_2$? Should the sample size penalty n^* be based on the total number of observations minus the number of fixed effects be used ($NJK - p$, the dimension of ℓ in REML), or should one instead weight the penalties with the number of independent observational units (N)? Numerous studies have looked for answers to these questions in longitudinal studies[10, 14, 16, 17, 18] but not in a true spatiotemporal framework.

The details of the five information criteria are given in Table 2, presented such that the smaller information criterion is the 'better' model[14, 17]. In a broad sense, each information criterion could have six forms: ML with $n^* = NJK$, ML with $n^* = N$, $REML_1$ with $n^* = NJK - p$, $REML_1$ with $n^* = N$, $REML_2$ with $n^* = NJK - p$, and $REML_2$ with $n^* = N$. However, since we ultimately used these information criteria to choose between parametric covariance structures and not between models with different fixed effects, making the X matrices identical between prospective models, $REML_1$ and $REML_2$ would give identical results. The estimation procedure used was purely based on REML and did not support maximum likelihood estimation, so we just concerned ourselves

with the full restricted likelihood $REML_1$ for ℓ. Thus, we investigated nine information criteria: AIC; and forms of $n^* = N$ or $NJK - p$ for AICC, BIC, CAIC, and HQIC.

SIMULATION STUDY

In order to assess the accuracy of these information criteria as well as the effects of covariance structure misspecification on statistical inference, we performed a simulation study which observed those properties under a variety of conditions. Specifically, we estimated the Type I and Type II error rates for testing a treatment-by-time fixed effect when the covariance structure used to fit the model either matched or did not match the covariance structure used to generate the data. Under the same conditions, we estimated how often different information criteria would correctly select the covariance structure used to generate the data when presented with a variety of possible structures. Furthermore, the effects of sample size and the degree of spatial or temporal correlation were investigated for how they changed IC accuracy or Type I and Type II error rates. In the design and reporting of these simulations, we endeavored to adhere to the guidelines proposed by Burton *et al.*[42] but for the sake of space many details have been reserved for the supplementary materials.

Simulation Details

The generated data is based on the structure of the data collected in the cardiac imaging study reported on by Schiros et al.[21]. The outcome variable is multivariate normal and was collected from two treatment groups at five time points, evenly spaced six months apart (J=5). At each observation, cardiac imaging was done where the outcome was observed at 16 points within the left ventricle (K=16). Thus, each subject had 80 observed outcomes. The layout of the spatially observed locations was modeled using the AHA's 17 segment model presented by Cerqueria et al.[22], where the left ventricle is laid out

in a circular pattern with concentric rings corresponding to the different levels of the left ventricle (base, mid, and apex), much as if one is looking down from the left atrium. The coordinates were defined as the center of each segment laid out on a unit circle shown in Figure 1 and are specifically defined in the supplementary materials[27].

The outcome was calculated as the result of a linear model with correlated errors. Specifically,

$$Y_{ijk} = \beta_0 + \beta_1 Time_{ij} + \beta_2 Group_i + \beta_3 Mid_k + \beta_4 Apex_k + \beta_5 Time_{ij} * Group_i + \epsilon_{ijk} \quad (6)$$

where $Time_{ij}$ was the continuous time of subject i's j^{th} observation; $Group_i$ was the treatment group for subject i; Mid_k and $Apex_k$ were indicator variables for whether the ijk^{th} observation was from the mid or apex of the left ventricle, respectively. The collected error terms for subject i were independent and identically distributed for all subjects and were distributed $\boldsymbol{\epsilon}_{i\bullet\bullet} = [\epsilon_{i,1,1}, \epsilon_{i,1,2}, ..., \epsilon_{i,1,16}, \epsilon_{i,2,1}, ..., \epsilon_{i,5,16}] \sim MVN_{80}(0, \sigma^2\boldsymbol{\Sigma})$, where σ^2 was the variance of the outcome (assumed to be equal across all observations) and $\boldsymbol{\Sigma}$ is an 80-by-80 separable parametric spatiotemporal correlation matrix. For the sake of simplicity we assumed $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ and $\sigma^2 = 1$.

The parts of the linear model itself that were changed between conditions include β_5 and $\boldsymbol{\Sigma}$. Note that β_5 represents the time-by-treatment interaction and if $\beta_5 \neq 0$ then there exists a difference in the time course between the two groups, which was the original research hypothesis. $\beta_5 = 0$ allowed for the examination of the Type I error rate, and the other values were chosen to create sigmoidal power curves. The generating $\boldsymbol{\Sigma}$ was varied between different parametric structures and degrees of correlation, as shown in Figure 2. Total sample size was set to either 50 or 100 subjects. The complete listing of simulation conditions is given in Table 3.

For each condition described in Table 3, 5000 independent simulated datasets were run and 12 separable covariance structures corresponding to the combinations of three spa-

tial (exponential, spherical, and Matérn) and four temporal (compound symmetric, autoregressive-1, Toeplitz, and unstructured) working correlation structures. From each model fit, we looked at the p-value for the Wald's test against the null hypothesis of $H_0 : \beta_5 = 0$ as well as the maximum restricted likelihood. A conditional F-test with a Kenward-Roger adjustment for denominator degrees of freedom was considered but did not alter the conclusions. We calculated the values of the 9 information criteria for each of the 12 covariance models using the restricted log-likelihood, and defined the 'chosen' model for that dataset as the one with the smallest criterion. The selection probabilities for each information criterion for a certain condition was calculated as the proportion of the 5000 datasets where the information criterion chose the covariance structure that had been used to generate the dataset.

The Type I error rates were calculated as the proportion of the 5000 datasets where a model with a certain covariance structure rejected the null hypothesis of $\beta_5 = 0$ at an $\alpha = 0.05$ level when the data was generated with β_5 truly equal to zero. Using the convention suggested by Bradley[43], we considered the Type I error rate to be conserved when it fell between 0.75α and 1.25α, or 0.0375 and 0.625 for our chosen α level. The value of 1.25α was chosen to be moderate, neither highly conservative (1.1α) nor liberal (1.5α) as Bradley described them. Thus, we classified the fit of a certain covariance structure on data generated under a given covariance structure to be overconservative if the empirical Type I error rate $\hat{\alpha}$ was below 0.0375 and inflated if the error rate was above 0.0625.

The Type II error rates (the power curves) were found by checking the proportion of times the Wald's test for $\beta_5 = 0$ was rejected on datasets generated with $\beta_5 > 0$. Note that the combinations of fitted and generating covariance structures that were deemed to have inflated Type I error rates ($\hat{\alpha} > 0.0625$) did not have their power investigated as they were deemed to be unreliable.

Results

Selection Accuracy of Information Criteria

The five different information criteria (AIC, AICC, BIC, CAIC, and HQIC), along with their different forms for the sample size penalties, were evaluated for their accuracy in selecting the true covariance structure that was used to generate simulated data when presented with a variety of possible structures to choose from; the summarized results are given in Figure 3 and the full results are in the supplementary materials. The information criteria were highly accurate as a whole, with the lowest value for any condition being 68.4% accuracy and the highest being over 99.9%.

In general, the consistent information criteria (BIC, CAIC, HQIC) seemed to have a higher accuracy than the efficient criteria (AIC, AICC). CAIC seemed to be the most accurate information criterion, followed closely by BIC then HQIC, with AICC further behind that. AIC was the least accurate of the information criteria evaluated in terms of covariance structure selection. As for the sample size adjustment, AICC seemed to be more accurate using the number of subjects (N) while the consistent criteria were slightly more accurate using the total number of observations minus the number of fixed effects ($NJK - p$).

When choosing between temporal correlation structures, all information criteria were able to perfectly distinguish between CS and AR-1 structures. Incorrect choices were the result of picking a more complex temporal function (Toeplitz, unstructured) which was more common among the efficient criteria. The degree of temporal correlation had no noticeable effect on selection accuracy. Conversely, the degree of spatial correlation had a large effect on accuracy of choosing the true spatial structure. In particular, accuracy dropped for all criteria under a high degree of spatial correlation compared to a low degree. The reason for this behavior can be guessed at from Figure 2; the spatial correlation structures were much more distinct and different under the low-correlation condition than the

high-correlation condition.

The larger sample size had a subtle effect on selection accuracy. Increasing the total number of subjects from 50 to 100 did not simply result in an across-the-board increase in accuracy. Under low spatial correlation, the larger sample size had a slight improvement for consistent criteria, no noticeable effect on AIC, and an apparent decrease in accuracy for AICC. When the spatial correlation was high, increasing the sample size to 100 seemed to mitigate some of the accuracy loss seen in the $N = 50$ groups, especially for exponential or spherical simulated data.

Effects of Correlation Structure Specification on Type I Error Rates

In addition to examining the accuracy of information criteria, we looked at the empirical Type I error rates for unadjusted Wald's tests under different combinations of generating and fitted covariance structures. The chosen α level for these tests was 0.05. The summarized results of these simulations are given in Figures 4 and 5, with the full results given in supplementary tables. As stated before, we considered the Type I error rate to be inflated if it was greater than 0.0625 (1.25α) and overly conservative if it was less than 0.0375 (0.75α).

We found that when the spatiotemporal covariance structure used to fit the data was the same as the one used to generate the data the Type I error rate is conserved. This result held regardless of the degree of spatial or temporal correlation, whether the sample size was 50 or 100, or what the actual structures were. Specifically, all of the Type I error rates for the conditions with matching generating and fitted covariance structures fell between 0.0448 and 0.0598 with a median of 0.0516. This conservation of the error rate held when more complex correlation functions were fit to special cases such as Matérn fit to exponential data and Toeplitz or unstructured fit to compound symmetric or autoregressive-1 data.

The simulation showed that misspecification of the correlation structure can cause

the Type I error rate to be inflated or overly conservative, sometimes extremely so. In cases where the fitted temporal correlation structure was correctly specified, the effects of spatial structure misspecification are shown in Figure 4. When the degree of spatial correlation was low, the mismatch of spherical structures and exponential/Matérn structures resulted in the Type I error not being conserved. However, when the degree of spatial correlation was high there was no problem with the Type I error rate with the mismatch of those structures. We also found that fitting an exponential structure to Matérn data can result in the Type I rate being overly conservative. On the temporal side, when the spatial structure was correctly specified, the results were much more straightforward as seen in Figure 5. Fitting compound symmetric to autoregressive-1 data greatly inflated the Type I error, while fitting autoregressive-1 to compound symmetric data resulted in a very conservative Type I error rate.

One very important observation from the simulations is that the sample size increase from 50 to 100 did not noticeably change situations where the Type I error rate was not conserved, suggesting that increasing sample size is an ineffective way to conserve the Type I error rate when the spatiotemporal structure is misspecified. A more effective strategy of conserving the Type I error rate seems to be proper selection of the covariance structure. As we have seen, commonly used information criteria are very accurate when it comes to picking the true parametric spatiotemporal covariance structure. To test this approach, we looked at the empirical Type I error rate when the covariance structure chosen by the information criterion was used; the detailed results are given in a supplementary table. We found that when the working covariance structure was chosen by one of the nine information criteria the Type I error rate was very well conserved ([0.0446,0.0600] for all IC and conditions; median=0.0516), even when using the least accurate AIC ([0.0448,0.0600], median=0.0514). This suggests that the strategy of using an information criterion to choose the working correlation structure is one that, on average, will conserve the α-level of a Wald's test of fixed effects.

Effects of Correlation Structure Misspecification on Power

In order to explore the relationship between fitted and actual spatiotemporal correlation structures with regard to their effect on Type II error rates, more simulation studies were run. As with the Type I error rate, the test of interest is an unadjusted Wald's test for a group-by-time interaction (β_5 in Equation 6). An example of the results is given in Figure 6 for data generated under spherical-by-compound symmetric, and the results for other generating structures are shown in the supplementary materials.

The first thing to note is that statistical power seems to be maximized when the true correlation structure is used to fit the data. That is, when the working correlation structure is the same as the one used to generate the data the resulting power is as large or larger than any other working correlation structure we considered. Other working covariance structures were able to match that level of power, though. In particular, the more complex working covariance structures that contained the true covariance structure (Matérn containing exponential, Toeplitz and unstructured containing both autoregressive-1 and compound symmetry) matched the true correlation structure in terms of power despite requiring additional degrees of freedom.

When the correlation structure was misspecified in a pairing that had an overly conservative Type I error rate, the resulting power was lower than that of the true correlation structure. For example, when an autoregressive-1 structure was fit to compound symmetric data the power was greatly reduced, reflecting the extremely low $\hat{\alpha}$ that was observed. Mismatched spatial structures that were overly conservative when spatial correlation was low had similarly lower power. Those pairings whose Type I error rate returned to normal for a high degree of spatial correlation had their power increase to roughly the maximum as well.

We did observe that power was higher for a high degree of temporal correlation compared to a low degree of temporal correlation, and that power seemed to be lower when

spatial correlation was high versus low. Increased spatial correlation means that there is less unique information amongst the multiple observations from a subject, which may explain the power loss. As for the gain in power under high temporal correlation, it is probable that this result only holds for inference on parameters related to time such as the treatment-by-time interaction we tested.

ILLUSTRATING THE USE OF COVARIANCE STRUCTURES IN A LONGITUDINAL CARDIAC IMAGING STUDY

The data used in this example analysis was first reported in the paper by Ahmed et al.[23], which contains the details of data collection and study design. In short, this study was a randomized controlled phase IIb trial for the use of medical therapy in the treatment of patients with chronic degenerative mitral regurgitation. The intent was that treatment would prevent the adverse left ventricular remodeling typically seen in MR patients. Over time the left ventricle balloons outward in these patients resulting in a more spherical chamber compared to the 'bullet' shape seen in healthy patients; a sample MRI scans of the hearts of a healthy subject and a MR patient, both enrolled in the study, are given in Figure 7. In the previous work global or summary measures were taken from the MRI scans, while this analysis makes use of outcomes from all 16 segments. The particular outcome analyzed here is the end-systolic (ES) wall thickness of the left ventricle.

A linear model similar to the one used in the simulation study was fitted to the data, where the treatment-by-time interaction was of interest. In this model, shown below, higher order interaction terms were included up through the three-way time/level/treatment interaction. These terms represent how the treatment would affect the time course of LV remodeling, but it would affect the different levels in different ways. The parentheses

represent the terms relating to ventricular level.

$$Y_{ijk} = \beta_0 + \beta_1 Time_{ij} + \beta_2 Group_i + (\beta_3 Mid_k + \beta_4 Apex_k) + \beta_5 Time_{ij} * Group_i$$
$$+ (\beta_6 Time_{ij} * Mid_k + \beta_7 Time_{ij} * Apex_k) + (\beta_8 Group_i * Mid_k + \beta_9 Group_i * Apex_k)$$
$$+ (\beta_{10} Group_i * Time_{ij} * Mid_k + \beta_{11} Group_i * Time_{ij} * Apex_k) + \epsilon_{ijk} \quad (7)$$

Like before, Y_{ijk} represents the wall thickness of the k^{th} segment of the i^{th} subject at the j^{th} time point and the error terms for subject i were independent and identically distributed for all subjects and were distributed $\epsilon_{i\bullet\bullet} = [\epsilon_{i,1,1}, \epsilon_{i,1,2}, ..., \epsilon_{i,1,16}, \epsilon_{i,2,1}, ..., \epsilon_{i,5,16}] \sim MVN_{80}(0, \sigma^2\mathbf{\Sigma})$, where σ^2 was the variance of the outcome (assumed to be equal across all observations) and $\mathbf{\Sigma}$ is an 80-by-80 separable parametric spatiotemporal correlation matrix. Twelve models were fit where $\mathbf{\Sigma}$ was varied to the structures discussed earlier. Note that only complete cases were used, giving a sample size of 26.

The log-likelihood of the fitted models was then used to calculate the nine information criteria. The BIC and CAIC with $n^* = NJK - p$ chose MAT$\otimes$CS to be the best covariance structure, while the others all picked EXP$\otimes$UN. When there is a discrepancy between criteria, it is best to look at how the estimated correlation functions compare to the observed correlation; the easiest way is to plot the correlation versus distance, as seen in Figure 8. The unstructured model works best for comparing the estimated functions to the 'true' correlation, but this approach is limited by the number of correlated outcomes. Here, five temporal observations made fitting an unstructured model simple while the 16 spatial locations meant that a spatial unstructured model could not be estimated. The values in Figure 8 were calculated by taking pairs of segments from every time point and calculating the correlation with an unstructured model in space and time; this strategy was inspired by the method used in Simpson et al.[26].

From the plot we can see that the compound symmetric model fits the temporal correlation well, but there is a bit of heterogeneity that could give the unstructured model

the nod when additional parameters are not too harshly penalized. On the other hand, the correlation between the 16 segments is extremely heterogeneous when considered in terms of distance between segments. None of the three functions fit the correlations particularly well, but with such a high amount of heterogeneity it is doubtful any one parametric function could fit the data closely. The exponential model may fit the middle distances better, but Matérn seems to fit the far distances better; the discrepancy between the information criteria in choosing a spatial structure can likely be explained by this.

Since we have chosen a covariance structure, we are now able to draw inferences about the fixed effects in the model. The results of the Wald's and conditional F-tests for the three-way interaction terms is given in Table 4. The simple Wald's χ^2 test and the conditional F-tests with the Kenward-Roger adjustment did not appear to give noticeably different results. Another thing to note is that the choice of covariance structure did affect the test statistics and p-values, although the effect was minor in this case. However, if one considers the adjusted residual degrees of freedom in the F-test it is clear that choosing a spatiotemporal covariance structure with more parameters can substantially reduce the degrees of freedom available for testing.

We can also see the importance of selecting a proper correlation structure. From the simulation results, it would appear that if the data was truly exponential- or Matérn-by-compound symmetric then a spherical-by-autoregressive-1 model would be entirely inappropriate and be extremely conservative. This result is borne out when such a SPH$\otimes$AR-1 model is fitted, as the test statistic was much smaller and the p-value much larger than when the chosen structures were fitted.

Since the three-way interaction was not statistically significant at an $\alpha = 0.05$ level, the next logical step is to remove it and test the two-way treatment-by-time interaction in a reduced model. Here we can see how the choice of correlation structures can be dependent on the fixed effects. When the twelve models were fitted and the information

criteria compared, all nine criteria chose the exponential-by-unstructured model. This is a different result than the conflicting choices from the full model, suggesting that the chosen structure should be reevaluated when the fixed effects change.

DISCUSSION

In this paper, we have presented a spatiotemporal linear model for data correlated through space and time and have considered the relationship separable parametric covariance structure choice has on statistical inference.

We have found that when the chosen covariance structure matches the true structure of the data (or has the true structure as a special case), statistical inference works 'best' as expected. Specifically, the Type I error rate is conserved and tests about fixed effects have their power maximized. When the chosen structure is misspecified, however, there is the potential for the Type I error rate to be inflated or too conservative which makes statistical inference unreliable or inefficient, respectively. Furthermore, misspecifications that are overly conservative in Type I error rate were found to also have noticeably reduced power compared to the true structure.

The best determinant of whether a given working covariance structure will have a negative effect on inference is whether or not it can closely approximate the data's structure. When one tries to use an exponentially decreasing function like AR-1 with a flat function like compound symmetry, the resulting statistical properties are irredeemably bad. Similarly, fitting a mostly linear function like the spherical structure to exponential data generally causes problems with inference; however, if all of the correlations are high (such as in Figure 2) then there is little difference in the functions and statistical inferences are reasonable. In practice it is advisable to plot the estimated correlation functions versus the unstructured values[25], or to consult a semivariogram[5]. This finding is similar to the one by Gurka *et al*.[19] that found that underfitting the correlation often led to the Type I error

rate not being conserved.

Overall, it seems that the use of more flexible and complex spatial and temporal correlation structures result in reliably high power regardless of the degree of correlation, assuming that the complex model can fully or at least reasonably approximate the true model. Note that this conclusion may be entirely reliant on the data having a high dimension per subject: despite the low sample size of 50, when one considers the effective degrees of freedom to be at most 3994 ($NJK - p = 50 * 5 * 16 - 6$) then it makes sense that spending a few more degrees of freedom to estimate additional covariance structure parameters would have a negligible effect in terms of power loss.

In addition, we have found that information criteria are highly accurate at choosing between separable spatiotemporal covariance models. Even the simplest AIC was at least 68% accurate. The consistent information criteria (BIC, CAIC, and HQIC) seemed to be more accurate than efficient (AIC, AICC) criteria, with CAIC being the most accurate. The CAIC's accuracy was generally over 90%, and often went over 99%. Although with REML estimation the sample size adjustment of N worked much better for AICC, HQIC performed better with $NJK - p$ and CAIC/BIC were largely invariant to which of the two were used. As a whole, we would recommend the use of CAIC for choosing a spatiotemporal covariance structure; if it is unavailable, though, BIC should work admirably as well. One possible caveat to this high accuracy was that only one variance was considered in our simulation study; simulations reported by Gurka[14] suggest that accuracy of information criteria can sharply decrease at higher variances, particularly for efficient criteria. Gurka also found that $REML_2$ was better for efficient criteria and $REML_1$ for consistent criteria for choosing fixed effects, but in our simulations the fixed effects were constant so there was inherently no difference between the two formulations; Additional work would be needed to address the appropriateness of using information criteria to choose fixed effects and covariance structures simultaneously.

In this study we only considered three spatial and four temporal correlation structures for model fitting, and only two temporal for data generation. The actual number of structures one may wish to choose from will likely be much larger in practice, and previous studies suggest that increasing the number of choices reduces the accuracy of information criteria monotonically[12, 13, 17]. We suggest that an analyst should consider the goodness-of-fit of spatial and temporal correlation structures when making a decision about which model to use for drawing inferences, and that graphical checks can be very useful as a supplement to information criteria. It should also be remembered that choosing a covariance structure should be redone if the predictors of the model change as the 'best' correlation function may change as well. This was observed in the above applied example, and is a direct consequence of the statistical theory; changing the fixed effects results in the restricted likelihood being projected into a different space which directly alters the estimation of the covariance parameters. The amount which correlation structure selection affects overall goodness-of-fit was not quantified in our simulation study, but would be useful to the field and a logical next step. A practical approach would be to employ backwards selection from a model with all potential predictors, choosing a new covariance model at each step back.

One of the most important assumptions in our model was that the temporal and spatial correlation were separable. It is an extremely useful assumption that we feel is reasonable, but still one that must be checked. Simpson et al.[26] recently proposed a likelihood ratio test for whether a separable parametric structure should be rejected when compared to a fully unstructured covariance matrix. Such an unstructured matrix was estimated by looking at subsets of temporal and spatial observations to fill out the area around the main diagonal, with the assertion that correlation quickly decays to zero with distant observations. Neither spatial nor temporal correlation quickly decayed with distance in our example, which agrees with Simpson's suggestion that a more general yet still computationally efficient method is needed. It should be noted that they did report that for their application

in caudate morphology the separability assumption was found to be invalid, so one cannot simply state that separable structures are always valid for longitudinal imaging data as there may be an interaction between space and time. Unfortunately, as mentioned before the vast majority of structured nonseparable correlation models rely on Taylor's hypothesis, which does not seem valid for such an application. Cressie and Huang[45] developed a framework for a wide variety of nonseparable parametric spatiotemporal functions, but work has not been done as to the appropriateness of those for longitudinal imaging studies. Work also remains to be done to examine the impact of violating the separability assumption in our model as it pertains to inference.

There are other considerations that fell outside the scope of this study that merit further investigation. For drawing statistical inferences, we looked at an unadjusted Wald's test as well as an F-test with a Kenward-Roger adjustment to the denominator degrees of freedom; one could argue for the use of an additional Kenward-Roger adjustment to the covariance matrix[4, 12, 13, 20] or of the REML-friendly likelihood ratio test[34] which would necessitate similar inquiries as to how covariance structure specification affects statistical inference. Another possible direction would be to consider the sensitivity of our model to the assumption of multivariate normality, and to look at how it behaves when faced with skewness or kurtosis in the outcome.

Lastly, it is worth considering how our model and an information criteria-based covariance structure selection approach handle some challenges seen in the analysis of real studies, such as unbalanced designs or missing data. These are issues that may encourage the use of summary methods which have seen much discussion and research[46, 47, 48, 49]. Therefore a direct comparison of such summary methods with our model is currently planned.

ACKNOWLEDGEMENTS

We wish to thank Drs. Louis Dell'Italia, Tom Denney, Jr., Himanshu Gupta, and Chun Schiros of the SCCOR study for their support and for the cardiac MRI data and images they provided. Predoctoral funding was provided by NHLBI T32HL079888. The SCCOR study was supported by National Institutes of Health Specialized Center of Clinically Oriented Research in Cardiac Dysfunction P50-HL077100.

REFERENCES

[1] Gelfand AE, Diggle PJ, Guttorp P, Furntes M. *Handbook of Spatial Statistics*. CRC Press: Boca Raton, 2010.

[2] Littell RC, Pendergast J, Natarajan R. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 2000;**19**:1793-1819.

[3] Diggle PJ. An approach to the analysis of repeated measurements. *Biometrics* 1988;**44**(4):959-971.

[4] Wolfinger RD. Covariance Structure Selection in General Mixed Models. *Communications in Statistics - Simulation and Computation* 1993;**22**(4):1079-1106.

[5] Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. *Journal of the Royal Statistical Society* 1998;**47**(3):299-350.

[6] Schaalje B, Zhang J, Pantula SG, Pollock KH. Analysis of Repeated-Measurements Data from Randomized Block Experiments. *Biometrics* 1991;**47**(3):813-824.

[7] Grady JA, Helms RW. Model Selection Techniques for the Covariance Matrix for Incomplete Longitudinal Data. *Statistics in Medicine* 1995;**14**:1397-1416.

[8] Liu S, Rovine MJ, Molenaar P. Selecting a linear mixed model for longitudinal data: repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods* 2012;**17**(1):15-30, DOI: 10.1037/a0026971.

[9] Keselman HJ, Algina J, Kowalchuk RK, Wolfinger RD. A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology* 1999;**52**(1):63-78, DOI: 10.1348/000711099158964.

[10] Wang J, Schaalje B. Model Selection for Linear Mixed Models Using Predictive Criteria. *Communications in Statistics - Simulation and Computation* 2009;**38**(4):788-801, DOI: 10.1080/03610910802645362.

[11] Ibrahim JG, Zhu H, Garcia RI, Guo R. Fixed and random effects selection in mixed effects models. *Biometrics* 2011;**67**(2):495-503, DOI: 10.1111/j.1541-0420.2010.01463.x.

[12] Gomez EV, Schaalje GB, Fellingham GW. Performance of the Kenward-Roger Method when the Covariance Structure is Selected Using AIC and BIC. *Communications in Statistics - Simulation and Computation* 2005;**34**(2):377-392, DOI: 10.1081/SAC-200055719.

[13] Guerin L, Stroup WW. A Simulation Study to Evaluate PROC MIXED Analysis of Repeated Measures Data. *Proceedings of the 2000 Conference on Applied Statistics in Agriculture* 2000; Kansas State University:170-203.

[14] Gurka MJ. Selecting the Best Linear Mixed Model Under REML. *The American Statistician* 2006;**60**(1):19-26, DOI: 10.1198/000313006X90396.

[15] Keselman HJ, Algina J, Kowalchuk RK, Wolfinger RD. A Comparison of Two Approaches For Selecting Covariance Structures in The Analysis of Repeated Measurements. *Communications in Statistics - Simulation and Computation* 1998;**27**(3):591-604.

[16] Vallejo G, Ato M, Valdés T. Consequences of Misspecifying the Error Covariance Structure in Linear Mixed Models for Longitudinal Data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 2008;**4**(1):10-21, DOI: 10.1027/1016-9040.12.1.10.

[17] Vallejo G, Fernández MP, Livacic-Rojas PE, Tuero-Herrero E. Selecting the best unbalanced repeated measures model. *Behavior Research Methods* 2011;**43**(1):18-36, DOI: 10.3758/s13428-010-0040-1.

[18] Simpson SL, Edwards LJ, Muller KE, Sen PK, Styner MA. A linear exponent AR(1) family of correlation structures. *Statistics in Medicine* 2010;**29**:1825-1838.

[19] Gurka MJ, Edwards LJ, Muller KE. Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine* 2011;**30**(22):2696-2707, DOI: 10.1002/sim.4293.

[20] Schaalje B, McBride JB, Fellingham GW. Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics* 2002;**7**(4):512-524.

[21] Schiros CG, Dell'Italia LJ, Gladden JD, Clark D, Aban I, ... Ahmed MI. Magnetic resonance imaging with 3-dimensional analysis of left ventricular remodeling in isolated mitral regurgitation: implications beyond dimensions. *Circulation* 2012;**125**(19):2334-2342, DOI: 10.1161/CIRCULATIONAHA.111.073239.

[22] Cerqueria MD, Weissman NJ, Dilsizian V, Jacobs AK, Kaul S, Laskey WK, ... Verani MS. Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart: A Statement for Healthcare Professionals From the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. *Circulation* 2002;**105**:539-542, DOI: 10.1161/hc0402.102975.

[23] Ahmed MI, Aban I, Lloyd SG, Gupta H, Howard G, Inusah S, ..., Dell'Italia LJ. A Randomized Controlled Phase IIb Trial of Beta-1-Receptor Blockade for Chronic

Degenerative Mitral Regurgitation. *Journal of the American College of Cardiology* 2012;**60**(9):833-838, DOI: 10.1016/j.jacc.2012.04.029.

[24] Bowman FD, Waller LA. Modeling of cardiac imaging data with spatial correlation. *Statistics in Medicine* 2004;**23**(6):965-985, DOI: 10.1002/sim.1741.

[25] Simpson SL, Edwards LJ, Muller KE, Styner MA. Kronecker Product Exponent AR(1) Correlation Structures for Multivariate Repeated Measures. *PLoS ONE* 2014; 9:e88864.

[26] Simpson, SL, Edwards, LJ, Styner, MA, Muller, KE. Separability tests for high-dimensional, low-sample size multivariate repeated measures data. *Journal of Applied Statistics* 2014; DOI: 10.1080/02664763.2014.919251.

[27] Seals, S. Spatial analysis of cardiovascular MRI data (dissertation). Birmingham, AL: University of Alabama at Birmingham; 2013.

[28] Jennrich RI, Schluchter MD. Models with Unbalanced Structured Covariance Matrices. *Biometrics* 1986;**42**(4):805-520.

[29] Harville DA. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* 1977;**72**(358):320-338.

[30] Laird NM. Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982;**38**(4):963-974.

[31] Warnes JJ, Ripley BD. Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika* 1987; **74**(3):640-642.

[32] Lindstrom MJ, Bates DM. Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association* 1988;**83**(404):1014-1022.

[33] Butler D, Cullis BR, Gilmour AR, Gogel BJ. ASReml-R reference manual (Release 2.00, 2007) [Software]. Accessed 4/11.2014. Available from http://www.vsni.co.uk/software/asreml

[34] Welham SJ, Thompson R. Likelihood Ratio Tests for Fixed Model Terms using Residual Maximum Likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1997;**59**(3):701-714, DOI: 10.1111/1467-9868.00092.

[35] Galecki AT. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics - Theory and Methods* 1994;**23**(11):3105-3119.

[36] Parks RW. Efficient Estimation of a System of Regression Equations when Disturbances are Both Serially and Contemporaneously Correlated. *Journal of the American Statistical Association* 1967;**62**(318):500-509.

[37] Akaike H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 1974;**19**(6):716-723.

[38] Hurvich CM, Tsai C. Regression and time series model selection in small samples. *Biometrika* 1989;**76**(2):297-307.

[39] Bozdogan H. Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika* 1987;**52**(3):345-370.

[40] Schwarz G. Estimating the Dimension of a Model. *Annals of Statistics* 1978;**6**(2):461-464.

[41] Hannan EJ, Quinn BG. The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, Series B* 1979;**41**(2):190-195.

[42] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006;**25**(24):4279-4292, DOI: 10.1002/sim.2673.

[43] Bradley JV. Robustness? *British Journal of Mathematical and Statistical Psychology* 1978;**31**:144-152.

[44] Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*. Wiley: Hoboken, 2004.

[45] Cressie, N, Huang, H. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* 1999;**94**:1330-1340.

[46] Albert PS. Longitudinal Data Analysis (Repeated Measures) in Clinical Trials. *Statistics in Medicine* 1999;**18**:1707-1732.

[47] Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. *Longitudinal Data Analysis*. CRC Press: Boca Raton, 2008.

[48] Gibbons R, Hedeker D, Waternaux C, Davis JM. Random Regression Models: A Comprehensive Approach to the Analysis of Longitudinal Psychiatric Data. *Psychopharmacology Bulletin* 1988;**24**(3):438-443.

[49] Zucker DM, Manor O, Gubman Y. Power comparison of summary measure, mixed model, and survival analysis methods for analysis of repeated-measures trials. *Journal of Biopharmaceutical Statistics* 2012;**22**(3):519-534, DOI: 10.1080/10543406.2010.550702.

Table 1: Table of the spatial and temporal correlation functions examined in this paper. Temporally, consider t_a to be the discrete time of the a^{th} observation in time, with $a = 1, ..., J$, and $Y(t_a)$ be the observed outcome at that time point. Let us assume that all observations are evenly spaced, so that $t_a - t_{a-1} = t_b - t_{b-1}$ for any $a, b \geq 2$. Spatially, consider d_{ab} as the Euclidian spatial distance between observations Y_a and Y_b.[2, 44]

Type	Correlation Fcn.	$Corr(Y_a, Y_b)$	Parameter Space	$Length(\boldsymbol{\theta})$
Spatial	Exponential	$\rho^{d_{ab}}$	$\rho \in (0, 1)$	1
	Spherical	$1 - \frac{3}{2}\left(\frac{d_{ab}}{\phi}\right) + \frac{1}{2}\left(\frac{d_{ab}}{\phi}\right)^3$	$\phi \in (0, max(d_{ab}))$	1
	Matérn	$\frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{d_{ab}}{\phi}\right)^{\nu} K_{\nu}\left(\frac{d_{ab}}{\phi}\right)$	$\phi \geq 0, \nu \geq 0$	2
Temporal	Comp. Symm.	ρ	$\rho \in (0, 1)$	1
	Autoregressive-1	$\rho^{\lvert t_b - t_a \rvert}$	$\rho \in (0, 1)$	1
	Toeplitz	$\rho_c, \quad c = \lvert a - b \rvert$	$\rho_c \in (0, 1)$	$J - 1$
	Unstructured	ρ_{ab}	$\rho_{ab} \in (0, 1)$	$\frac{J(J-1)}{2}$

Table 2: Table of the function forms of the different information criteria (IC) under REML, presented so that 'smaller is better.' The ℓ is the log-likelihood, s is the number of parameters in the covariance structure, and n^* is the sample size penalty of either the number of subjects (N) or the dimension of the log-likelihood in REML ($NJK - p$, number of observations minus the number of fixed effects).

IC	Source		Form with $n^* = N$	Form with $n^* = NJK - p$
AIC	Akaike	(1974)	$-2\ell + 2s$	$-2\ell + 2s$
BIC	Schwarz	(1978)	$-2\ell + s\log(N)$	$-2\ell + s\log(NJK - p)$
HQIC	Hannan & Quinn	(1979)	$-2\ell + 2s\log(\log(N))$	$-2\ell + 2s\log(\log(NJK - p))$
CAIC	Bozdogan	(1987)	$-2\ell + s(\log(N) + 1)$	$-2\ell + s(\log(NJK - p) + 1)$
AICC	Hurvich & Tsai	(1989)	$-2\ell + 2s\left(\frac{N}{N-s-1}\right)$	$-2\ell + 2s\left(\frac{NJK-p}{NJK-s-1}\right)$

Table 3: Total number of conditions for the simulation study which included varying the sample size, temporal and spatial structures used to generate the data, the degree of spatial and temporal correlation used in data generation, and the values of β_5 used.

Sample Size	N=50	N=100
Σ_T Structures	2 (CS,AR-1)	2 (CS,AR-1)
Degree of Σ_T	2 (High, Low)	2 (High, Low)
Σ_S Structures	3 (Exp,Sph,Mat)	3 (Exp,Sph,Mat)
Degree of Σ_S	2 (High, Low)	2 (High, Low)
Values of β_5	$\beta_5 = \{0, 0.05, 0.10, 0.15, 0.20\}$	$\beta_5 = \{0\}$
Total Number of Conditions	120	24
	144	

Table 4: Table of the statistical inferences for the treatment-by-time-by-level and treatment-by-time interactions for end-systolic wall thickness as the outcome.

Model and Variable Tested	Chosen $\Sigma_S \otimes \Sigma_T$	IC Choosing Σ	Wald χ^2 Test		Conditional F-Test	
			Test stat.	p-value	Test statistic	p-value
Full Model, Time*Treatment*Level	MAT⊗CS	BIC, CAIC	χ^2_2 =1.66	0.4370	$F_{2,1113.0} = 0.828$	0.4373
	EXP⊗UN	All others	χ^2_2 =1.60	0.4593	$F_{2,413.2} = 0.778$	0.4599
	MAT⊗UN	None	$\chi^2_2 = 1.64$	0.4397	$F_{2,366.4} = 0.822$	0.4406
	SPH⊗AR-1	None	$\chi^2_2 = 0.70$	0.7057	$F_{2,1231.4} = 0.348$	0.7059
Reduced Model, Time*Treatment	EXP⊗UN	All IC	$\chi^2_1 = 0.40$	0.5364	$F_{1,133.1} = 0.382$	0.5375
	MAT⊗UN	None	$\chi^2_1 = 0.34$	0.5593	$F_{1,111.1} = 0.341$	0.5605

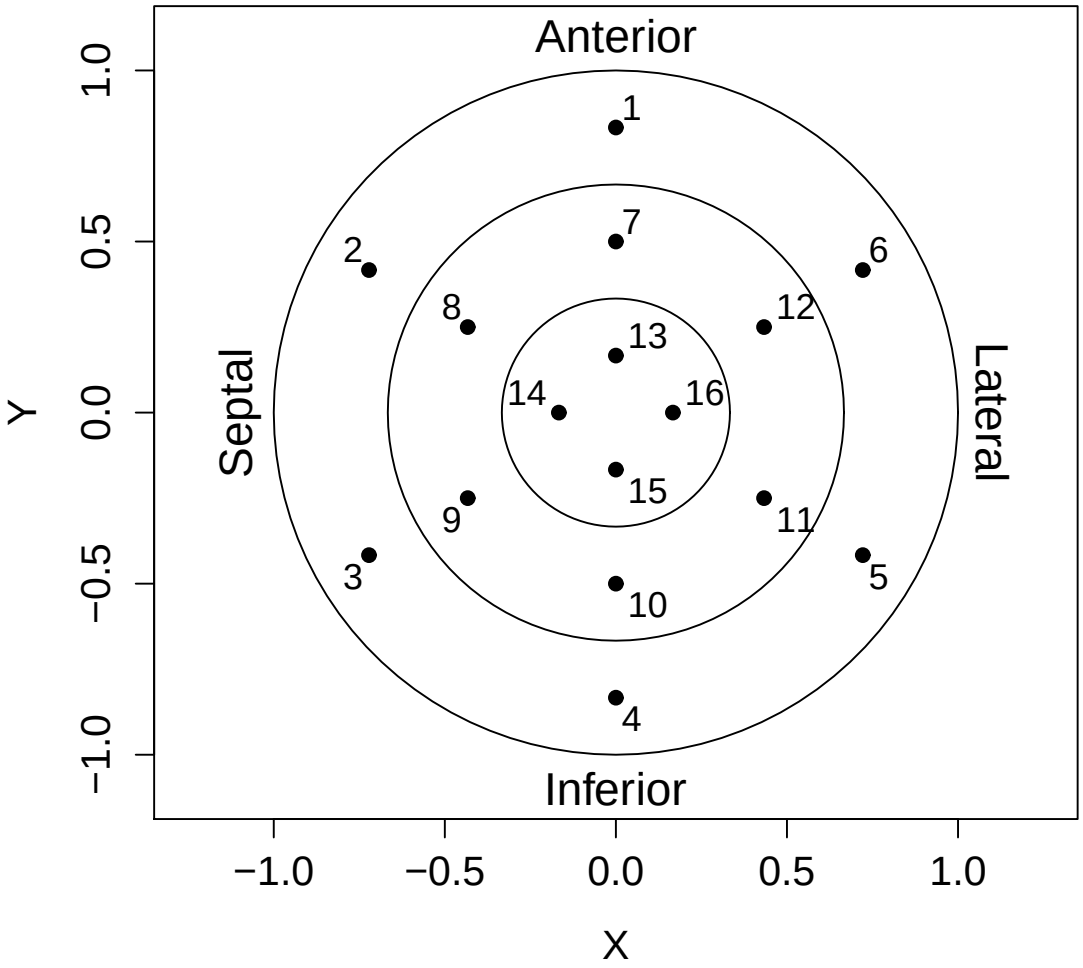

Figure 1: Plot of the 16 segments of the left ventricle. The outer ring corresponds to the base, the middle ring to the mid, and the inner circle to the apex[22, 27]. The numbers correspond to the segment's index as defined in Table A.1.

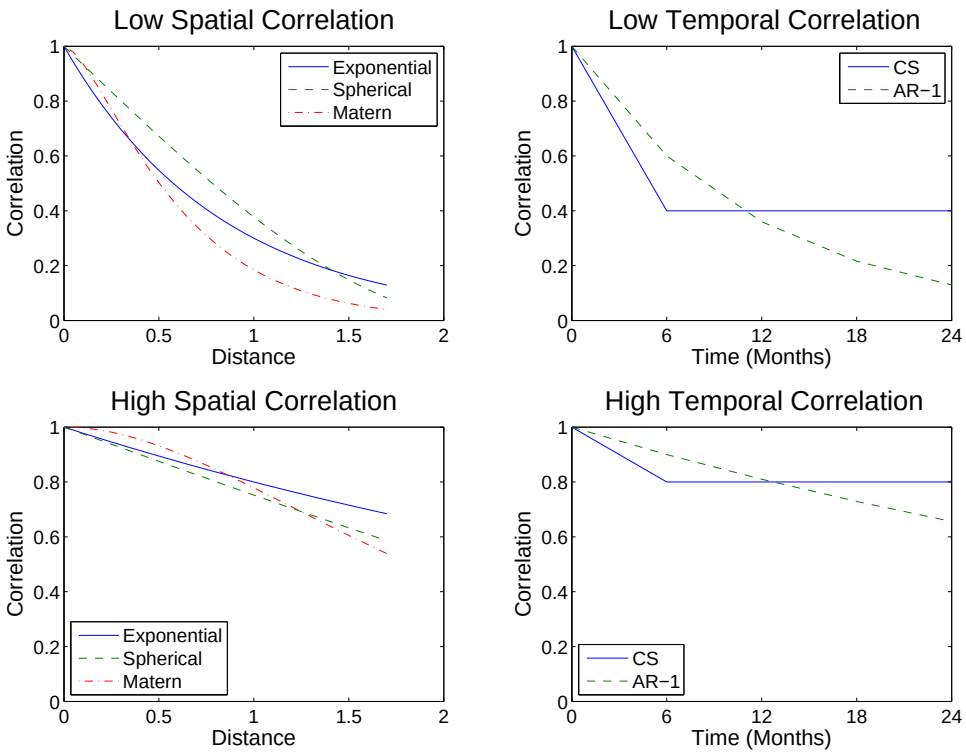

Figure 2: Plots of the covariance functions used to generate the spatiotemporal data. The spatial structures (exponential, spherical, and Matérn) are on the left and the temporal structures (compound symmetric and autoregressive-1) are on the right. The functions used to generate data with a low degree of correlation are on the top, and those generating a high degree are on the bottom.

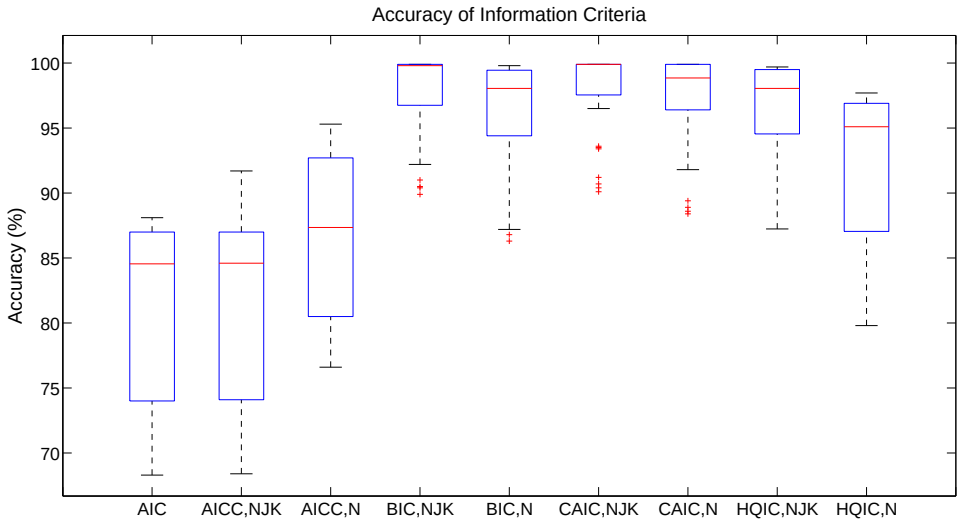

Figure 3: Box plots of the accuracies of the nine information criteria for selecting the generating correlation structure from 12 working structures. The values plotted are aggregated from 48 simulation conditions.

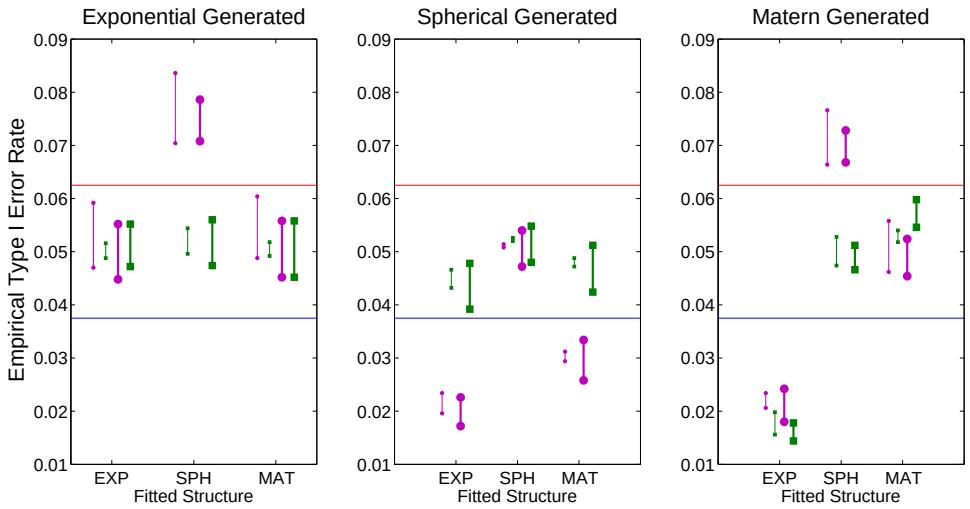

Figure 4: Plot of the empirical Type I error rates for all combinations of generating and working spatial and the correct temporal correlation structures. The endpoints denote the minimum and maximum observed rates when the temporal and spatial structures, respectively, were correctly specified. Magenta circles denote a low degree of spatial correlation and green squares a high degree. Smaller markers denote a sample size of N=50, while the larger ones denote N=100.

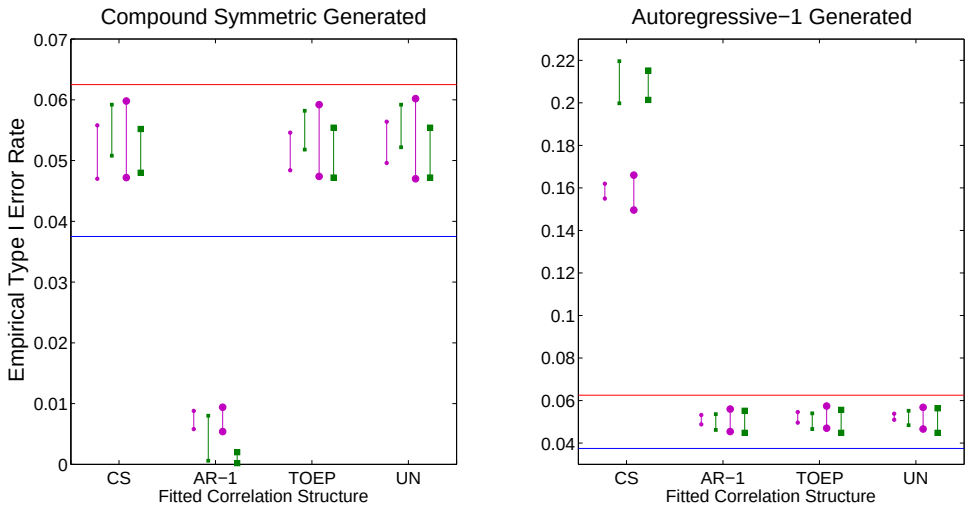

Figure 5: Plot of the empirical Type I error rates for all combinations of generating and working temporal and correct spatial correlation structures. The endpoints denote the minimum and maximum observed rates when the temporal and spatial structures, respectively, were correctly specified. Magenta circles denote a low degree of temporal correlation and green squares a high degree. Smaller markers denote a sample size of N=50, while the larger ones denote N=100.

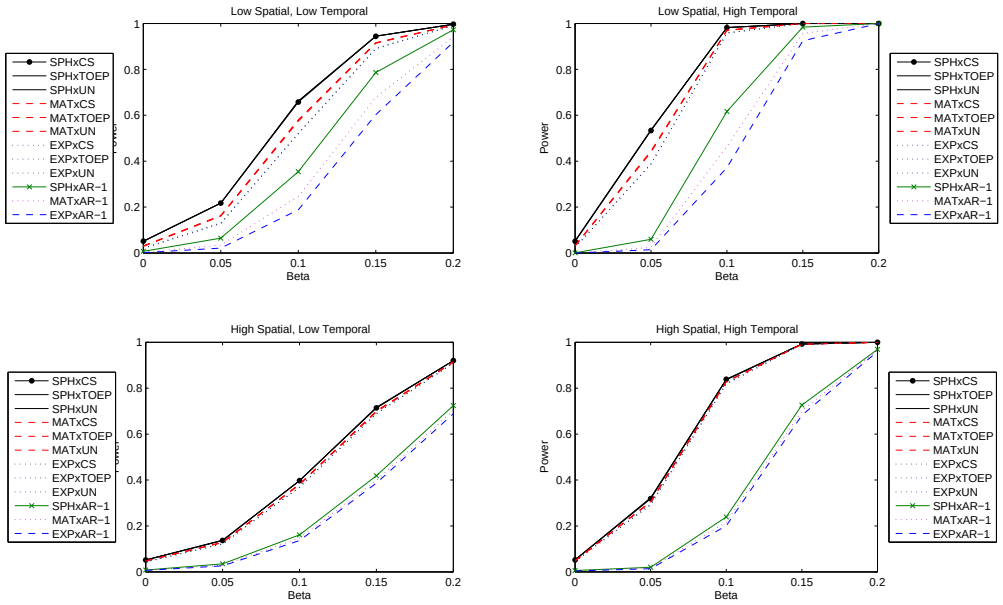

Figure 6: Plot of power curves for a generating covariance structure of SPH⊗CS under high and low degrees of correlation in space and time. The working covariance structure corresponding to the true structure is denoted with a dot and solid black line. Note that working covariance structures with approximately equivalent power are grouped in the legend and share a common line style and color.

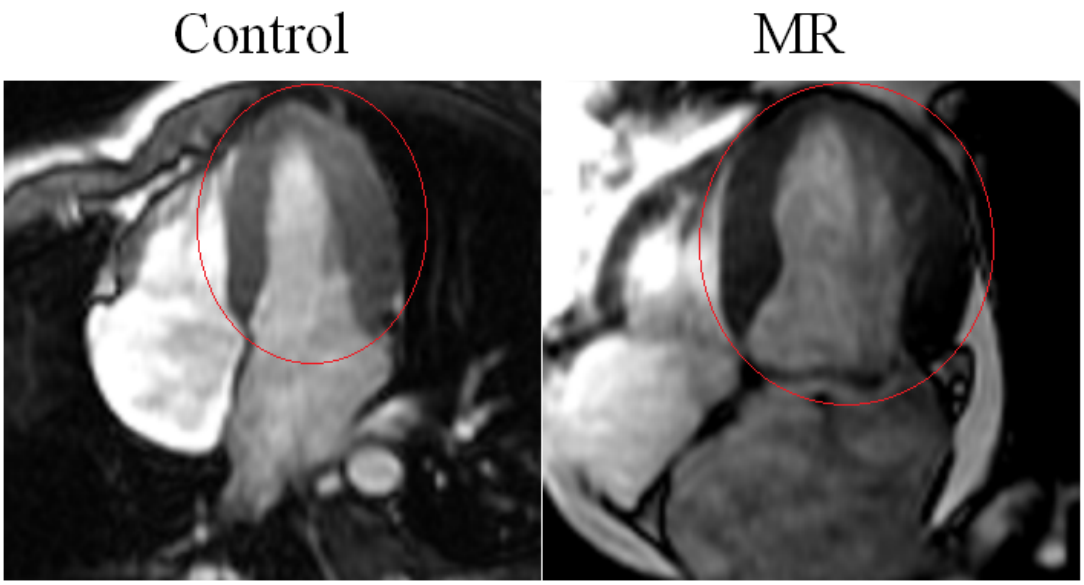

Figure 7: Representative end-systolic (ES) cardiac MRI scan with the four chamber view from a control subject and a MR patient. Note how remodeling has caused the MR patient to have a more spherical left ventricle (circled) compared with the healthy subject.

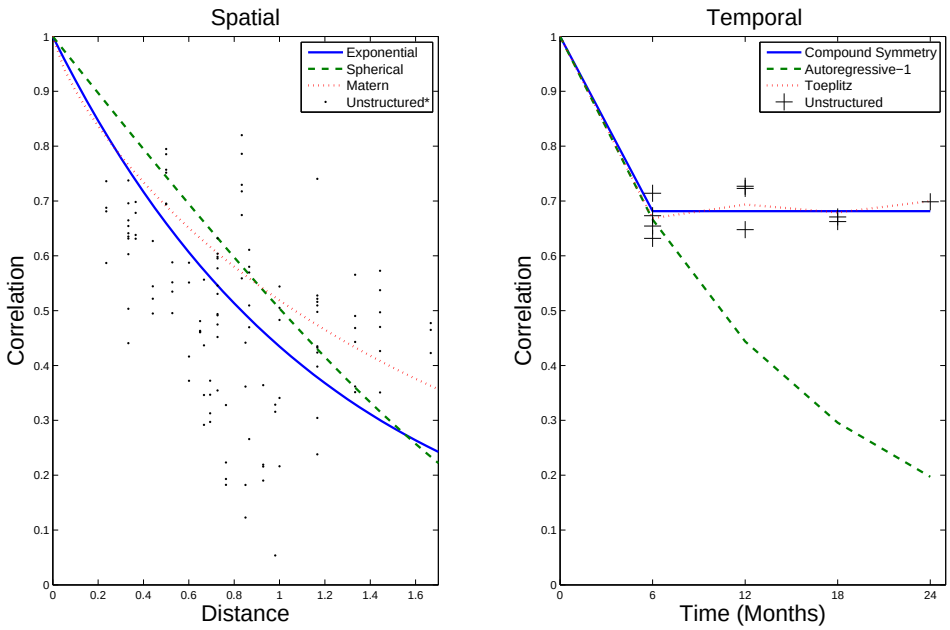

Figure 8: Plot of the estimated correlation functions for the full model of the end-systolic wall thickness in the SCCOR data. On the left are the spatial functions when the temporal correlation is unstructured, while on the right are the temporal structures for a Matérn spatial structure. The unstructured spatial correlations were estimated from subset pairs of the 16 segments.

APPENDIX

Simulation Details

In this simulation study, for the sake of simplicity we assumed $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ and $\sigma^2 = 1$. No loss of generality was expected, as σ^2 simply scales the outcome and inferences about β_5 should not depend on the values of the other βs. The parts of the linear model itself that were changed between conditions include β_5 and $\mathbf{\Sigma}$. Note that β_5 represents the time-by-treatment interaction and if $\beta_5 \neq 0$ then there exists a difference in the time course between the two groups, which was the original research hypothesis. $\beta_5 = 0$ allowed for the testing of the Type I error rate, and the other values were chosen to create sigmoidal power curves.

As mentioned earlier, we assumed that the spatiotemporal correlation structure $\mathbf{\Sigma}$ was separable in space and time. Thus, $\mathbf{\Sigma} = \mathbf{\Sigma}_T \otimes \mathbf{\Sigma}_S$ for parametric temporal ($\mathbf{\Sigma}_T$) and spatial ($\mathbf{\Sigma}_S$) correlation functions. We generated data from combinations of compound symmetric and autoregressive-1 temporal structures and exponential, spherical, and Matérn spatial structures, for a total of 6 possible parametric structures of $\mathbf{\Sigma}$. When choosing the values of θ for $\mathbf{\Sigma}(\theta)$, we looked at values that produced high and low correlation in both space and time for four combinations of degree of correlation. The parameters for the correlation functions that produced the data are given in Table A.2.

We also varied the total number of subjects (N) so that $N = 50$ or 100. The treatment groups were considered to be balanced, with the number per group being 25 and 50, respectively. Note that since there are 80 observations per subject, these numbers of subjects corresponds to 4000 and 8000 total observations, respectively. Due to the computational burden from fitting models with complex covariance structures on 8000 observations, the N=100 sample size was only evaluated for Type I error rate as it can be assumed that, for a given working covariance structure, increased sample size will just increase power and thus not provide enough novel conclusions to justify the lengthy computation time.

To achieve the desired precision in the estimates of Type I errors and selection probabilities, we used a simulation size of 5000 for each condition so that the 95% confidence interval on the estimate of the Type I error rate would have a width of about 1%. The data was generated using the 'mvrnorm' function in the MASS (v. 7.3-29) package of R (v. i368 3.0.2) where each subject's 80 observations were drawn at once, independently from the other subjects. The random number generator used in the 'mvrnorm' function is the Mersenne-Twister generator. The linear model fitting, along with the inference on the fixed effects, was done using the ASReml-R package (v. 3.0, VSN International, Hemel Hempstead, UK)[33]. The 12 possible covariance models were fitted on the same dataset to increase the comparability of the results. The seeds were changed between each simulation iteration so that each run would be generated independently. Care was taken to ensure all twelve models in each iteration converged, typically by increasing the number of iterations or by re-running troublesome datasets with improved initial values. The most problematic structures to fit were ones based off of the spherical spatial model fit to non-spherical data; these sometimes required over 50 iterations to converge, while most models were fit in less than 15 iterations.

Legend of Supplementary Material

- Table A.1: Spatial coordinates of the 16 segments of the AHA model of the left ventricle.
- Table A.2: Parameters for the correlation structures used to generate the simulated data.
- Table A.3: Accuracy of the nine information criteria in selecting the true model for a variety of generating correlation functions and conditions.
- Table A.4: Empirical Type I error rates for the treatment-by-time interaction for various working correlation structures fitted to simulated data generates with a compound

symmetric temporal correlation function.

- Table A.5: Empirical Type I error rates for the treatment-by-time interaction for various working correlation structures fitted to simulated data generates with an autoregressive-1 temporal correlation function.
- Table A.6: Empirical Type I error rates for the treatment-by-time interaction for the working correlation structures chosen by the given information criterion when fit to the given type of generated data.
- Figure A.1: Plot of power curves for a generating covariance structure of EXP$\otimes$CS under high and low degrees of correlation in space and time.
- Figure A.2: Plot of power curves for a generating covariance structure of MAT$\otimes$CS under high and low degrees of correlation in space and time.
- Figure A.3: Plot of power curves for a generating covariance structure of EXP$\otimes$AR-1 under high and low degrees of correlation in space and time.
- Figure A.4: Plot of power curves for a generating covariance structure of SPH$\otimes$AR-1 under high and low degrees of correlation in space and time.
- Figure A.5: Plot of power curves for a generating covariance structure of MAT$\otimes$AR-1 under high and low degrees of correlation in space and time.

Table A.1: Spatial coordinates of the 16 segments in the model of the left ventricle[27, 22]. They are denoted as their level (base, mid, apex), orientation (anterior, septal, inferior, lateral), and index number.

Base, Ant. (1)	$(0,\frac{5}{6})$	Mid, Ant. (7)	$(0,\frac{1}{2})$	Apex, Ant. (13)	$(0,\frac{1}{6})$
Base, Ant.Sep. (2)	$(\frac{-5\sqrt{3}}{12},\frac{5}{12})$	Mid, Ant.Sep. (8)	$(\frac{-\sqrt{3}}{4},\frac{1}{4})$	Apex, Sep. (14)	$(\frac{-1}{6},0)$
Base, Inf.Sep. (3)	$(\frac{-5\sqrt{3}}{12},\frac{-5}{12})$	Mid, Inf.Sep. (9)	$(\frac{-\sqrt{3}}{4},\frac{-1}{4})$	Apex, Inf. (15)	$(0,\frac{-1}{6})$
Base, Inf. (4)	$(0,\frac{-5}{6})$	Mid, Inf. (10)	$(0,\frac{-1}{2})$	Apex, Lat. (16)	$(\frac{1}{6},0)$
Base, Inf.Lat. (5)	$(\frac{5\sqrt{3}}{12},\frac{-5}{12})$	Mid, Inf.Lat. (11)	$(\frac{\sqrt{3}}{4},\frac{-1}{4})$		
Base, Ant.Lat (6)	$(\frac{5\sqrt{3}}{12},\frac{5}{12})$	Mid, Ant.Lat (12)	$(\frac{\sqrt{3}}{4},\frac{1}{4})$		

Table A.2: Parameters used to produce high and low degrees of correlation in the parametric correlation structures.

Correlation Type	Correlation Function	Low Correlation	High Correlation
Spatial	Exponential	$\rho = 0.3$	$\rho = 0.8$
	Spherical	$\phi = 2.25$	$\phi = 6$
	Matérn	$\phi = 0.4, \nu = 1$	$\phi = 0.9, \nu = 2$
Temporal	Compound Symmetry	$\rho = 0.4$	$\rho = 0.8$
	Autoregressive-1	$\rho = 0.6$	$\rho = 0.9$

Table A.3: Accuracy (in %) of the information criteria (IC) with different versions of the sample size penalty (n^*) for data generated from six separate correlation structures under multiple combinations of degree of correlation (high and low in both space [$\mathbf{\Sigma}_S$] and time[$\mathbf{\Sigma}_T$]) and total sample size (N).

Generating			N=50				N=100			
			Low $\mathbf{\Sigma}_S$		High $\mathbf{\Sigma}_S$		Low $\mathbf{\Sigma}_S$		High $\mathbf{\Sigma}_S$	
$\mathbf{\Sigma}$ Structure	IC	n^*	Low $\mathbf{\Sigma}_T$	High $\mathbf{\Sigma}_T$	Low $\mathbf{\Sigma}_T$	High $\mathbf{\Sigma}_T$	Low $\mathbf{\Sigma}_T$	High $\mathbf{\Sigma}_T$	Low $\mathbf{\Sigma}_T$	High $\mathbf{\Sigma}_T$
Exp⊗CS	AIC		72.7	73.8	68.7	69.0	72.7	74.5	72.5	72.8
	AICC	NJK-p	72.7	73.8	68.7	69.1	72.8	74.5	72.5	72.8
	AICC	N	81.9	82.3	77.0	77.8	77.6	79.1	77.1	77.7
	BIC	NJK-p	99.8	99.4	90.5	91.0	99.8	99.6	96.5	96.9
	BIC	N	94.4	94.0	87.2	87.4	96.4	96.3	94.0	94.4
	CAIC	NJK-p	99.9	99.7	90.7	91.2	99.9	99.7	96.5	96.9
	CAIC	N	97.3	97.0	88.9	89.4	97.9	97.8	95.3	95.8
	HQIC	NJK-p	95.5	95.1	87.8	88.2	95.8	95.8	93.4	94.0
	HQIC	N	85.7	86.4	80.4	80.8	88.8	89.0	97.5	88.4
Sph⊗CS	AIC		86.7	86.3	76.3	76.4	87.2	87.0	82.2	82.6
	AICC	NJK-p	86.8	86.3	76.3	76.4	87.2	87.0	82.2	82.6
	AICC	N	93.4	93.1	83.3	83.4	91.4	91.3	85.8	87.1
	BIC	NJK-p	99.9	99.9	93.2	93.6	> 99.9	> 99.9	92.2	98.2
	BIC	N	98.9	98.9	90.5	90.4	99.7	99.6	96.7	97.1
	CAIC	NJK-p	99.9	> 99.9	93.4	93.6	> 99.9	> 99.9	98.3	98.3
	CAIC	N	99.7	99.6	91.8	92.2	> 99.9	99.9	97.4	97.7
	HQIC	NJK-p	99.3	99.2	91.1	91.1	99.7	99.4	96.5	97.0
	HQIC	N	95.0	95.1	85.7	85.8	97.7	97.7	93.0	93.5
Mat⊗CS	AIC		88.1	86.5	87.2	87.2	87.5	87.5	87.5	87.0
	AICC	NJK-p	88.2	86.5	87.3	87.2	91.7	87.6	87.6	87.1
	AICC	N	95.1	94.7	95.0	95.1	87.6	92.3	91.8	91.5
	BIC	NJK-p	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9
	BIC	N	99.3	99.2	99.0	99.3	99.6	99.7	99.6	99.6
	CAIC	NJK-p	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9	> 99.9
	CAIC	N	99.8	99.7	99.8	99.9	99.9	> 99.9	99.9	99.9
	HQIC	NJK-p	99.6	99.4	99.4	99.6	99.5	99.7	99.6	99.6
	HQIC	N	95.9	95.4	95.7	95.8	97.6	97.2	97.3	96.9
Exp⊗AR-1	AIC		74.7	74.2	68.3	68.5	74.8	73.2	72.0	73.0
	AICC	NJK-p	74.8	74.4	68.4	68.6	74.8	73.3	72.1	73.0
	AICC	N	83.0	82.9	76.6	76.6	79.1	78.8	76.8	77.7
	BIC	NJK-p	99.6	99.6	90.4	89.9	99.8	99.7	96.6	96.6
	BIC	N	94.5	94.7	86.8	86.3	96.5	96.7	94.0	94.4
	CAIC	NJK-p	99.8	99.8	90.4	90.1	99.9	99.9	96.7	96.7
	CAIC	N	97.0	97.3	88.6	88.4	98.2	98.2	95.5	95.4
	HQIC	NJK-p	95.5	95.7	87.6	87.24	96.0	96.2	93.5	93.8
	HQIC	N	86.6	86.5	80.1	79.8	90.2	89.8	97.5	87.5
Sph⊗AR-1	AIC		86.9	86.9	76.3	75.9	87.0	86.3	82.0	83.0
	AICC	NJK-p	86.9	86.9	76.3	75.9	87.0	86.3	82.1	83.1
	AICC	N	93.6	93.5	83.4	82.5	91.0	90.6	86.3	86.9
	BIC	NJK-p	99.9	99.8	93.3	93.5	> 99.9	> 99.9	98.4	98.2
	BIC	N	98.9	98.9	90.8	90.6	99.6	99.6	97.0	97.2
	CAIC	NJK-p	> 99.9	99.9	93.5	93.5	> 99.9	> 99.9	98.5	98.2
	CAIC	N	99.5	99.6	92.0	92.3	99.9	99.9	97.7	97.7
	HQIC	NJK-p	99.1	99.2	91.4	91.4	99.5	99.5	96.8	97.0
	HQIC	N	95.2	95.1	85.8	84.8	96.8	97.2	93.4	94.0
Mat⊗AR-1	AIC		87.2	86.9	87.0	86.6	87.0	86.8	86.1	86.9
	AICC	NJK-p	87.3	86.9	87.0	86.6	87.1	86.9	86.1	86.9
	AICC	N	94.7	94.3	95.3	94.5	91.7	91.2	90.9	91.9
	BIC	NJK-p	> 99.9	> 99.9	> 99.9	99.9	> 99.9	> 99.9	> 99.9	> 99.9
	BIC	N	99.1	99.3	99.3	99.0	99.6	99.8	99.7	99.7
	CAIC	NJK-p	> 99.9	> 99.9	> 99.9	99.9	> 99.9	> 99.9	> 99.9	> 99.9
	CAIC	N	99.7	99.9	99.9	99.6	99.9	99.9	99.9	> 99.9
	HQIC	NJK-p	99.4	99.6	99.5	99.3	99.5	99.6	99.6	99.6
	HQIC	N	95.4	95.1	95.9	95.3	97.2	97.1	96.9	97.4

Table A.4: Empirical Type I error rate for inferences on the treatment-by-time interaction where the fitted covariance structure is one of twelve separable spatiotemporal structures. The results are for data generated from three separate spatial correlation structures crossed with a compound symmetric temporal correlation structure under multiple combinations of degree of correlation (high and low in both space [$\boldsymbol{\Sigma}_S$] and time[$\boldsymbol{\Sigma}_T$]) and total sample size (N). Results from the models fitted with the covariance structure used to generate the data are denoted by highlighting. Error rates greater than 0.0625 are considered to be inflated and are in **red bold** while those less than 0.0375 are considered to be too conservative and both are listed in *blue italics*.

Generating $\boldsymbol{\Sigma}$	Fitted $\boldsymbol{\Sigma}$	N=50				N=100			
		Low $\boldsymbol{\Sigma}_S$		High $\boldsymbol{\Sigma}_S$		Low $\boldsymbol{\Sigma}_S$		High $\boldsymbol{\Sigma}_S$	
		Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$
Exp⊗CS	Exp⊗CS	0.0470	0.0592	0.0490	0.0516	0.0510	0.0552	0.0472	0.0550
	Sph⊗CS	**0.0704**	**0.0836**	0.0520	0.0544	**0.0728**	**0.0786**	0.0474	0.0560
	Mat⊗CS	0.0488	0.0604	0.0492	0.0518	0.0514	0.0558	0.0472	0.0548
	Exp⊗AR-1	*0.0076*	*0.0014*	*0.0058*	*0.0012*	*0.0070*	*0.0012*	*0.0076*	*0.0020*
	Sph⊗AR-1	*0.0164*	*0.0022*	*0.0070*	*0.0014*	*0.0138*	*0.0030*	*0.0082*	*0.0020*
	Mat⊗AR-1	*0.0078*	*0.0014*	*0.0064*	*0.0014*	*0.0070*	*0.0010*	*0.0076*	*0.0020*
	Exp⊗Toep	0.0484	0.0582	0.0486	0.0528	0.0494	0.0544	0.0474	0.0552
	Sph⊗Toep	**0.0708**	**0.0844**	0.0512	0.0556	**0.0732**	**0.0786**	0.0476	0.0558
	Mat⊗Toep	0.0488	0.0608	0.0498	0.0532	0.0498	0.0550	0.0468	0.0550
	Exp⊗UN	0.0500	0.0592	0.0496	0.0540	0.0500	0.0554	0.0470	0.0534
	Sph⊗UN	**0.0704**	**0.0864**	0.0506	0.0552	**0.0734**	**0.0798**	0.0480	0.0564
	Mat⊗UN	0.0500	0.0608	0.0494	0.0546	0.0500	0.0558	0.0468	0.0536
Sph⊗CS	Exp⊗CS	*0.0200*	*0.0234*	0.0432	0.0456	*0.0226*	*0.0200*	0.0478	0.0392
	Sph⊗CS	0.0514	0.0508	0.0520	0.0522	0.0540	0.0500	0.0538	0.0480
	Mat⊗CS	*0.0294*	*0.0310*	0.0472	0.0486	*0.0306*	*0.0276*	0.0502	0.0424
	Exp⊗AR-1	*0.0014*	*<0.0002*	*0.0064*	*0.0040*	*0.0018*	*0.0002*	*0.0040*	*0.0004*
	Sph⊗AR-1	*0.0074*	*0.0006*	*0.0082*	*0.0060*	*0.0076*	*0.0010*	*0.0074*	*0.0006*
	Mat⊗AR-1	*0.0032*	*<0.0002*	*0.0072*	*0.0060*	*0.0038*	*0.0002*	*0.0056*	*0.0004*
	Exp⊗Toep	*0.0212*	*0.0234*	0.0446	0.0456	*0.0216*	*0.0202*	0.0484	0.0380
	Sph⊗Toep	0.0504	0.0518	0.0520	0.0530	0.0542	0.0512	0.0546	0.0472
	Mat⊗Toep	*0.0292*	*0.0302*	0.0480	0.0486	*0.0310*	*0.0272*	0.0512	0.0424
	Exp⊗UN	*0.0208*	*0.0238*	0.0426	0.0462	*0.0218*	*0.0206*	0.0480	0.0388
	Sph⊗UN	0.0516	0.0536	0.0516	0.0522	0.0560	0.0500	0.0554	0.0472
	Mat⊗UN	*0.0312*	*0.0324*	0.0472	0.0486	*0.0316*	*0.0280*	0.0508	0.0420
Mat⊗CS	Exp⊗CS	*0.0232*	*0.0234*	*0.0198*	*0.0156*	*0.0180*	*0.0242*	*0.0144*	*0.0178*
	Sph⊗CS	**0.0766**	**0.0754**	0.0494	0.0474	**0.0728**	**0.0716**	0.0512	0.0480
	Mat⊗CS	0.0558	0.0522	0.0518	0.0540	0.0516	0.0524	0.0598	0.0550
	Exp⊗AR-1	*0.0016*	*<0.0002*	*0.0016*	*<0.0002*	*0.0012*	*<0.0002*	*0.0020*	*<0.0002*
	Sph⊗AR-1	*0.0166*	*0.0016*	*0.0100*	*0.0080*	*0.0136*	*0.0012*	*0.0070*	*0.0004*
	Mat⊗AR-1	*0.0084*	*0.0008*	*0.0088*	*0.0080*	*0.0054*	*0.0002*	*0.0094*	*0.0006*
	Exp⊗Toep	*0.0230*	*0.0252*	*0.0200*	*0.0158*	*0.0192*	*0.0242*	*0.0146*	*0.0178*
	Sph⊗Toep	**0.0768**	**0.0744**	0.0506	0.0512	**0.0734**	**0.0710**	0.0520	0.0478
	Mat⊗Toep	0.0546	0.0544	0.0512	0.0556	0.0506	0.0528	0.0592	0.0554
	Exp⊗UN	*0.0230*	*0.0246*	*0.0202*	*0.0162*	*0.0190*	*0.0244*	*0.0152*	*0.0182*
	Sph⊗UN	**0.0776**	**0.0764**	0.0518	0.0502	**0.0732**	**0.0722**	0.0530	0.0478
	Mat⊗UN	0.0564	0.0552	0.0534	0.0536	0.0506	0.0526	0.0602	0.0544

Table A.5: Empirical Type I error rate for inferences on the treatment-by-time interaction where the fitted covariance structure is one of twelve separable spatiotemporal structures. The results are for data generated from three separate spatial correlation structures crossed with an autoregressive-1 temporal correlation structure under multiple combinations of degree of correlation (high and low in both space [$\boldsymbol{\Sigma}_S$] and time[$\boldsymbol{\Sigma}_T$]) and total sample size (N). Results from the models fitted with the covariance structure used to generate the data are denoted by highlighting. Error rates greater than 0.0625 are considered to be inflated and are in **red bold** while those less than 0.0375 are considered to be too conservative and both are listed in *blue italics*.

Generating $\boldsymbol{\Sigma}$	Fitted $\boldsymbol{\Sigma}$	N=50				N=100			
		Low $\boldsymbol{\Sigma}_S$		High $\boldsymbol{\Sigma}_S$		Low $\boldsymbol{\Sigma}_S$		High $\boldsymbol{\Sigma}_S$	
		Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$
Exp⊗AR-1	Exp⊗CS	**0.1566**	**0.2124**	**0.1550**	**0.2132**	**0.1572**	**0.2088**	**0.1604**	**0.2096**
	Sph⊗CS	**0.1930**	**0.2580**	**0.1592**	**0.2176**	**0.1948**	**0.2526**	**0.1604**	**0.2116**
	Mat⊗CS	**0.1578**	**0.2132**	**0.1546**	**0.2136**	**0.1560**	**0.2076**	**0.1598**	**0.2088**
	Exp⊗AR-1	0.0510	0.0484	0.0488	0.0496	0.0510	0.0448	0.0506	0.0552
	Sph⊗AR-1	**0.0746**	**0.0740**	0.0516	0.0496	**0.0718**	**0.0708**	0.0522	0.0560
	Mat⊗AR-1	0.0508	0.0488	0.0490	0.0500	0.0508	0.0452	0.0502	0.0552
	Exp⊗Toep	0.0506	0.0486	0.0496	0.0486	0.0510	0.0448	0.0508	0.0542
	Sph⊗Toep	**0.0760**	**0.0734**	0.0510	0.0500	**0.0718**	**0.0708**	0.0522	0.0566
	Mat⊗Toep	0.0506	0.0486	0.0492	0.0484	0.0510	0.0456	0.0514	0.0538
	Exp⊗UN	0.0538	0.0500	0.0510	0.0494	0.0512	0.0448	0.0522	0.0548
	Sph⊗UN	**0.0768**	**0.0734**	0.0510	0.0520	**0.0718**	**0.0704**	0.0536	0.0562
	Mat⊗UN	0.0536	0.0494	0.0506	0.0498	0.0514	0.0456	0.0526	0.0538
Sph⊗AR-1	Exp⊗CS	**0.0900**	**0.1460**	**0.1488**	**0.2060**	**0.0910**	**0.1458**	**0.1398**	**0.1946**
	Sph⊗CS	**0.1572**	**0.2128**	**0.1620**	**0.2196**	**0.1540**	**0.2152**	**0.1498**	**0.2100**
	Mat⊗CS	**0.1102**	**0.1720**	**0.1538**	**0.2122**	**0.1136**	**0.1716**	**0.1438**	**0.2010**
	Exp⊗AR-1	*0.0196*	*0.0210*	0.0450	0.0466	*0.0172*	*0.0220*	0.0428	0.0476
	Sph⊗AR-1	0.0508	0.0512	0.0526	0.0526	0.0472	0.0536	0.0512	0.0548
	Mat⊗AR-1	*0.0308*	*0.0312*	0.0488	0.0488	*0.0258*	*0.0334*	0.0462	0.0512
	Exp⊗Toep	*0.0196*	*0.0216*	0.0458	0.0454	*0.0172*	*0.0222*	0.0428	0.0474
	Sph⊗Toep	0.0518	0.0510	0.0546	0.0516	0.0480	0.0516	0.0506	0.0556
	Mat⊗Toep	*0.0320*	*0.0318*	0.0488	0.0480	*0.0266*	*0.0336*	0.0456	0.0508
	Exp⊗UN	*0.0216*	*0.0218*	0.0462	0.0472	*0.0178*	*0.0234*	0.0436	0.0474
	Sph⊗UN	0.0518	0.0518	0.0536	0.0530	0.0472	0.0522	0.0502	0.0564
	Mat⊗UN	*0.0314*	*0.0322*	0.0494	0.0498	*0.0270*	*0.0344*	0.0460	0.0504
Mat⊗AR-1	Exp⊗CS	**0.0986**	**0.1268**	**0.0788**	**0.1296**	**0.0930**	**0.1374**	**0.0916**	**0.1236**
	Sph⊗CS	**0.1962**	**0.2370**	**0.1540**	**0.2100**	**0.1894**	**0.2396**	**0.1576**	**0.2034**
	Mat⊗CS	**0.1592**	**0.1998**	**0.1574**	**0.2126**	**0.1496**	**0.2014**	**0.1660**	**0.2112**
	Exp⊗AR-1	*0.0206*	*0.0214*	*0.0166*	*0.0178*	*0.0216*	*0.0210*	*0.0182*	*0.0156*
	Sph⊗AR-1	**0.0742**	**0.0664**	0.0474	0.0528	**0.0688**	**0.0668**	0.0486	0.0466
	Mat⊗AR-1	0.0512	0.0462	0.0532	0.0536	0.0454	0.0458	0.0560	0.0546
	Exp⊗Toep	*0.0216*	*0.0206*	*0.0176*	*0.0180*	*0.0214*	*0.0204*	*0.0182*	*0.0162*
	Sph⊗Toep	**0.0752**	**0.0656**	0.0480	0.0530	**0.0690**	**0.0668**	0.0488	0.0486
	Mat⊗Toep	0.0514	0.0466	0.0524	0.0540	0.0470	0.0458	0.0574	0.0552
	Exp⊗UN	*0.0214*	*0.0220*	*0.0166*	*0.0192*	*0.0220*	*0.0212*	*0.0188*	*0.0158*
	Sph⊗UN	**0.0764**	**0.0674**	0.0484	0.0550	**0.0678**	**0.0668**	0.0486	0.0492
	Mat⊗UN	0.0520	0.0484	0.0534	0.0552	0.0466	0.0472	0.0568	0.0552

Table A.6: Empirical Type I error rate for inferences on the treatment-by-time interaction where the covariance structure used was chosen by the given information criterion (IC) with a certain version of the sample size penalty (n^*). The results are for data generated from six separate correlation structures under multiple combinations of degree of correlation (high and low in both space [$\boldsymbol{\Sigma}_S$] and time[$\boldsymbol{\Sigma}_T$]) and total sample size (N). Error rates greater than 0.0625 are considered to be inflated and are in **red bold** while those less than 0.0375 are considered to be too conservative and both are listed in *blue italics*.

Generating			N=50				N=100			
			Low $\boldsymbol{\Sigma}_S$		High $\boldsymbol{\Sigma}_S$		Low $\boldsymbol{\Sigma}_S$		High $\boldsymbol{\Sigma}_S$	
$\boldsymbol{\Sigma}$ Structure	IC	n^*	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$	Low $\boldsymbol{\Sigma}_T$	High $\boldsymbol{\Sigma}_T$
Exp⊗CS	AIC		0.0474	0.0596	0.0494	0.0522	0.0512	0.0554	0.0476	0.0548
	AICC	NJK-p	0.0474	0.0596	0.0494	0.0522	0.0512	0.0554	0.0476	0.0548
	AICC	N	0.0470	0.0596	0.0496	0.0528	0.0512	0.0556	0.0476	0.0552
	BIC	NJK-p	0.0468	0.0592	0.0498	0.0526	0.0510	0.0552	0.0472	0.0550
	BIC	N	0.0466	0.0598	0.0498	0.0528	0.0516	0.0552	0.0472	0.0548
	CAIC	NJK-p	0.0470	0.0592	0.0498	0.0526	0.0510	0.0552	0.0472	0.0550
	CAIC	N	0.0466	0.0596	0.0498	0.0526	0.0512	0.0552	0.0472	0.0550
	HQIC	NJK-p	0.0466	0.0598	0.0498	0.0528	0.0516	0.0552	0.0472	0.0548
	HQIC	N	0.047	0.0596	0.0498	0.0528	0.0514	0.0558	0.0474	0.0550
Sph⊗CS	AIC		0.0514	0.0510	0.0510	0.0516	0.0546	0.0506	0.0536	0.0484
	AICC	NJK-p	0.0514	0.0510	0.0510	0.0516	0.0546	0.0506	0.0536	0.0484
	AICC	N	0.0516	0.0512	0.0514	0.0520	0.0544	0.0504	0.0536	0.0486
	BIC	NJK-p	0.0514	0.0508	0.0520	0.0520	0.0540	0.0500	0.0536	0.0480
	BIC	N	0.0516	0.0510	0.0516	0.0522	0.0540	0.0500	0.0536	0.0480
	CAIC	NJK-p	0.0514	0.0508	0.0520	0.0520	0.0540	0.0500	0.0536	0.0480
	CAIC	N	0.0514	0.0508	0.0518	0.0522	0.0540	0.0500	0.0536	0.0480
	HQIC	NJK-p	0.0516	0.0508	0.0520	0.0522	0.0540	0.0500	0.0536	0.0480
	HQIC	N	0.0516	0.0510	0.0516	0.0520	0.0538	0.0500	0.0538	0.0484
Mat⊗CS	AIC		0.0558	0.0528	0.0516	0.0548	0.0514	0.0526	0.0600	0.0550
	AICC	NJK-p	0.0558	0.0528	0.0516	0.0548	0.0514	0.0526	0.0600	0.0550
	AICC	N	0.0556	0.0526	0.0520	0.0544	0.0516	0.0524	0.0598	0.0550
	BIC	NJK-p	0.0558	0.0522	0.0518	0.0540	0.0516	0.0524	0.0598	0.0550
	BIC	N	0.0558	0.0522	0.0520	0.0540	0.0516	0.0524	0.0600	0.0550
	CAIC	NJK-p	0.0558	0.0522	0.0518	0.0540	0.0516	0.0524	0.0598	0.0550
	CAIC	N	0.0558	0.0522	0.0518	0.0540	0.0516	0.0524	0.0598	0.0550
	HQIC	NJK-p	0.0558	0.0522	0.0518	0.0540	0.0516	0.0524	0.0600	0.0550
	HQIC	N	0.0556	0.0526	0.0520	0.0544	0.0516	0.0522	0.0596	0.0552
Exp⊗AR-1	AIC		0.0506	0.0492	0.0498	0.0496	0.0510	0.0448	0.0506	0.0546
	AICC	NJK-p	0.0506	0.0492	0.0498	0.0496	0.0510	0.0448	0.0506	0.0546
	AICC	N	0.0510	0.0492	0.0492	0.0494	0.0510	0.0448	0.0508	0.0550
	BIC	NJK-p	0.0510	0.0486	0.0490	0.0498	0.0510	0.0448	0.0510	0.0552
	BIC	N	0.0510	0.0486	0.0494	0.0496	0.0512	0.0448	0.0508	0.0552
	CAIC	NJK-p	0.0510	0.0484	0.0490	0.0498	0.0510	0.0448	0.0510	0.0552
	CAIC	N	0.0510	0.0484	0.0492	0.0498	0.0512	0.0448	0.0510	0.0552
	HQIC	NJK-p	0.0510	0.0488	0.0492	0.0496	0.0512	0.0446	0.0508	0.0552
	HQIC	N	0.0506	0.0491	0.0494	0.0494	0.0508	0.0446	0.0510	0.0550
Sph⊗AR-1	AIC		0.0506	0.0510	0.0526	0.0524	0.0476	0.0534	0.0512	0.0546
	AICC	NJK-p	0.0506	0.0510	0.0526	0.0524	0.0476	0.0534	0.0512	0.0546
	AICC	N	0.0506	0.0512	0.0520	0.0524	0.0474	0.0534	0.0514	0.0546
	BIC	NJK-p	0.0508	0.0512	0.0520	0.0526	0.0472	0.0536	0.0510	0.0546
	BIC	N	0.0508	0.0512	0.0520	0.0526	0.0474	0.0536	0.0512	0.0546
	CAIC	NJK-p	0.0508	0.0512	0.0522	0.0526	0.0472	0.0536	0.0510	0.0546
	CAIC	N	0.0508	0.0512	0.0520	0.0526	0.0472	0.0536	0.0510	0.0546
	HQIC	NJK-p	0.0508	0.0512	0.0520	0.0526	0.0472	0.0536	0.0512	0.0546
	HQIC	N	0.0508	0.0512	0.0518	0.0524	0.0472	0.0536	0.0512	0.0546
Mat⊗AR-1	AIC		0.0504	0.0460	0.0528	0.0540	0.0460	0.0464	0.0566	0.0548
	AICC	NJK-p	0.0504	0.0460	0.0528	0.0540	0.0460	0.0464	0.0566	0.0548
	AICC	N	0.0508	0.0458	0.0530	0.0538	0.0456	0.0458	0.0566	0.0548
	BIC	NJK-p	0.0512	0.0462	0.0532	0.0536	0.0454	0.0458	0.0560	0.0546
	BIC	N	0.0512	0.0462	0.0532	0.0538	0.0456	0.0458	0.0560	0.0546
	CAIC	NJK-p	0.0512	0.0462	0.0532	0.0536	0.0454	0.0458	0.0560	0.0546
	CAIC	N	0.0512	0.0462	0.0532	0.0536	0.0456	0.0458	0.0560	0.0546
	HQIC	NJK-p	0.0512	0.0462	0.0532	0.0536	0.0456	0.0458	0.0560	0.0546
	HQIC	N	0.0510	0.0458	0.0530	0.0536	0.0454	0.0458	0.0564	0.0548

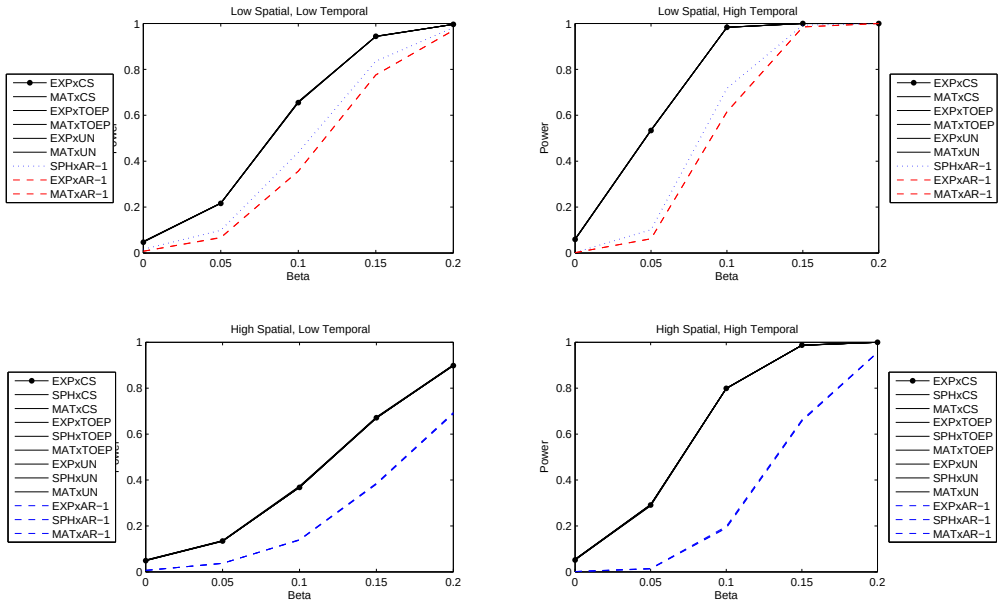

Figure A.1: Plot of power curves for a generating covariance structure of EXP⊗CS under high and low degrees of correlation in space and time. The working covariance structure corresponding to the true structure is denoted with a dot and solid black line. Note that working covariance structures with approximately equivalent power are grouped in the legend and share a common line style and color.

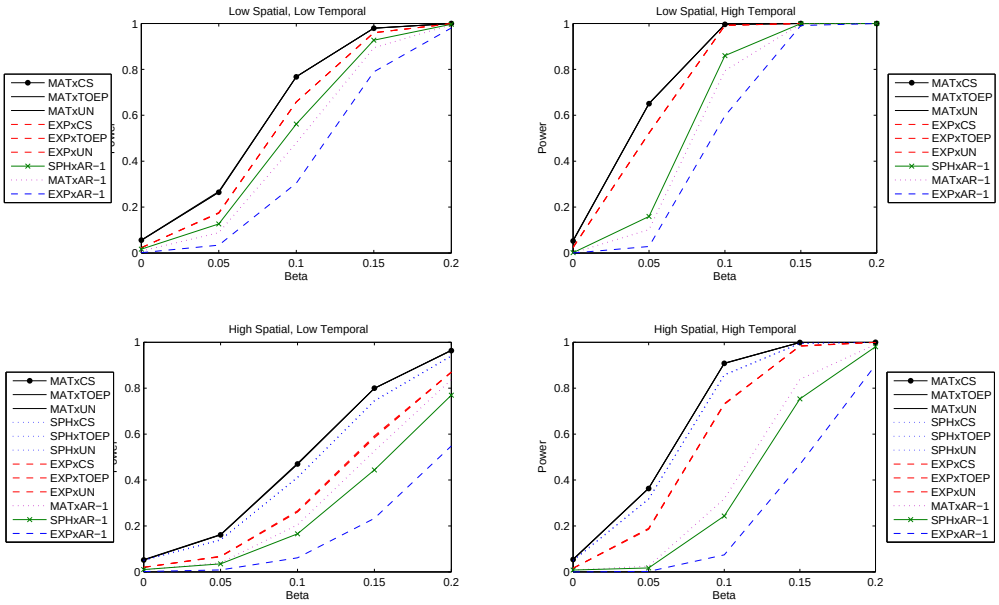

Figure A.2: Plot of power curves for a generating covariance structure of MAT$\otimes$CS under high and low degrees of correlation in space and time. The working covariance structure corresponding to the true structure is denoted with a dot and solid black line. Note that working covariance structures with approximately equivalent power are grouped in the legend and share a common line style and color.

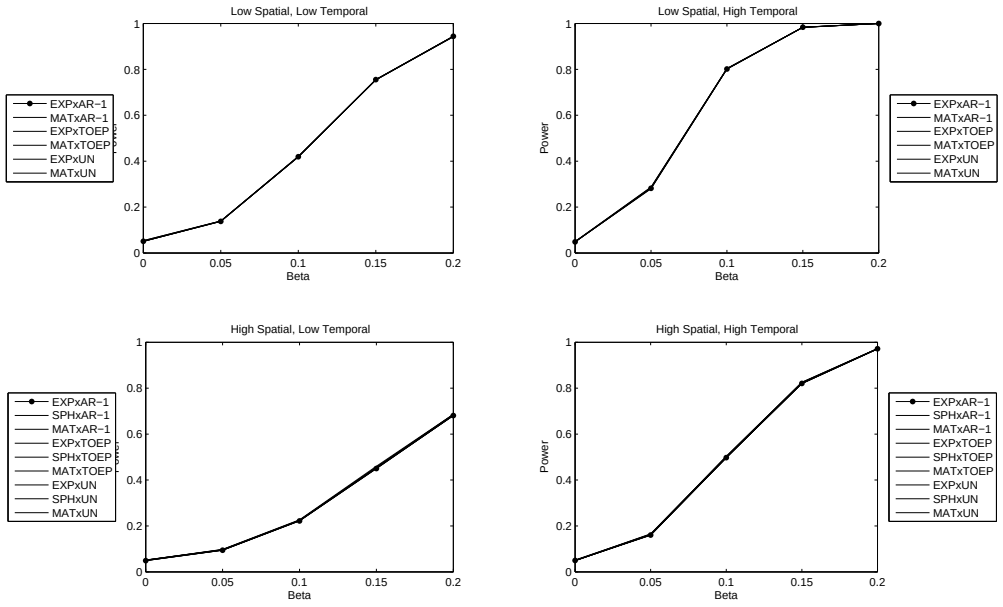

Figure A.3: Plot of power curves for a generating covariance structure of EXP$\otimes$AR-1 under high and low degrees of correlation in space and time. The working covariance structure corresponding to the true structure is denoted with a dot and solid black line. Note that working covariance structures with approximately equivalent power are grouped in the legend and share a common line style and color.

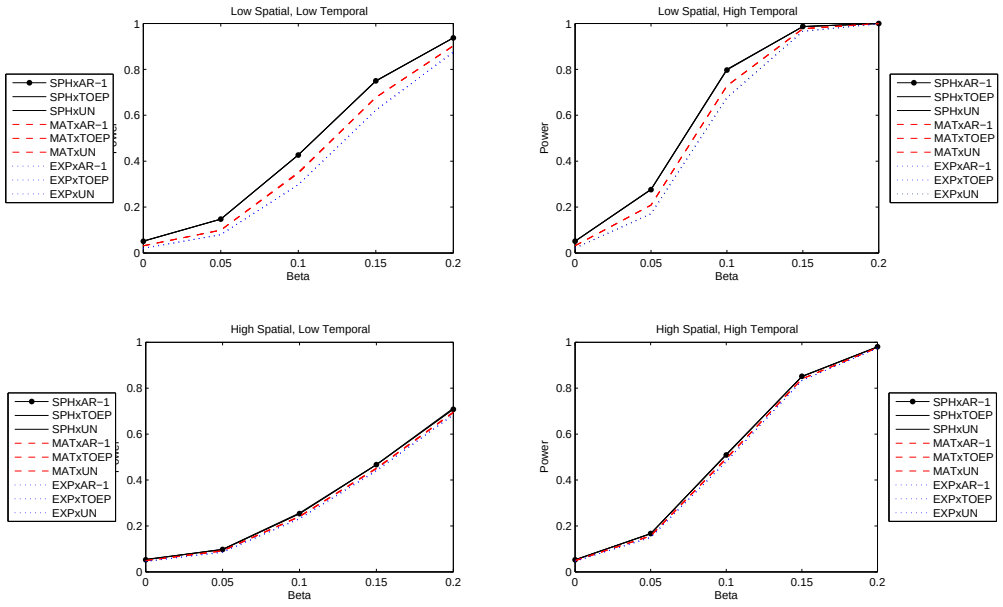

Figure A.4: Plot of power curves for a generating covariance structure of SPH⊗AR-1 under high and low degrees of correlation in space and time. The working covariance structure corresponding to the true structure is denoted with a dot and solid black line. Note that working covariance structures with approximately equivalent power are grouped in the legend and share a common line style and color.

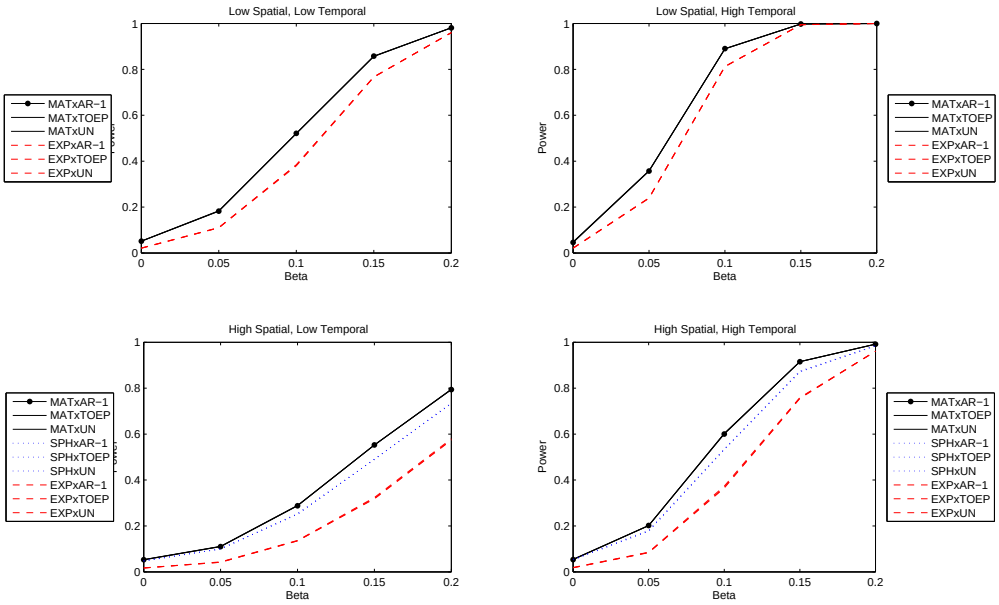

Figure A.5: Plot of power curves for a generating covariance structure of MAT$\otimes$AR-1 under high and low degrees of correlation in space and time. The working covariance structure corresponding to the true structure is denoted with a dot and solid black line. Note that working covariance structures with approximately equivalent power are grouped in the legend and share a common line style and color.

COMPARING SUMMARY METHODS AND A SPATIOTEMPORAL MODEL IN THE ANALYSIS OF LONGITUDINAL IMAGING DATA

BRANDON GEORGE, INMACULADA ABAN

In preparation for the *Journal of Statistical Computation and Simulation*

Format adapted for dissertation

ABSTRACT

Summary methods of both spatial and temporal data have been used previously in the analysis of longitudinal imaging data. This approach has been poorly received by clinical and statistical investigators alike due to the potential loss of power associated with the reduction of data; the objection holds particularly true when considering data obtained with magnetic resonance imaging (MRI) which is extremely expensive to obtain. To address this concern, we have previously proposed the use of a linear model with a separable parametric correlation structure for the error terms.

A simulation study, whose structure was modeled after a longitudinal cardiac imaging study, was done to compare the statistical properties of our proposed method and several common summary measures in time (slope, endpoint, and area-under-the-curve analysis) and space (regional averages analyzed together or separately, and a global average). We found that when testing for a treatment-by-time effect, our model with the separable parametric covariance structure more reliably conserved the Type I error rate and had greater statistical power than the summary methods. Of the temporal summary methods, slope analysis performed best on data generated with a linear time course. A notable finding was that the common practice of analyzing spatial regions separately without a multiple testing correction greatly inflated the Type I error, while the use of a Bonferroni correction resulted in a loss of power. The effects of missing data were considered, and the summary measures were not found to improve in relation to our proposed model.

INTRODUCTION

In longitudinal clinical studies where the outcome is measured at multiple locations via an imaging modality, statistical analysis is complicated by the presence of spatial and temporal correlation among a subject's outcomes. In order to directly model this correlation, we have previously defined and investigated a linear model with a separable, parametric correlation structure for the analysis of longitudinal imaging data. We found that when the parametric structures (both spatial and temporal) are properly specified the Type I error rate is conserved and power for testing a treatment-by-time interaction is maximized. We also found that when the structures were misspecified the Type I error rate could be inflated or overly conservative, the latter leading to reduced statistical power[17]. Overall, it seems that the closer a fitted structure can approximate the true underlying correlation structure, the better the properties of resulting statistical inference. In making the decision for which structure to use in the model, information criteria proved to be highly accurate at selecting the true parametric structure, or at least one that was sufficiently similar to produce approximately the same inference.

Fitting the spatiotemporal model requires specialized software either bought commercially or coded by hand as it cannot be fit using the standard packages in R or procedures in SAS. (Although PROC MIXED can accommodate some crossed structures, in SAS 9.4 they are limited to at least one being unstructured which may be unacceptable for a large number of spatial or temporal observations.) While clinical researchers have relied on simpler methods for analyzing this sort of data[2, 27], statisticians have modeled the data with correlation structures ranging from simple[5] to complex[29, 30, 31].

Another approach is to consolidate the data using summary methods, where summary statistics are used in place of multiple observations to make the resulting correlation simpler or nonexistent. There have been many different summary measures used and refined over the years. The most common summary measure is a slope or average growth

rate[13, 3]; if the data does not follow a linear trend but is at least monotonic, much work has been done for how the data can be weighted or transformed to be linear[6, 25, 24]. Other approaches to summarizing change over time is the total change between the final and baseline observations (also called endpoint analysis), or the use of the parameters of higher order polynomials as an individual's outcome variable[33, 26, 18]. Other methods that may be better for peaked curves rather than monotonic growth include the area under the curve (AUC), a weighted average, the maximum observed value, and the time to the maximum value[23].

Summary measures been used for a very long time[33], as they do have some key benefits. One is that consolidating the data allows for the application of basic statistical methods (t-test, ANOVA) to the summarized data, which was extremely helpful in the era before computers. The elimination of correlation from the observational units is also helpful when the sample size may be too small to properly estimate covariance parameters[14]. Another benefit is that the decision of what summary measure to use can be scientifically useful as it focuses the analysis on the true purpose of the study and is usually very easy to interpret[23, 13]. For example, if the growth rate between treatment groups is of interest then the individual growth rates can be modeled directly.

There are several downsides to the use of summary measures, however. It involves 'throwing away' data, which investigators loathe due the expense of data collection, particularly in longitudinal imaging studies and even more so in those that use MRI[15]. It can seem quite wasteful to spend millions collecting data only to not use all of it. Summary measures may also be misleading as two different response profiles can produce the same summary value; it may also be the case that a summary measure simply does not meaningfully capture an individual's responses[24, 14]. A further limitation is that summarizing data over time or space precludes the use of time- or space-varying covariates in the analysis[14].

Summary measures also have an interesting relationship with missing data, as it can be both a benefit and limitation to the approach. On the positive side, for certain measures a summary measure can be calculated for a subject with some missing data; this prevents the subject from being excluded in complete case analysis or eliminates the need to handle missing data in the model. The downside is that just because the summary measure can be calculated, it may have different statistical properties that can violate model assumptions. For instance, most summary measures are calculated as a weighted sum or difference of the observed values; the number of observations being summed or subtracted will directly affect the variance of the measure which can introduce heteroscedasticity into the dataset[13, 14]. This can lead to bias in estimates and inflation of the Type I error rate[18, 16].

A possible justification for summary measures is that if correlation between observations is high then the data can be summarized without a meaningful loss of information, since in correlated data there is less unique information than if the observations were all independent. However, very little work has been done to quantify how much power is lost by using summary measures compared to a method that utilized all the data and directly modeled the correlation. Many examples have been given that have compared how the two approaches differ for a sample dataset[6, 26, 18, 16, 13], but to our knowledge the only simulation study comparing them was done by Zucker, Manor, and Gubman (2012). Their study compared mixed models with random effects for both the intercept and slope to averages, both simple and time-weighted, in their Type I and II error rates for simulated data. The conditions were straightforward with no missing data and a small (≤ 5) number of repeated measurements, and they found that in these cases the summary measures tended to perform as well as if not better than the mixed models[34]. Although there seems to be evidence that summary measures should be considered as a serious alternative to directly modeling correlation, it is currently unknown how the added complications of spatial correlation and missing data affect that conclusion.

Our research into the comparison of summary measures and spatiotemporal models has been motivated by longitudinal imaging studies, specifically those in cardiology. In the past, these studies have either avoided outcomes with spatiotemporal correlation in favor of 'global' outcomes such as ejection fraction[1], or have averaged the segments within different levels (base, mid, and apex) of the left ventricle and compared the levels through pairwise comparisons at set time points or through a longitudinal mixed model[2, 4, 27]. These are examples of spatial summary measures, in contrast to the temporal summary measures discussed above. The core idea behind spatial summary measures is that the outcomes from a given spatial area are summed or averaged over either subregions or the entire observed area. The choice of subregions is entirely specific to a given application, but as an example the imaging data from the 16 segments of the left ventricle could be averaged according to the level it resides in or by the coronary artery that feeds it[11].

The pros and cons of spatial summary measures are fairly similar to those of the temporal kind. In exchange for simplifying the model and removing the spatial correlation, spatial summary measures result in a loss of information. In particular, there is a loss of spatial resolution, or the ability to identify how the outcome changes over space. For example, summarizing over the levels of the left ventricle precludes the ability to identify whether the outcome changes going from the anterior to the inferior side of the heart. In addition, summarizing over space introduces the modifiable area unit problem which is where the significance level of a treatment or exposure can vary based on how the total observed area is subdivided. Looking at a larger number of subregions helps prevent this problem at the cost of introducing more complexity into the model. In practice, it may be best to find an intermediate number of subregions to break the data into[32].

The particular scenarios of interest were inspired by the arm of the UAB SCCOR study which looked at how medical therapy affected the progression of mitral regurgitation[2]. It did so by looking at functional (strain rate, rotation) and geometric (radius of curvature-

to-wall thickness ratio) measures of disease progression at the segment-level, averaged over the levels or the entirety of the left ventricle. Unfortunately, that initial analysis failed to find a significant treatment-by-time interaction for any of the segment-level outcomes[2]. It is the hope of the clinical researchers that the use of all of the data through a spatiotemporal covariance structure will increase statistical power and give the ability to find a significant treatment effect.

There has been no prior work that compared spatial and temporal summary measures to a full spatiotemporal model with regard to statistical inference. It is the goal of this paper to determine how the use of summary measures and correlation structures affect the Type I and Type II error rates for testing a treatment effect measured by a treatment-by-time interaction. We will consider the effects of sample size, degree of correlation, generating correlation function, and the presence of missing data on inference from these models. Through the use of simulation studies, we hope to answer the question of whether our previously proposed spatiotemporal model is worth the added complexity.

STATISTICAL MODELS

In this section, we shall outline the statistical models used to incorporate the spatial and temporal aspects of the data, either through a correlation structure or through a summary measure. In all cases, consider a longitudinal imaging study with N subjects observed at J points in time, and at K locations at each time point. We consider Y_{ijk} to be the observed outcome of the i^{th} subject at the j^{th} time point and k^{th} location for $i = 1, ..., N$, $j = 1, ..., J$, and $k = 1, ..., K$. This means that, for complete data, each subject has JK observed outcomes, which can be summarized as the JK-length vector $\mathbf{Y}_i$. The set $\{t_1, ..., t_j, ..., t_J\}$ are the times of the J observations. Also consider the $JK \times p$ matrix $\mathbf{X}_i$ to be (possibly space- and time-varying) covariates for subject i, so that $E[\mathbf{Y}_i] = \mathbf{X}_i\boldsymbol{\beta}$. In this case, we will also assume normality and homoscedasticity such that $\mathbf{Y}_i \sim MNV(\boldsymbol{\mu}_i, \sigma^2\boldsymbol{\Sigma})$ for mean vector $\boldsymbol{\mu}_i$ and correlation matrix $\boldsymbol{\Sigma}$.

Separable Parametric Spatiotemporal Model

In our previous paper, the properties of a spatiotemporal linear model with a separable parametric correlation structure were explored. This model was defined as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{1}$$

where $\boldsymbol{\beta}$ are the p parameters corresponding to the covariates and $\boldsymbol{\epsilon}_i \sim MVN(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$. For applications where a treatment group and treatment-by-time interaction is of main interest, we can consider a simple model for our clinical application to be

$$E[Y_{ijk}] = \beta_0 + \beta_1 Time_j + \beta_2 Group_i + \beta_3 Mid_k + \beta_4 Apex_k + \beta_5 Time_j Group_i \tag{2}$$

where $Time_j$ is the time of the j^{th} visit, $Group_i = 1$ if subject i is in the main treatment group and equal to 0 for placebo, and Mid_k and $Apex_k$ are indicator variables for whether segment k is in the mid or apex of the left ventricle.

For our model we assume that the correlation matrix $\boldsymbol{\Sigma}$ is separable and parametric such that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_T \otimes \boldsymbol{\Sigma}_S$, where $\boldsymbol{\Sigma}_T$ and $\boldsymbol{\Sigma}_S$ are parametric temporal and spatial correlation matrices of dimension $J \times J$ and $K \times K$, respectively, with parameters $\boldsymbol{\theta}$. To represent its parametric nature the correlation matrix could be referred to as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. There are near-endless choices for parametric correlation functions, but in our previous work we considered exponential, spherical, and Matérn for spatial correlation and compound symmetry, autoregressive-1, Toeplitz, and unstructured for temporal correlation. We found that when fitting the true correlation structure to the data, or at least one that could closely approximate the true structure, the Type I error rate was conserved and the Type II error rate was minimized. When the fitted correlation structure did not properly match the data, we found that it could inflate the Type I error rate or markedly reduce the statistical power for hypothesis tests about $\boldsymbol{\beta}$. Fortunately, we found that information criteria were highly accurate

at choosing the true structure, or at least one that matched the data well enough to reliably conserve the Type I error rate[17]. Note that the consistent (CAIC[7], BIC[28], HQIC[19]) information criteria tended to perform better, with the most accurate being CAIC with a sample size adjustment of $NJK - p$.

Due to the potentially large number of parameters to estimate in $\boldsymbol{\theta}$, it is recommended that restricted maximum likelihood (REML) estimation be used in lieu of standard maximum likelihood (ML)[20]. This involves maximizing the following restricted log-likelihood according to σ^2 and $\boldsymbol{\theta}$, which can be done using iterative algorithms such as those described by Jennrich and Schluchter (1986) and Lindstrom and Bates (1988)[21, 22]. Note that this estimation can still be done if there is missing data in subject i's outcomes by taking a submatrix of $\mathbf{X}_i$ and $\boldsymbol{\Sigma}$ corresponding to the non-missing values in $\mathbf{Y}_i$.

$$\begin{aligned} REML_1(\sigma^2, \boldsymbol{\theta}|Y) = &- \frac{NJK - p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{N}\mathbf{X}_i'[\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1}\mathbf{X}_i\right| - \frac{1}{2}\sum_{i=1}^{N}\log|\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})| \\ &- \frac{1}{2}\sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})'[\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) \end{aligned} \tag{3}$$

General Notes Regarding Summary Methods

As mentioned previously, summary methods are a common way researchers consolidate their data in order to remove correlations between observations. Removing correlation not only makes the model simpler but may be necessary if the sample size is too small to estimate correlation parameters[14, 15]. This can be seen in Table 1, where most of the summary methods reduce the J or K observations through time or space into a single observation with no correlation parameters. Much like we previously assumed spatial and temporal correlation functions can be chosen separately, spatial and temporal sum-

mary methods can be mixed and matched as well; combining two summary methods on spatiotemporal data could reduce the data to a single observation per subject, which could be analyzed using a test as simple as the t-test. A major drawback of this simplicity is that it prevents the addition of time- or space-varying covariates[14, 15], as will be discussed in the simulation section.

The specifics of the various summary measures will be detailed in the next two sections, but there are some common traits that all of them share. One is that they are all essentially weighted summations of the observed outcomes, although the weights vary greatly. One implication of this is that when the outcome is assumed to be normally distributed the summary measure will also have a normal distribution, albeit with different mean and variance. Another consequence is that the variance of the summary measure is a function of the weighted sum of the correlations of the included observations. Thus, although summary measures may eliminate correlated outcome variables, the correlation inherent in the original data can have a large impact on model efficiency and assumptions. Of particular concern is the assumption of homoscedasticity, since missing or unevenly spaced data can make variances different between subjects. This is due to the summation weights changing as well as the number of terms in the summation differing between subjects. Although in theory one could add weights to the summary measures to correct for the missingness or unevenness of observations, those weights would be a function of the unknown correlation between observations and thus not be suitable in practice[14, 15].

Spatial Summary Methods

Regional Summary Measures

In order to more easily analyze data collected at a variety of spatial locations, a common strategy is to take an average over part of the total observed area. This approach has the benefit of reducing the dimensionality of the problem to a more manageable level.

These averages between subregions can then be analyzed concurrently using a parametric correlation function or, if the number of subregions is sufficiently small, an unstructured correlation model. In our cardiac MRI example, this would be reducing the sixteen segments to three levels which reduces the number of pairs in the unstructured model from 120 to 3, easily estimable even in a small sample size. In practice, the subdivision of the total observed area is entirely specific to a particular application. There may even be multiple ways to subdivide the same data; for instance, the sixteen segments could be divided by level or by which coronary artery feeds them. The division should be a function of the science of the study rather than p-values, as it is well known that redrawing boundaries can alter the conclusions of statistical inference[32].

An alternative to modeling the correlation between subregion averages would be to analyze the subregions in separate models. This approach has seen use in clinical research[12, 27], but has unknown statistical validity. Performing multiple tests could inflate the Type I error rate, and a p-value correction (such as Bonferroni) has the potential to be overly conservative. Due to the multiple tests being correlated, it is unclear how inflated or overly conservative the two approaches would be. It is one of the goals of this paper to quantify how analyzing subregions separately affects inference, both with and without a correction.

To generalize the approach of taking an average within a region, consider the K spatial observations as being divided into M subregions, with a location belonging to exactly one subregion. To define this division, we define the set K_m, $m = 1, ..., M$, as the set containing the indices of all the spatial locations corresponding to the m^{th} division; $|K_m|$ denotes the number of locations within that set. Therefore, the spatial summary measures corresponding to the averages within these regions are defined as

$$Y_{ij}^{m} = \frac{1}{|K_m|} \sum_{k \in K_m} Y_{ijk} \tag{4}$$

which has variance

$$Var(Y_{ij}^m) = \frac{\sigma^2}{|K_m|^2}\left[|K_m| + 2\sum_{k<l;\ k,l\in K_m}\sum Corr(Y_{ijk}, Y_{ijl})\right]. \tag{5}$$

As a result of summarizing the spatial data in this way, the data vector for a given subject will have only M rows per temporal observation instead of the previous K. If the subregions are analyzed in the same model, then $\mathbf{Y}_i$ will have length JM, while separate analysis of the regions would involve M outcome vectors each having length J. It is important to note that the variance of the M outcomes may be heterogeneous, as both the number of locations being averaged ($|K_m|$) and the sum of correlations between locations may vary from one region to the next. If this is the case, then the assumption of homoscedasticity may be violated which could have adverse effects on statistical inference. In addition, the kinds of predictors that one can be fit are limited so that the space-varying covariates are at the region-level rather than the level of the observed locations.

A Global Summary Measure

Another approach to summarizing the data in space is to take the average of all of the locations, reducing the spatial dimension from K to 1. Correspondingly, the length of $\mathbf{Y}_i$ is reduced from JK to just J. This approach has also seen clinical use due to how it totally eliminates the spatial correlation from the modeling of the data[2], although by doing so all of the spatial resolution is lost from the analysis and any space-varying covariates are excluded from modeling. The calculation of the single outcome per time point is quite simple,

$$Y_{ij}^{Global} = \frac{1}{K}\sum_{k=1}^{K} Y_{ijk}, \tag{6}$$

but when the variance is considered it becomes apparent that the spatial correlation does

affect the variance of the outcome.

$$Var(Y_{ij}^{Global}) = \frac{\sigma^2}{K^2}\left[K + 2\sum\sum_{1\leq k<l\leq K} Corr(Y_{ijk}, Y_{ijl})\right] \tag{7}$$

Specifically, when the correlation increases the variance of the average increases.

Temporal Summary Methods

Due to longitudinal studies having a longer history than imaging studies, many more summary measures have been proposed to handle longitudinal data than spatial data. There are enough different types of summary measures to fill an entire review article, possibly more, so here we will focus on three of the most popular ones: endpoint analysis, the slope over time, and the area under the curve. Note that the choice of a temporal summary measure should be motivated by the research question pertaining to the time course of the outcome[23], since in practice they all reduce the temporal dimension from J to 1 and prevent the use of time-varying covariates.

In addition to the properties of the measures themselves, we will also consider how they perform in the presence of missing data. Missing data is a near certainty in longitudinal studies, especially when the subjects are human, and it tends to take the form of missing observations in time. Closely related is the issue of uneven follow-up, where subjects may have the same number of observations but they were taken at different times from baseline. Note that we are only considering temporal missing rather than spatial; this is due to missing data in longitudinal imaging studies being far more likely to come from a missed visit or dropout than from part of an image being missing. Low-quality images do happen, but one can generally assume that the imaging technician would be aware of when that occurred and take additional images until an acceptable one has been obtained.

Endpoint Analysis

Endpoint analysis is where the baseline observation is subtracted from the final observation to give the total change over the study's duration. The calculation is extremely simple as is its variance,

$$Y_{ik}^{Endpoint} = Y_{iJk} - Y_{i1k} \tag{8}$$

$$Var\left(Y_{ik}^{Endpoint}\right) = 2\sigma^2[1 - Corr(Y_{iJk}, Y_{i1k})], \tag{9}$$

but the result has the potential to directly answer certain research questions. Endpoint analysis is best suited to answer the question of how an exposure or treatment changes an outcome after a certain period of time[18]. In the context of cardiac imaging in patients with mitral regurgitation, it could be used to quantify how much the left ventricular ejection fraction had decreased after a certain period of time.

In practice, endpoint analysis is best when the total change is of interest rather than the rate of change. If the time course is linear, then the slope and endpoint analysis should give the same answer but the slope will typically have a smaller variance. Therefore, endpoint analysis is often more suited for when the time course is non-linear, such as when change levels off with time[23]. Similar options to endpoint analysis exist that may be more efficient, such as an ANCOVA where the final observation is used as the outcome and the baseline is used as a covariate[16, 13].

Endpoint analysis discards the observations between the initial and final observation, which involves throwing away a large amount of information. On the bright side, this means that it does not matter if those observations are missing or unevenly spaced. However, if the J^{th} observation is missing it means that the subject cannot be used. One may think to use the last observed value, but this can bias the outcome measure towards zero. In practice, endpoint analysis can produce a great deal of bias and inflate the Type I error rate when the missingness is not completely at random[18]. If there is a large range in

when the J^{th} observations were taken, endpoint analysis may also be inappropriate since the subjects would not have had the same amount of time for change to have occurred. This unevenness can also result in heteroscedasticity, since if the time between Y_{iJk} and Y_{i1k} is not uniform then the correlation (and also $Var\left(Y_{ik}^{Endpoint}\right)$) will probably be different for each subject[14].

Slope Analysis

One of the more common approaches to summarizing longitudinal data is to fit a parametric growth curve to each subject's time course and use one or more of the estimated parameters as that subject's outcome variable. The nature of the parametric curve is dependent on the 'typical' shape of subjects' change over time, but the most basic would be to fit a line using ordinary least squares estimation and take the slope as the outcome variable.

$$Y_{ik}^{Slope} = \frac{\sum_{j=1}^{J}\left(t_j - \bar{t}\right)\left(Y_{ijk} - \frac{1}{J}\sum_{l=1}^{J} Y_{ilk}\right)}{\sum_{j=1}^{J}\left(t_j - \bar{t}\right)^2} \tag{10}$$

This approach is best suited for cases where the average change over time is of scientific interest to a researcher[23]. In practice, using the slope as the outcome has many similarities to a mixed model with random slopes for the individuals[24], though it has the potential to be highly efficient due to not having to estimate covariate parameters affecting the intercept[34].

The variance of the slope is much more complicated to compute than the variance of the change score; the details of the derivation are given in the Supplementary Material.

The variance's full form is

$$
\begin{aligned}
Var\left(Y_{ik}^{Slope}\right) = & \frac{\sigma^2}{\left[J\sum_{j=1}^{J} t_j^2 - \left(\sum_{j=1}^{J} t_j\right)^2\right]^2} \left[J^2 \left(\sum_{j=1}^{J} \left[t_j \sum_{l\in[1,J],l\neq j} t_l Corr(Y_{ijk}, Y_{ilk}) \right] \right) \right. \\
& + \left(\sum_{j=1}^{J} t_j \right)^2 \left(2 \sum_{1\leq j<l\leq J} \sum Corr(Y_{ijk}, Y_{ilk}) \right) \\
& - J \left(\sum_{j=1}^{J} t_j \right) \left(\sum_{j=1}^{J} \left[t_j \sum_{l\in[1,J],l\neq j} Corr(Y_{ijk}, Y_{ilk}) \right] \right) \\
& - J \left(\sum_{j=1}^{J} t_j \right) \left(\sum_{j=1}^{J} \sum_{l\in[1,J],l\neq j} t_l Corr(Y_{ijk}, Y_{ilk}) \right] \\
& \left. + J \left(J \sum_{j=1}^{J} t_j^2 - \left[\sum_{j=1}^{J} t_j \right]^2 \right) \right] .
\end{aligned}
\tag{11}
$$

If we make some simplifying assumptions it is possible to reduce Equation 11 to a form that allows for easy understanding of how the study and model parameters affect the variance. If we assume that the times between follow-up visits are fixed at t for all subjects with no loss to follow-up, and that the temporal correlation is compound symmetric with $Corr(Y_{ijk}, Y_{ilk}) = \rho$, then we can compute the variance of an individual's slope to be

$$
Var\left(Y_{ik}^{Slope}\right) = \frac{\sigma^2 12(1-\rho)}{t^2 J(J+1)(J-1)} \tag{12}
$$

under these assumptions. Note that as the number of observations (J) or the spacing between them (t) increase, the variance of the slope decreases. In addition, the variance decreases as the temporal correlation ρ approaches 1, suggesting that slope analysis is even more efficient under cases of high temporal correlation.

If the growth is nonlinear, several options exist. One would be to fit higher order polynomials and analyze the different coefficients separately[33]. Polynomials up to the order $J-1$ could be fit (assuming no missing data), and although higher order terms will

improve the fit of the curves there is a potential loss of interpretability as well as diminishing returns for the additional degrees to the polynomial[26]. Another option would be to transform the observed outcome variables so they are approximately linear in time; this could be done using a log transformation[6, 25] or by the use of weighting[24]. It has been noted that for small sample sizes, such as those in imaging studies, it may be impossible to use statistical metrics to choose between growth curve models, so it is important to let the research question and good judgment guide the decision[25].

Assuming that the time courses are truly linear (or that the growth curve or transformation is appropriate), the slope is fairly robust to missing data or unevenly spaced observations. At least, given that the model is correct one would not expect either case to cause bias. Missing data and unevenly spaced observations can affect the efficiency of the approach[3], however, as $Var(Y_{ik}^{Slope})$ is a function of both the number of observed time points and the spread of those points through time ($\sum_{j=1}^{J}(t_j - \bar{t})^2$).

Area Under the Curve Analysis

Another option for summarizing longitudinal data is to take the area under the curve (AUC). The AUC may be useful when the time courses are not monotonic as it provides a measure of whether the observations tended to be higher for a longer duration over the study period. Such situations may occur when the outcome is highly variable over time or if the 'standard' time course is naturally peaked. The calculation is a fairly simple integration, typically done using the method of trapezoids, and is essentially a time-weighted summation of the observed values[23].

$$\begin{aligned} Y_{ik}^{AUC} &= \frac{1}{2}\sum_{j=1}^{J-1}(t_{j+1} - t_j)(Y_{ijk} + Y_{i,j+1,k}) \\ &= \frac{1}{2}\left[(t_2 - t_1)Y_{i1k} + (t_J - t_{J-1})Y_{iJk} + \sum_{j=2}^{J-1}(t_{j+1} - t_{j-1})Y_{ijk}\right] \end{aligned} \tag{13}$$

The variance of the AUC is much more complicated due to the sum of sums in Y_{ik}^{AUC}.

$$\begin{aligned} Var\left(Y_{ik}^{AUC}\right) = \frac{\sigma^2}{4}\Bigg[&\left((t_2-t_1)^2+(t_J-t_{J-1})^2+\sum_{j=2}^{J-1}(t_{j+1}-t_{j-1})^2\right) \\ &+2(t_J-t_{J-1})(t_2-t_1)Corr(Y_{iJk},Y_{i1k}) \\ &+2\sum_{j=2}^{J-1}(t_2-t_1)(t_{j+1}-t_{j-1})Corr(Y_{i1k},Y_{ijk}) \\ &+2\sum_{j=2}^{J-1}(t_J-t_{J-1})(t_{j+1}-t_{j-1})Corr(Y_{iJk},Y_{ijk}) \\ &+2\sum_{j=2}^{J-2}\sum_{l=j+1}^{J-1}(t_{j+1}-t_{j-1})(t_{l+1}-t_{l-1})Corr(Y_{ijk},Y_{ilk})\Bigg] \end{aligned} \tag{14}$$

In this form, some of the nuances of the variance can be lost in the complexity. Assuming the time between follow-up visits is fixed at t for all subjects, the variance simplifies to

$$\begin{aligned} Var\left(Y_{ik}^{AUC}\right) = \frac{\sigma^2t^2}{4}\Bigg[&(4J-6)+2Corr(Y_{iJk},Y_{i1k})+4\sum_{j=2}^{J-1}Corr(Y_{i1k},Y_{ijk}) \\ &+4\sum_{j=2}^{J-1}Corr(Y_{iJk},Y_{ijk})+8\sum_{j=2}^{J-2}\sum_{l=j+1}^{J-1}Corr(Y_{ijk},Y_{ilk})\Bigg]. \end{aligned} \tag{15}$$

A further simplifying assumption to make the variance of the AUC easier to understand would be to assume a compound symmetric correlation of ρ between all observations through time. After some algebra, the variance in this case becomes

$$\begin{aligned} Var\left(Y_{ik}^{AUC}\right) &= \frac{\sigma^2t^2}{4}\left[(4J-6)+(4J^2-12J+10)\rho\right] \\ &= \frac{\sigma^2t^2}{2}\left[(2J-3)+(2J^2-6J+5)\rho\right] \end{aligned} \tag{16}$$

which shows that the variance of the AUC increases quadratically with both the time between observations (t) and the number of observations (J).

Due to this property, it is advisable to instead use the ratio of the AUC to the time of the final observed outcome. This adjustment results in the summary measure being akin to a time-weighted average of the observed Y values. For complete data, the adjusted Y^{AUC} would be

$$\begin{aligned}\frac{Y_{ik}^{AUC}}{t_J} &= \frac{1}{2t_J}\sum_{j=1}^{J-1}(t_{j+1}-t_j)(Y_{ijk}+Y_{i,j+1,k}) \\ &= \frac{1}{2t_J}\left[(t_2-t_1)Y_{i1k}+(t_J-t_{J-1})Y_{iJk}+\sum_{j=2}^{J-1}(t_{j+1}-t_{j-1})Y_{ijk}\right]\end{aligned} \tag{17}$$

with variance

$$Var\left(\frac{Y_{ik}^{AUC}}{t_J}\right) = \frac{Var\left(Y_{ik}^{AUC}\right)}{t_J^2}. \tag{18}$$

This adjustment also has the benefit of stabilizing the variance of the summary measure. Under the previous simplifying assumptions of equally spaced observations (making $t_J = (J-1)t$) and compound symmetric correlation, the variance of the adjusted AUC is

$$Var\left(\frac{Y_{ik}^{AUC}}{(J-1)t}\right) = \frac{\sigma^2}{2(J-1)^2}\left[(2J-3)+(2J^2-6J+5)\rho\right] \rightarrow \rho\sigma^2 \text{ as } J\rightarrow\infty \tag{19}$$

which is no longer an increasing function of t or J. The AUC is still limited, though, since as long as $\rho > 0$ the variance of the adjusted AUC will not go to zero even as the length of follow-up increases. In addition, the variance increases with ρ which suggests that AUC analysis may become less efficient is scenarios with high temporal correlation.

It is apparent that the presence of unevenly spaced observations can drastically affect both the value of Y_{ik}^{AUC} and its variance. Missing data from the second to $J-1^{st}$ observations may not impact the calculation of the AUC since the curve can just be drawn between the adjacent observed outcome values, though like all summary methods it can affect the variance. The largest concern when using AUC as a summary measure is when the last observed value was taken. Ideally, each subject will have their J^{th} observation taken at

time t_J. If the duration of observation is different between subjects, however, the AUC will be larger or smaller for longer or shorter follow-up times, respectively. This could occur for uneven follow-up times for the J^{th} observation or if the J^{th} observation is missing. Adjusting by t_J may not fix this problem, as reduced follow-up duration implies that there is less time for a time-weighted average to move away from Y_{i1k}. Consequently, it can be very dangerous to use the AUC (adjusted or not) when the total follow-up time is different between subjects.

SIMULATION STUDY DESIGN

The goal of this simulation study was to assess how our proposed spatiotemporal model compares with summary measures in both space and time with regard to statistical inference (Type I and II error rates) about a treatment-by-time interaction. These error rates were investigated under several different conditions such as generating correlation function, degree of correlation, and sample size. Additionally, the effects of missing data were considered as were whether the inference was made with a simple Wald's test or an F-test with a denominator degrees of freedom correction. In the design and reporting of these simulations, we endeavored to adhere to the guidelines proposed by Burton et al. (2006)[9].

Basic Structure of Generated Data

The generated data is based on the structure of the data collected in the cardiac imaging study reported on by Schiros et al. (2012)[27]. The outcome variable is continuous and was collected from two treatment groups at five time points, evenly spaced six months apart (J=5). At each observation, cardiac imaging was done where the outcome was observed at 16 points within the left ventricle (K=16). Thus, each subject had 80 observed outcomes. The layout of the spatially observed locations was modeled using the AHA's

17 segment model presented by Cerqueria et al. (2002), where the left ventricle is laid out in a circular pattern with concentric rings corresponding to the different levels of the left ventricle (base, mid, and apex), much as if one is looking down from the left atrium[11]. The coordinates were defined as the center of each segment laid out on a unit circle, defined in Table A.1 and shown in Figure A.1 with the spatial lag defined as the Euclidian distance between the centers[29].

The outcome was calculated as the result of a linear model with correlated errors. Specifically,

$$Y_{ijk} = \beta_0 + \beta_1 Time_{ij} + \beta_2 Group_i + \beta_3 Mid_k + \beta_4 Apex_k + \beta_5 Time_{ij} * Group_i + \epsilon_{ijk} \quad (20)$$

where $Time_{ij}$ was the continuous time of subject i's j^{th} observation; $Group_i$ was the treatment group for subject i; Mid_k and $Apex_k$ were indicator variables for whether the ijk^{th} observation was from the mid or apex of the left ventricle, respectively; and the collected error terms for subject i were independent and identically distributed for all subjects with distribution

$$\boldsymbol{\epsilon}_i = [\epsilon_{i,1,1}, \epsilon_{i,1,2}, ..., \epsilon_{i,1,16}, \epsilon_{i,2,1}, ..., \epsilon_{i,5,16}] \sim MVN_{80}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}),$$

where σ^2 was the variance of the outcome (assumed to be equal across all observations) and $\boldsymbol{\Sigma}$ is an 80-by-80 parametric spatiotemporal correlation matrix. In this simulation study, for the sake of simplicity we assumed $\beta_0 = \beta_1 = \beta_3 = \beta_4 = 1$, $\beta_2 = 0$, and $\sigma^2 = 1$. No loss of generality was expected, as σ^2 simply scales the outcome.

Model Parameters Changed Between Conditions

The parts of the linear model itself that were changed between conditions include β_5 and $\boldsymbol{\Sigma}$. Note that β_5 represents the time-by-treatment interaction and if $\beta_5 \neq 0$ then

there exists a difference in the time course between the two groups, which was the original research hypothesis. We generated data from the linear model with five different values for $\beta_5 : \{0, 0.05, 0.10, 0.15, 0.20\}$. Setting $\beta_5 = 0$ allowed us to assess the Type I error rate, while the other values allowed us to construct a power curve to compare Type II error rates for different conditions.

As mentioned earlier, we assumed that the spatiotemporal correlation structure $\mathbf{\Sigma}$ was separable in space and time. Thus, $\mathbf{\Sigma} = \mathbf{\Sigma}_T \otimes \mathbf{\Sigma}_S$ for parametric temporal ($\mathbf{\Sigma}_T$) and spatial ($\mathbf{\Sigma}_S$) correlation functions. We generated data from combinations of compound symmetric and autoregressive-1 temporal structures and exponential, spherical, and Matérn spatial structures, for a total of 6 possible parametric structures of $\mathbf{\Sigma}$. When choosing the values of $\boldsymbol{\theta}$ for $\mathbf{\Sigma}(\boldsymbol{\theta})$, we looked at values that produced high and low correlation in both space and time. The parameters that produced these correlation structures are given in Table A.2 and the correlation functions themselves are plotted in Figure A.2. We considered 4 combinations of parameters (low spatial/low temporal, low spatial/high temporal, high spatial/low temporal, and high spatial/high temporal correlation) to represent the degree of correlation, resulting in 24 (6×4) different covariance structures used to generate the simulated data.

When fitting the models with a parametric correlation structure we fitted only the true model. This choice was motivated by how information criteria were previously shown to be highly accurate at choosing the true structure, and that the effects of misspecification were well explored in our past work[17].

We also varied the total number of subjects (N) so that $N = 50$ or 100. This was done to examine the convergence of the asymptotic Wald and F-tests when the sample size is small, so that we could determine if the sample size of a typical longitudinal imaging study is enough to employ large-sample methods of inference. The treatment groups were considered to be balanced, with the number per group being 25 and 50, respectively. Note

that since there are 80 observations per subject, these numbers of subjects corresponds to 4000 and 8000 total observations, respectively. Due to the computational burden from fitting models with complex covariance structures on 8000 observations, the N=100 sample size was only evaluated for information criterion accuracy and Type I error rate, but not the four values of β_5 to evaluate power; this is because it can be safely assumed that, for a given working covariance structure and the cases considered here, increased sample size will increase power and thus not provide enough novel conclusions to justify the lengthy computation time of additional simulations.

The effect of missingness in the data was also considered as well. The method of missingness was assumed to be completely at random. The extent of missing data was derived from the dataset inspiring this work: data is only missing from absent follow-up visits at a 10% rate per observation time, with the missingness for each temporal observation independent of one another. In this case, there is temporal missingness but not spatial missingness. The baseline visit is always observed, so each subject can have K, $2K$, ..., $(J-1)K$, or JK observed outcomes.

The complete listing of conditions is given in Table 2.

Simulation Details

To achieve the desired precision in the estimates of Type I error rates, we used a simulation size of 5000 for each condition. Specifically, we wished for the 95% confidence interval on the estimate of the Type I error rate to have a width of about 1%. The data was generated using the 'mvrnorm' function in the MASS (v. 7.3-29) package of R (v. i368 3.0.2) where each subject's 80 observations were drawn at once, independently from the other subjects. The random number generator used in the 'mvrnorm' function is the Mersenne-Twister generator. When present in the simulation condition, missing data was assessed independently for each time point for each individual; a Uniform(0,1) variable

was generated and if it was less than 0.1 that follow-up visit was marked as missing and excluded from the analysis.

The linear model fitting, along with the inference on the fixed effects, was done using the ASReml-R package (v. 3.0, VSN International, Hemel Hempstead, UK)[10]. Results for a Wald's test and an F-test with a Kenward-Roger's correction for denominator degrees of freedom, built into the ASReml package, were recorded at each simulation run. The seeds were changed between each simulation iteration so that each run would be generated independently. Care was taken to ensure all sixteen models in each iteration converged, typically by re-running troublesome datasets with improved initial values. Non-convergence was most prominent in the averages of left ventricular levels analyzed together, as the parameters in the unstructured spatial correlation matrix were extremely close to 1; use of proper initial values obtained through pair-wise Pearson correlations between the level averages proved to be the most reliable way of getting convergence.

Simulation Output

For each condition described in Table 2, and each of the 5000 independent simulated datasets, we fit sixteen linear models corresponding to different approaches to correlation in space and time, listed in Table 3. All combinations of four spatial and four temporal methods were considered: modeling the correlation directly with a parametric correlation function or with a summary measure in space (regional averages analyzed jointly with M=3, regional averages analyzed separately, and a global average) or time (endpoint, slope, and t_J-corrected AUC analysis). When a correlation function was employed the true function was used. When a correlation function was used for both space and time, the Kronecker product of the two was used (i.e. correlation was separable in the fitted model). Table 3 also details what predictors were used when fitting each model; note that certain summary measures precluded the use of time- or space-varying covariates. Note that due to their Type I error rates not being conserved when complete data was used, the level averages analyzed

separately (with or without Bonferroni correction) were excluded from the simulations with MCAR data resulting in only twelve models being fit for those conditions.

From each model fit, we looked at the p-value for the hypothesis test of whether the treatment group affected the change over time; when the temporal correlation was used this was a test for treatment-by-time interaction ($H_0 : \beta_{Time_j Group_i} = 0$) and when a temporal summary measure was used it was a test for a group effect on the summary value ($H_0 : \beta_{Group_i} = 0$); these βs refer to the fixed effects associated with the predictors in Table 3. As mentioned above, the p-values for these hypothesis tests were obtained through a Wald's test and a corrected F-test. For the level averages analyzed separately, the effect of correction for multiple testing was considered. The uncorrected approach declared significance when any of the three levels rejected for $p < \alpha$, while the other approach used a Bonferroni correction such that a significant finding was declared when $p < (\alpha/3)$ for any of the three levels.

The Type I error rates were calculated as the proportion of the 5000 datasets where a fitted model rejected the relevant null hypothesis (see Table 3) at an $\alpha = 0.05$ level when the data was generated with β_5 truly equal to zero. Using the convention suggested by Bradley (1978), we considered the Type I error rate to be conserved when it fell between 0.75α and 1.25α, or 0.0375 and 0.625 for our chosen α level. The value of 1.25α was chosen to be moderate, neither highly conservative (1.1α) nor liberal (1.5α) as Bradley described them[8]. Thus, we classified the fit of a certain covariance structure on data generated under a given covariance structure to be overconservative if the empirical Type I error rate $\hat{\alpha}$ was below 0.0375 and inflated if the error rate was above 0.0625. The Type II error rates (the power curves) were found by calculating the proportion of times the hypothesis test for a treatment-over-time effect was rejected on datasets generated with $\beta_5 > 0$ (Equation 20).

SIMULATION RESULTS

Type I Error Rates - No Missing Data

The results of the Type I error rate simulations for complete data are given in Figures 1, 2, and 3. The figures include the Bradley-inspired bounds of (0.0375,0.0625) and the stricter 99% confidence bounds of (0.04206,0.05794); both are included to provide a scale to evaluate what constitutes a mild or severe deviation from a Type I error rate of 0.05. For the twelve models from Figures 1 and 2, there were no clear relationships between the observed Type I error rate and the generating correlation structure and degree of spatial/temporal correlation. Since there was not an apparent pattern, the Type I error rates were aggregated over these 24 conditions to produce the box plots, which thus gives an idea for how a certain approach behaves for data with various properties. For models where level averages were analyzed separately (Figure 3) a strong relationship existed between degree of spatial correlation and whether or not a Bonferroni correction was used.

As was seen in our previous work[17], when the true separable parametric correlation structure was fitted our proposed spatiotemporal model consistently conserved the Type I error rate under all conditions and sample sizes, and both tests demonstrated convergence to their asymptotic α-level. As seen in the first panel of Figure 1, they are even conserved with regards to the stricter 99% confidence interval.

We did observe trends across different spatial and temporal methods, although there did not appear to be an interaction between the effects of different spatial methods and temporal methods. Therefore for the sake of brevity (as 1,920 Type I error rates were collected) the spatial and temporal methods will have their statistical properties discussed separately.

For the spatial methods, several results are apparent. First is that, keeping the temporal method the same, there seemed to be little difference between modeling spatial correlation between 16 segments, analyzing averages of the three levels together via an unstruc-

tured correlation matrix, and a global average as all three approaches did a decent job of conserving the Type I error rate regardless of sample size or test (Wald vs. F-test). This is not surprising as there was no spatial component to the treatment-over-time effect, so with regards to the model used to generate the data the spatial aspect was a nuisance.

However, the approach of analyzing the three level averages separately (Figure 3) seemed to be a poor approach, since a lack of correction caused inflated Type I error rates even though only three tests were done. A Bonferroni correction causes a different problem of an overly conservative Type I error rate. The degree of spatial correlation had a large effect on how inflated/overly conservative this approach was: the Bonferroni correction was less conservative under low spatial correlation while the uncorrected method was less inflated under high spatial correlation. These results are reasonable due to how the three tests are correlated. When spatial correlation is low the correlation between the tests will also be low, so a Bonferroni correction would be reasonable while a lack of correction would result in a greatly inflated Type I error (up to 14.3% for three uncorrelated 0.05-level tests). Conversely, if the tests are highly correlated (as is the case when spatial correlation is high) then the Type I error rate would not be very inflated for multiple tests, meaning that a Bonferroni correction is "overkill" and would result in a very conservative overall test.

Among the temporal methods, it seems that the use of the true temporal correlation structure reliably conserved the Type I error regardless of sample size or type of test. The temporal summary measures had the potential for the Type I error rate to be inflated when the sample size was small (N=50) and the Wald's test was used, especially for AUC and endpoint analysis. Slope analysis was slightly better, but still had some potential inflation for the Wald's test with a small sample size. A sample size of N=100 generally did a better job of conserving the Type I error rate for the Wald's test than for N=50 but still had some error rates outside the 99% confidence interval for the temporal summary measures. However the F-test with corrected denominator degrees of freedom had a conserved Type

I error rate for all temporal methods regardless of sample size. The F-test also seemed to result in a median Type I error rate from different conditions being closer to 0.05 than the Wald test. Interestingly, the F-test performed better for temporal summary measures even when the global average was used and all observations were independent. These results suggest that a corrected F-test should be used when temporal summary measures are employed on longitudinal data, especially if the sample size is small.

Power Curves - No Missing Data

In addition to the Type I error rate, the correlation models and summary methods were compared by their Type II error rate when making inference about a treatment-by-time interaction. Although there were some differences in the overall power between the different generating correlation structures, the relative power of the spatial and temporal methods was largely the same between them. Therefore, representative power curves associated with the simulation results for a MAT$\times$CS structure are given in Figures 4 (Wald) and 5 (F-test) while the curves associated with the other five correlation structures are given in Figures A.3 to A.12 in the Supplementary Materials. In these figures, each of the sixteen models is identified by a combination of line color (spatial method) and style (temporal method). The multiple figures report the observed power when a Wald's or corrected F-test were used, and for what spatiotemporal correlation structure was used to generate the data. The power curves under different generating correlation structures are given to evaluate whether the relative power of the methods vary with the underlying correlation structure; note that in practice one does not know the data's true correlation structure and thus we want our conclusions to be generalizable to whatever correlation structure is encountered in practice. In addition, the Wald and F-test did not seem to differ in power which suggests that there would be no reason not to use the more reliable corrected F-test in practice.

In all of the cases, we observed that for a given temporal method the use of a spatial correlation model had the highest power among the four spatial methods. The global

average and the level averages analyzed jointly with an unstructured correlation matrix (LVL-UN) proved to be roughly equivalent to one another; the pair had less power than the spatial correlation model, although the difference was slightly less pronounced in cases of high spatial correlation. The level averages analyzed separately with a Bonferroni correction (LVL-I) had the smallest power of the four, but was still competitive in cases of low spatial correlation. When spatial correlation was high, which was where the Bonferroni correction made the Type I error rate overly conservative, the independent level averages had greatly reduced power compared to the other spatial methods. Recall that the uncorrected independent level averages had an inflated Type I error rate, and as such did not have their power examined. It should be noted that regardless of spatial method, the overall statistical power was reduced in the presence of high spatial correlation; this result is not surprising, as higher correlation implies that there is less unique information per subject.

As for the temporal methods for a given spatial method, the two most powerful methods seemed to be the temporal correlation model and slope analysis. The two had roughly equivalent power in most cases, which is not surprising given the model generating the data could be considered a 'best-case scenario' for slope analysis. Endpoint analysis was less powerful than the temporal correlation model or slope analysis, but could still be considered competitive. Given the scenarios had a constant rate of change over time and a set length of follow-up for each subject, one could consider these to be acceptable conditions to use endpoint analysis. AUC analysis, however, proved to be substantially less powerful than the other methods. The degree of temporal correlation had a profound impact on the power of these methods; higher correlation resulted in higher power for the temporal correlation model, slope analysis, and endpoint analysis, while higher correlation led to lower power in the AUC analysis. Considering the relationships seen in Equations 9, 12, and 16 this is not surprising, as we have seen that the variance of endpoint and slope summary measures decrease with increasing correlation while the variance of the AUC increases with it. These results seem to be evidence that AUC analysis should not be used

when a change over time is of interest and temporal correlation is not negligible.

Type I Error Rates - With Missing Data

The observed Type I error rates for the twelve models (recall, the level averages analyzed separately were excluded due to poor performance on complete data) fitted to data with missing observations are given in Figures 6 and 7. When aggregated over the 2,160,000 iterations across all conditions, the percent of missingness closely followed the expected amount and binomial distribution. The temporal correlation model had an average of 8.00% missingness, endpoint analysis had an average of 10.00% missingness, and slope and AUC analysis both had a mean missingness of 0.01%. The histograms of the percent missingness for each iteration are given in Figure A.13. In general, the results were similar to the Type I error rates of the complete datasets. There did not seem to be a difference in conservation between spatial summary methods, which is not surprising since the missingness was purely temporal. The use of a temporal correlation model seemed to conserve the Type I error rate under most conditions.

For temporal summary methods, the Wald's test with N=50 had a slightly inflated Type I error rate, although the degree of inflation did not seem to be much different from the non-missing conditions. However, in the presence of missing data the Type I error rate of the Wald's test was not conserved at N=100; this is in contrast to the complete data conditions where the increased sample size resulted in a conserved error rate. This suggests that, where the Wald's test is concerned, a sample size of N=100 is not sufficient to achieve convergence of asymptotic properties when there is missing data in longitudinal imaging studies. The F-test, conversely, conserved the Type I error rate for the temporal summary methods for N=50 and N=100; this reinforces the conclusion that the corrected F-test is preferred when temporal summary methods are used.

Power Curves - With Missing Data

The simulated power curves for the twelve methods under a MATxCS generating correlation structure are shown in Figures 8 (Wald) and 9 (F-test), with the other structures in Figures A.14 to A.23 in the Supplemental Materials. We found that the relative power of the twelve methods did not differ between the six generating correlation structures, much like the complete data case.

The power curves on complete data and those with missing data were similar, but some differences arose. The most powerful method was still our spatiotemporal model, and the Wald and corrected F-test performed approximately the same for all models. A spatial correlation structure making use of all sixteen segments was still more powerful than one taking a global average or analyzing level averages together with an unstructured working correlation matrix. On the temporal side, AUC analysis was still lacking in power and endpoint analysis was still fairly competitive.

The biggest difference is that all models lost power when missingness was introduced to the data generation step. This is expected, as less data naturally means less power. However, the gap between the temporal correlation model and the temporal summary methods seems to have widened due to the introduction of missing data. In particular, slope analysis was marginally less powerful than a temporal correlation model in the presence of missing data. Since the true time course was linear and the missingness was completely at random, the individual slope estimates are unbiased; the observed loss of power must therefore be due to changes with the variance, either a direct increase from fewer observation going into the slope estimation or the heteroscedasticity between subjects with different numbers of lost follow-ups.

CONCLUSIONS

Longitudinal imaging studies are characterized by a large amount of correlated data on a small number of independent subjects. It was therefore necessary to consider the Type I error rate for asymptotic tests such as Wald's and the F-test, as it was not known whether the repeated measures offset the concerns regarding convergence with a small number of subjects. We found that our previously proposed model with a separable parametric correlation structure consistently conserved the Type I error rate for both tests regardless of sample size, generating correlation structure, degree of correlation, and missing data. Conversely, summary measures occasionally produced inflated Type I error rates in the Wald test when the sample size was small (N=50) or even when it was larger (N=100) when some data was missing. Although the Type I error rate was better conserved when the corrected F-test was used, this suggests that summary methods may be unreliable to use on longitudinal imaging data. It should be noted that it is unclear from our results whether this is due to the underlying mechanics of summary measures or to how they reduce the dimension of the dataset; one could expect a model using 4000 observations to have better asymptotic properties than another using just 50, even if those 4000 observations were highly correlated within clusters of size 80.

Of particular note is the approach where regional averages were analyzed separately, creating a multiple testing problem. This analysis strategy was of particular interest because of how it is commonly used in practice when analyzing longitudinal imaging data. We found that the degree of spatial correlation had a large impact on how inflated or overly conservative the Type I error rate was. Without a correction for multiple testing, the Type I error rate was typically inflated with a higher error rate under lower degrees of correlation. With a Bonferroni correction, the Type I error rates were usually too conservative which led to a loss of power; the power loss was particularly substantial when the spatial correlation was high. Since a lack of correction inflates the Type I error and a Bonferroni approach

overcorrects and inflates the Type II error, one could suggest an intermediate correction. The problem with this reasoning is that the amount of correction needed is a direct function of the correlation between the regions, and if the inter-region correlation was known or estimated there would be little reason to not analyze the regions together with at least an unstructured correlation model. The rudimentary approach of analyzing regions separately appears to be misleading at worst and inefficient at best, and thus should not be used in practice.

When comparing the models based off of power, our spatiotemporal model had the highest of the approaches considered across all conditions. Despite the target of interest (the treatment-by-time interaction) having no spatial component in the generating model, the use of spatial correlation structures with data from all 16 segments routinely beat out the global average and the jointly-modeled level averages. The temporal summary measures also resulted in a loss of power, particularly for the area-under-the-curve analysis. Although endpoint analysis was competitive and slope analysis closely rivaled the temporal correlation model, the finding that the gap in power was as big or even larger when some data is missing was disconcerting as it implies that the use of summary measures to 'smooth over' missing data is likely to be counterproductive.

Our simulation study had several limitations that would be useful areas to expand upon in future research. The generating time course was linear which meant that our model's fixed effects were correctly specified and the summary measures (especially endpoint and slope analysis) were operating in a best-case scenario. Additional simulations under a non-linear time course or heterogeneous time courses between subjects would provide results that may be more applicable to practical applications. The simulation design also set the fitted correlation structures to match the true structure, which failed to capture the uncertainty of choosing the true correlation structure and presumed there even was a true separable parametric function underlying the data. Such scenarios where the difficul-

ties of modeling correlation in space and time are present could lend credibility to summary methods where correlation is simplified or removed.

There are other interesting directions for this line of research to take. It would be of interest to compare the models in the presence of a three-way interaction between treatment group, time, and region to quantify the power lost when smoothing out a target of inference with spatial and temporal components. Lastly, it would be a potential credit to our spatiotemporal model to quantify the power gained from controlling for time- and space-varying covariates compared to summary methods that cannot use them.

ACKNOWLEDGEMENTS

We wish to thank Drs. Louis Dell'Italia, Tom Denney, Jr., and Himanshu Gupta of the UAB SCCOR study for their support and for the cardiac MRI data they provided. Predoctoral funding was provided by NHLBI T32HL079888. The UAB SCCOR study was supported by National Institutes of Health Specialized Center of Clinically Oriented Research in Cardiac Dysfunction P50-HL077100.

REFERENCES

[1] Ahmed, M.I., Gladden, J.D., Litovsky, S.H., Lloyd, S.G., Gupta, H., Inusah, S., Denny Jr., T., Powell, P., McGiffin, D.C., Dell'Italia, L.J. (2010). Increased oxidative stress and cardiomyocyte myofibrillar degeneration in patients with chronic isolated mitral regurgitation and ejection fraction $> 60\%$. *Journal of the American College of Cardiology*, 55(7): 671-679.

[2] Ahmed, M.I., Aban, I., Lloyd, S.G., Gupta, H., Howard, G., Inusah, S., Peri, K., Robinson, J., Smith, P., McGiffin, D.C., Schiros, C.G., Denney, T., Dell'Italia, L.J. (2012). A Randomized Controlled Phase IIb Trial of Beta-1-Receptor Blockade for Chronic

Degenerative Mitral Regurgitation. *Journal of the American College of Cardiology* 60 (9): 833-838.

[3] Albert, P.S. (1999). Longitudinal Data Analysis (Repeated Measures) in Clinical Trials. *Statistics in Medicine* 18: 1707-1732.

[4] Beyar, R., Weiss, J.L., Shapiro, E.P., Graves, W.L., Rogers, W.J., Weisfeldt, M.L. (1993). Small apex-to-base heterogeneity in radius-to-thickness ratio by three-dimensional magnetic resonance imaging. *American Journal of Physiology*, 264, H133-H140.

[5] Bowman, F.D., Waller, L.A. (2004). Modeling of cardiac imaging data with spatial correlation. *Statistics in Medicine* 23 (6): 965-985.

[6] Box, G.E.P. (1950). Problems in the analysis of growth and wear curves. *Biometrics* 6(4): 362-189.

[7] Bozdogan, H. (1987). Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika* 52 (3): 345-370.

[8] Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology* 31: 144-152.

[9] Burton, A., Altman, D.G., Royston, P., Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine* 25 (24): 4279-4292.

[10] Butler, D., Cullis, B.R., Gilmour, A.R., Gogel, B.J. (2007). ASReml-R reference manual (Release 2.00) [Software]. Available from http://www.vsni.co.uk/software/asreml

[11] Cerqueria, M.D., Weissman, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., Pennell, D.J., Rumberger, J.A., Ryan, T., Verani, M.S. (2002). Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart: A State-

ment for Healthcare Professionals From the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. *Circulation* 105: 539-542.

[12] Denney, T.S. Jr., Nagaraj, H.M., Lloyd, S.G., Aban, I., Corros, C., Seghatol-Eslami, F., McGiffin, D.C., Dell'Italia, L.J., Gupta, H. (2007). Effect of Primary Mitral Regurgitation of Left Ventricular Synchrony. *American Journal of Cardiology* 100(4): 707-711.

[13] Everitt, B.S. (1995) The Analysis of Repeated Measures: A Practical Review with Examples. *Statistician* 44(1): 113-135.

[14] Fitzmaurice, G., Laird, N., Ware, J. (2004). *Applied Longitudinal Analysis*. Hoboken: Wiley.

[15] Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. (2008). *Longitudinal Data Analysis*. Boca Raton: CRC Press.

[16] Frison, L., Pocock, S.J. (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 11(13): 1685-1704.

[17] George, B., Aban, I. (2014). Selecting a Separable Parametric Spatiotemporal Covariance Structure for Longitudinal Imaging Data. Submitted to *Statistics in Medicine*.

[18] Gibbons, R., Hedeker, D., Waternaux, C., Davis, J.M. (1988). Random Regression Models: A Comprehensive Approach to the Analysis of Longitudinal Psychiatric Data. *Psychopharmacology Bulletin* 24 (3): 438-443.

[19] Hannan, E.J., Quinn, B.G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, Series B* 41 (2): 190-195.

[20] Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* 72 (358): 320-338.

[21] Jennrich, R.I., Schluchter, M.D. (1986). Models with Unbalanced Structured Covariance Matrices. *Biometrics* 42 (4): 805-520.

[22] Lindstrom, M.J., Bates, D.M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association* 83 (404): 1014-1022.

[23] Matthews, J.N.S., Altman, D.G., Campbell, M.J., Royston, P. (1990). Analysis of Serial Measurements in Medical Research. *BMJ* 300(6719):230-235.

[24] Matthews, J.N.S. (1993). A refinement to the analysis of serial data using summary measures. *Statistics in Medicine* 12(1): 27-37.

[25] Rao, C.R. (1958). Some statistical methods for comparison of growth curves. *Biometrics* 14: 1-17.

[26] Rowell, J.G., Walters, D.R. (1976). Analysing data with repeated observations on each experimental unit. *The Journal of Agricultural Science* 87(2): 423-432.

[27] Schiros, C.G., Dell'Italia, L.J., Gladden, J.D., Clark, D., Aban, I., Gupta, H., Lloyd, S.G., McGiffin, D.C., Perry, G., Denney, T.S., Ahmed, M.I. (2012). Magnetic resonance imaging with 3-dimensional analysis of left ventricular remodeling in isolated mitral regurgitation: implications beyond dimensions. *Circulation* 125 (19): 2334-2342.

[28] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* 6 (2): 461-464.

[29] Seals, S. (2013) Spatial analysis of cardiovascular MRI data (dissertation). Birmingham, AL: University of Alabama at Birmingham.

[30] Simpson, S.L., Edwards, L.J., Muller, K.E., Sen, P.K., Styner, M.A. (2010). A linear exponent AR(1) family of correlation structures. *Statistics in Medicine* 29: 1825-1838.

[31] Simpson, S.L., Edwards, L.J., Muller, K.E., Styner, M.A. (2014). Kronecker Product Exponent AR(1) Correlation Structures for Multivariate Repeated Measures. *PLoS ONE* 9: e88864.

[32] Waller, L.A., Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley and Sons.

[33] Wishart, J. (1938). Growth-Rate Determinations in Nutrition Studies with the Bacon Pig and Their Analysis. *Biometrika* 30 (1-2): 16-28.

[34] Zucker, D.M., Manor, O., Gubman, Y. (2012). Power comparison of summary measure, mixed model, and survival analysis methods for analysis of repeated-measures trials. *Journal of Biopharmaceutical Statistics* 22 (3): 519-534.

Table 1: Dimensions and correlation matrices for several approaches to spatial and longitudinal data. The value M refers to a number of distinct regions within the observed area such that $M \ll K$. Note that d_{uv} is the distance between locations u and v with $f(d_{1K}|\boldsymbol{\theta})$ being a parametric correlation function. The terms ρ_{uv} are parameters for an unstructured correlation matrix.

Type of Data	Method	Dim($\boldsymbol{\Sigma}_{S/T}$)	$\boldsymbol{\Sigma}_{S/T}$
Spatial	Correlation	$K \times K$	$\boldsymbol{\Sigma}_S = \begin{bmatrix} 1 & f(d_{12}\|\boldsymbol{\theta}) & \cdots & f(d_{1K}\|\boldsymbol{\theta}) \\ f(d_{12}\|\boldsymbol{\theta}) & 1 & \cdots & f(d_{2K}\|\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ f(d_{1K}\|\boldsymbol{\theta}) & f(d_{2K}\|\boldsymbol{\theta}) & \cdots & 1 \end{bmatrix}$
	Regional Average, Unstructured	$M \times M$	$\boldsymbol{\Sigma}_S = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1M} \\ \rho_{12} & 1 & \cdots & \rho_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1M} & \rho_{2M} & \cdots & 1 \end{bmatrix}$
	Regional Average, Separate	1×1	$\boldsymbol{\Sigma}_S = [1]$
	Global Average	1×1	$\boldsymbol{\Sigma}_S = [1]$
Temporal	Correlation	$J \times J$	$\boldsymbol{\Sigma}_T = \begin{bmatrix} 1 & f(t_1,t_2\|\boldsymbol{\theta}) & \cdots & f(t_1,t_J\|\boldsymbol{\theta}) \\ f(t_1,t_2\|\boldsymbol{\theta}) & 1 & \cdots & f(t_2,t_J\|\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ f(t_1,t_J\|\boldsymbol{\theta}) & f(t_2,t_J\|\boldsymbol{\theta}) & \cdots & 1 \end{bmatrix}$
	Endpoint Analysis	1×1	$\boldsymbol{\Sigma}_T = [1]$
	Slope Analysis	1×1	$\boldsymbol{\Sigma}_T = [1]$
	AUC Analysis	1×1	$\boldsymbol{\Sigma}_T = [1]$

Table 2: Total number of conditions for the simulation study which included varying the sample size, temporal and spatial structures used to generate the data, the degree of spatial and temporal correlation used in data generation, the values of β_5 used, and whether or not there was missing data. β_5 refers to the size of the treatment-by-time effect in Equation 20.

Sample Size	N=50	N=100
Σ_T Structures	2 (CS,AR-1)	2 (CS,AR-1)
Degree of Σ_T	2 (High, Low)	2 (High, Low)
Σ_S Structures	3 (Exp,Sph,Mat)	3 (Exp,Sph,Mat)
Degree of Σ_S	2 (High, Low)	2 (High, Low)
Values of β_5	$\beta_5 = \{0, 0.05, 0.10, 0.15, 0.20\}$	$\beta_5 = \{0\}$
Missing Data	2 (Yes/No)	2 (Yes/No)
Total Number of Conditions	240	48
	288	

Table 3: Sixteen linear models made from combination of spatial and temporal methods, the form of their fitted predictors, and what predictor is the focus of hypothesis testing for a treatment-by-time. These are also the sixteen models considered in our simulation study.

Spatial Method	Temporal Method	Fitted Model	Tested Predictor
Correlation	Correlation	$E[Y_{ijk}] \sim 1 + Time_j + Group_i + Mid_k + Apex_k + Time_j Group_i$	$Time_j Group_i$
	Endoint	$E[Y_{ik}] \sim 1 + Group_i + Mid_k + Apex_k$	$Group_i$
	Slope	$E[Y_{ik}] \sim 1 + Group_i + Mid_k + Apex_k$	$Group_i$
	AUC	$E[Y_{ik}] \sim 1 + Group_i + Mid_k + Apex_k$	$Group_i$
Level Average, Together	Correlation	$E[Y_{ijm}] \sim 1 + Time_j + Group_i + Mid_m + Apex_m + Time_j Group_i$	$Time_j Group_i$
	Endoint	$E[Y_{im}] \sim 1 + Group_i + Mid_m + Apex_m$	$Group_i$
	Slope	$E[Y_{im}] \sim 1 + Group_i + Mid_m + Apex_m$	$Group_i$
	AUC	$E[Y_{im}] \sim 1 + Group_i + Mid_m + Apex_m$	$Group_i$
Level Average, Separately	Correlation	$E[Y_{ijm}] \sim 1 + Time_j + Group_i + Time_j Group_i$	$Time_j Group_i$
	Endoint	$E[Y_{im}] \sim 1 + Group_i$	$Group_i$
	Slope	$E[Y_{im}] \sim 1 + Group_i$	$Group_i$
	AUC	$E[Y_{im}] \sim 1 + Group_i$	$Group_i$
Global Average	Correlation	$E[Y_{ij}] \sim 1 + Time_j + Group_i + Time_j Group_i$	$Time_j Group_i$
	Endoint	$E[Y_i] \sim 1 + Group_i$	$Group_i$
	Slope	$E[Y_i] \sim 1 + Group_i$	$Group_i$
	AUC	$E[Y_i] \sim 1 + Group_i$	$Group_i$

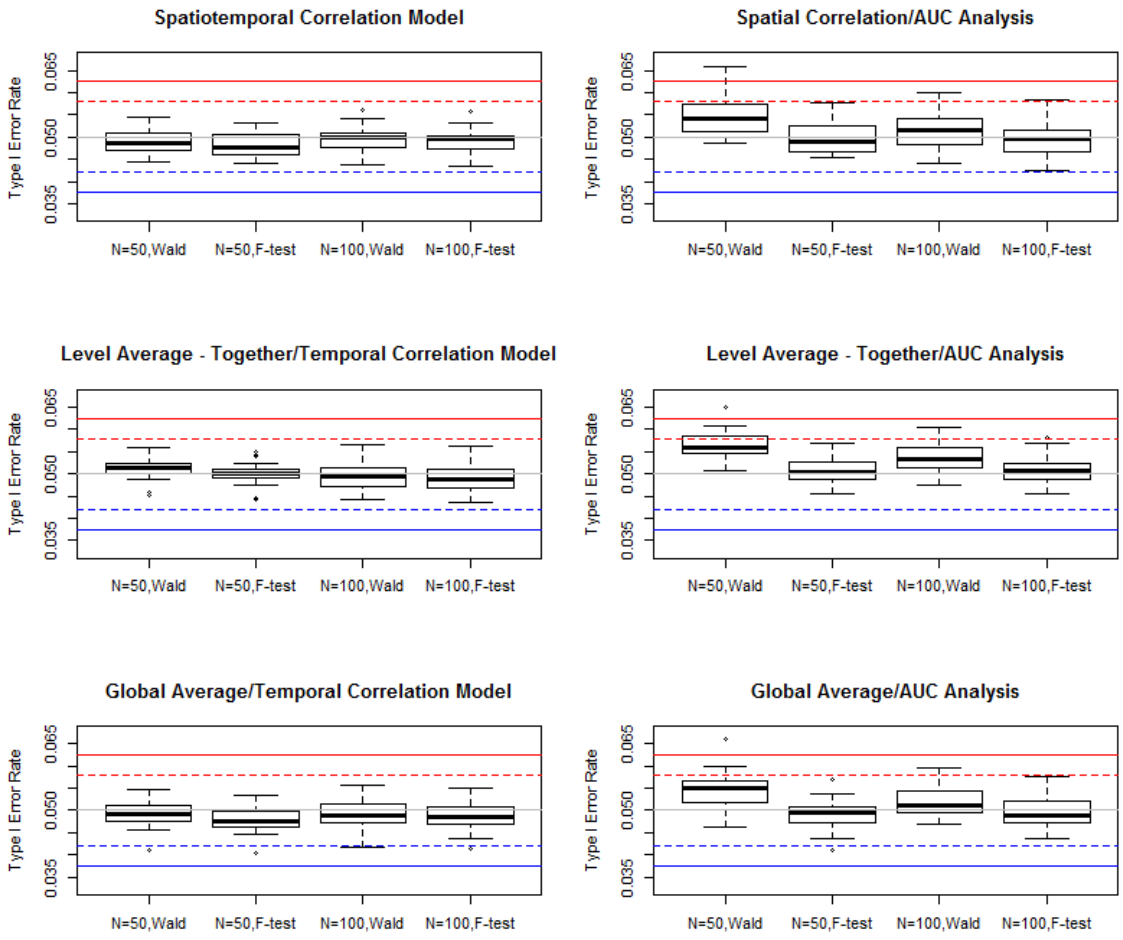

Figure 1: Plot of the empirical Type I error rates for the methods using temporal correlation or AUC analysis, for different sample sizes and tests. The dashed lines correspond to the theoretical 99% confidence interval for $\alpha = 0.05$, and the solid blue and red lines refer to the (0.0375,0.0625) bounds, respectively. The solid grey line denotes $\alpha = 0.05$. Each box plot is the aggregate of the four degrees of correlation and six generating correlation structures.

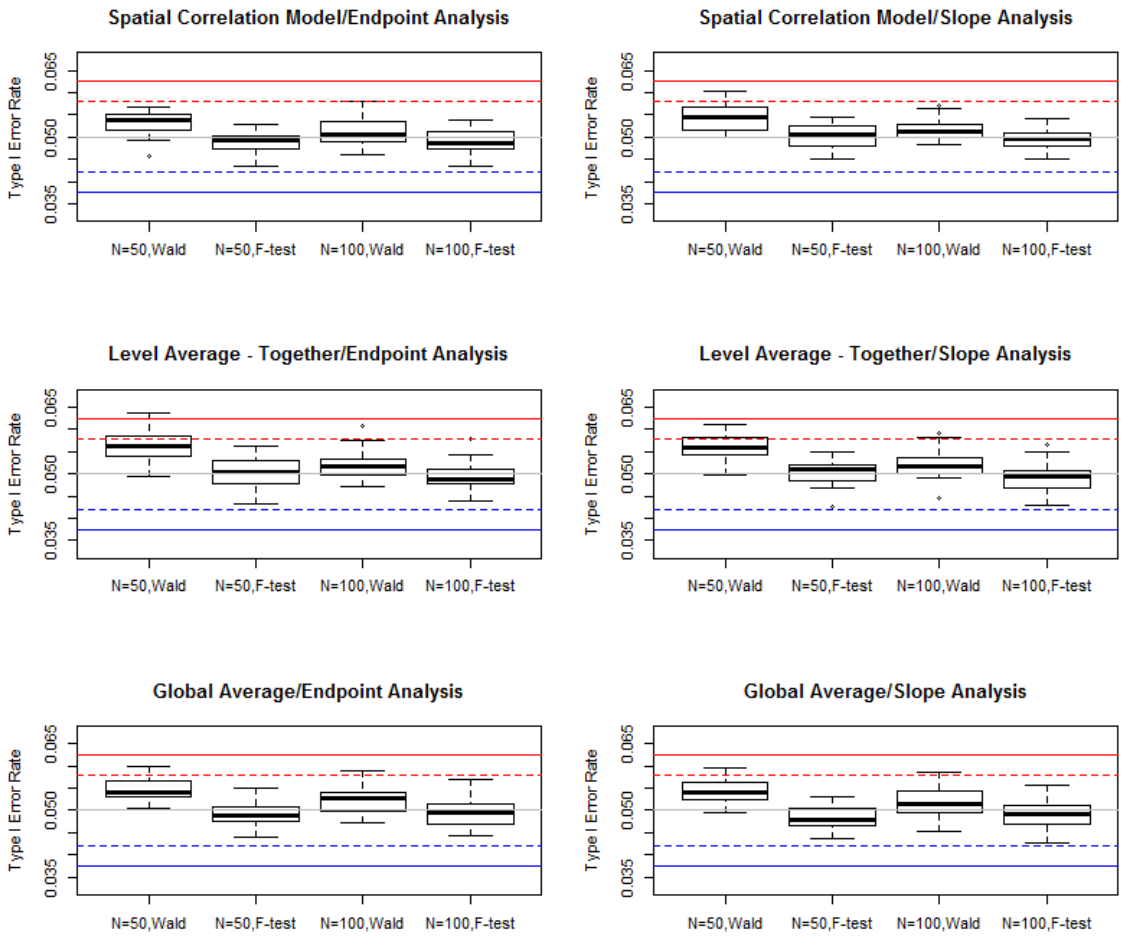

Figure 2: Plot of the empirical Type I error rates for the methods using endpoint analysis or slope analysis, for different sample sizes and tests. The dashed lines correspond to the theoretical 99% confidence interval for $\alpha = 0.05$, and the solid blue and red lines refer to the (0.0375,0.0625) bounds, respectively. The solid grey line denotes $\alpha = 0.05$. Each box plot is the aggregate of the four degrees of correlation and six generating correlation structures.

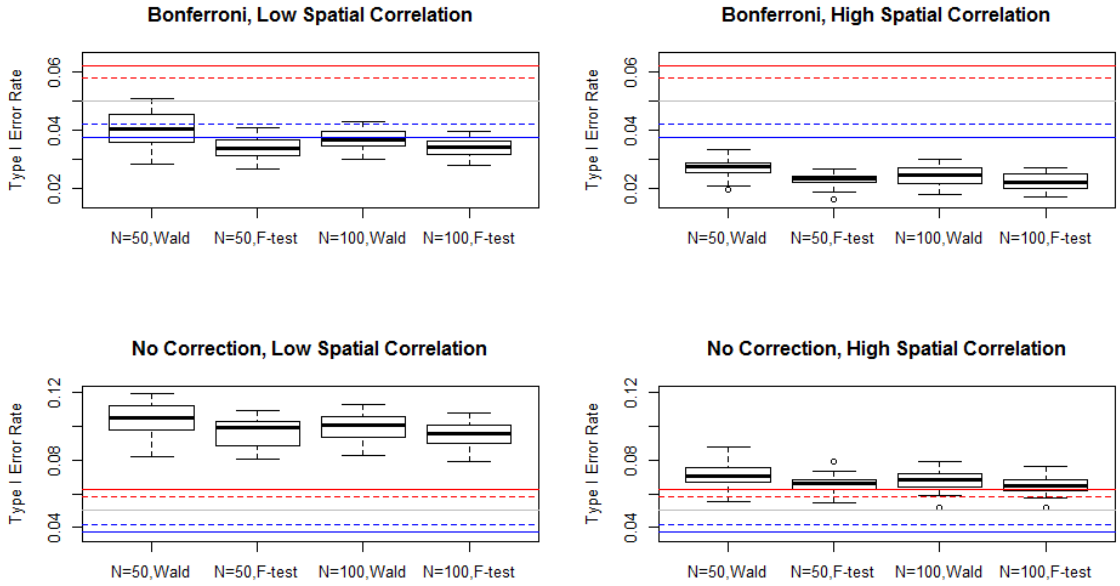

Figure 3: Plot of the empirical Type I error rates for the methods using regional averages over the three levels analyzed separately, for different sample sizes and tests. The dashed lines correspond to the theoretical 99% confidence interval for $\alpha = 0.05$, and the solid blue and red lines refer to the (0.0375,0.0625) bounds, respectively. The solid grey line denotes $\alpha = 0.05$. Each box plot is the aggregate of the four degrees of correlation and six generating correlation structures.

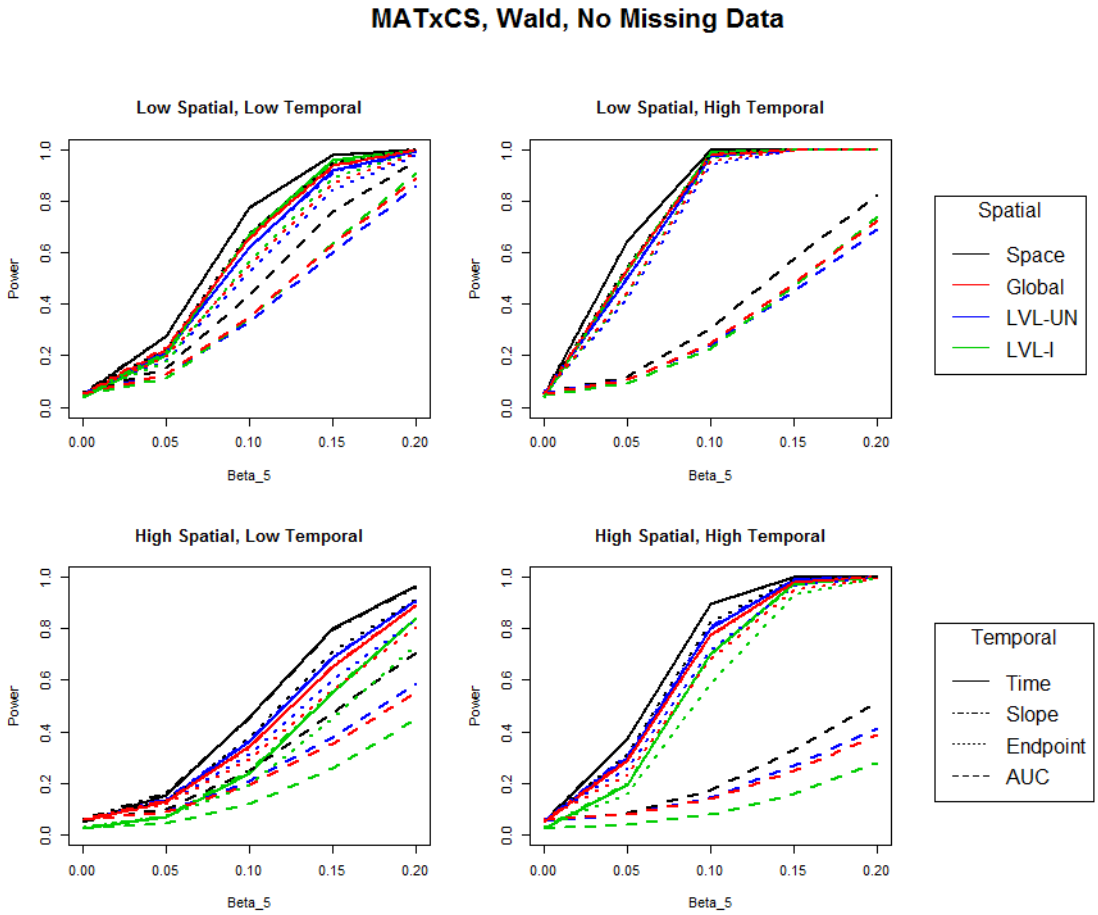

Figure 4: Plot of the empirical power curves for the sixteen models under a Matérn-by-compound symmetric generating correlation structure where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

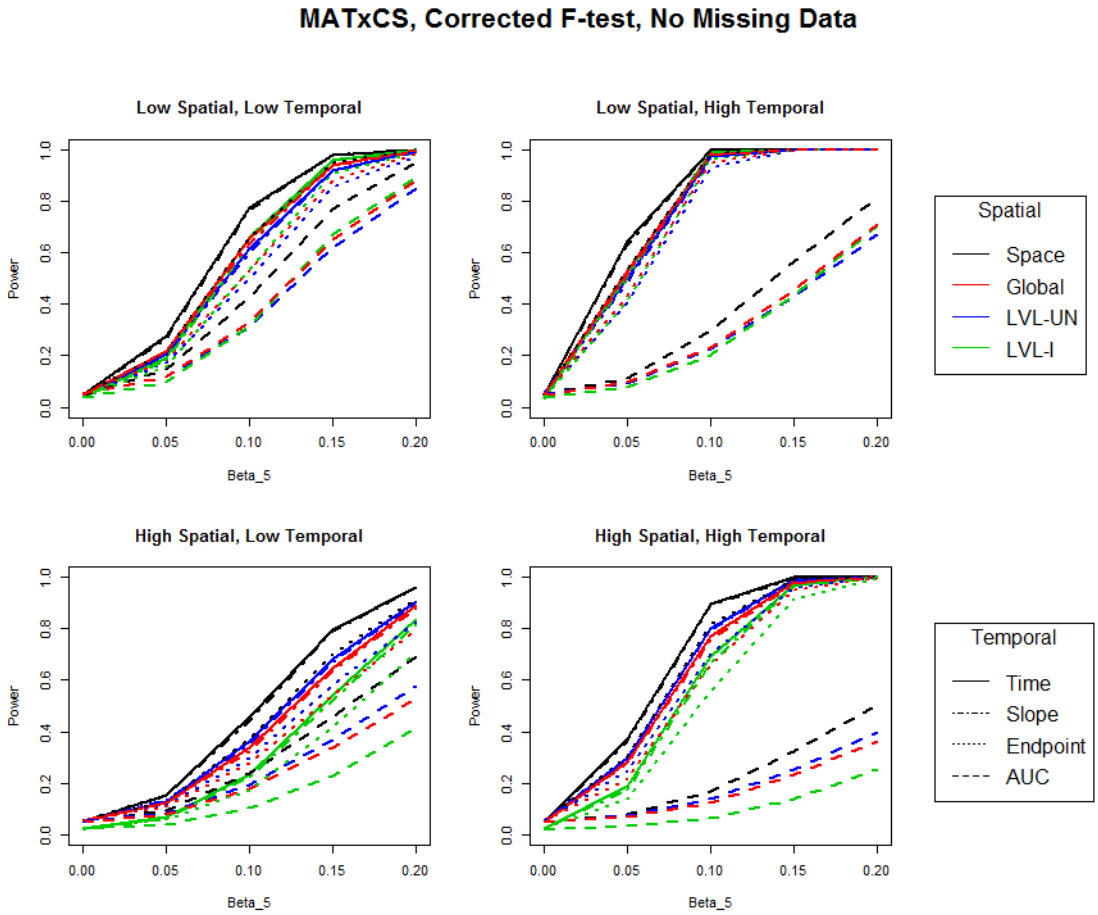

Figure 5: Plot of the empirical power curves for the sixteen models under a Matérn-by-compound symmetric generating correlation structure where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

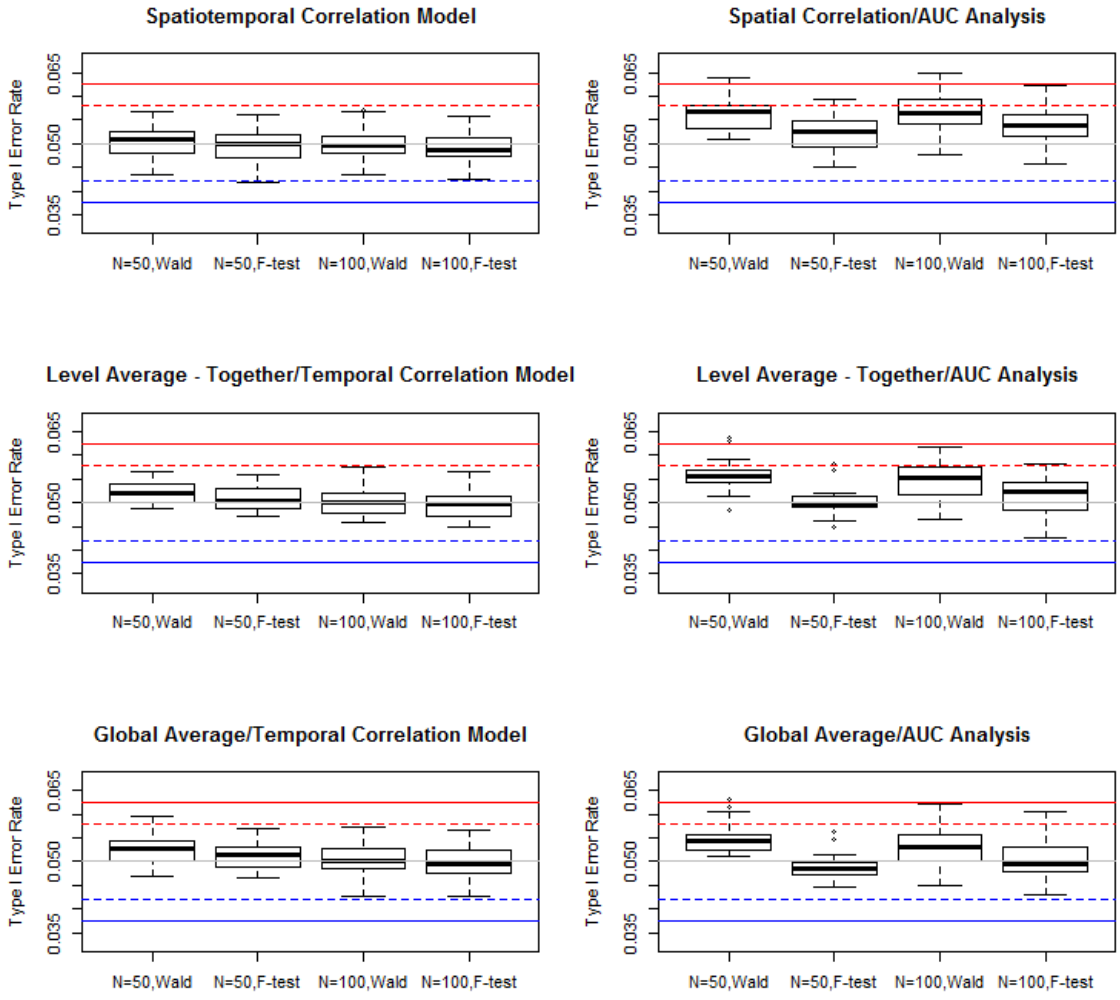

Figure 6: Plot of the empirical Type I error rates for the methods using temporal correlation or AUC analysis, for different sample sizes and tests on data with missing observations. The dashed lines correspond to the theoretical 99% confidence interval for $\alpha = 0.05$, and the solid blue and red lines refer to the (0.0375,0.0625) bounds, respectively. The solid grey line denotes $\alpha = 0.05$. Each box plot is the aggregate of the four degrees of correlation and six generating correlation structures.

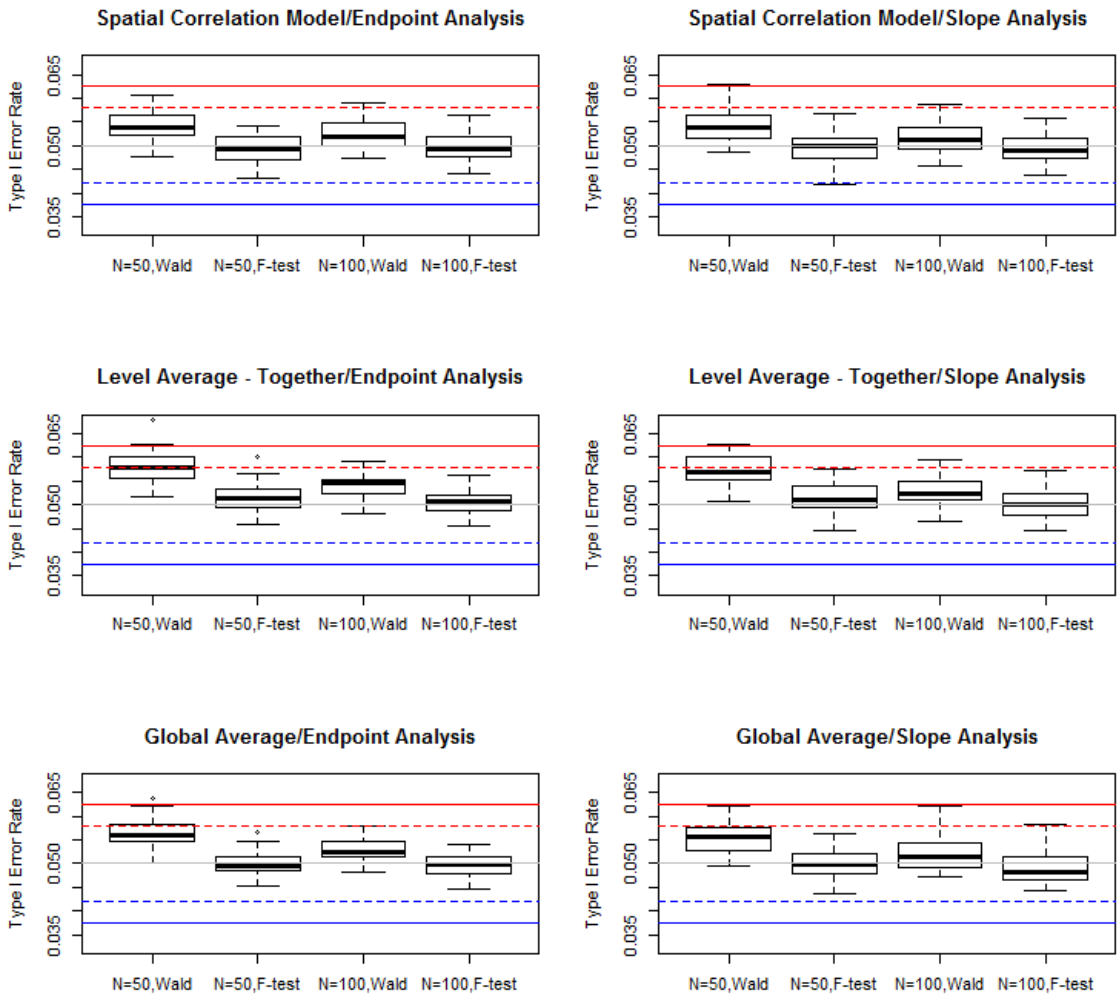

Figure 7: Plot of the empirical Type I error rates for the methods using endpoint analysis or slope analysis, for different sample sizes and tests on data with missing observations. The dashed lines correspond to the theoretical 99% confidence interval for $\alpha = 0.05$, and the solid blue and red lines refer to the (0.0375,0.0625) bounds, respectively. The solid grey line denotes $\alpha = 0.05$. Each box plot is the aggregate of the four degrees of correlation and six generating correlation structures.

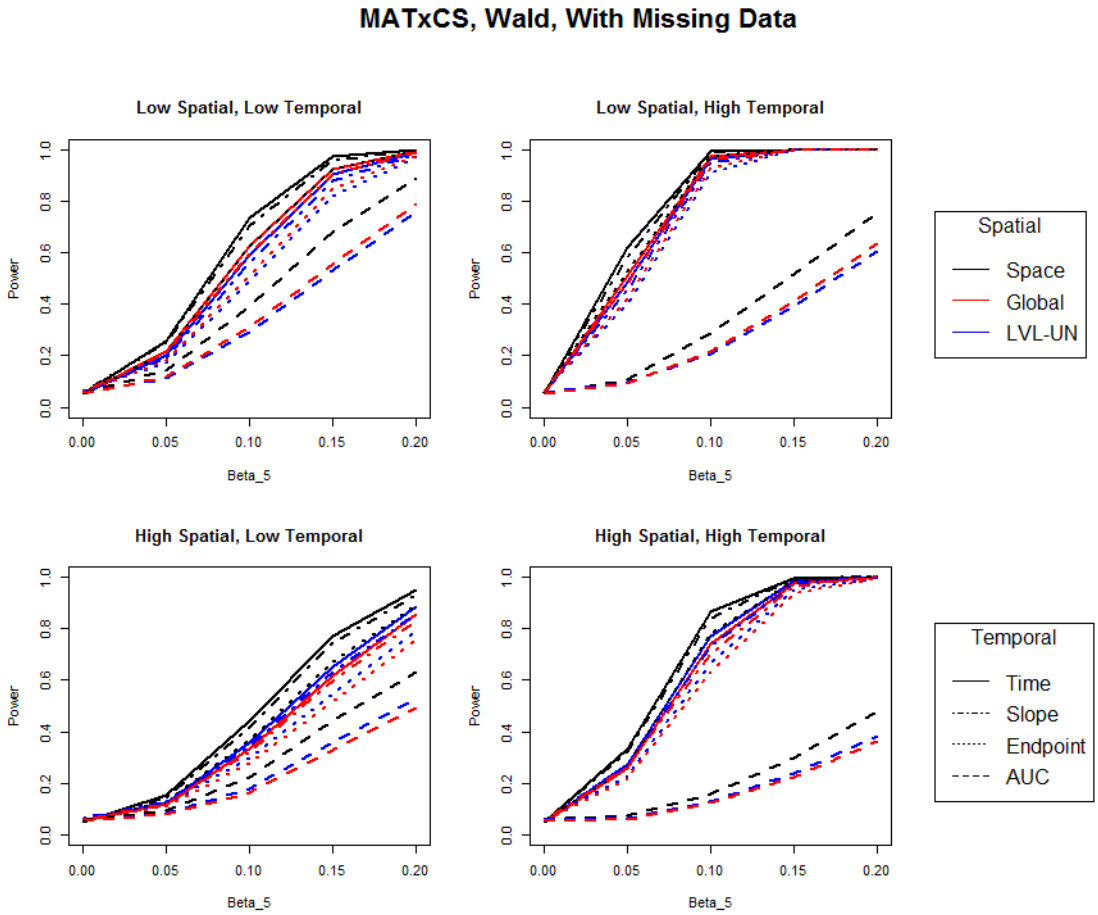

Figure 8: Plot of the empirical power curves for the twelve models under a Matérn-by-compound symmetric generating correlation structure with longitudinally missing data where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

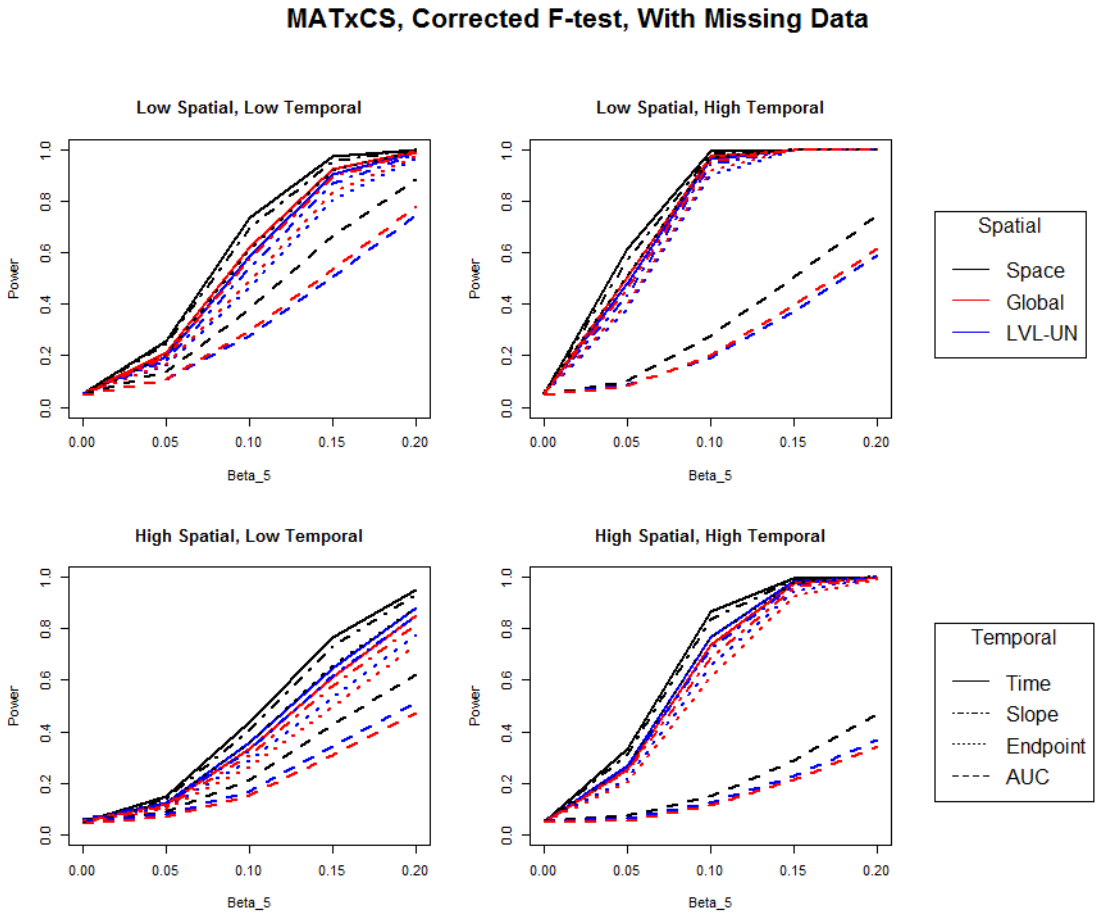

Figure 9: Plot of the empirical power curves for the twelve models under a Matérn-by-compound symmetric generating correlation structure with longitudinally missing data where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

SUPPLEMENTARY MATERIAL

Derivation of the Slope Variance

Consider $Y_{i1k}, Y_{i2k}, ..., Y_{iJk}$ to be the J observations from subject i at location k. One can calculate the slope through those J observations (assuming no missing data) using ordinary least squares estimation, which is $\hat{b}_1$ in the estimator

$$\hat{b} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix} = (X'X)^{-1}X'y \tag{A.1}$$

where

$$X = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_J \end{bmatrix} \tag{A.2}$$

is a $J \times 2$ matrix with t_j being the observed time of observation j with $j = 1, ..., J$. It is of interest to determine how the variance of $\hat{b}_1$ is influenced by the follow-up design and the temporal correlation between points. By defining the covariance of the J observations as the $J \times J$ matrix V,

$$Var(\hat{b}) = (X'X)^{-1}X'VX(X'X)^{-1}. \tag{A.3}$$

By assuming homogeneity of variance we can consider V in terms of a temporal correlation matrix Σ_T with $\rho_{jl} = Corr(Y_{ijk}, Y_{ilk})$.

$$Var(\{Y_{i1k}, ..., Y_{iJk}\}) = V = \sigma^2\Sigma_T = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1J} \\ \rho_{12} & 1 & \cdots & \rho_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1J} & \rho_{2J} & \cdots & 1 \end{bmatrix} \tag{A.4}$$

The calculations for the terms in Equation A.3 are as follows.

$$(X'X)_{2x2} = \begin{bmatrix} J & \sum_{j=1}^{J} t_j \\ \sum_{j=1}^{J} t_j & \sum_{j=1}^{J} t_j^2 \end{bmatrix} \tag{A.5}$$

$$|X'X| = J\sum_{j=1}^{J} t_j^2 - \left(\sum_{j=1}^{J} t_j\right)^2 \tag{A.6}$$

$$(X'X)^{-1} = \frac{1}{J\sum_{j=1}^{J} t_j^2 - \left(\sum_{j=1}^{J} t_j\right)^2} \begin{bmatrix} \sum_{j=1}^{J} t_j^2 & -\sum_{j=1}^{J} t_j \\ -\sum_{j=1}^{J} t_j & J \end{bmatrix} \tag{A.7}$$

$$(X'VX)_{2x2} = \sigma^2 \begin{bmatrix} J + 2\sum\sum_{1\le j<l\le J} \rho_{jl} & \sum_{j=1}^{J} t_j + \sum_{j=1}^{J} t_j \left(\sum_{l\in[1,J],l\neq j} \rho_{jl}\right) \\ \sum_{j=1}^{J} t_j + \sum_{j=1}^{J}\sum_{l\in[1,J],l\neq j} t_l\rho_{jl} & \sum_{j=1}^{J} t_j^2 + \sum_{j=1}^{J} t_j \left(\sum_{l\in[1,J],l\neq j} t_l\rho_{jl}\right) \end{bmatrix} \tag{A.8}$$

$$(X'X)^{-1}X'VX = \frac{\sigma^2}{|X'X|} \begin{bmatrix} |X'X| + \left(\sum_{j=1}^{J} t_j^2\right)\left(2\sum\sum_{1\le j<l\le J}\rho_{jl}\right) - \left(\sum_{j=1}^{J} t_j\right)\left(\sum_{j=1}^{J}\sum_{l\in[1,J],l\neq j} t_l\rho_{jl}\right), \\ \left(\sum_{j=1}^{J} t_j^2\right)\left(\sum_{j=1}^{J} t_j\left(\sum_{l\in[1,J],l\neq j}\rho_{jl}\right)\right) - \left(\sum_{j=1}^{J} t_j\right)\left(\sum_{j=1}^{J} t_j\left(\sum_{l\in[1,J],l\neq j} t_l\rho_{jl}\right)\right); \\ J\left(\sum_{j=1}^{J}\sum_{l\in[1,J],l\neq j} t_l\rho_{jl}\right) - 2\left(\sum_{j=1}^{J} t_j\right)\left(\sum\sum_{1\le j<l\le J}\rho_{jl}\right), \\ |X'X| + J\left(\sum_{j=1}^{J} t_j\left(\sum_{l\in[1,J],l\neq j} t_l\rho_{jl}\right)\right) - \left(\sum_{j=1}^{J} t_j\right)\left(\sum_{j=1}^{J} t_j\left(\sum_{l\in[1,J],l\neq j}\rho_{jl}\right)\right) \end{bmatrix} \tag{A.9}$$

We are only interested in $Var(\hat{b}_1)$, element (2,2) in $(X'X)^{-1}X'VX(X'X)^{-1}$, so for the sake of space let us consider the matrix multiplication of the second row of $(X'X)^{-1}X'VX$ (Equation A.9) and the second column of $(X'X)^{-1}$. We then find that the variance is

$$\begin{aligned} Var\left(Y_{ik}^{Slope}\right) = & \frac{\sigma^2}{\left[J\sum_{j=1}^{J} t_j^2 - \left(\sum_{j=1}^{J} t_j\right)^2\right]^2} \left[J^2\left(\sum_{j=1}^{J}\left[t_j \sum_{l\in[1,J],l\neq j} t_l\rho_{jl}\right]\right)\right. \\ & + \left(\sum_{j=1}^{J} t_j\right)^2\left(2\sum\sum_{1\le j<l\le J}\rho_{jl}\right) - J\left(\sum_{j=1}^{J} t_j\right)\left(\sum_{j=1}^{J}\left[t_j \sum_{l\in[1,J],l\neq j}\rho_{jl}\right]\right) \\ & \left. -J\left(\sum_{j=1}^{J} t_j\right)\left(\sum_{j=1}^{J}\left[\sum_{l\in[1,J],l\neq j} t_l\rho_{jl}\right]\right) + J\left(J\sum_{j=1}^{J} t_j^2 - \left[\sum_{j=1}^{J} t_j\right]^2\right)\right]. \end{aligned} \tag{A.10}$$

In order to make the results more easily understood, one can make simplifying assumptions about the correlation and the follow-up times. By making the follow-up times

be t units of time apart, we get the simpler forms of the X matrices.

$$X_{Jx2} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_J \end{bmatrix} == \begin{bmatrix} 1 & 0 \\ 1 & t \\ \vdots & \vdots \\ 1 & (J-1)t \end{bmatrix} \tag{A.11}$$

$$(X'X)_{2x2} = \begin{bmatrix} J & \sum_{j=1}^{J} t_j \\ \sum_{j=1}^{J} t_j & \sum_{j=1}^{J} t_j^2 \end{bmatrix} = \begin{bmatrix} J & \frac{tJ(J-1)}{2} \\ \frac{tJ(J-1)}{2} & \frac{t^2J(J-1)(2J-1)}{6} \end{bmatrix} \tag{A.12}$$

$$|X'X| = J\sum_{j=1}^{J} t_j^2 - \left(\sum_{j=1}^{J} t_j\right)^2 = \frac{t^2J^2(J+1)(J-1)}{12} \tag{A.13}$$

$$(X'X)^{-1} = \frac{1}{J\sum_{j=1}^{J} t_j^2 - \left(\sum_{j=1}^{J} t_j\right)^2} \begin{bmatrix} \sum_{j=1}^{J} t_j^2 & -\sum_{j=1}^{J} t_j \\ -\sum_{j=1}^{J} t_j & J \end{bmatrix} = \begin{bmatrix} \frac{2(2J-1)}{J(J+1)} & \frac{-6}{tJ(J+1)} \\ \frac{-6}{tJ(J+1)} & \frac{12}{t^2J(J+1)(J-1)} \end{bmatrix} \tag{A.14}$$

We can also make the assumption of compound symmetry, such that $\rho_{jl} = \rho$ for all observations.

$$\begin{aligned} (X'VX)_{2x2} &= \sigma^2 \begin{bmatrix} J + J(J-1)\rho & \frac{tJ(J-1)}{2} + t\rho\frac{J(J-1)^2}{2} \\ \frac{tJ(J-1)}{2} + t\rho\frac{J(J-1)^2}{2} & t^2\frac{J(J-1)(2J-1)}{6} + \frac{t^2\rho J(J-1)(J-2)(3J-1)}{12} \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} J[1+\rho(J-1)] & \frac{tJ(J-1)}{2}[1+\rho(J-1)] \\ \frac{tJ(J-1)}{2}[1+\rho(J-1)] & \frac{t^2J(J-1)}{12}[(4J-2)+\rho(J-2)(3J-1)] \end{bmatrix} \end{aligned} \tag{A.15}$$

$$(X'X)^{-1}X'VX == \sigma^2 \begin{bmatrix} 1+\rho(J-1) & \frac{t\rho J(J-1)}{2} \\ 0 & 1-\rho \end{bmatrix} \tag{A.16}$$

$$(X'X)^{-1}X'VX(X'X)^{-1} == \sigma^2 \begin{bmatrix} \frac{(4J+2)+\rho(J^2-3J+2)}{J(J+1)} & \frac{-6(1-\rho)}{tJ(J+1)} \\ \frac{-6(1-\rho)}{tJ(J+1)} & \frac{12(1-\rho)}{t^2J(J+1)(J-1)} \end{bmatrix} \tag{A.17}$$

Again, the formula of interest is in the lower right cell.

Supplementary Tables

Table A.1: Spatial coordinates of the 16 segments in the model of the left ventricle[29]. They are denoted as their level (base, mid, apex), orientation (anterior, septal, inferior, lateral), and index number.

Base, Ant. (1)	$(0,\frac{5}{6})$	Mid, Ant. (7)	$(0,\frac{1}{2})$	Apex, Ant. (13)	$(0,\frac{1}{6})$
Base, Ant.Sep. (2)	$(\frac{-5\sqrt{3}}{12},\frac{5}{12})$	Mid, Ant.Sep. (8)	$(\frac{-\sqrt{3}}{4},\frac{1}{4})$	Apex, Sep. (14)	$(\frac{-1}{6},0)$
Base, Inf.Sep. (3)	$(\frac{-5\sqrt{3}}{12},\frac{-5}{12})$	Mid, Inf.Sep. (9)	$(\frac{-\sqrt{3}}{4},\frac{-1}{4})$	Apex, Inf. (15)	$(0,\frac{-1}{6})$
Base, Inf. (4)	$(0,\frac{-5}{6})$	Mid, Inf. (10)	$(0,\frac{-1}{2})$	Apex, Lat. (16)	$(\frac{1}{6},0)$
Base, Inf.Lat. (5)	$(\frac{5\sqrt{3}}{12},\frac{-5}{12})$	Mid, Inf.Lat. (11)	$(\frac{\sqrt{3}}{4},\frac{-1}{4})$		
Base, Ant.Lat (6)	$(\frac{5\sqrt{3}}{12},\frac{5}{12})$	Mid, Ant.Lat (12)	$(\frac{\sqrt{3}}{4},\frac{1}{4})$		

Table A.2: Parameters used to produce high and low degrees of correlation in the parametric correlation structures.

Correlation Type	Correlation Function	Low Correlation	High Correlation
Spatial	Exponential	$\rho = 0.3$	$\rho = 0.8$
	Spherical	$\phi = 2.25$	$\phi = 6$
	Matérn	$\phi = 0.4, \nu = 1$	$\phi = 0.9, \nu = 2$
Temporal	Compound Symmetry	$\rho = 0.4$	$\rho = 0.8$
	Autoregressive-1	$\rho = 0.6$	$\rho = 0.9$

Simulation Design

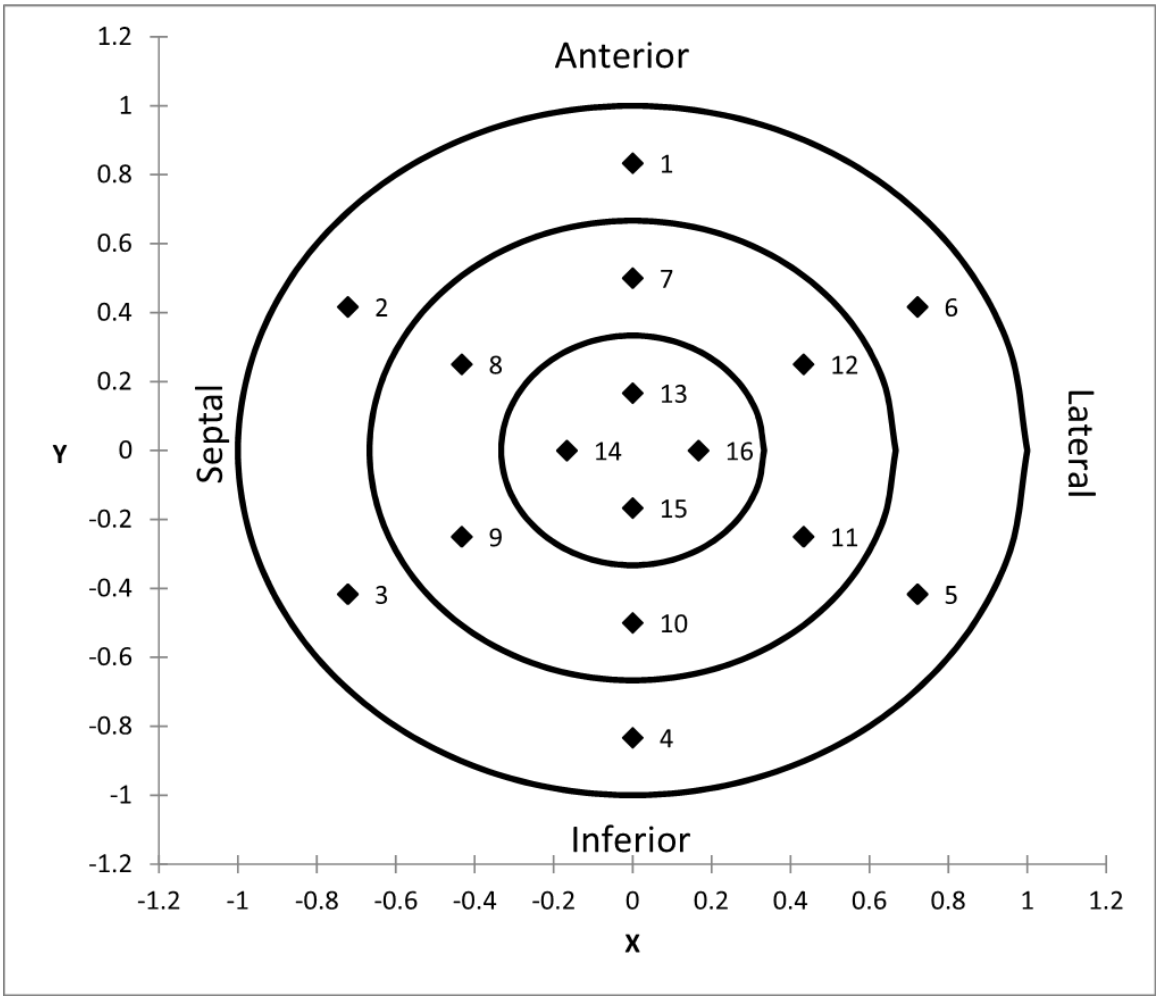

Figure A.1: Plot of the 16 segments of the left ventricle. The outer ring corresponds to the base, the middle ring to the mid, and the inner circle to the apex[11, 29]. The numbers correspond to the segment's index as defined in Table A.1.

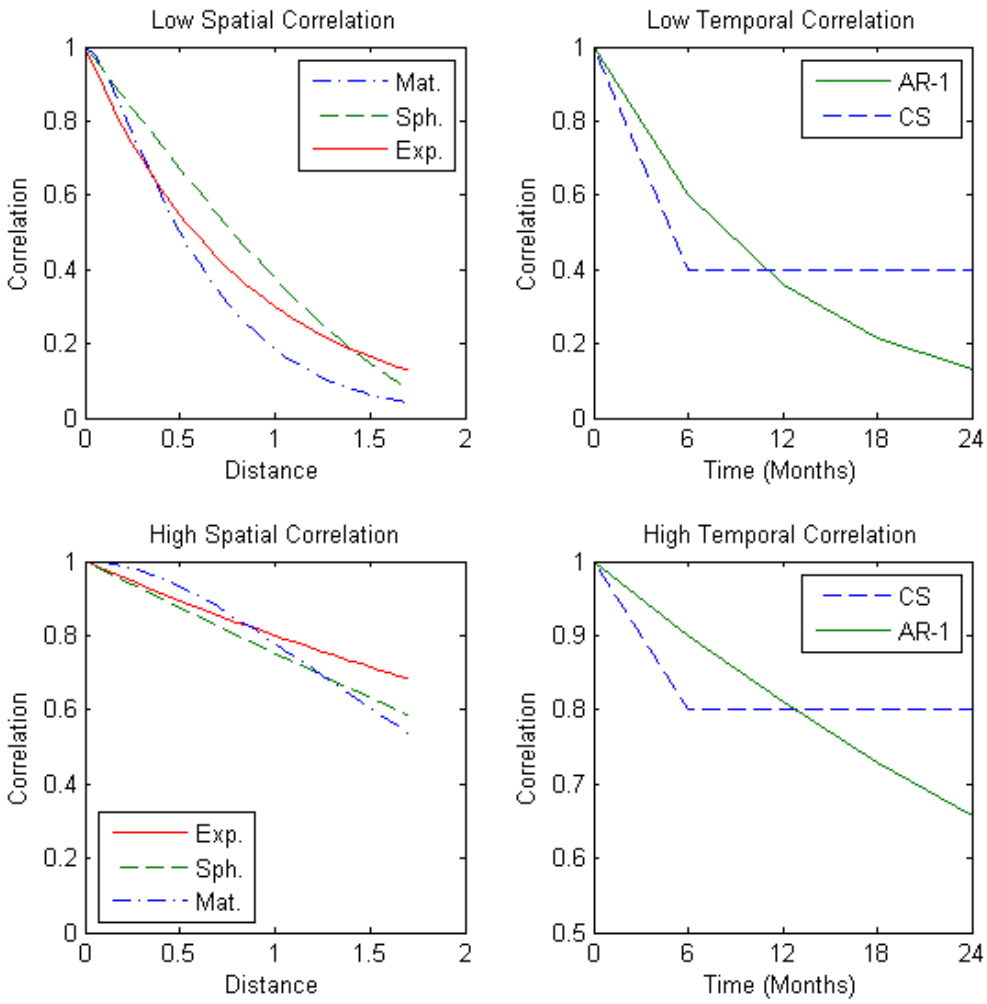

Figure A.2: Plots of the covariance functions used to generate the spatiotemporal data. The spatial structures (exponential, spherical, and Matérn) are on the left and the temporal structures (compound symmetric and autoregressive-1) are on the right. The functions used to generate data with a low degree of correlation are on the top, and those generating a high degree are on the bottom.

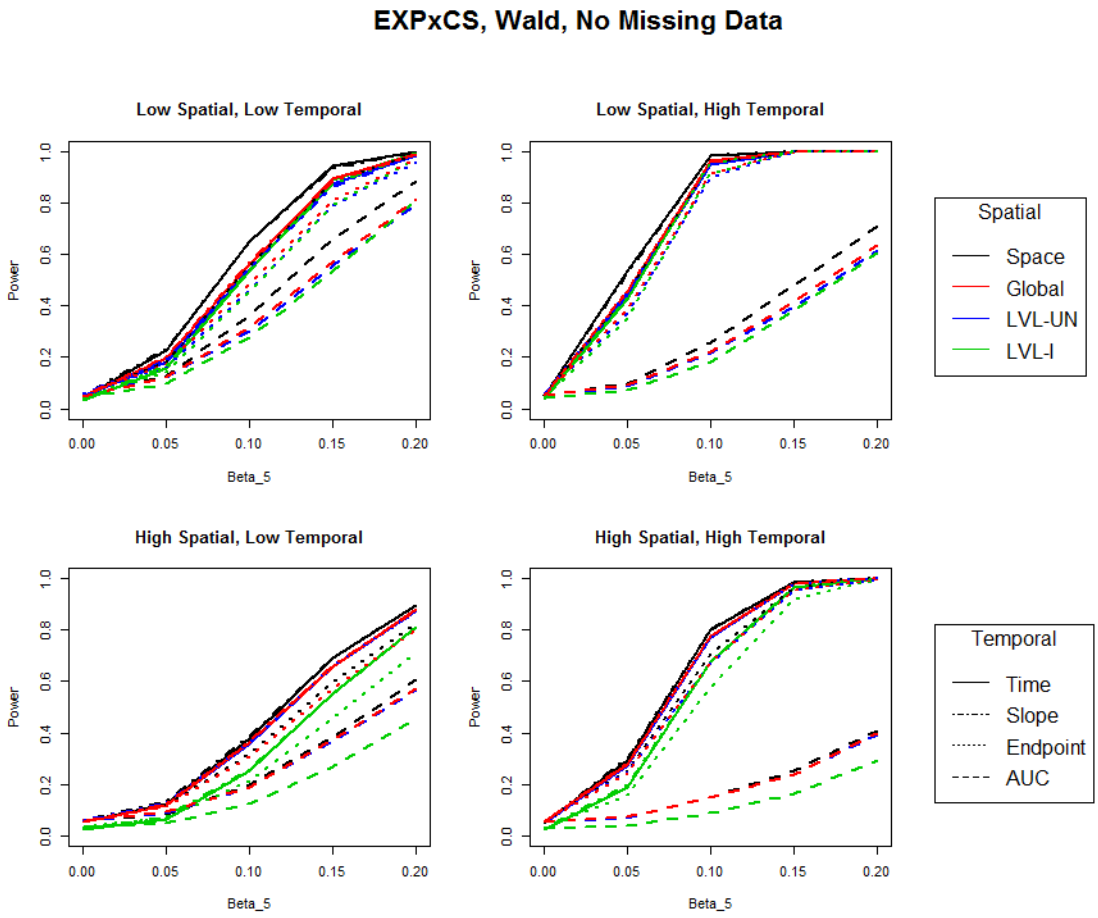

Figure A.3: Plot of the empirical power curves for the sixteen models under an exponential-by-compound symmetric generating correlation structure where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

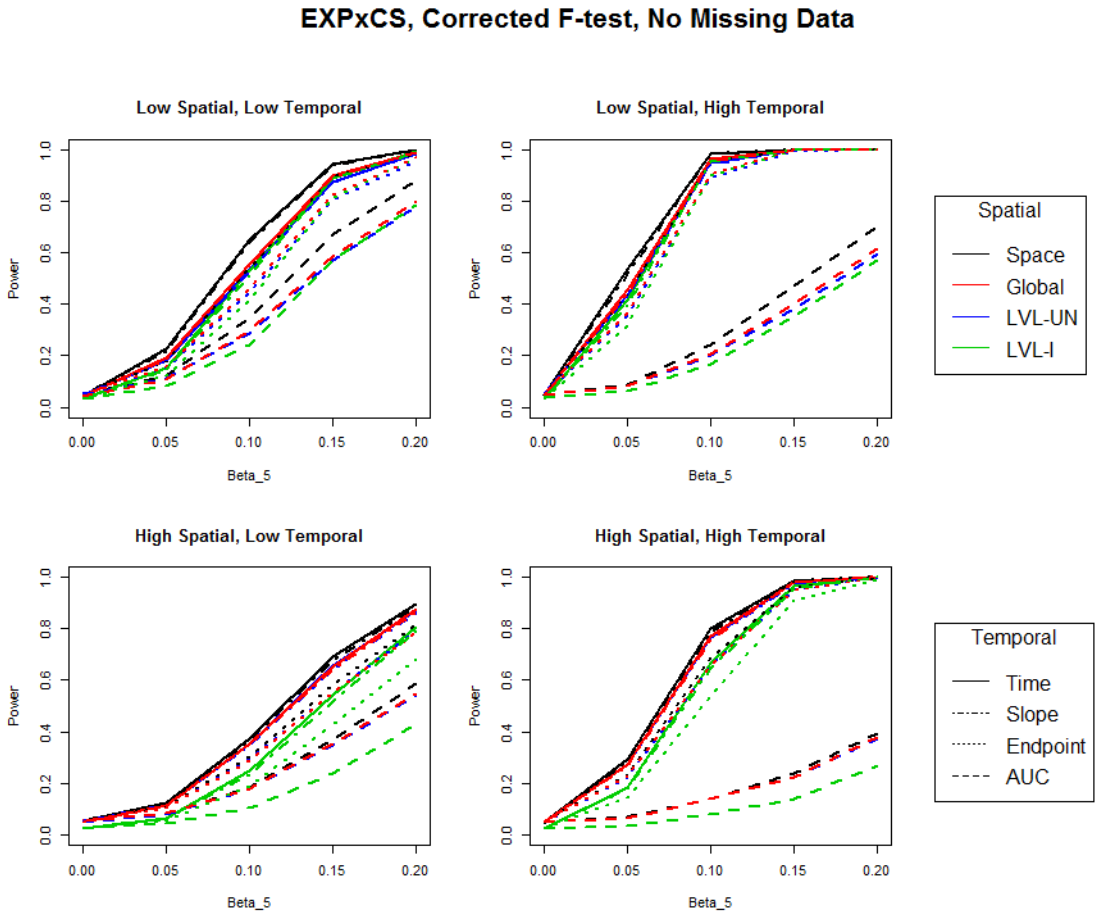

Figure A.4: Plot of the empirical power curves for the sixteen models under an exponential-by-compound symmetric generating correlation structure where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

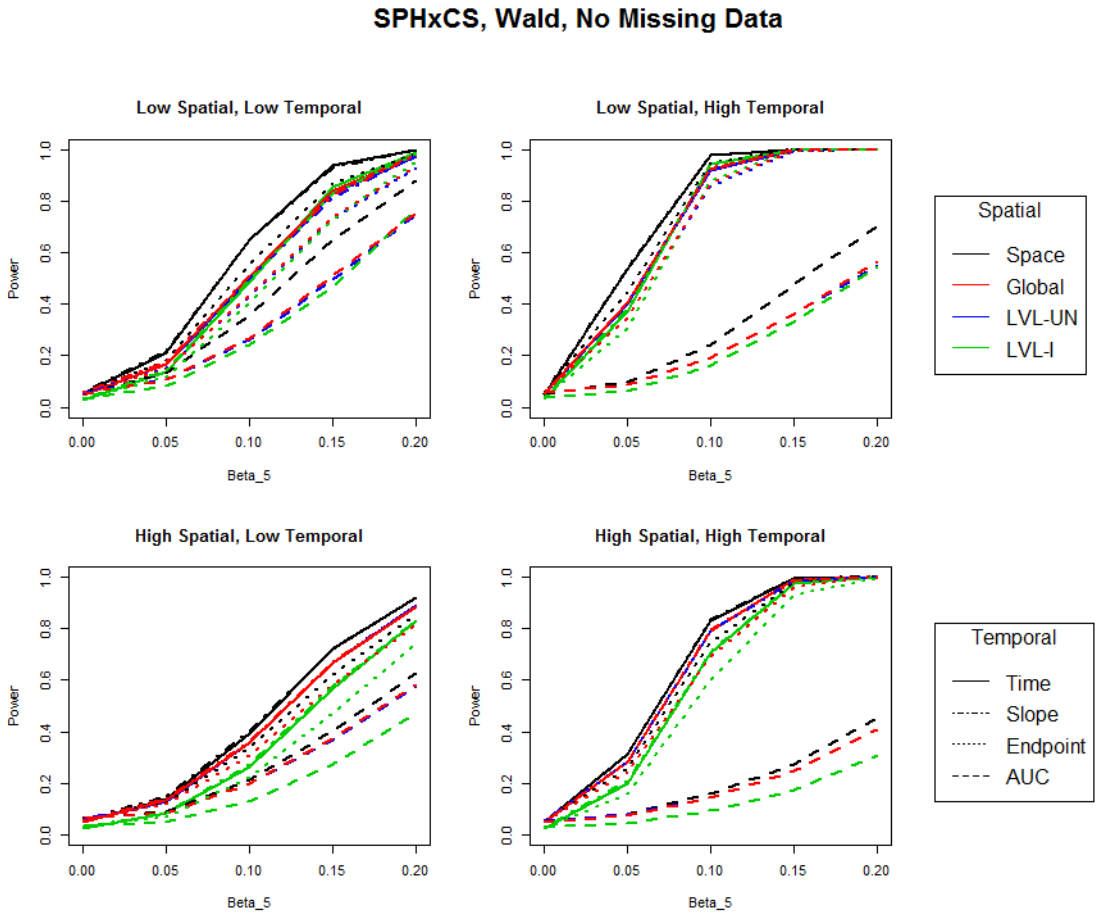

Figure A.5: Plot of the empirical power curves for the sixteen models under a spherical-by-compound symmetric generating correlation structure where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

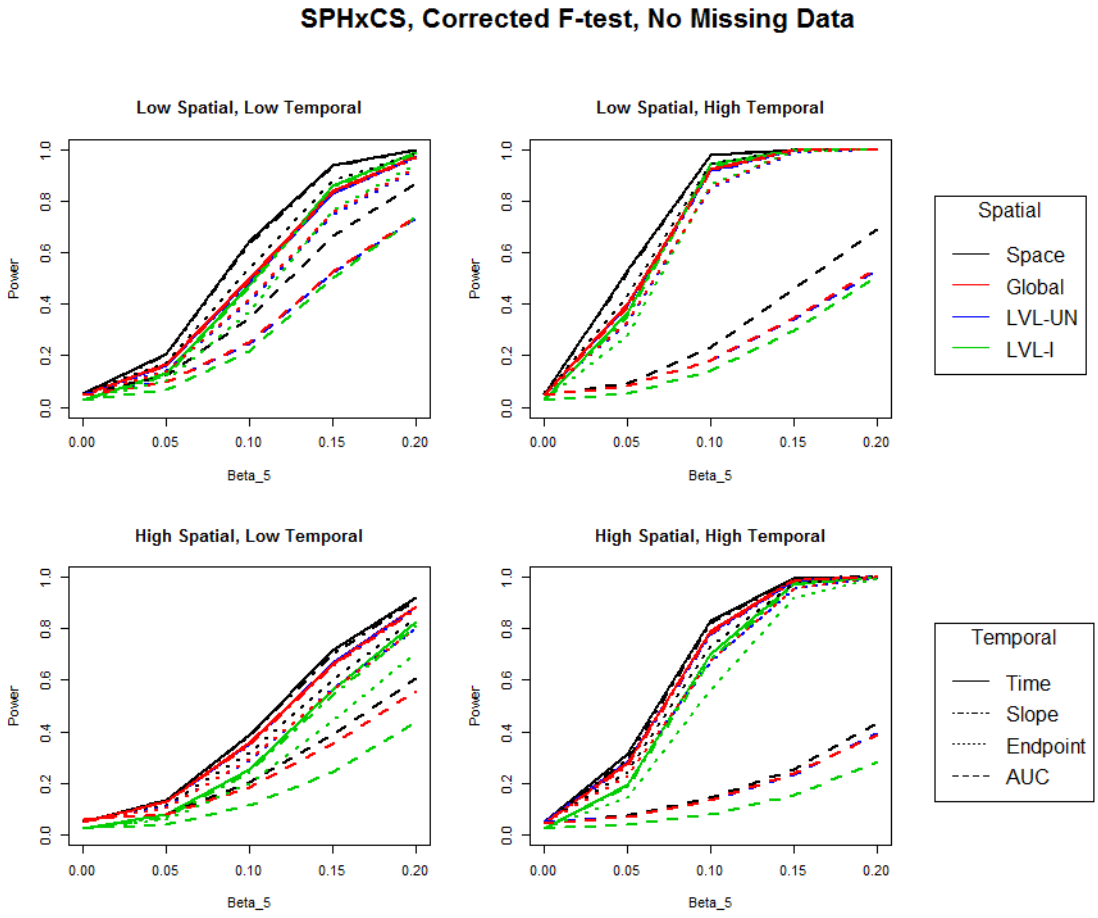

Figure A.6: Plot of the empirical power curves for the sixteen models under a spherical-by-compound symmetric generating correlation structure where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

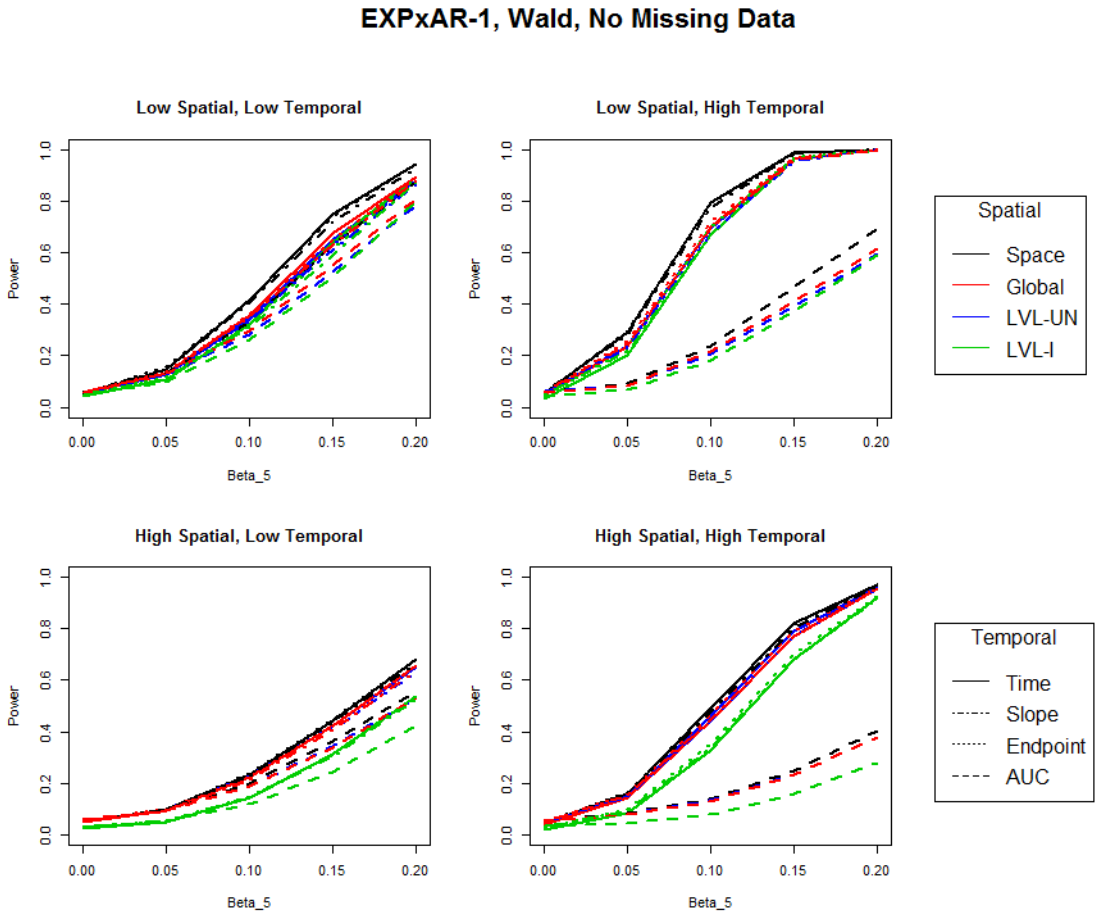

Figure A.7: Plot of the empirical power curves for the sixteen models under an exponential-by-autoregressive-1 generating correlation structure where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

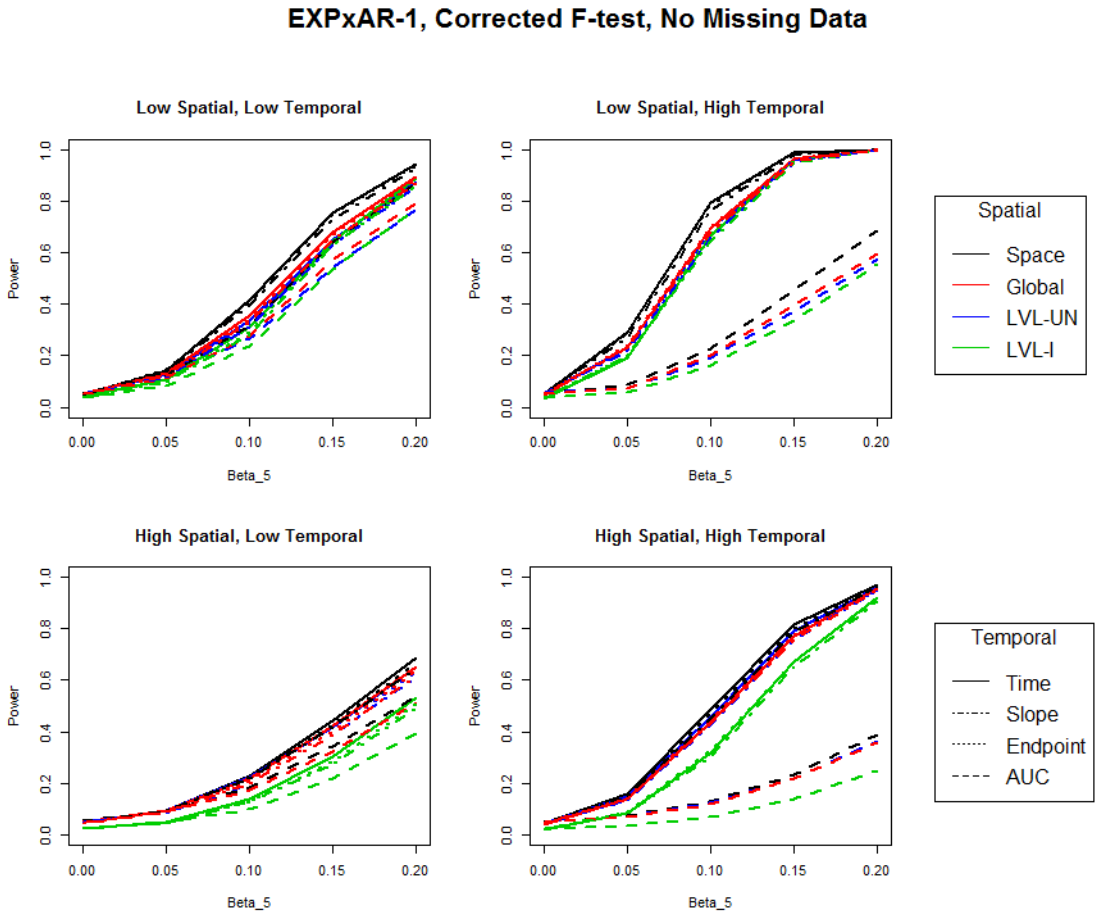

Figure A.8: Plot of the empirical power curves for the sixteen models under an exponential-by-autoregressive-1 generating correlation structure where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

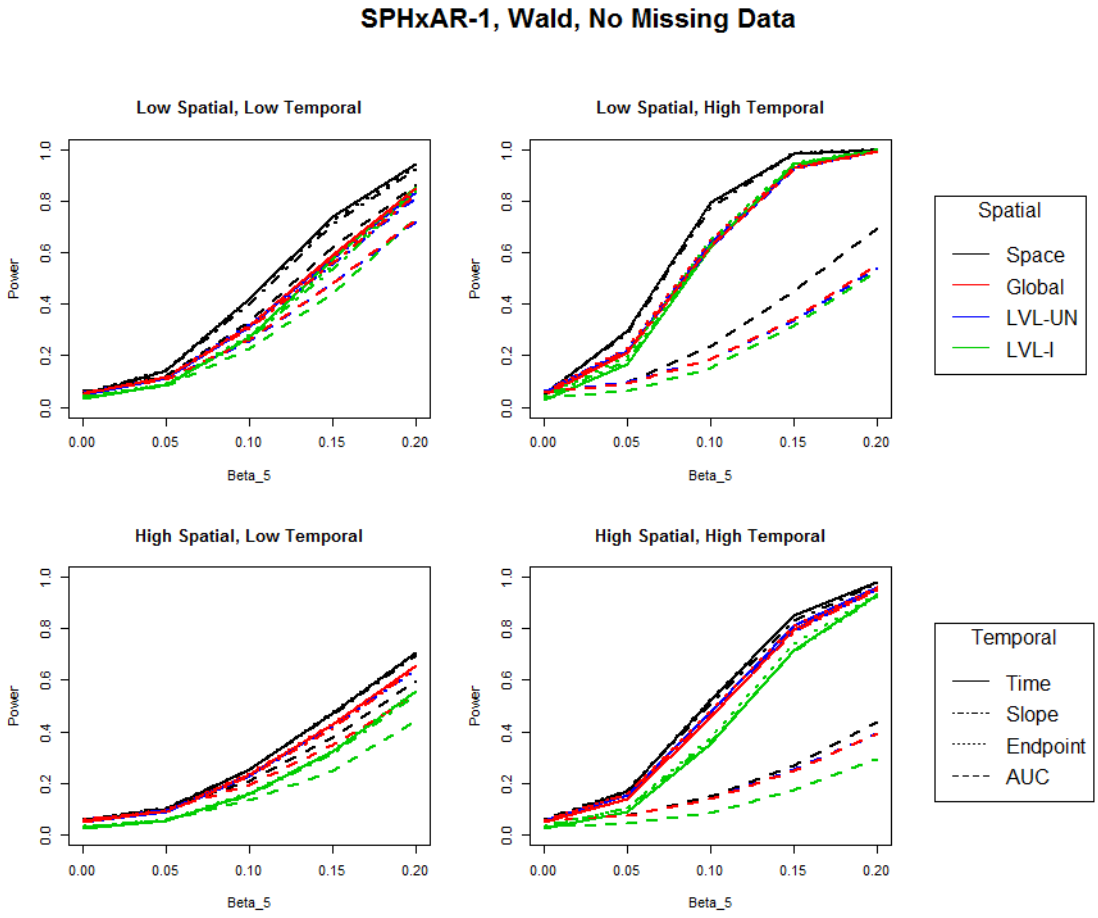

Figure A.9: Plot of the empirical power curves for the sixteen models under a spherical-by-autoregressive-1 generating correlation structure where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

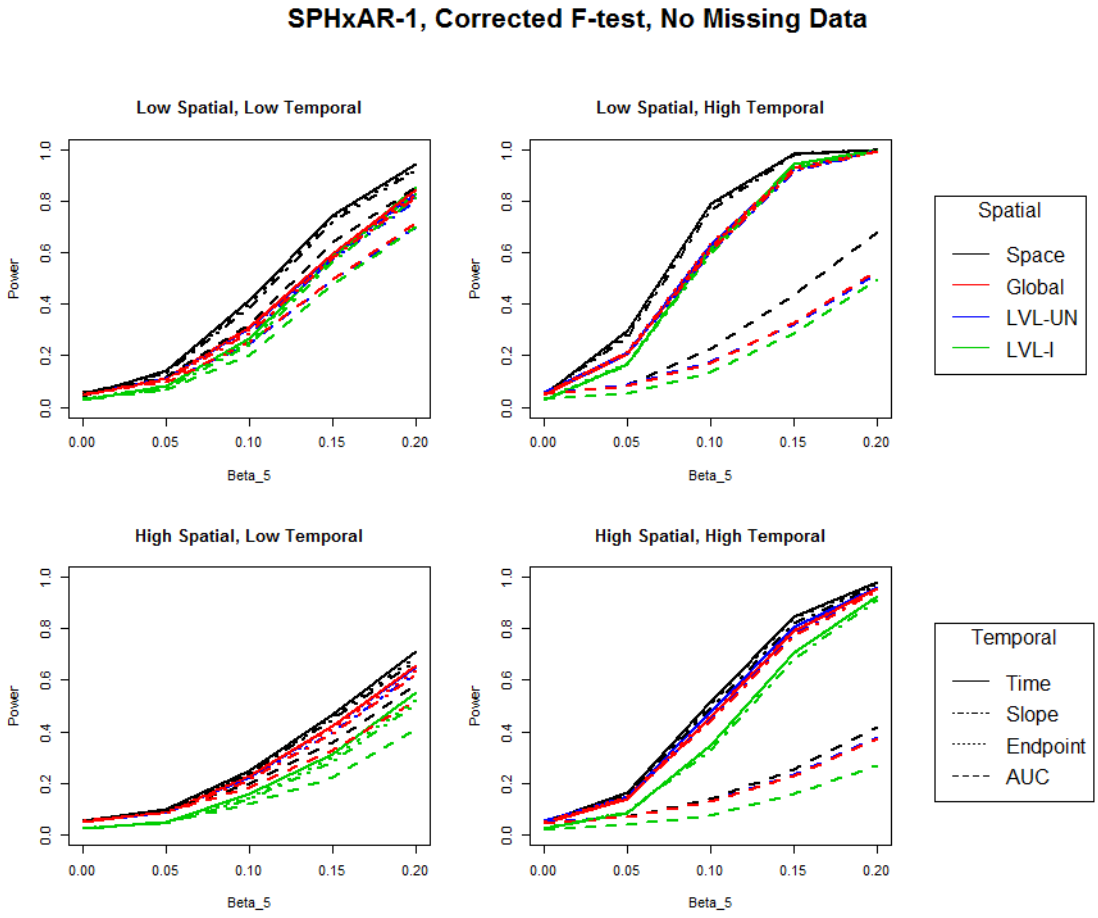

Figure A.10: Plot of the empirical power curves for the sixteen models under a spherical-by-autoregressive-1 generating correlation structure where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

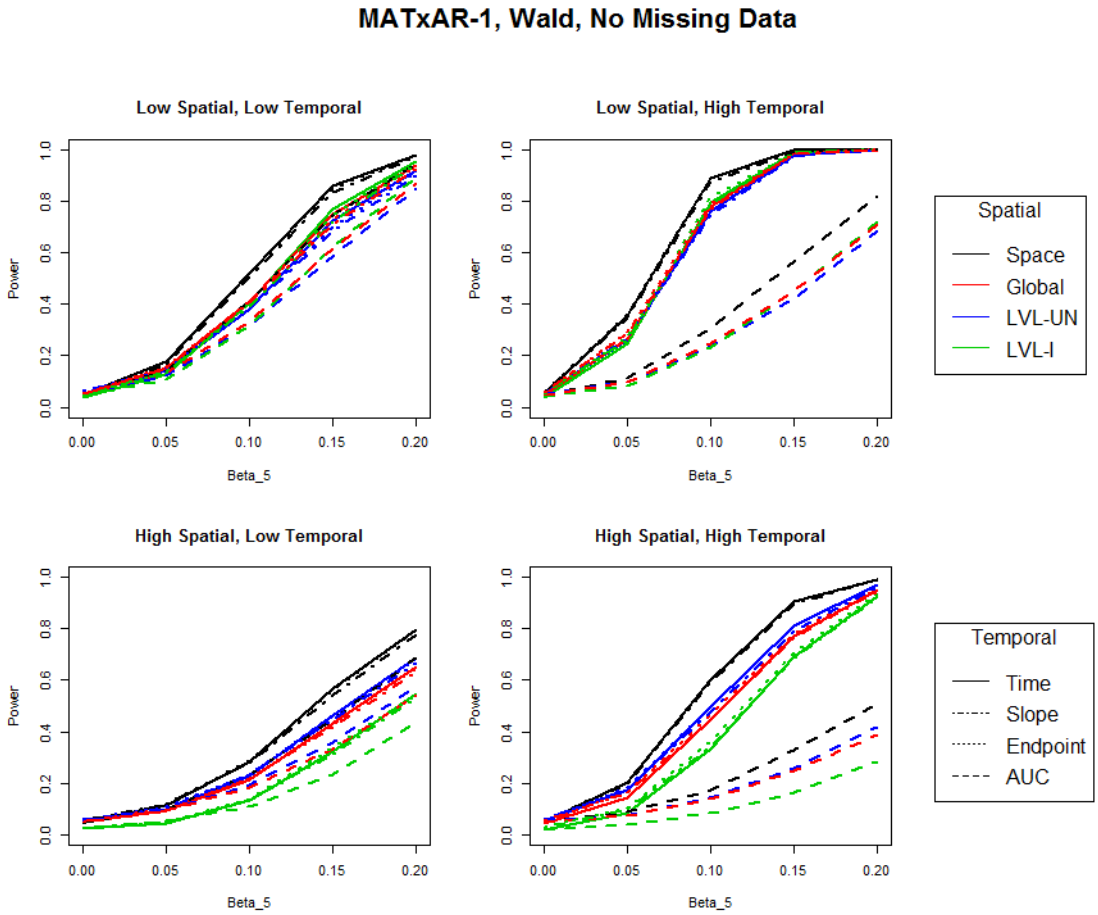

Figure A.11: Plot of the empirical power curves for the sixteen models under a Matérn-by-autoregressive-1 generating correlation structure where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

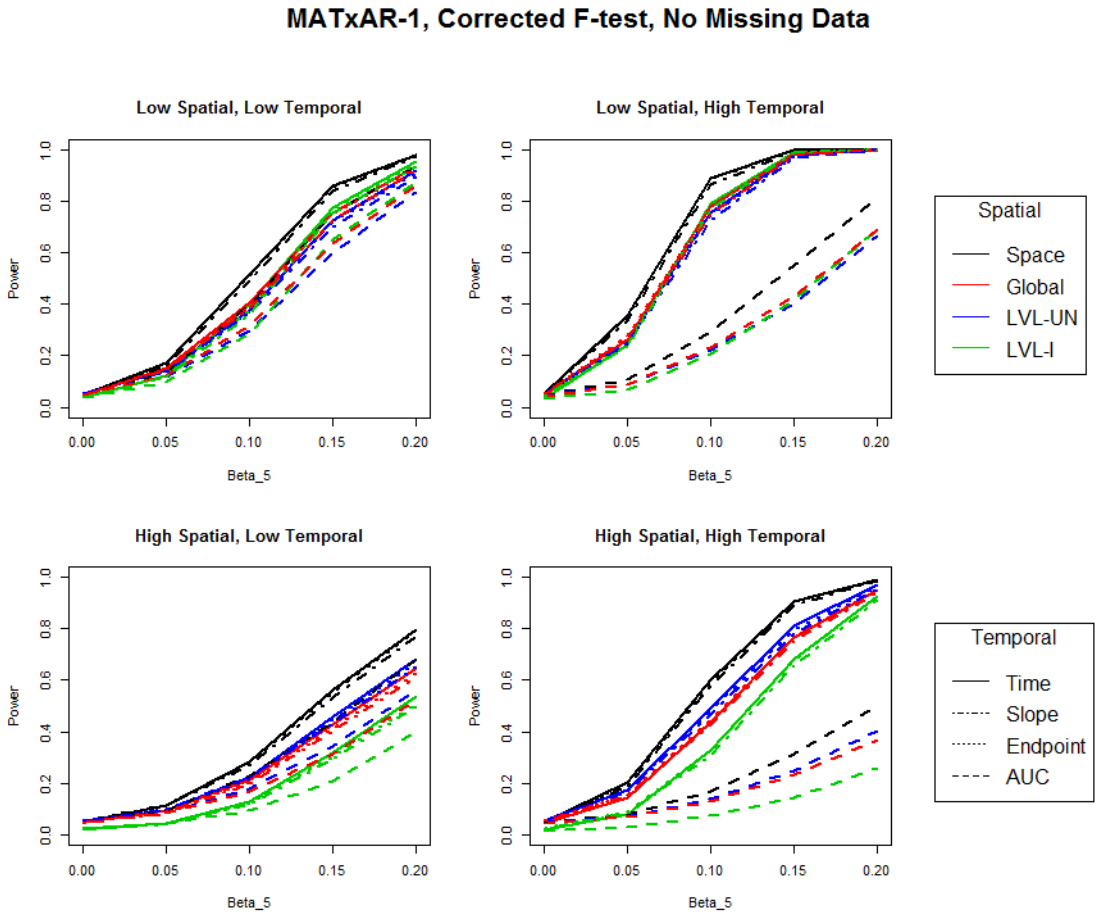

Figure A.12: Plot of the empirical power curves for the sixteen models under a Matérn-by-autoregressive-1 generating correlation structure where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

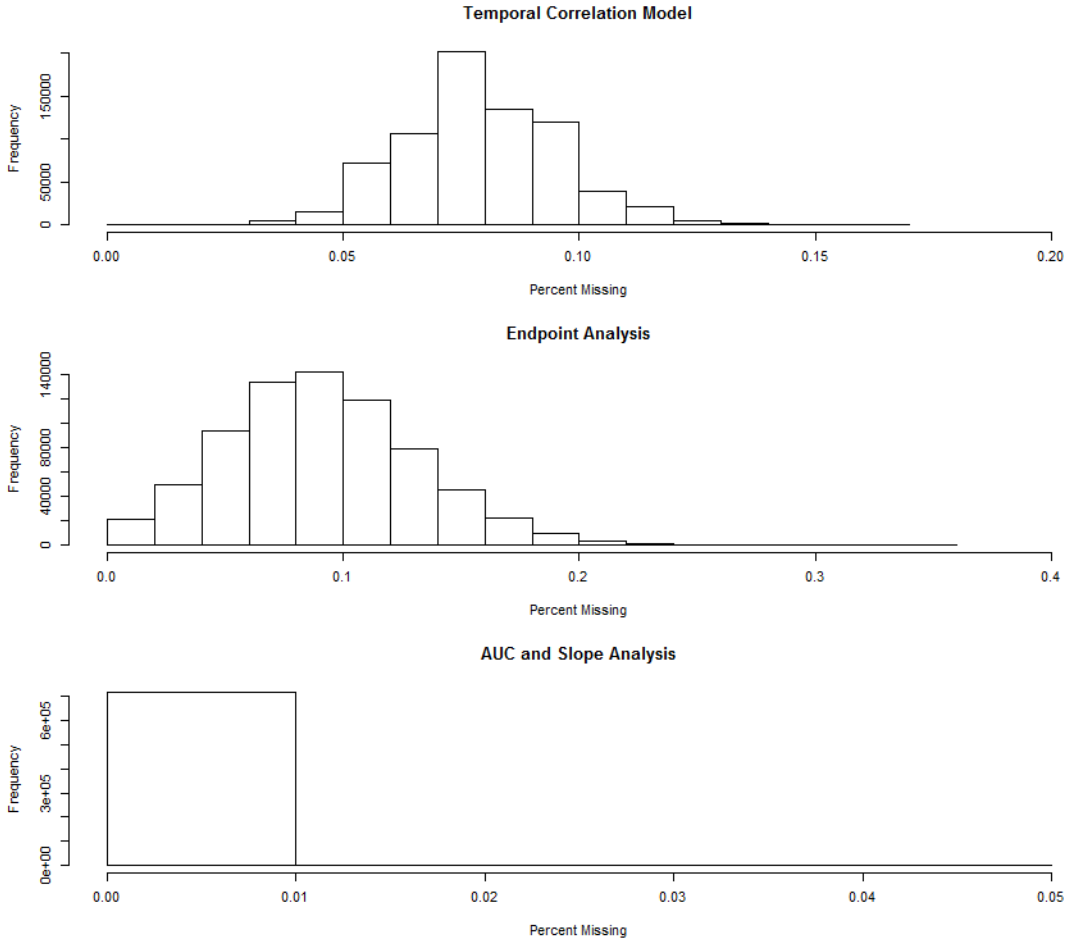

Figure A.13: Histograms of the percent of missing data for the three approaches to longitudinally missing data: a correlation model that uses all of a subject's non-missing observations, endpoint analysis that only uses subjects with a first and last observation, and slope and AUC analysis that use individual estimates as long as the subject has at least two observations in time. The means of the three distributions are 8%, 10%, and 0.01%.

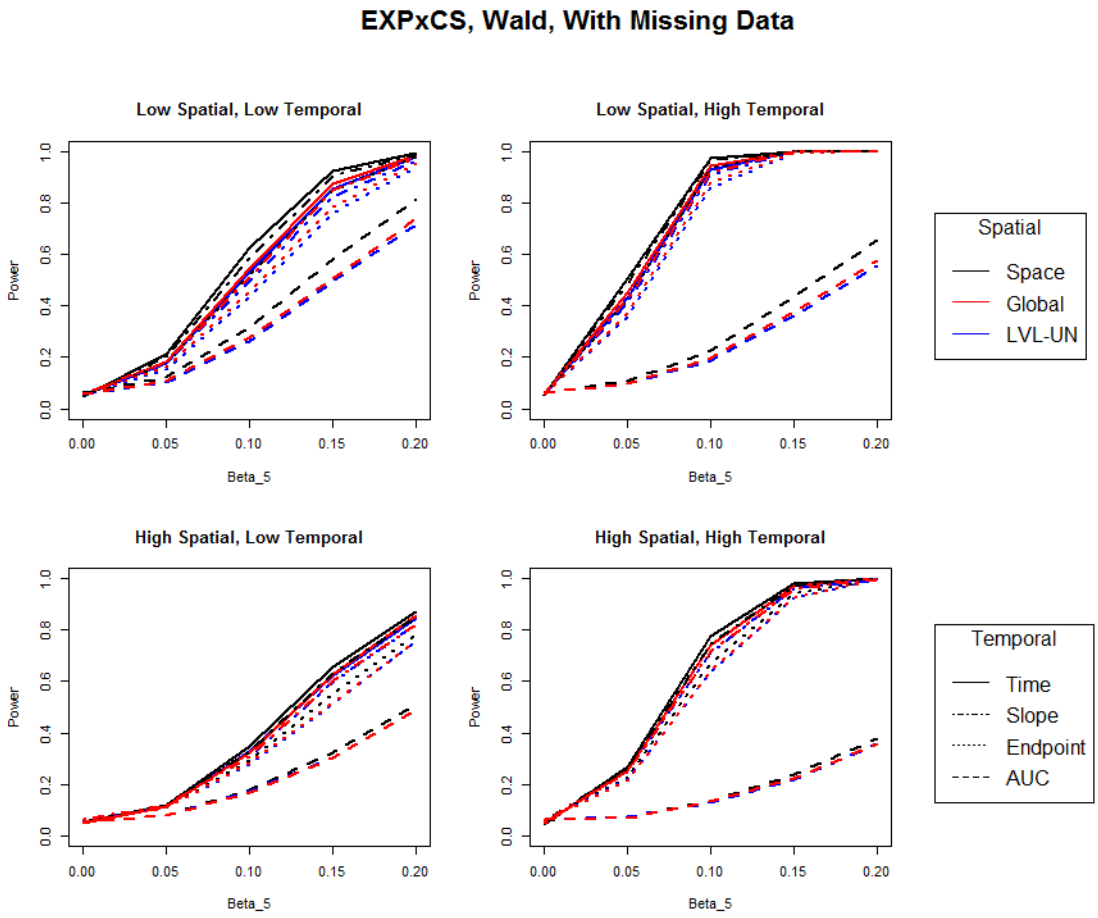

Figure A.14: Plot of the empirical power curves for the twelve models under an exponential-by-compound symmetric generating correlation structure with longitudinally missing data where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

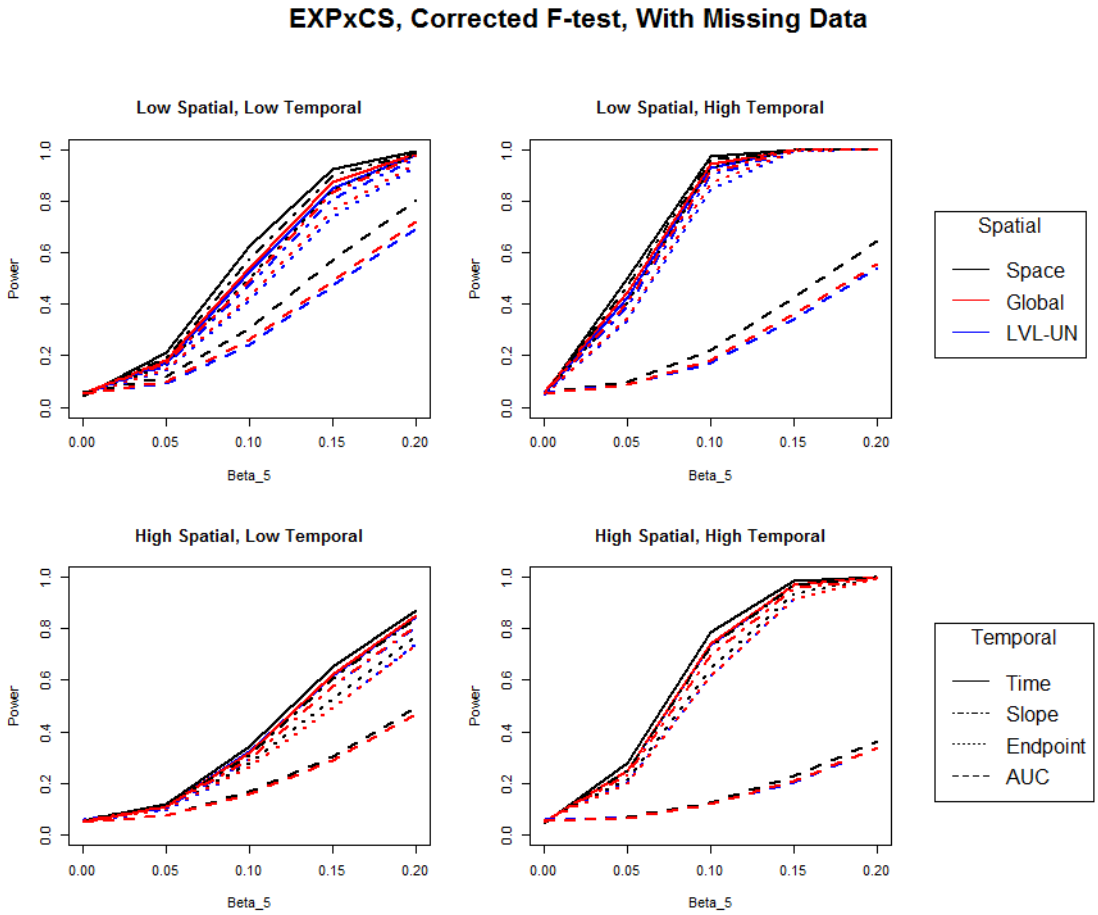

Figure A.15: Plot of the empirical power curves for the twelve models under an exponential-by-compound symmetric generating correlation structure with longitudinally missing data where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

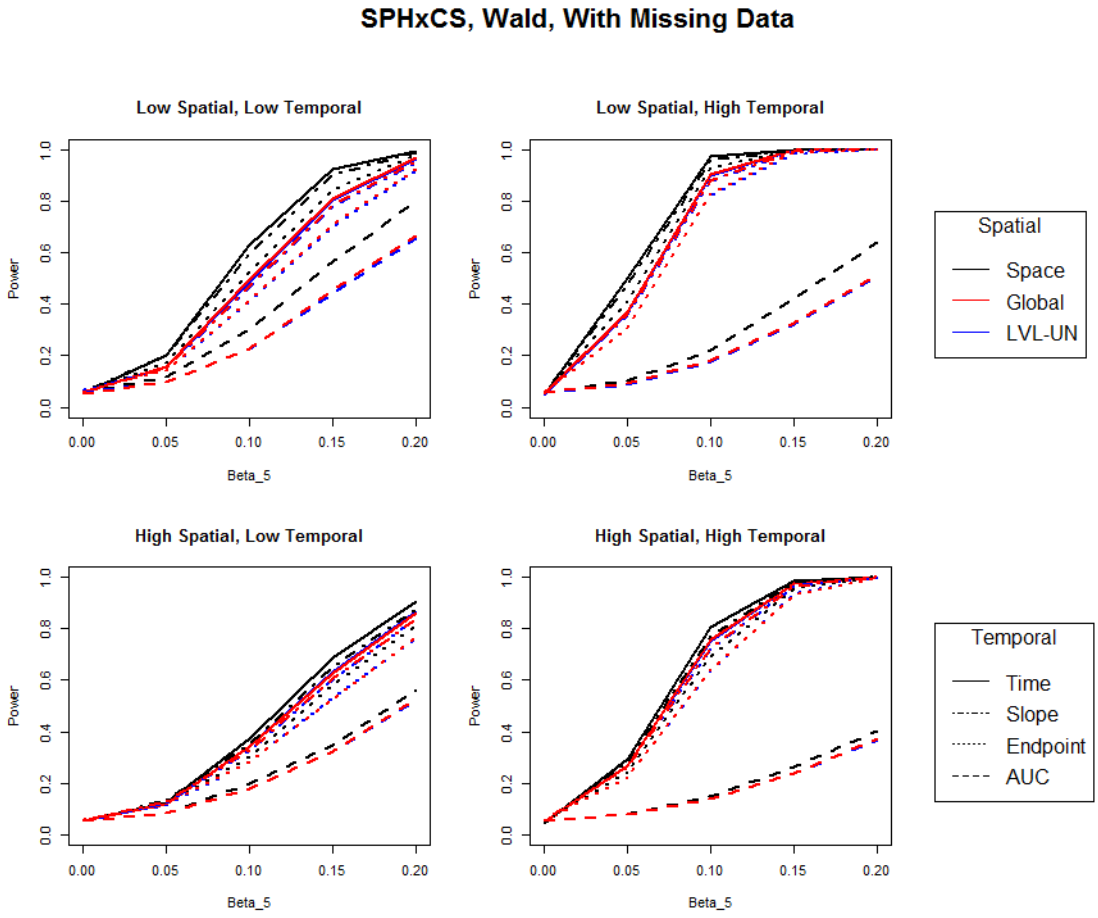

Figure A.16: Plot of the empirical power curves for the twelve models under a spherical-by-compound symmetric generating correlation structure with longitudinally missing data where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

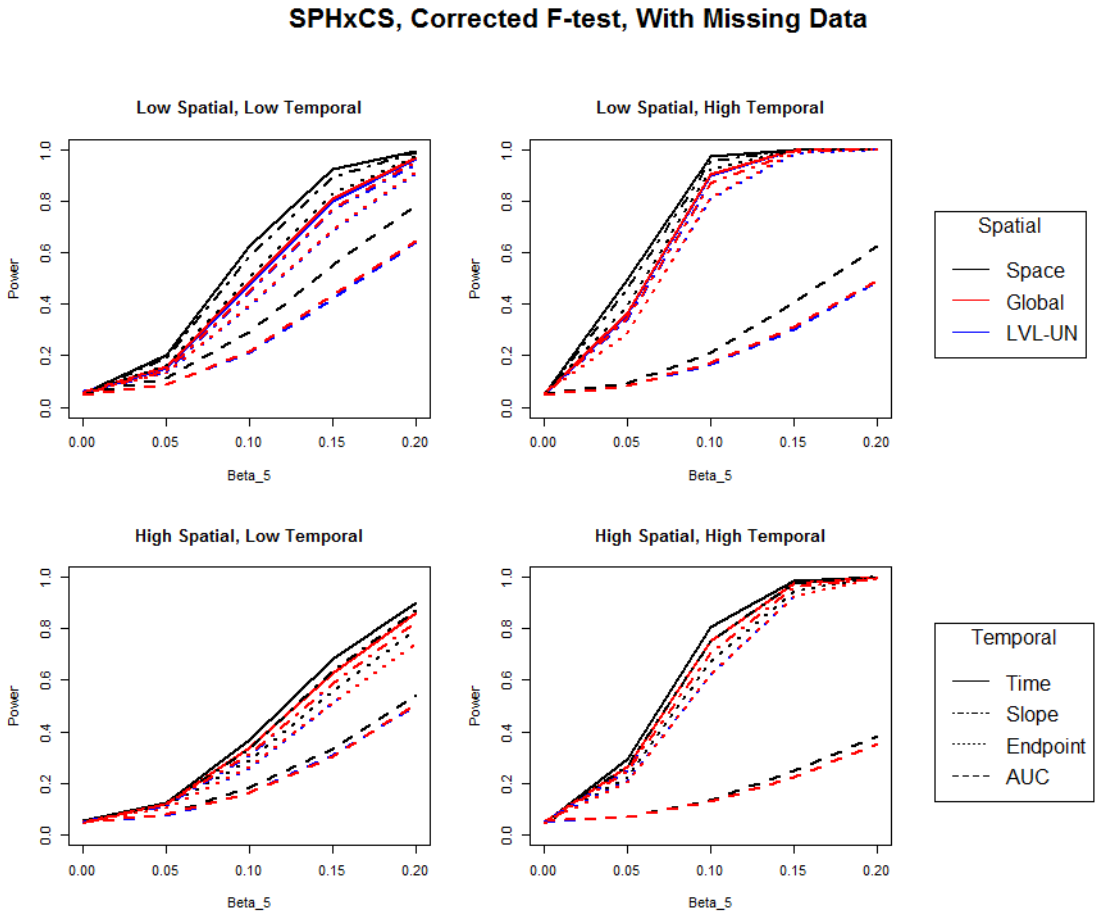

Figure A.17: Plot of the empirical power curves for the twelve models under a spherical-by-compound symmetric generating correlation structure with longitudinally missing data where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

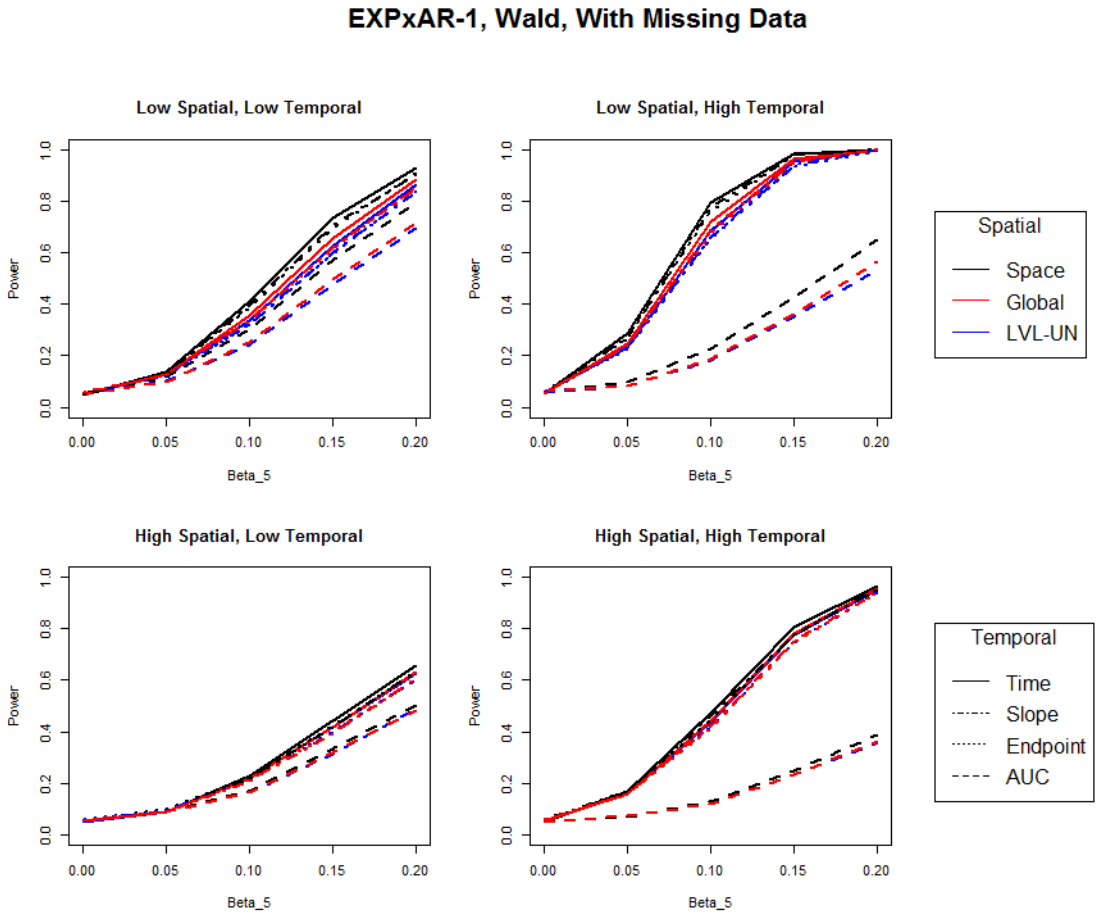

Figure A.18: Plot of the empirical power curves for the twelve models under an exponential-by-autoregressive-1 generating correlation structure with longitudinally missing data where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

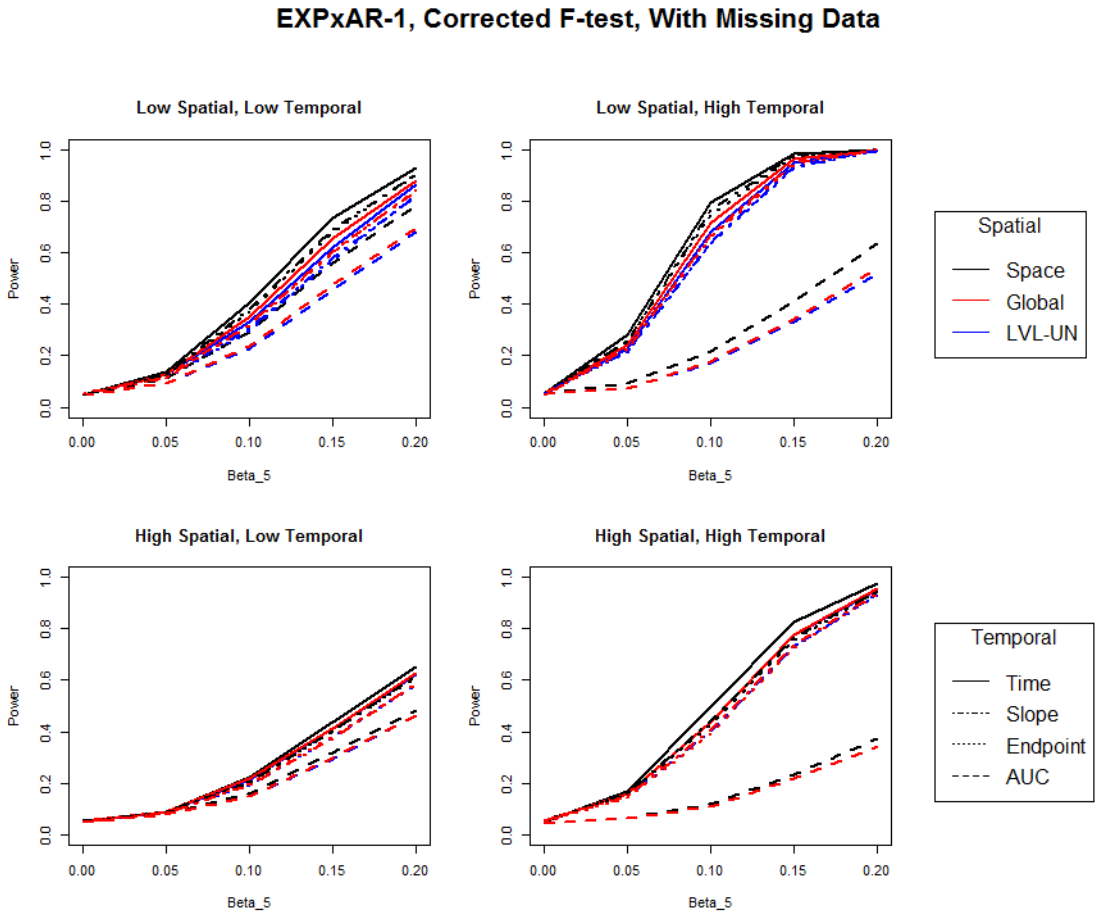

Figure A.19: Plot of the empirical power curves for the twelve models under an exponential-by-autoregressive-1 generating correlation structure with longitudinally missing data where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

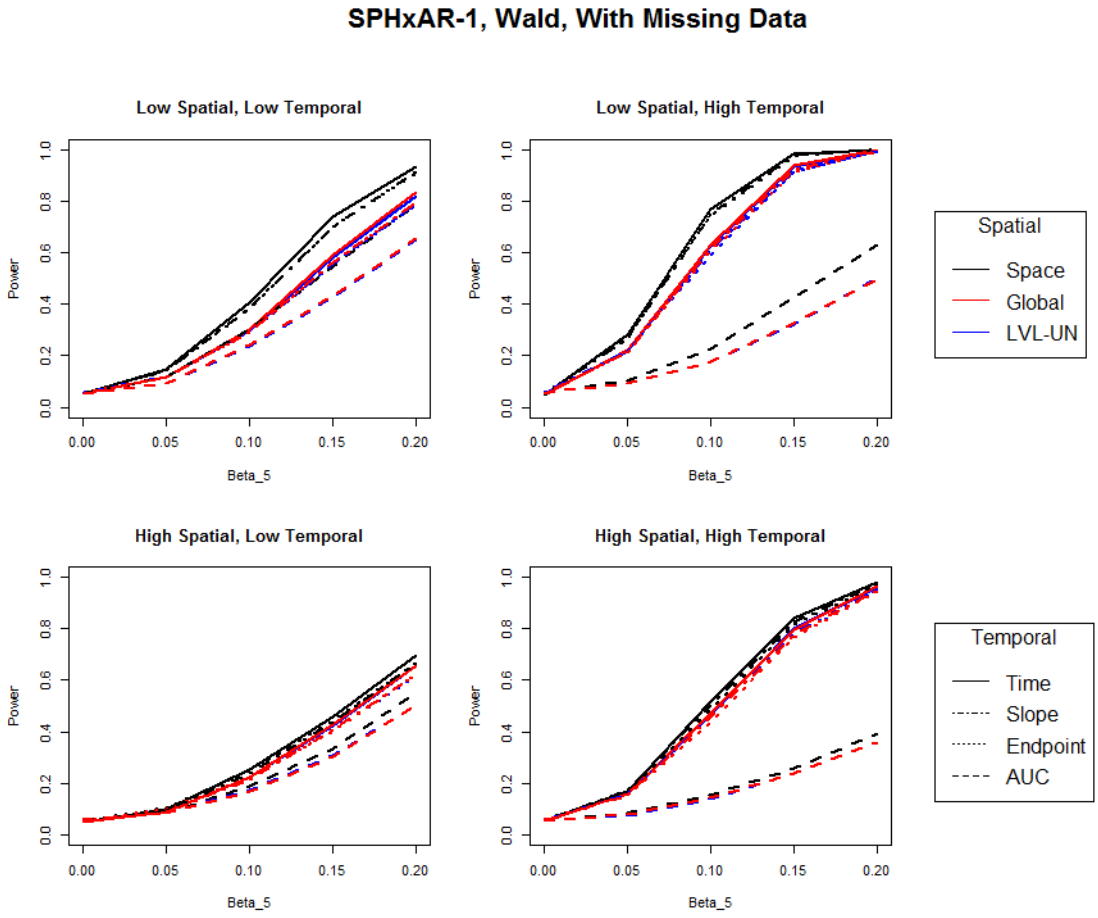

Figure A.20: Plot of the empirical power curves for the twelve models under a spherical-by-autoregressive-1 generating correlation structure with longitudinally missing data where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

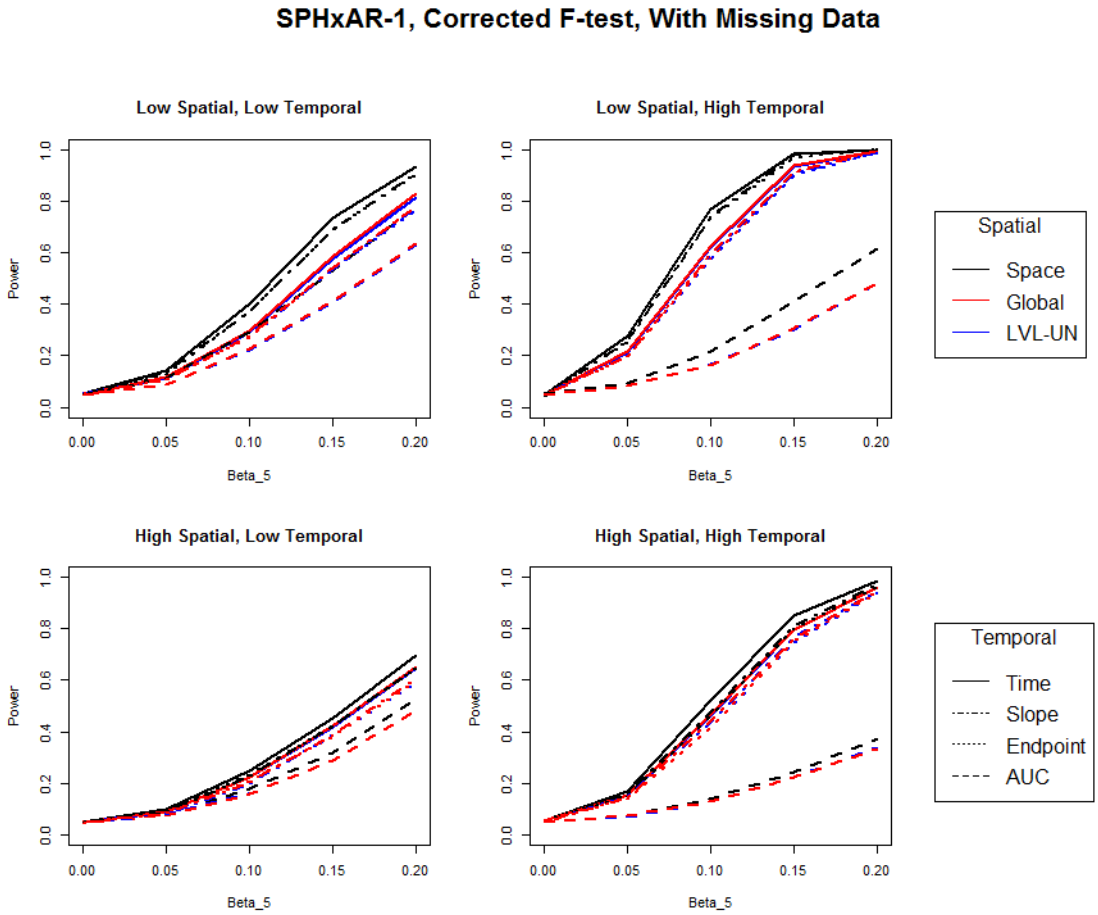

Figure A.21: Plot of the empirical power curves for the twelve models under a spherical-by-autoregressive-1 generating correlation structure with longitudinally missing data where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

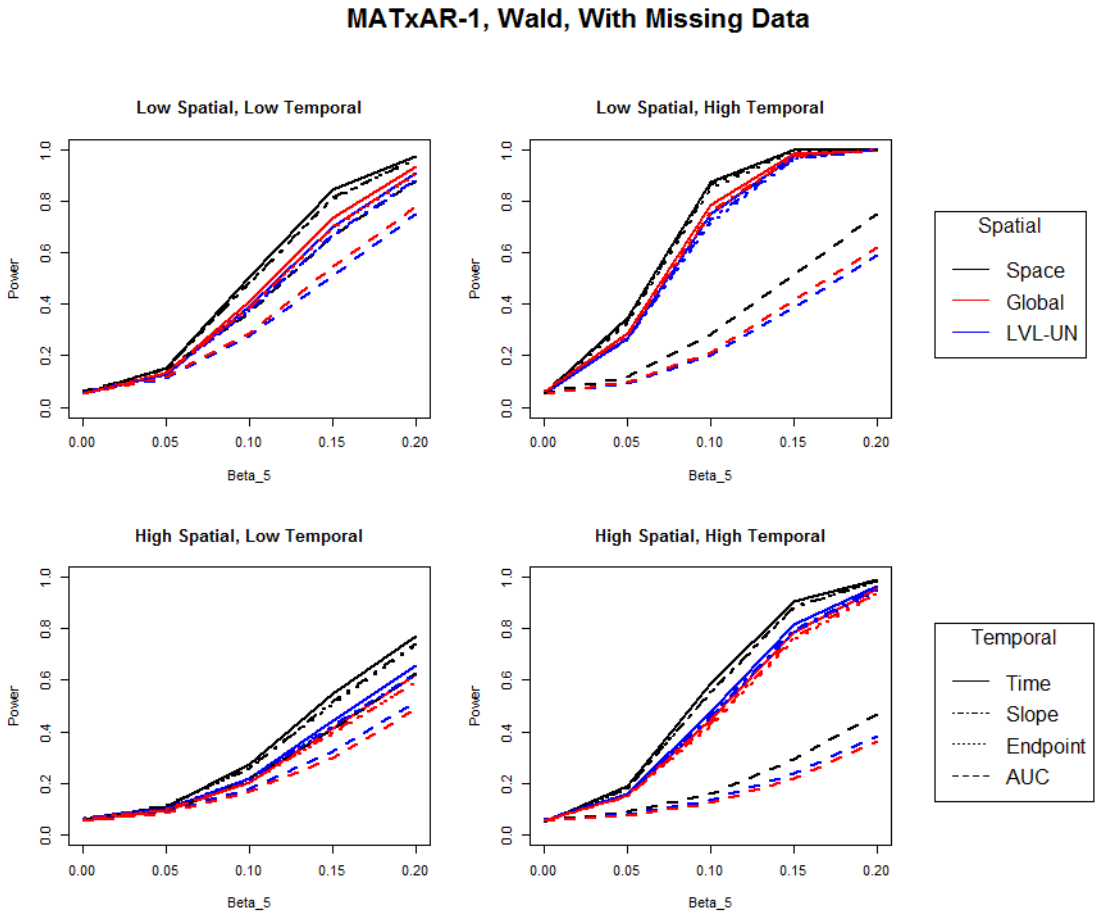

Figure A.22: Plot of the empirical power curves for the twelve models under a Matérn-by-autoregressive-1 generating correlation structure with longitudinally missing data where a Wald's test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

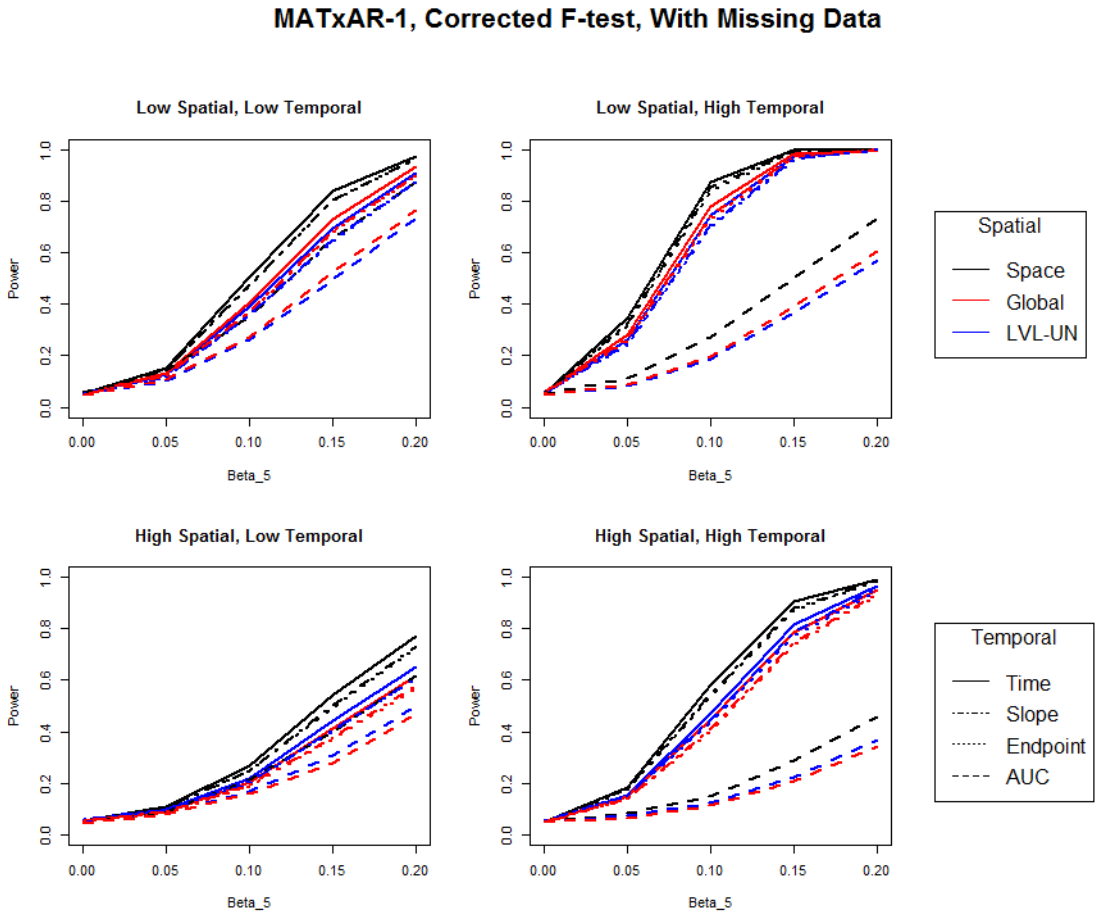

Figure A.23: Plot of the empirical power curves for the twelve models under a Matérn-by-autoregressive-1 generating correlation structure with longitudinally missing data where a corrected F-test was used for inference about the treatment-by-time interaction. Results from the different combinations of degrees of spatial and temporal correlation are given in the different panels. Note that for a certain combination the spatial method is denoted by color and the temporal method by line style.

APPLYING A SPATIOTEMPORAL CORRELATION MODEL FOR LONGITUDINAL IMAGING DATA

BRANDON GEORGE, INMACULADA ABAN

In preparation for *Annals of Applied Statistics*

Format adapted for dissertation

ABSTRACT

Longitudinal imaging studies have both spatial and temporal correlation among the multiple outcome measurements from a subject. Statistical methods of analysis must properly account for this autocorrelation. In this work we discuss how a linear model with a separable parametric correlation structure could be used to analyze data from such a study. The goal of this paper is to provide an easily understood description of how such a model works and discuss how it can be applied to real data. Model assumptions are discussed and the process of selecting a working correlation structure is thoroughly discussed. The steps necessitating collaboration between statistical and scientific investigators have been highlighted, as have considerations for missing data or uneven follow-up.

The results from a completed longitudinal imaging study were considered for illustration purposes. The data comes from a clinical trial for medical therapy for patients with mitral regurgitation, with repeated measurements taken at sixteen locations from the left ventricle to measure disease progression. The spatiotemporal correlation model was compared to previously used summary measures to demonstrate improved power as well as increased flexibility in the use of time- and space-varying predictors.

INTRODUCTION

Imaging studies have grown in popularity in recent years as clinical investigators are making use of the ability of imaging modalities to accurately measure outcomes within the body. These modalities allow quantification of internal anatomical or physiological properties that would have previously required invasive surgery or autopsy, which provides investigators with new understanding of how the body works. It thus makes perfect sense

for imaging studies to be joined with longitudinal studies, which allow for observation of how an outcome changes over time.

Although these longitudinal imaging studies offer wonderful knowledge of how the inside of the body changes over time, there are challenges in the statistical analysis of such datasets. The repeated measures of an individual introduces temporal correlation between observations that must be controlled for. In addition, imaging studies frequently take multiple outcome measures from a single image (for instance, functional MRI scans may measure oxygen consumption at thousands of points throughout the brain) which are spatially correlated.

In order to control for these two sources of correlation, we have proposed the use of a linear model with a separable parametric spatiotemporal correlation structure[1]. We have shown that information criteria are highly accurate at choosing an appropriate combination of spatial and temporal correlation functions that conserve the Type I error rate and maximize statistical power[1]. We have also demonstrated that such a model is better at conserving the Type I error rate and has higher power compared to certain summary methods (i.e. regional averages, endpoint and slope analysis) which have previously been used to analyze longitudinal imaging studies[2, 3, 4].

The goal of this paper is to demonstrate how such a model would be applied to real-world longitudinal imaging data. The demonstration includes how a separable correlation structure can be chosen and evaluated, what kinds of inferences can be made using this model, and how the ever-present concern of missing data can be addressed.

The data used in this example analysis was first reported in the paper by Ahmed et al. (2012) [3]. This study was a randomized controlled phase IIb trial for the use of Toprol, a beta-blocker, in the treatment of patients with chronic degenerative mitral regurgitation. This approach of beta-blockers is done due to evidence of an elevated adrenergic response in MR patients[5] and how hyperactivation of the β-adrenergic pathway leads to a decrease

in viability in myocardial cells[6], possibly via oxidative stress[7]. Supported by promising results in canine models[8], the intent was that treatment with β-blockers would prevent the stress and subsequent decompensation typically seen in MR patients.

Our example will consider the radius of curvature-to-wall thickness ratio (R/T ratio) as an outcome measured repeatedly at multiple spatial locations. The R/T ratio is meant to be a measure of the sphericity of the left ventricle, which is relevant to mitral regurgitation as part of the natural progression of the disease is the left ventricle becoming more spherical rather than the healthy 'bullet' shape. The R/T ratio had been previously considered but only at the level of a global average or level-based averaged analyzed separately[4], while this paper considers it for multiple segments with spatial correlation explicitly modeled. Thus, we shall consider whether the treatment has an effect on the R/T ratio over time, such that a clinically important finding would be that it significantly reduces the increase in sphericity over time relative to placebo.

STATISTICAL MODEL

In order to analyze the longitudinal imaging data, we look to use a linear model with a separable parametric correlation structure. The theoretical and technical details have been described in our previous work[1, 2], but it is worth breaking down the different parts of the model. The base is a *linear model*, where the mean response is a sum of predictor variables and their estimated coefficients (such as the commonly used simple linear and logistic regression models). This structure allows for the use of subject-specific predictors (i.e age, sex, race), time-varying predictors (i.e. systolic blood pressure at each visit), space-varying predictors (such as what part of the heart an outcome is from), and linear (or higher order) trends over time or space. We assume a multivariate normal distribution with constant variance for a subject's responses.

The other parts refer to how the spatial and temporal correlation is modeled. Unlike

simple linear regression, where all of the observations are independent and uncorrelated, our model directly quantifies the correlation between the outcome measurements. The *separable* nature of the correlation structure means that the spatial and temporal correlation can be handled separately so that how correlation changes over space does not influence how it changes over time. The assumption of separability not only makes interpretation easy but it makes the math simpler as well: the correlation between two observations is the product of their spatial and temporal correlations.

Parametric refers to how the correlation over space or time is modeled as a parametric function. For example, an exponential or autoregressive-1 function implies that the correlation decreases exponentially at a constant rate over space or time, respectively. The benefit of parametric functions is that they allow for the modeling of correlation between a large number of observations while only estimating a small number of parameters. This is in contrast to an unstructured correlation model which has no functional form and increases with the square of the number of locations; 16 spatial locations results in an unstructured matrix requiring the estimation of 120 parameters which may be inefficient (if not impossible) for an imaging study where the number of subjects is often limited. If the number of repeated measures is small, such as for a limited number of follow-up visits, then the unstructured correlation model may be viable.

Estimation of our model is best done using restricted maximum likelihood estimation (REML), as discussed previously[1]. Unfortunately, the estimation of a linear model with the kind of separable parametric correlation structure we proposed is not currently supported by common statistical software such as SAS (v9.4)[9] or R (lme4 package v1.1-7)[10]. Therefore the estimation would need to be done by hand-coding an estimation algorithm or by utilizing specialized commercial statistical software such as ASReml[11].

One challenge in implementing this model is the selection of parametric functions for the correlation structure. In some ways this can be seen as needing to be done twice

since a spatial function and a temporal function must be chosen, but it is best to examine many potential combinations concurrently. In practice, it may be necessary to fit a large number of combinations to be comfortable that one models the correlation well. Numerous resources exist that define spatial[12] and temporal[13, 14, 15] correlation functions. It is often but not always the case that functions with a larger number of parameters offer a better fit to the observed correlation. In general, one wishes to balance the number of parameters in the correlation structure with the goodness-of-fit of the model. To this end, we found that, in the scenarios we considered, information criteria were highly (over 90% in some cases) accurate at choosing the true correlation structure, and that at the least they reliably choose a structure that will conserve the Type I error rate and maximize power when given a sufficient number of structures to choose from[1]. Information criteria are simple to use, as each structure is given a score and the one with the smallest score is the 'best' model. Although there are a large number of information criteria that one could use, the most popular two (AIC and BIC) appear in most statistical programs; note that of the two the BIC may be much more accurate at correlation structure selection than the AIC[1].

CLINICAL TRIAL DESIGN AND DATA STRUCTURE

The data used in this paper comes from the UAB SCCOR (Specialized Centers of Clinically Oriented Research) study, specifically Project 1-Aim 1. In order for MR patients to be eligible they had to have moderate to severe mitral regurgitation characterized by mitral valve prolapse and thickening of its leaflets (assessed by an echocardiograph), left ventricular end-systolic dimension under 40mm, and left ventricular ejection fraction over 55%. Exclusion criteria included heart failure, prior myocardial infarction, coronary artery disease, kidney failure (assessed via creatinine levels), hypertension, and other valvular disorders. In other words, the patient cohort had normal cardiovascular health with the exception of having advanced mitral regurgitation. Patient allocation began with 19 in the treatment group and 19 in the placebo group. The baseline characteristics of the cohort

are detailed in Ahmed et al. (2012) where the study was first reported, but a brief version is that the two treatment groups were balanced in size and did not significantly differ in demographics (age, sex, race) or baseline MRI-derived outcomes (end-diastolic volume, ejection fraction, peak early filling rate) and physical exam findings (blood pressure, pulse, New York Heart Association class).

After randomization, patients were dosed daily with either Toprol XL (a β_1-adrenergic receptor blocker) or placebo and followed for two years. Per protocol, cardiac MRI scans were taken at baseline and every six months after, giving cardiac imaging data for five discrete time points. The 3D MRI scans gave information regarding the geometry and structure of the myocardium, and employed tissue tagging and harmonic phase analysis to quantify functional parameters such as wall stress and maximal strain. The image was then mapped to the standard 17-segment AHA model[16], where structural and functional outcomes were taken by averaging over each segment. Segment 17 at the tip of the apex was excluded, giving us outcomes measured at 16 spatial locations from a given imaging session. The 16 segment model was fit to a unit circle with the intersegment distances being the Euclidian distance between the centroids of each segment, shown in Figure 1 and quantified in Table 1. The figure also denotes the levels and which coronary arteries feed each segment. Considering both spatial and temporal points, in the complete case each subject had 80 observations (16 spatial observations at each of 5 time points). The mean R/T ratio for each group at each segment and time point is given in Figure 2; note that this figure is meant to be descriptive of the time courses for each segment and that the error bars should not be used for inference as doing 80 simultaneous correlated tests without correction is statistically unsound.

In this analysis, we initially considered the 38 randomized subjects who had longitudinal imaging data: 29 subjects had complete data with the other 9 subjects having missed one or more follow-up visits (3 on Toprol, 2 on placebo) or attended but were miss-

ing some MRI data (2 Toprol, 2 placebo). A total of 9 follow-up visits were missed for a loss-to-follow-up rate of 5%. In addition, 8 subjects (6 placebo, 2 Toprol) had their mitral regurgitation progress far enough during the study to make surgical intervention necessary; two of the placebo group underwent surgery immediately following randomization but had all five visits recorded as part of a separate arm of the SCCOR study. The initial analysis was performed using an intent-to-treat design that included all of the observations from subjects undergoing surgery (including the two who had immediate surgery), but a secondary sensitivity analysis was performed with those subjects' post-surgery observations excluded from the analysis as the patients were removed from Toprol or placebo after surgery. Along the lines of intent-to-treat, the planned visit times were used in the analysis as the time for the subjects' visits.

STATISTICAL ANALYSIS

One possible measure of the left ventricular remodeling and associated increase in sphericity is the radius of curvature-to-wall thickness (R/T) ratio. Previous work has shown that in healthy mammalian hearts, the R/T ratio is approximately constant from the base to the apex of the left ventricle[18]. Thus, the R/T ratio may be a better indicator of departure from the normal ventricular structure as it does not vary between segments to the extent of wall thickness or radius of curvature alone.

Linear Model with a Separable Parametric Correlation Structure

In order to compare statistical methodologies, we analyzed the SCCOR data with our proposed model as well as with a battery of summary measures. Our linear model has the form

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{1}$$

where $\mathbf{X}_i\boldsymbol{\beta}$ are predictor variables and their associated parameters and $\mathbf{Y}_i$ are the observed

end-diastolic R/T ratios for subject i. In our application, we are looking to fit the model

$$\begin{aligned}\mathbf{X}_i\boldsymbol{\beta} = \beta_0 &+ \beta_1 Sex_i + \beta_2 Group_i + \beta_3 Time_{ij} + (\beta_4 Mid_k + \beta_5 Apex_k) \\ &+ (\beta_6 RCA_k + \beta_7 LCX_k) + \beta_8 Time_{ij} Sex_i + \beta_9 Time_{ij} * Group_i \\ &+ (\beta_{10} Time_{ij} * Mid_k + \beta_{11} Time_{ij} * Apex_k) \\ &+ (\beta_{12} Time_{ij} * RCA_k + \beta_{13} Time_{ij} * LCX_k) \end{aligned} \quad (2)$$

where $Time_{ij}$ was the time of subject i's j^{th} observation; Sex_i was an indicator variable for whether the subject was male(=1) or female (=0); $Group_i$ was an indicator variable for the treatment group for subject i, be it placebo (=0) or beta-blocker (=1); Mid_k and $Apex_k$ were indicator variables for whether the ijk^{th} observation was from the mid or apex of the left ventricle, respectively, with the base as the reference group; RCA_k and LCX_k were indicator variables for whether the ijk^{th} observation was from the a segment supplied by the right coronary artery (RCA) or the left circumflex(LCX), respectively, with the left anterior descending (LAD) as the reference group. Note that the coronary artery designation should *not* be interpreted as cardiac perfusion necessarily being related to ventricular remodeling in mitral regurgitation patients; we simply use it as a convenient way to spatially subdivide the left ventricle circumferentially as the six segments in the base and mid do not neatly subdivide into non-overlapping anterior/inferior and lateral/septal groups. Thus, dividing the segments into three circumferential regions allows us to test differences between the sides, as the radius and wall thickness are known to vary circumferentially. The parentheses denote terms corresponding to the level of the left ventricle or the side of the heart, such that inferences made about the terms inside are done together in a two degree of freedom test. A quadratic time effect was considered but dropped as it was not found to be significant.

Our model also assumes that $\boldsymbol{\epsilon}_i$ follows a multivariate normal distribution with mean zero and a separable parametric correlation structure. For the correlation structure,

we considered all twelve combinations of three spatial correlation functions (exponential, spherical, and Matérn) crossed with four temporal correlation functions (compound symmetric[CS], autoregressive-1[AR-1], Toeplitz, and unstructured[UN]). We chose a working correlation structure via BIC with a sample size adjustment of the total number of subjects in the dataset, which we found to be reliably accurate at choosing the true correlation structure[1]. To confirm our choice, we also plotted the estimated spatial and temporal correlation functions versus the observed correlation between pairs of observations. To calculate the 120 unstructured spatial correlation parameters, we fit the model on pairs of segments from all subjects at all observed time points with an unstructured temporal correlation model.

Although the correlation structure is very important, the assumption of normality is also essential to check. Despite the correlation, when looked at individually the residuals in our model are assumed to have a normal distribution with mean 0 and variance σ^2. Unfortunately, initial diagnostics suggest that the end-diastolic R/T ratio does not follow a normal distribution and is in fact skewed (Figure 3). The histogram suggests that the R/T ratio is skewed upwards, a common concern in ratios when the denominator gets small but appeared to be corrected by a log transform. The lack of normality is easily seen in the QQ plot of the R/T ratio, which has a distinct curved pattern that does not follow along the line. The QQ plot also implies that the log transform 'fixed' the skewness, as it closely follows a straight line indicating a normal distribution is a good fit.

The linear models were fit using the ASReml-R package (v. 3.0, VSN International, Hemel Hempstead, UK)[11]. The model with the chosen covariance structure then had its fixed effects tested with a conditional F-test with a Kenward-Roger adjustment for the denominator degrees of freedom at an α-level of 0.05[19]. The corrected F-test is generally considered to be better to use than a simple Wald's test when there are correlation parameters in the model and the sample size is small; this assertion supported by simulation

studies which found the corrected F-test to conserve the Type I error rate better without a meaningful loss in power for sample sizes comparable to typical longitudinal imaging studies[2].

Summary Methods

We also implemented various summary measures in space and time in order to demonstrate the benefits of our model compared to previous methods for analyzing longitudinal imaging data. The spatial summary measures considered included a global average of all 16 segments and averages with the levels (base, mid, and apex) analyzed jointly resulting in one and three spatial observations per subject per time, respectively. The temporal summary measures used were slope and endpoint analysis. Endpoint analysis considered the change between the baseline and two-year observations, and was coded as missing if either of those two observations were missing. Slope analysis considered the slope estimated via simple linear regression on the relevant spatial unit (segment, level, or ventricle) and was only missing if the subject had only one visit. Note that when the temporal summary measures were used, the fixed effects relate whether a predictor (such as treatment group) affects the overall change over time or estimated slope in endpoint and slope analysis, respectively. All combinations of these summary measures along with the direct modeling of spatial and temporal correlation led to a total of nine models to compare. When one of the models needed a parametric correlation function chosen, the BIC was used. Convergence of the model estimation was only an issue for the level averages as the inter-level correlations were near 1; these problems were overcome by setting initial values of the correlation parameters to be the Pearson correlation estimates between levels.

The fixed effects used in the nine models are detailed in Table 2, and were chosen for uniformity in what the models controlled for. All of the models allowed for the use of subject-level predictors such as the treatment group and the subject's sex. Sex was chosen as it may relate to the geometry of the heart through a subject's size or shape. The

summary measures may restrict the use of other predictors, however. The level averages prevent the use of space-varying covariates besides those identifiable at the subject level; in our example the side of the heart cannot be modeled when the level averages are used. The temporal summary methods similarly prevent the use of time-varying covariates or even the use of the subject's exact visit time as a predictor. The exact times of observation are used in the calculation of each subject's slope in slope analysis, a summary measure previously found to be highly competitive with the temporal correlation model when the data is truly linear[2], but not in endpoint analysis.

RESULTS

Linear Model with a Separable Parametric Spatiotemporal Covariance Structure

Let us first focus on the application of our previously proposed spatiotemporal model. The first step when using it is to choose appropriate spatial and temporal correlation functions. Using the BIC to pick a model, we found that a Matérn-by-unstructured correlation model provided the 'best' balance of goodness-of-fit and simplicity for the observed data (Table 3). This was the most complex correlation structure we considered, so it warranted investigation of how well the model truly fit the data.

Figure 4 shows the estimated correlation functions along with the unstructured correlation estimates between time points and segments. The unstructured temporal correlation provides the best possible fit, and we can see that the simpler parametric functions do not necessarily fit the observed correlation very well. We observed that the unstructured spatial correlation seemed to have a random scatter when plotted versus the distance between segments, which would make parametric modeling extremely difficult. However, of the three functions considered it did seem that Matérn gave the best fit, as the others more severely underestimated the true correlation between far apart segments. For completeness, a compound symmetric model was tried but its associated BIC (-9313.348) was not supe-

rior to a Matérn function, suggesting there is indeed a slight downward trend of correlation over distance. Note that this approach of graphically examining the fits of the estimated correlation functions can also be highly useful when multiple information criteria provide conflicting answers for which model is 'best.'

After the correlation structure has been chosen and the assumption of normality has been checked, the next step is to review the estimates and inferences about the predictors in the model. The estimates of the parameters given in Equation 2 are given in Table 4 along with their associated test statistics and p-values. As mentioned, we used a F-test with a Kenward-Roger correction for the denominator degrees of freedom; the effect of the correction can be seen in the reported test statistics as the denominators are not integers. We have seen in previous work that the corrected F-test is more reliable at conserving the Type I error rate than the simpler Wald's test when analyzing highly correlated data such as in longitudinal imaging studies[2].

The parameter of interest, the treatment-by-time interaction β_9, was not significantly different from zero (p=0.2073). The linear trend over time was also not significant (p=0.1365) which suggests that the R/T ratio is simply not changing very much over the twenty four months of observation. The effect of sex was strongly significant ($p< 0.0001$), with men having slightly smaller R/T ratios than women; this is likely due to men being larger and having larger hearts such that the larger wall thickness overrides the larger radius of curvature. The level of the left ventricle was also highly significant ($p< 0.0001$) with the R/T ratio increasing from the base down to the apex. We expected to see this trend as it matches what is seen in the natural progression of left ventricular remodeling due to mitral regurgitation. The side of the left ventricle was also significantly associated with the R/T ratio ($p< 0.0001$) such that the lateral side (LCX) had the highest ratio, followed by the anterior/septal side (LAD), with the smallest R/T ratio on the inferior/septal side (RCA) It is possible that this trend is due to the balancing forces from other chambers of

the heart, especially the right ventricle opposite the septum, that are lacking in the lateral side which only has the pericardium restraining the myocardium. Although the pericardium is not elastic and typically acts to prevent ballooning of the left ventricle, it can weaken and stretch over time with chronic pressure from an overloaded left ventricle. However, there is a natural difference in the R/T ratio between the septal and lateral ventricular walls seen in healthy patients that may be the source of this statistically significant difference. Assuming the R/T ratio is a valid measure of sphericity, this may suggest that the highest sphericity in mitral regurgitation patients is in their left ventricles' lateral and apical region. These results can be seen in Figure 2.

The sensitivity of the intent-to-treat analysis strategy was considered by a secondary analysis where all observations taken after surgery were excluded from the dataset. Eight subjects had surgery, two from the medical therapy group and six from the placebo group. Two subjects from the placebo group had surgery after randomization but before the baseline MRI scan. The results of the analysis did not noticeably change, as the BIC still chose a Matérn-by-unstructured model (Table 3) and none of the inferences changed at a $\alpha = 0.05$ level (Table 4). The estimates themselves changed slightly, suggesting that surgery may not be independent of the R/T ratio.

Summary Methods

Now that we have considered the results of applying our proposed model to the UAB SCCOR data, it is of interest to compare it to the results from summary methods commonly used in longitudinal imaging data analysis. For the sake of brevity, we shall only consider inference upon the effect of medical therapy on the time course of the end-diastolic R/T ratio, log-transformed to correct for skewness. The test statistics for the corrected F-test and associated p-values for the nine models are given in Table 5 along with what correlation structure was chosen (if applicable) via BIC to fit the data.

The first thing to note is that our model that used a spatiotemporal correlation structure had the second smallest p-value and test statistic, passed only by slope analysis using all 16 segments, although as mentioned before it was not significant at a α-level of 0.05. We can also note that the denominator degrees of freedom are the largest for our model, and much smaller for the summary methods. In every case the corrected degrees of freedom are smaller than the number of observations used in the analysis, although the difference varies with the summary measure used. The spatial summary measures only slightly reduced the degrees of freedom despite a large reduction in the number of observations, while the temporal summary measures reduced the degrees of freedom to around the number of independent subjects. This reduction in the degrees of freedom is expected in a Kenward-Roger adjustment, and reflects how the information is condensed and possibly lost. It may also reflect how despite summary measures simplifying the dimension of correlation among the observations, they may increase the correlation between the remaining measures. For example, the correlations between the level averages were around 0.9, which suggests that the information from those three observations is far less than three independent observations and closer to a single observation.

Comparing the different methods, it seems that the use of spatial correlation estimated a larger treatment-by-time effect than spatial summary measures, despite the target of inference not having a spatial component. On the temporal side, slope analysis estimated a larger treatment effect than a temporal correlation model resulting in a smaller p-value, though the two models are fairly similar which is not surprising based on the results of simulation studies[2] and how both models assume a linear time course. There are many possible reasons the endpoint analysis lagged behind slightly: endpoint analysis induces more missing observations, it uses fewer temporal observations than slope analysis and may have a higher variance[2], and it measures changes over time in a slightly different way. Regardless of the reason, it seems that a temporal correlation model or slope analysis are preferable for longitudinal data when a linear time course is assumed. Endpoint anal-

ysis needs a valid scientific reason to be preferable to the other methods, such as a time course that is known to be non-linear but a total change over time is of interest.

The results of the summary method comparison were slightly different when the post-surgery observations were excluded, as seen in Table 6. As mentioned above, there were four from the placebo group and two from the treatment group who underwent surgery before the end of the two-year follow-up. The spatial comparison did not change, as the use of a spatial correlation model still provided greater estimates and smaller p-values than spatial summary measures. The temporal methods changed, as they were not equally affected by the loss of data. The temporal correlation approach lost the relevant observations but still retained the six subjects' pre-surgery data. The nature of the cutoff meant that endpoint analysis had to count the eight as missing, dropping the number of included subjects from 36 to a mere 28. On the other hand, slope analysis was mostly unimpeded by the exclusion as it lost only the subjects who had only post-surgery observations, although the slopes from the other six subjects certainly changed.

Of the nine methods, all still had p-values greater than 0.05. The temporal correlation model had slightly smaller estimates and only a small loss in the denominator degrees of freedom. Slope analysis had much larger estimates and essentially the same degrees of freedom, resulting in smaller p-values. However, endpoint analysis had much smaller estimates and noticeably reduced degrees of freedom which led to much larger p-values. Some of these changes could be due to how the methods handle missing data, but there is also the concern regarding this type of missingness. Since the excluded observations are from subjects who underwent valve repair surgery and surgery is only done on patients whose mitral regurgitation has progressed far enough, then any measure of disease progression (such as the R/T ratio) could not be missing completely at random if post-surgery observations are excluded. The missingness also did not affect the two groups equally; six versus two may seem trivial, but when the original group sizes were nineteen and the missingness

is not completely at random it is plausible that it could affect inference to the extent we have observed.

It should also be noted that although in this study the qualitative results (all $p > 0.05$) did not change between the methods, the numbers themselves did. It would certainly be possible for significance to change between the methods when they are applied to a different dataset. As always, scientific justification should be used to choose a method instead of 'cherry picking' the one that gives the most favorable p-value.

DISCUSSION

In this paper we have described how a linear model with a separable parametric correlation structure could be used in practice, and have illustrated the method using data from a longitudinal imaging study. Only general guidelines can be given, as each application has its own nuances. A general strategy for implementing our proposed model on spatiotemporal data could be considered as such:

1. Decide on all of the predictors in the analysis that would be of interest to scientific investigators.

2. Decide on a number of spatial and temporal correlation functions to try to fit to the data. Functions with different properties should be considered, such as different shapes and a mixture of simpler functions and more complex functions which may have greater flexibility. This step should also be a collaboration between statistician and investigator, as the functions should be able to model the correlation behavior expected by prior scientific knowledge.

3. Fit linear models with all of those predictors included in the fixed effects for a wide variety of combinations of spatial and temporal correlation functions. If the number of combinations considered is too small, it is possible that none of them will model the correlation sufficiently well.

4. Choose between models using information criteria. One should also compare the estimated correlation to the observed correlation; graphical methods are highly useful to assess goodness-of-fit. If none of the fitted structures seem appropriate, additional approaches to modeling the covariance should be considered.

5. Perform inferences upon the fixed effects using the model with the chosen correlation structure.

6. If some predictors are not significant, a more parsimonious model can be obtained with backwards selection. Note that the above steps for choosing a correlation structure must be repeated for each new set of fixed effects.

The greatest challenge to this approach is finding a correlation structure that fits the data well. There has been an immense amount of work done to define valid parametric correlation functions (too many to list exhaustively here) so one option would be to simply try more structures, such as the flexible linear exponent autoregressive (LEAR) function[20]. This may require statistical programming to augment or develop the model estimation software if the desired functions are not already supported. Another point that should be considered is the assumption of separability; if there is an interaction between spatial and temporal correlation then no pairing of separate functions will properly model the true correlation. Much work has been done to test this assumption of separability, but a good starting point would be recent likelihood ratio test proposed by Simpson et al. that was designed with longitudinal imaging studies in mind[21]. More statistical research needs to be done to determine how sensitive a model like ours is to violations of separability and what nonseparable methods are appropriate to use in our given application. The assumption of multivariate normality is also highly important and should be checked; deviations can possibly be helped by a transformation to the outcome values.

One option would be to consider summary methods, but as we have seen they can produce mixed results. We have seen that spatial summary methods may be a poor choice in

practice even if the predictor of interest is not space-varying. Previous work has shown that the common practice of analyzing multiple spatial summary measures separately has very poor statistical properties and should be avoided[2]. It should be noted that the AHA's model is a case of a spatial summary measure, where the thousands of voxels from a three-dimensional image are mapped to the 16 segments of the heart[16]. One justification for this approach is that many of the functional and geometric outcomes (such as wall thickness) are not voxel-specific and in fact require several voxels to calculate; since some aggregation is required, it is reasonable to aggregate them to a manageable number of subregions while still maintaining lateral and circumferential resolution of the ventricle. Some outcomes, such as perfusion and oxygen consumption in fMRI studies, do have voxel-level outcomes which allows for greater complexity in the analysis of the imaging data. It is up to the statistician and scientist to jointly decide whether the voxel-level data should be used in these studies, or if the loss of resolution is worth the gain in interpretability by summarizing over cerebral structures while still controlling for spatial correlation.

Temporal summation may be valid when a change over time is of interest, but it is unknown how they fare when drawing inference about other types of predictors. Slope analysis may provide good power for testing changes over time and can be resistant to uneven follow-up times or data missing completely at random. However, it assumes a linear time course and precludes the use of time-varying covariates and the ability to test for anything not related to a linear interaction with time. Endpoint analysis is worse, as it is extremely sensitive to missing data or uneven follow-up times along with the same limitations on predictors as slope analysis. In general, it should only be used when a total change over time is of interest and the time course is non-linear.

Another issue to consider when doing the analysis is how to handle the imperfections of real data. Missing values can be handled reasonably well by using a submatrix of the full correlation structure that pertains to each subject's observed outcomes. Uneven

follow-up times are more difficult, as they preclude the use of many temporal correlation functions that assume there are a finite number of evenly spaced observations. One option (which we have used in this paper) is to use the planned observation times; this allows the use of an unstructured temporal correlation model which is desirable but does involve ignoring information that was collected. One possibility would be to utilize correlation functions that were traditionally considered to be spatial, using the true observation time as the distance, but such an approach needs statistical validation before it can be recommended.

ACKNOWLEDGEMENTS

We wish to thank Drs. Louis Dell'Italia, Tom Denney, Jr., and Himanshu Gupta of the UAB SCCOR study for their support and for the cardiac MRI data they provided. Predoctoral funding was provided by NHLBI T32HL079888. The UAB SCCOR study was supported by National Institutes of Health Specialized Center of Clinically Oriented Research in Cardiac Dysfunction P50-HL077100.

REFERENCES

[1] George, B., Aban, I. (2014). Selecting a Separable Parametric Spatiotemporal Covariance Structure for Longitudinal Imaging Data. Accepted by *Statistics in Medicine*.

[2] George, B., Aban, I. (2014). Comparing Summary Methods and a Spatiotemporal Model in the Analysis of Longitudinal Imaging Data. Manuscript in preparation.

[3] Ahmed, M.I., Aban, I., Lloyd, S.G., Gupta, H., Howard, G., Inusah, S., ... Dell'Italia, L.J. (2012). A Randomized Controlled Phase IIb Trial of Beta-1-Receptor Blockade for Chronic Degenerative Mitral Regurgitation. *Journal of the American College of Cardiology* 60 (9): 833-838.

[4] Schiros, C.G., Dell'Italia, L.J., Gladden, J.D., Clark, D., Aban, I., ... Ahmed, M.I. (2012). Magnetic resonance imaging with 3-dimensional analysis of left ventricular remodeling in isolated mitral regurgitation: implications beyond dimensions. *Circulation* 125 (19): 2334-2342.

[5] Nagatsu, M., Zile, M.R., Tsutsui, H., Schmid, P.S., DeFreyte, D., Cooper, G., Carabello, B.A. (1994). Native β-Adrenergic Support for Left Ventricular Dysfunction in Experimental Mitral Regurgitation Normalizes Indexes of Pump and Contractile Function. *Circulation*, 89(2), 818-826.

[6] Mann, D.L., Kend, R.L., Parsons, B., Cooper G. (1992). Adrenergic effects on the biology of the adult mammalian cardiocyte. *Circulation*, 85, 790-804.

[7] Ahmed, M.I., Gladden, J.D., Litovsky, S.H., Lloyd, S.G., Gupta, H., Inusah, S., Denny Jr., T., Powell, P., McGiffin, D.C., Dell'Italia, L.J. (2010). Increased oxidative stress and cardiomyocyte myofibrillar degeneration in patients with chronic isolated mitral regurgitation and ejection fraction $> 60\%$. *Journal of the American College of Cardiology*, 55(7), 671-679.

[8] Tsutsui, H., Spinale, F.G., Nagatsu, M., Schmid, P.S., Ishihara, K., DeFreyte, G., Cooper, G., Carabello, B.A. (1994). Effects of Chronic β-Adrenergic Blockade on the Left Ventricular and Cardiocyte Abnormalities of Chronic Canine Mitral Regurgitation. *The Journal of Clinical Investigation, Inc.*, 93, 2639-2648.

[9] SAS Institute Inc. *SAS/STAT 13.2 User's Guide*. SAS Institute Inc.: Cary, NC, 2014.

[10] Bates, D., Maechler, M., Bolker, B., Walker, S. (2014). Package 'lme4' [Software]. Available from http://lme4.r-forge.r-project.org/

[11] Butler, D., Cullis, B.R., Gilmour, A.R., Gogel, B.J. (2007). ASReml-R reference manual (Release 2.00) [Software]. Available from http://www.vsni.co.uk/software/asreml

[12] Waller, L.A., Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley and Sons.

[13] Littell, R.C., Pendergast, J., Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 19: 1793-1819.

[14] Schaalje, B., McBride, J.B., Fellingham, G.W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics* 7 (4): 512-524.

[15] Vallejo, G., Fernández, M.P., Livacic-Rojas, P.E., Tuero-Herrero, E. (2011). Selecting the best unbalanced repeated measures model. *Behavior Research Methods* 43 (1): 18-36.

[16] Cerqueria, M.D., Weissman, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., ... Verani, M.S. (2002). Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart: A Statement for Healthcare Professionals From the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. *Circulation* 105: 539-542.

[17] Seals, S. (2013) Spatial analysis of cardiovascular MRI data (dissertation). Birmingham, AL: University of Alabama at Birmingham.

[18] Beyar, R., Weiss, J.L., Shapiro, E.P., Graves, W.L., Rogers, W.J., Weisfeldt, M.L. (1993). Small apex-to-base heterogeneity in radius-to-thickness ratio by three-dimensional magnetic resonance imaging. *American Journal of Physiology* 264: H133-H140.

[19] Kenward, M.G., Roger, J.H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics* 53: 983-997.

[20] Simpson, S.L., Edwards, L.J., Muller, K.E., Styner, M.A. (2014). Kronecker Product Exponent AR(1) Correlation Structures for Multivariate Repeated Measures. *PLoS ONE* 9: e88864.

[21] Simpson, SL, Edwards, LJ, Styner, MA, Muller, KE. (2014). Separability tests for high-dimensional, low-sample size multivariate repeated measures data. *Journal of Applied Statistics* DOI: 10.1080/02664763.2014.919251.

Table 1: Spatial coordinates of the 16 segments in the model of the left ventricle[17]. They are denoted as their level (base, mid, apex), orientation (anterior, septal, inferior, lateral), and index number.

Base, Ant. (1)	$(0,\frac{5}{6})$	Mid, Ant. (7)	$(0,\frac{1}{2})$	Apex, Ant. (13)	$(0,\frac{1}{6})$
Base, Ant.Sep. (2)	$(\frac{-5\sqrt{3}}{12},\frac{5}{12})$	Mid, Ant.Sep. (8)	$(\frac{-\sqrt{3}}{4},\frac{1}{4})$	Apex, Sep. (14)	$(\frac{-1}{6},0)$
Base, Inf.Sep. (3)	$(\frac{-5\sqrt{3}}{12},\frac{-5}{12})$	Mid, Inf.Sep. (9)	$(\frac{-\sqrt{3}}{4},\frac{-1}{4})$	Apex, Inf. (15)	$(0,\frac{-1}{6})$
Base, Inf. (4)	$(0,\frac{-5}{6})$	Mid, Inf. (10)	$(0,\frac{-1}{2})$	Apex, Lat. (16)	$(\frac{1}{6},0)$
Base, Inf.Lat. (5)	$(\frac{5\sqrt{3}}{12},\frac{-5}{12})$	Mid, Inf.Lat. (11)	$(\frac{\sqrt{3}}{4},\frac{-1}{4})$		
Base, Ant.Lat (6)	$(\frac{5\sqrt{3}}{12},\frac{5}{12})$	Mid, Ant.Lat (12)	$(\frac{\sqrt{3}}{4},\frac{1}{4})$		

Table 2: Nine linear models made from combinations of spatial and temporal methods, the form of their fitted predictors, and what predictor is the focus of hypothesis testing for whether the treatment (Trt_i) changes the time-course of disease progression.

Spatial Method	Temporal Method	Fitted Model	Tested Predictor
Correlation	Correlation	$E[Y_{ijk}] \sim 1 + Sex_i + Time_{ij} + Trt_i + Mid_k + Apex_k + RCA_k + LCX_k + Time_{ij}Sex_i + Time_{ij}Trt_i + Time_{ij}(Mid_k + Apex_k) + Time_{ij}(RCA_k + LCX_k)$	$Time_{ij}Trt_i$
	Endpoint	$E[Y_{ik}] \sim 1 + Sex_i + Trt_i + Mid_k + Apex_k + RCA_k + LCX_k$	Trt_i
	Slope	$E[Y_{ik}] \sim 1 + Sex_i + Trt_i + Mid_k + Apex_k + RCA_k + LCX_k$	Trt_i
Level Average, Together	Correlation	$E[Y_{ijm}] \sim 1 + Sex_i + Trt_i + Time_{ij} + Mid_m + Apex_m + Time_{ij}Sex_i + Time_{ij}Trt_i + Time_{ij}(Mid_m + Apex_m)$	$Time_{ij}Trt_i$
	Endpoint	$E[Y_{im}] \sim 1 + Sex_i + Trt_i + Mid_m + Apex_m$	Trt_i
	Slope	$E[Y_{im}] \sim 1 + Sex_i + Trt_i + Mid_m + Apex_m$	Trt_i
Global Average	Correlation	$E[Y_{ij}] \sim 1 + Sex_i + Trt_i + Time_{ij} + Time_{ij}Trt_i + Time_{ij}Sex_i$	$Time_{ij}Trt_i$
	Endpoint	$E[Y_i] \sim 1 + Sex_i + Trt_i$	Trt_i
	Slope	$E[Y_i] \sim 1 + Sex_i + Trt_i$	Trt_i

Table 3: Table of the BIC for each of the twelve fitted correlation structures, for the datasets with all observations or with the follow-up visits post-surgery excluded from the model. The smallest value in each column is the chosen model for that given criterion and is denoted in **bold**.

Correlation Structure	Length of θ	BIC	
		All Data	Post-Surgery Excluded
EXP⊗CS	3	-9433.689	-7966.974
SPH⊗CS	3	-9234.837	-7799.790
MAT⊗CS	4	-9505.525	-8027.130
EXP⊗AR-1	3	-9249.962	-7837.511
SPH⊗AR-1	3	-9062.495	-7681.949
MAT⊗AR-1	4	-9305.186	-7880.888
EXP⊗TOE	6	-9467.324	-7994.755
SPH⊗TOE	6	-9269.310	-7828.571
MAT⊗TOE	7	-9537.101	-8052.622
EXP⊗UN	12	-9480.570	-8013.066
SPH⊗UN	12	-9285.138	-7850.161
MAT⊗UN	13	**-9544.087**	**-8063.303**

Table 4: The parameter estimates and associated statistical inference for each of the predictors in the model for the natural log of the R/T ratio with a Matérn-by-unstructured correlation matrix, both with and without post-surgery observations.

	All Data			Post-Surgery Excluded		
Predictor	Parameter Estimate	F	p-Value	Parameter Estimate	F	p-Value
Intercept	1.44647	$F_{1,162} = 4864.00$	< 0.0001	1.41065	$F_{1,145.6} = 4697.00$	< 0.0001
Sex (Male)	-0.18809	$F_{1,162} = 22.30$	< 0.0001	-0.16339	$F_{1,145.8} = 16.22$	< 0.0001
Time	-0.00313	$F_{1,159.9} = 2.24$	0.1365	-0.00225	$F_{1,139.7} = 0.11$	0.7400
Treatment	0.01022	$F_{1,162} = 0.92$	0.3398	0.03049	$F_{1,145.9} = 1.81$	0.1786
Mid	0.15700	$F_{2,532.5} = 152.50$	< 0.0001	0.15847	$F_{2,498} = 152.90$	< 0.0001
Apex	0.20666			0.21457		
RCA	-0.06461	$F_{2,566.6} = 56.62$	< 0.0001	-0.06755	$F_{2,529.5} = 55.21$	< 0.0001
LCX	0.07954			0.08349		
Sex*Time	0.00182	$F_{1,159.9} = 1.12$	0.2917	0.00162	$F_{1,139.6} = 0.76$	0.3850
Treatment*Time	0.00217	$F_{1,159.9} = 1.60$	0.2073	0.00198	$F_{1,139.8} = 1.12$	0.2909
Mid*Time	0.00012	$F_{2,548.1} = 0.07$	0.9337	0.00055	$F_{2,476.8} = 0.60$	0.5485
Apex*Time	-0.00005			0.00039		
RCA*Time	-0.00038	$F_{2,550.5} = 0.19$	0.8277	-0.00015	$F_{2,474.2} = 0.02$	0.9774
LCX*Time	-0.00009			0.00001		

Table 5: Inferences about a treatment effect over time on the log of the end-diastolic R/T ratio from the nine combinations of spatial and temporal methods from the UAB SCCOR study using all observed outcomes.

Spatial Method	Temporal Method	Number of Observations	Fitted Σ Structure	Tested Predictor	Est. Effect per Month	Test Statistic	p-Value
Correlation	Correlation	2894	Matérn⊗UN	$Time_{ij}Group_i$	0.00217	$F_{1,159.9} = 1.60$	0.2073
	Endpoint	574	Matérn	$Group_i$	0.00204	$F_{1,37} = 1.17$	0.2862
	Slope	592	Matérn	$Group_i$	0.00310	$F_{1,35} = 1.83$	0.1848
Level Average, Together	Correlation	543	UN⊗CS	$Time_{ij}Group_i$	0.00147	$F_{1,154.0} = 0.83$	0.3646
	Endpoint	108	UN	$Group_i$	0.00151	$F_{1,32.9} = 0.57$	0.4674
	Slope	111	UN	$Group_i$	0.00205	$F_{1,34} = 0.99$	0.3267
Global Average	Correlation	181	CS	$Time_{ij}Group_i$	0.00148	$F_{1,141.2} = 0.91$	0.3416
	Endpoint	36	N/A	$Group_i$	0.00144	$F_{1,33} = 0.53$	0.4709
	Slope	37	N/A	$Group_i$	0.00215	$F_{1,34} = 1.01$	0.3213

Table 6: Inferences about a treatment effect over time on the log of the end-diastolic R/T ratio from the nine combinations of spatial and temporal methods from the UAB SCCOR study excluding all outcomes observed after surgery.

Spatial Method	Temporal Method	Number of Observations	Fitted Σ Structure	Tested Predictor	Est. Effect per Month	Test Statistic	p-Value
Correlation	Correlation	2558	Matérn⊗UN	$Time_{ij}Group_i$	0.00198	$F_{1,139.8} = 1.12$	0.2909
	Endpoint	446	Matérn	$Group_i$	0.00138	$F_{1,26.1} = 0.40$	0.5326
	Slope	560	Matérn	$Group_i$	0.00416	$F_{1,31.9} = 2.20$	0.1482
Level Average, Together	Correlation	480	UN⊗CS	$Time_{ij}Group_i$	0.00101	$F_{1,134.6} = 0.34$	0.5618
	Endpoint	84	UN	$Group_i$	0.00041	$F_{1,24.9} = 0.03$	0.8589
	Slope	105	UN	$Group_i$	0.00373	$F_{1,31.9} = 2.44$	0.1285
Global Average	Correlation	160	AR-1	$Time_{ij}Group_i$	0.00161	$F_{1,152.9} = 0.36$	0.5488
	Enpdoint	28	N/A	$Group_i$	0.00065	$F_{1,25} = 0.09$	0.7681
	Slope	35	N/A	$Group_i$	0.00385	$F_{1,32} = 2.15$	0.1525

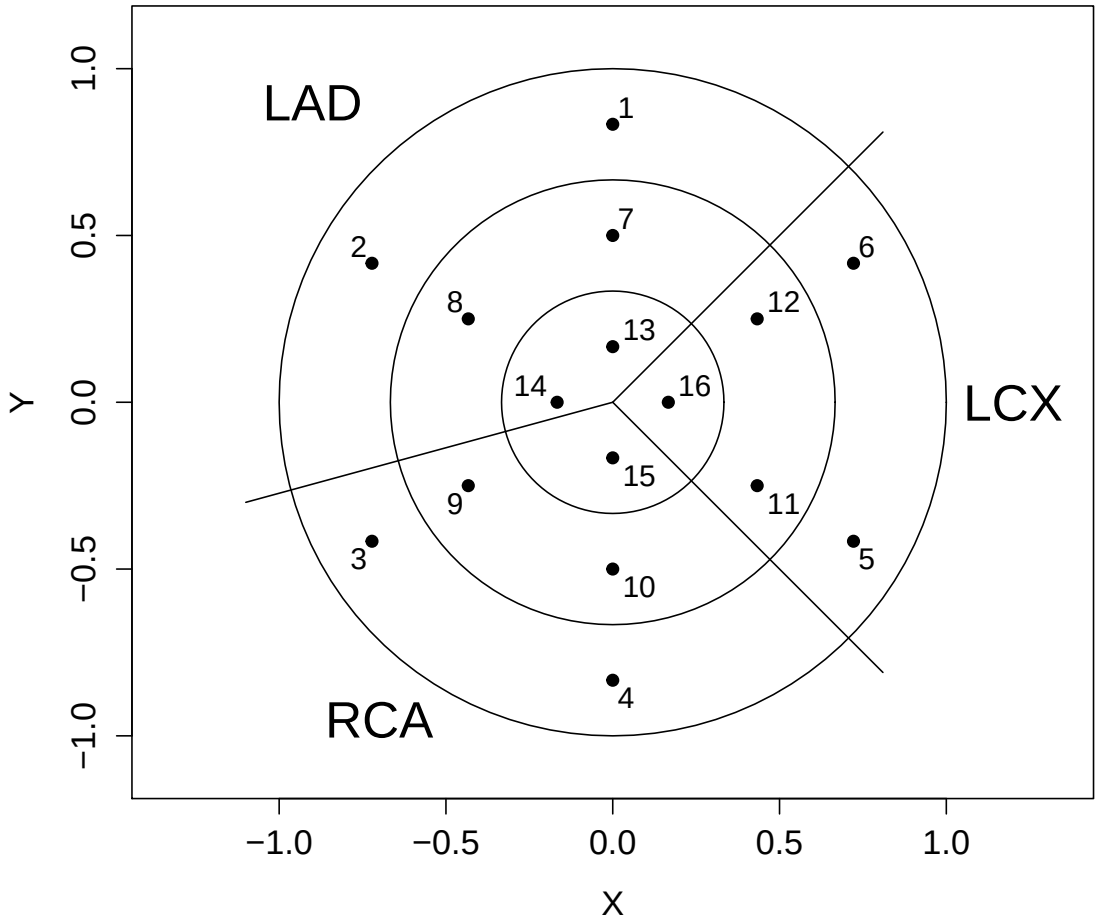

Figure 1: Plot of the 16 segments of the left ventricle. The outer ring corresponds to the base, the middle ring to the mid, and the inner circle to the apex[16, 17]. The right region corresponds to the segments supplied by the left circumflex (LCX), the upper left those supplied by the left anterior descending (LAD), and the bottom left those supplied by the right coronary artery (RCA). The numbers correspond to the segment's index as defined in Table 1.

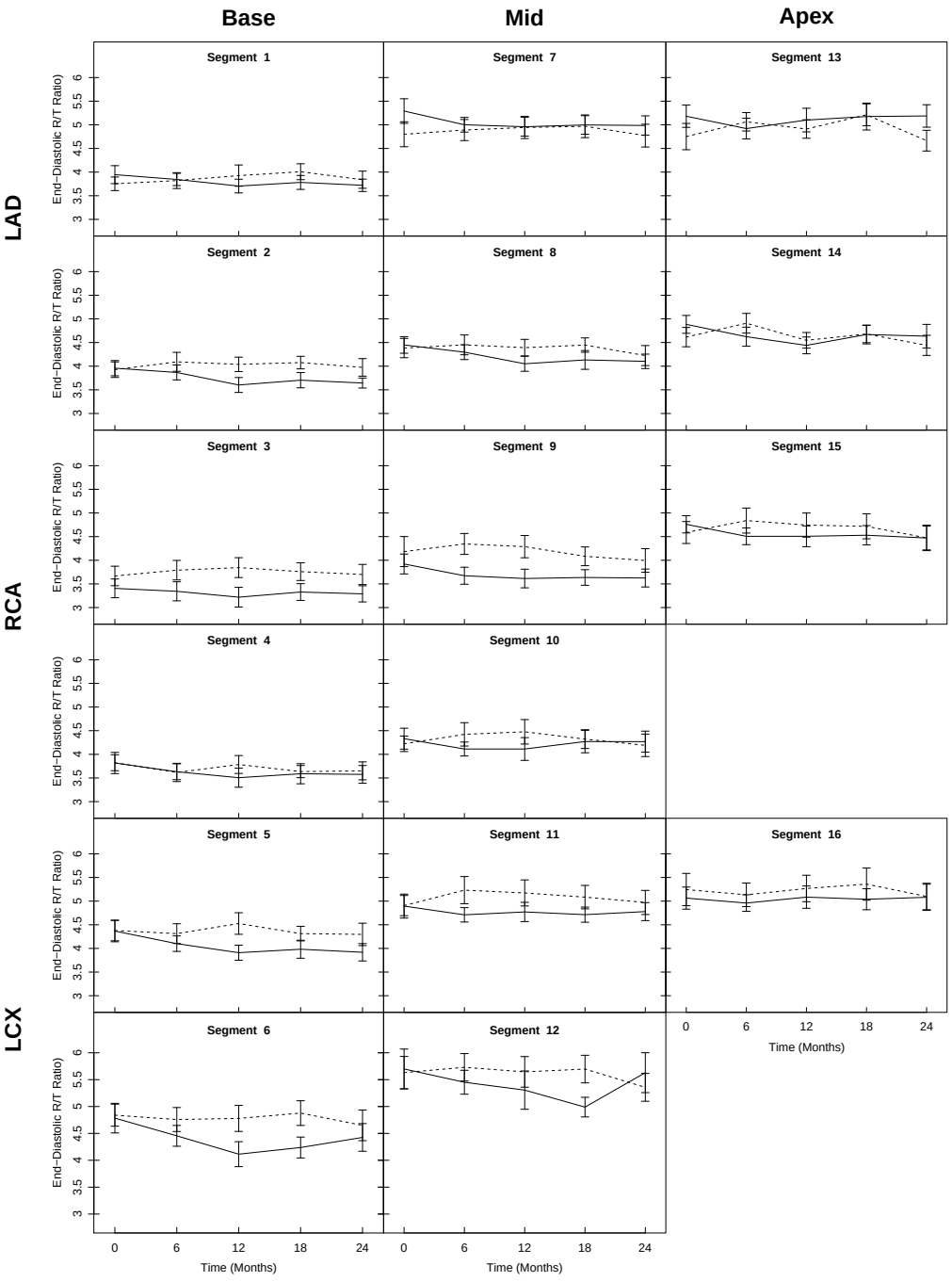

Figure 2: Plot of the time courses of the mean end-diastolic R/T ratios among all subjects in each group at the 16 segments of the left ventricle. The solid lines represent the average time courses of the placebo group, and the dashed line the courses of the treatment group. The error bars demark one standard error of the mean in each direction (total width is two standard errors) and are shown for the purpose of describing the spread of the R/T ratio among subjects and should not be taken as formal inference.

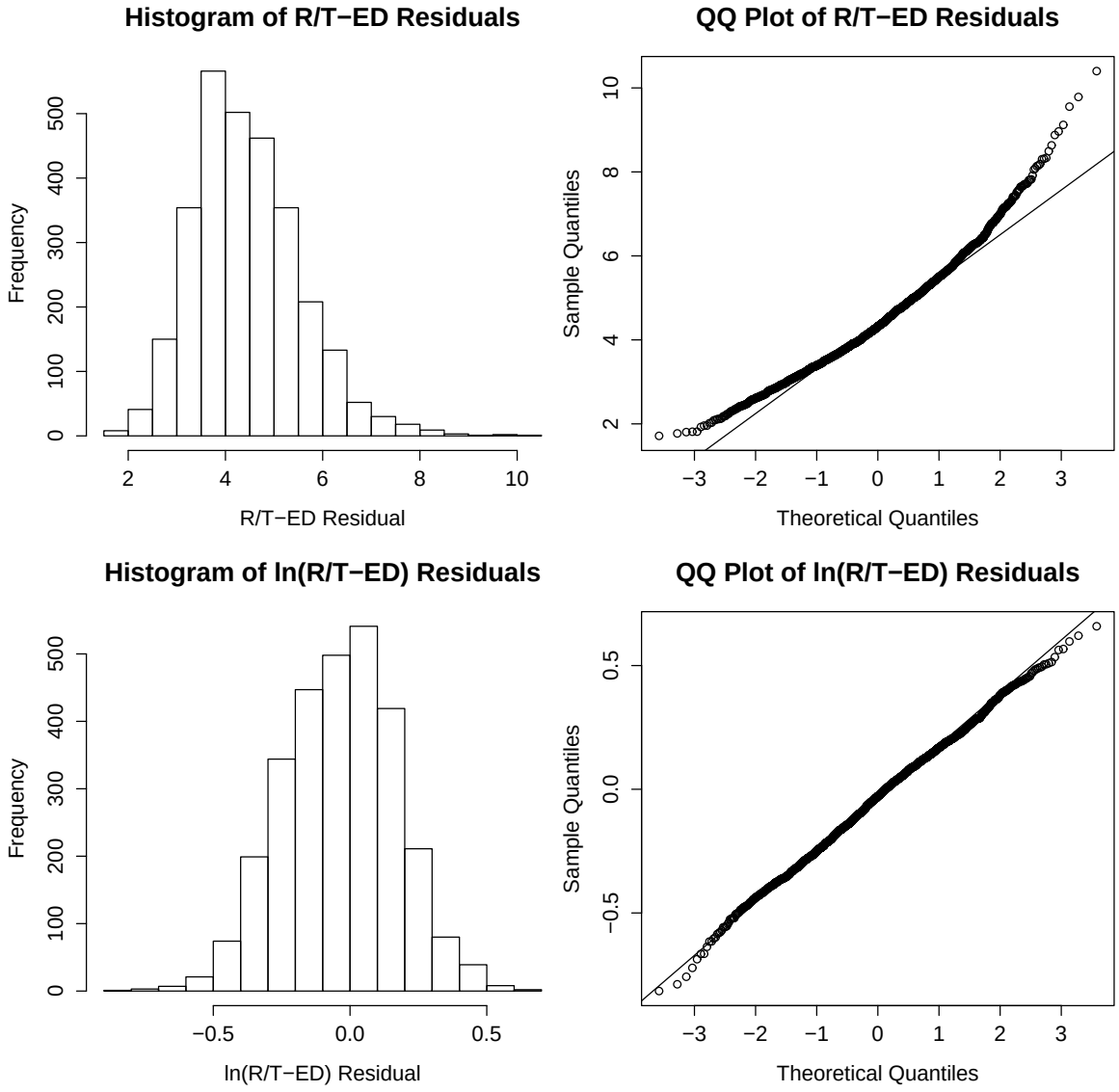

Figure 3: Histogram of the residuals of the end-diastolic R/T ratio and the log of the end-diastolic R/T ratio in the SCCOR study, for a fitted MAT⊗UN correlation structure, across all observations from Equation 2.

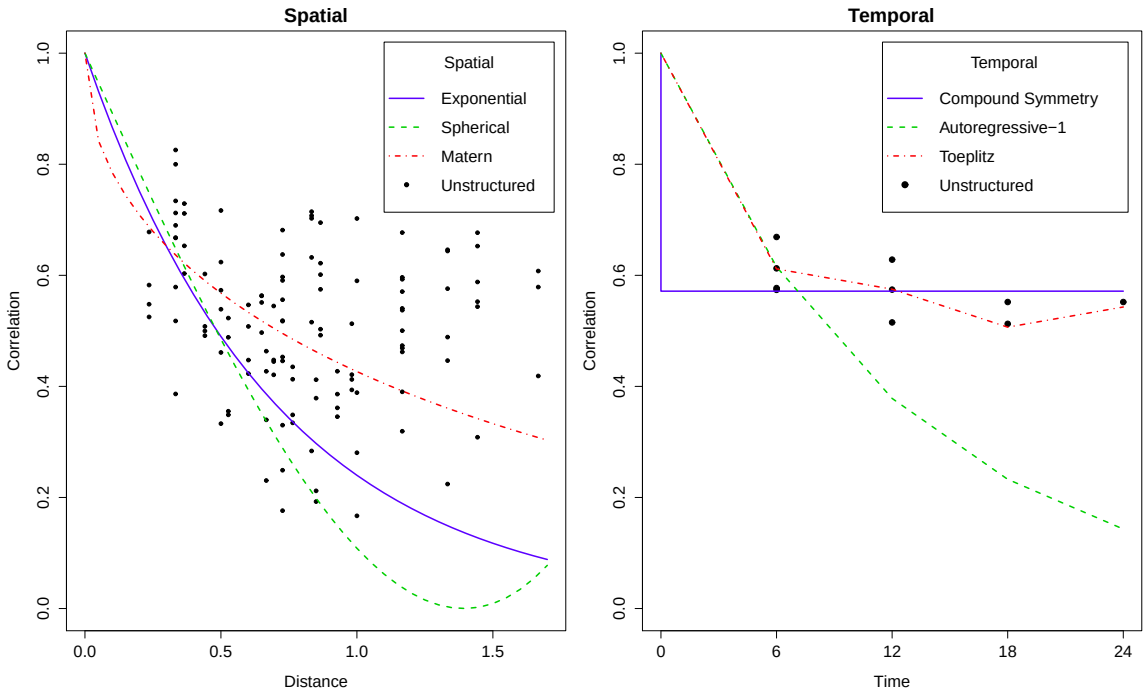

Figure 4: Plot of the unstructured spatial and temporal correlations and the associated estimated correlation functions for the natural log of the end-diastolic R/T ratio using the SCCOR data including post-surgery observations. On the left are the spatial functions when the temporal correlation is unstructured, while on the right are the temporal structures for a Matérn spatial structure.

CONCLUSION

Summary

The statistical model discussed in this dissertation allows longitudinal imaging studies to make greater use of the outcomes observed through imaging modalities. Instead of being reliant on summary methods to eliminate the correlation among observations, our model lets investigators take a multivariate approach and use outcomes from multiple locations in the image and quantify how they change over time. This allows for greater spatial resolution of anatomy and physiology in the analysis that could potentially lead to more powerful studies or more sensitive diagnoses and thus better patient outcomes.

Simulation studies showed that our model may have higher power than summary methods and be better at controlling the rate of false positives, justifying the added complexity. Our model also offers more flexibility in predictors than summary methods as it can utilize time- and space-varying covariates. The simulations also demonstrated that the practice of analyzing regions separately, ignoring their spatial correlation, has very poor statistical properties and should not be used.

The core of our model is the separable parametric spatiotemporal correlation structure. By assuming separability, one can consider how correlation changes over space and time independently which makes interpretation easier. The use of parametric correlation functions can improve the efficiency of the analysis and also improves the interpretability of the model as one can plot the estimate for how quickly correlation decreases with distance. Our research found that the properties of statistical inference about predictors in the linear model depend strongly on whether the correlation structure used is a good approximation for the true correlation; a good fit leads to high power and a reliable false positive

rate while poor modeling of the correlation may lose power or have an increased rate of false positives. We found that, for the scenarios we considered, information criteria are very good at choosing a working correlation structure that is 'close enough' to the data to have good statistical properties. However, the selection relies on having the proper predictors in the model and appropriate correlation functions to choose from which means that there must be collaboration between the statistical and clinical investigators when applying our model.

Future Directions

As always, there is still much work to be done in this research area. Some of it involves further validation of our proposed model under different scenarios to make it more generalizable, while other future directions involve taking our model further.

One of the biggest assumption of our proposed model is that of separability of spatial and temporal correlation. Although some work has been done on separability in longitudinal imaging data by groups such as Simpson *et al.*, there are still many questions left unanswered. A nonseparable unstructured model may be inefficient for applications with enough observations in space or time, while the practice of relating distance in space and time through a velocity may not be a reasonable assumption for anatomical outcomes such as wall thickness. It is important to understand what kinds of nonseparable parametric spatiotemporal correlation functions are appropriate for use on longitudinal imaging data, and to validate means of choosing between them and separable alternatives.

In addition to separability of the correlation, we assumed homoscedasticity of the outcomes. However, in practice the variance may change over time or over space. It is therefore highly relevant to quantify how sensitive our model is to heteroscedasticity, and to research how our model could be extended to directly model heterscedasticity in space and/or time.

One result of our work is that we quantified the effect on statistical inference when correlation structures are misspecified. However, we only looked at three spatial and four temporal correlation functions; it would be useful to know whether a wider variety of functions can approximate one another. Along these lines, it would be of interest to research how spatial structures can be used to model temporal correlation. Most temporal correlation functions assume observations to be evenly spaced which never occurs in practice; it is possible for the models to fit better (which could improve power) if the exact observation times are used. In our specific application of the 16 segment model to left ventricular MRI scans, it would also be of clinical interest to research new ways to define spatial distance; the anatomical structure is not actually two-dimensional so it may be more accurate to use a three-dimensional measure of distance.

In the simulation study of the second paper, missingness was designed to be completely at random. However, as we saw in the third paper, missing data may not occur at random. Although the likelihood-based approach offers some protection against data that is strictly missing at random (MAR), it may not perform well when the data is not missing at random (NMAR). Before our model can be deemed viable for widespread application, it is necessary to first understand how it fares when presented with MAR or NMAR data.

We know from the theory of REML and the demonstration in the first paper that the predictors can affect the selection of a parametric correlation structure. However, we do not know exactly how sensitive information criteria are to over- or underspecified mean structures. It could be useful to quantify how the observed and predicted correlation change with different sets of predictors.

In all of our simulations we simulated data with a linear time course and performed inference about a linear treatment-by-time effect. This was done in the context of a longitudinal clinical trial, where such an effect is the primary focus of the study. However, it would be of great interest to see how the findings of our simulation study change when the

target of inference is something different such as a space-varying predictor, or a three-way interaction. The effects of a non-linear time course would also be relevant.

Finally, it would be useful to explore how our model can be used in the design of a longitudinal imaging study. Specifically, how can our model be used to estimate sample size of a potential study? There are many factors to consider beyond the typical projections of effect size and variance of the outcome, such as different spatial and temporal correlation functions and degrees of correlation. It would also be extremely interesting to see how the degree of temporal correlation influences the tuning of more follow-up visits (J) versus more independent subjects (N) in the context of imaging-derived outcomes.

REFERENCES

[1] Ahmed, M.I., Aban, I., Lloyd, S.G., Gupta, H., Howard, G., Inusah, S., Peri, K., Robinson, J., Smith, P., McGiffin, D.C., Schiros, C.G., Denny Jr., T., Dell'Italia, L.J. (2012), "A Randomized Controlled Phase IIb Trial of β_1-Receptor Blockade for Chronic Degenerative Mitral Regurgitation," *Journal of the American College of Cardiology*, 60(9), 833-838.

[2] Ahmed, M.I., Gladden, J.D., Litovsky, S.H., Lloyd, S.G., Gupta, H., Inusah, S., Denny Jr., T., Powell, P., McGiffin, D.C., Dell'Italia, L.J. (2010), "Increased oxidative stress and cardiomyocyte myofibrillar degeneration in patients with chronic isolated mitral regurgitation and ejection fraction $> 60\%$," *Journal of the American College of Cardiology*, 55(7), 671-679.

[3] Albert, P.S. (1999), "Longitudinal Data Analysis (Repeated Measures) in Clinical Trials," *Statistics in Medicine*, 18, 1707-1732.

[4] Beyar, R., Weiss, J.L., Shapiro, E.P., Graves, W.L., Rogers, W.J., Weisfeldt, M.L. (1993), "Small apex-to-base heterogeneity in radius-to-thickness ratio by three-dimensional magnetic resonance imaging," *American Journal of Physiology*, 264, H133-H140.

[5] Bowman, F.D., Waller, L.A. (2004), "Modelling of Cardiac Imaging Data with Spatial Correlation," *Statistics in Medicine*, 23, 965-985.

[6] Burton, A., Altman, D.G., Royston, P., Holder, R.L. (2006), "The design of simulation studies in medical statistics," *Statistics in Medicine*, 25, 4279-4292.

[7] Castillo, E., Lima, J.A.C., Bluemke, D.A. (2003), "Regional Myocardial Function: Advances in MR Imaging and Analysis," *Radiographics*, 23, S127-S140.

[8] Cerqueira, M.D., Weissman, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., Pennell, D.J., Rumberger, J.A., Ryan, T., Verani, M.S., "Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart : A Statement for Healthcare Professionals From the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association," *Circulation*, 105, 539-542.

[9] Cressie, N., Huang, H. (1999), "Classes of nonseparable, spatio-temporal stationary covariance functions," *Journal of the American Statistical Association*, 94, 1330-1340.

[10] De Iaco, S., Myers, D.E., Posa, T. (2001), "Space-time analysis using a general product-sum model," *Statistics & Probabilitity Letters*, 52, 21-28.

[11] Enriquez-Sarano, M., Akins, C.W., Vahanian, A. (2009), "Mitral regurgitation," *Lancet*, 373, 1382-1394.

[12] Enriquez-Sarano, M., Tajik, A.J., Schaff, H.V., Orszulak, T.A., McGoon, M.D., Bailey, K.R., Frye, R.L. (1994), "Echocardiographic prediction of left ventricular function after correction of mitral regurgitation: results and clinical implications," *J Am Coll Cardiol*, 24, 1536-1543.

[13] Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. (ed.) (2008), *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, Boca Raton, FL: CRC Press.

[14] Gaasch, W.H., Meyer, T.E. (2008), "Left Ventricular Response to Mitral Regurgitation : Implications for Management," *Circulation*, 118, 2298-2303.

[15] Gelfand, A.E., Diggle, P., Guttorp, P., Furntes, M. (ed.) (2010), *Handbook of Spatial Statistics*, Boca Raton, FL: CRC Press.

[16] Genton, M.G. (2007), "Separable approximations of space-time covariance matrices," *Environmetrics*, 18, 681-695.

[17] Guerin, L., Stroup, W.W. (2000), "A Simulation Study to Evaluate PROC MIXED Analysis of Repeated Measures Data," *Proceedings of the 12th Kansas State University Conference onApplied Statistics in Agriculture*, April 30May 2, 2000 in Manhattan,KS, 170-203.

[18] Harville, D.A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72(358), 320-338.

[19] Jenrich, R.I., Schluchter, M.D. (1986), "Unbalanced Repeated-Measures Models with Structured Covariance Matrices," *Biometrics*, 42, 805-820.

[20] Laird, N.M., Ware, J.H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963-974.

[21] Lindstrom, M.J., Bates, D.M. (1988), "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83(404), 1014-1022.

[22] Liu, S., Rovine, M.J., Molenaar, P.C.M. (2012), "Selecting a Linear Mixed Model for Longitudinal Data: Repeated Measures Analysis of Variance, Covariance Pattern Model, and Growth Curve Approaches," *Psycological Methods*, 17(1), 15-30.

[23] Mann, D.L., Kend, R.L., Parsons, B., Cooper G. (1992), "Adrenergic effects on the biology of the adult mammalian cardiocyte," *Circulation*, 85, 790-804.

[24] Matthews, J.N.S., Altman, D.G., Campbell, M.J., Royston, P. (1990), "Analysis of Serial Measurements in Medical Research," *BMJ*, 300(6719):230-235.

[25] McVeigh, E. (2006), "Measuring mechanical function in the failing heart," *Journal of Electrocardiology*, 39, S24-S27.

[26] Mehta, R.H., Supiano, M.A., Oral, H., Grossman, M., Montgomery, D.S., Smith, M.J., Starling, M.R. (2003), "Compared with control subjects, the systemic sympathetic nervous system is activated in patients with mitral regurgitation," *American Heart Journal*, 145(6), 1078-1085.

[27] Nagatsu, M., Zile, M.R., Tsutsui, H., Schmid, P.S., DeFreyte, D., Cooper, G., Carabello, B.A. (1994), "Native β-Adrenergic Support for Left Ventricular Dysfunction in Experimental Mitral Regurgitation Normalizes Indexes of Pump and Contractile Function," *Circulation*, 89(2), 818-826.

[28] Nkomo, V.T., Gardin, J.M., Skelton, T.N., Gottdiener, J.S., Scott, C.G., Enriquez-Sarano, M. (2006), "Burden of valvular heart diseases: a population-based study," *Lancet*, 368, 1005-1011.

[29] Pat, B., Killingsworth, C., Denney, T., Zheng, J., Powell, P., Tillson, M., Dillon, A.R., Dell'Italia, L.J. (2008), "Dissociation between cardiomyocyte function and remodeling with β-adrenergic receptor blockade in isolated canine mitral regurgitation," *Am J Physiol Heart Circ Physiol*, 295, H2321-H2327.

[30] Schiros, C.G., Dell'Italia, L.J., Gladden, J.D., Clark III, D., Aban, I., Gupta, H., Lloyd, S.G., McGiffin, D.C., Perry, G., Denny Jr., T., Ahmed, M.I. (2012), "Magnetic resonance imaging with 3-dimensional analysis of left ventricular remodeling in isolated mitral regurgitation: implications beyond dimensions," *Circulation*, 125, 2334-2342.

[31] Schiros, C.G., Ahmed, M.I., Sanagala, T., Zha, W., McGiffin, D.C., Bamman, M.M., Gupta, H., Lloyd, S.G., Denny Jr., T., Dell'Italia, L.J. (2013), "Importance of three-

dimensional geometric analysis in the assessment of the athlete's heart," *American Journal of Cardiology*, 111(7), 1067-1072.

[32] Shehata, M.L., Cheng, S., Osman, N.F., Bluemke, D.A., Lima, J.A.C. (2009), "Myocardial tissue tagging with cardiovascular magnetic resonance," *Journal of Cardiovascular Magnetic Resonance*, 11(55), 1-12.

[33] Simpson S.L., Edwards L.J., Muller K.E., Sen P.K., Styner M.A. (2010), "A linear exponent AR(1) family of correlation structures," *Statistics in Medicine*,29,1825-1838.

[34] Simpson S.L., Edwards L.J., Muller K.E., Styner M.A. (2014), "Kronecker Product Exponent AR(1) Correlation Structures for Multivariate Repeated Measures," *PLoS ONE*, 9:e88864.

[35] Simpson, S.L., Edwards, L.J., Styner, M.A., Muller, K.E. (2014), "Separability tests for high-dimensional, low-sample size multivariate repeated measures data," *Journal of Applied Statistics*, DOI: 10.1080/02664763.2014.919251.

[36] Tsutsui, H., Spinale, F.G., Nagatsu, M., Schmid, P.S., Ishihara, K., DeFreyte, G., Cooper, G., Carabello, B.A. (1994), "Effects of Chronic β-Adrenergic Blockade on the Left Ventricular and Cardiocyte Abnormalities of Chronic Canine Mitral Regurgitation," *The Journal of Clinical Investigation, Inc.*, 93, 2639-2648.

[37] Waller, L.A., Gotway, C.A. (2004), *Applied Spatial Statistics for Public Health Data*, Hoboken, NJ: John Wiley and Sons, Inc.

[38] Wang, H., Amini, A.A. (2012), "Cardiac Motion and Deformation Recovery From MRI: A Review," *IEEE Transactions on Medical Imaging*, 31(2), 487-503.

[39] Wishart, J. (1938), "Growth-rate determinations in nutrition studies with the bacon pig, and their analysis," *Biometrika*, 30, 16-28.

[40] Zucker, D.M., Manor, O., Gubman, Y. (2012), "Power Comparison of Summary Measure, Mixed Model, and Survival Methods for Analysis of Repeated-Measures Trials," *Journal of Biopharmaceutical Statistics*, 22, 519-534.

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Institutional Review Board for Human Use

DATE: February 4, 2013

MEMORANDUM

TO: Brandon George
Principal Investigator

FROM: Cari Oliver, CIP
Assistant Director
Office of the Institutional Review Board (OIRB)

RE: Request for Determination—Human Subjects Research
IRB Protocol #N130128005– Spatio - Temporal Analysis of SCCOR Cardiac Imaging Data

A member of the Office of the IRB has reviewed your application for Designation of Not Human Subjects Research for above referenced proposal.

The reviewer has determined that this proposal is **not** subject to FDA regulations and is **not** Human Subjects Research. Note that any changes to the project should be resubmitted to the Office of the IRB for determination.

470 Administration Building
701 20th Street South
205.934.3789
Fax 205.934.1301
irb@uab.edu

The University of
Alabama at Birmingham
Mailing Address:
AB 470
1530 3RD AVE S
BIRMINGHAM AL 35294-0104