2006

A Model to Predict 24-Hour Urinary Creatinine Level Using Repeated Measurements

Donna S. Kroos
Virginia Commonwealth University

A Model to Predict 24-Hour Urinary Creatinine Level Using Repeated Measurements

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Mathematical Sciences (Statistics) at Virginia Commonwealth University.

By

Donna S. Kroos
BS (Mathematics), Pennsylvania State University, 1980
MBA (Quantitative Decision Making), Rochester Institute of Technology, 1996

Director: Dr. James E. Mays,
Associate Professor,
Department of Statistical Sciences & Operations Research

Virginia Commonwealth University
Richmond, Virginia
December, 2006

Acknowledgement

There are many individuals who have been instrumental in this thesis project and hence deserving of recognition. First, I would like to thank each of my thesis committee members—Dr. James E. Mays, Dr. Shelley A. Harris, and Dr. Edward L. Boone.

Dr. Mays' statistical knowledge, insight and guidance throughout my research were extremely valuable. I am honored to have him both as a mentor and as a friend.

I wish to thank Dr. Harris for granting me the opportunity to analyze the creatinine data from the "Pesticide Dose Monitoring in Turf Applicators" study. Her patience and candor in answering my many questions are greatly appreciated.

Dr. Boone's knowledge of mixed model methodologies provided much guidance to my usage of this modeling technique.

My appreciation also is extended to the National Institute of Occupational Safety and Health, Centers for Disease Control and Prevention, for the research grant that funded the "Pesticide Dose Monitoring in Turf Applicators" study.

I would like to thank Ms. Kristen Wells and Ms. Diane Bishop, Department of Epidemiology and Community Health, Virginia Commonwealth University. Their assistance, in extracting the creatinine and subject data from the pesticide dose study data base, was instrumental to my thesis efforts.

Lastly, I want to thank my husband Jim Kroos. I am, and always will be, grateful for his unwavering love and support throughout my academic pursuits and our life together.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike Information Criterion
ANTE	Ante-dependence
AR(1)	First Order Autoregressive
BIC	Schwarz's Bayesian Information Criterion
BMI	Body Mass Index
BSA	Body Surface Area
CAPD	Continuous Ambulatory Peritoneal Dialysis
CDC	Centers for Disease Control and Prevention
CS	Compound Symmetric
EBLUE	Estimated Best Linear Unbiased Estimator
EBLUP	Estimated Best Linear Unbiased Predictor
GEE	Generalized Estimating Equations
LBM	Lean Body Mass
ML	Maximum Likelihood
MSE	Mean Square Error
MSPR	Mean Squared Prediction Error
NIOSH	National Institute of Occupational Safety and Health

REML .....Restricted Maximum Likelihood

RLD.....Restricted Likelihood Distance

SD .....Standard Deviation

STL .....Scientific Testing Labs

TOEP.....Toeplitz

UN.....Unstructured

VCU .....Virginia Commonwealth University

Abstract

A MODEL TO PREDICT 24-HOUR URINARY CREATININE LEVEL
USING REPEATED MEASUREMENTS

By Donna S. Kroos

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2006

Thesis Director: Dr. James E. Mays,
Associate Professor, Department of Statistical Sciences & Operations Research

Creatinine is a metabolic waste product, removed from the blood by the kidneys, and excreted in the urine. The measurement of creatinine is used in the assessment and monitoring of many medical conditions as well as in the determination or adjustment of absorbed dosage of pesticides. Earlier models to predict 24-hour urinary creatinine used ordinary least squares regression and assumed that the subjects' observations were uncorrelated. However, many of these studies had repeated creatinine measurements for each of their subjects. Repeated measures on the same subject frequently are correlated. Using data from the NIOSH-CDC "Pesticide Dose Monitoring in Turf Applicators" study, this thesis project built a model to predict 24-hour urinary creatinine using the Mixed Model methodology. A covariance structure, that permitted multiple observations for any one individual to be correlated, was identified and utilized. The predictive capabilities of this model were then compared to the earlier models investigated.

Introduction

Creatinine (from muscle creatine) is a metabolic waste product removed from the blood by the kidneys and excreted in the urine (Katzung 1989). The concentration of creatinine in urine will vary based on a number of factors including hydration, kidney function, age, and gender (Heymsfield et al. 1983). Other factors thought to influence creatinine include muscle mass, height, weight, emotional stress, exercise, and protein consumption (Turner and Cohn 1975; Heymsfield et al. 1983; Kesteloot and Joossens 1993; Proctor et al. 1999). It also has been shown that there are large changes in urinary creatinine after trauma, infection, and inflammatory conditions (Fuller and Elia 1988).

The 24-hour level of creatinine in urine is frequently used in the assessment of many medical conditions, especially those relating to chronic kidney disease (Letteri et al. 1975; Johnson et al. 2004). Other uses include investigating the relationship between creatinine and dietary protein intake (Kesteloot and Jooseens 1993; Poortmans et al. 1997), using creatinine excretion in body composition studies (Welle et al. 1996) and estimating age related muscle loss (Proctor et al. 1999).

Creatinine concentration also is sometimes used in pesticide exposure research to provide a rough estimation of whether a urine sample is 'complete' (Harris et al. 2000). 'Complete' in this context implies that all individual urine voids for the specified time period (usually 24 hours) were captured in the study participant's container.

The National Institute of Occupational Safety and Health (NIOSH), Centers for Disease Control and Prevention (CDC) provided a four year grant to Virginia Commonwealth University (VCU) for a national study entitled "Pesticide Dose Monitoring in Turf Applicators." The overall goal of this study was to determine the absorbed dose, and factors that affect the absorbed dose, of various pesticides (including weed control products and insecticides) by the evaluation of the amount of parent compound or metabolites excreted in the urine of professional lawn care workers who were exposed to these pesticides over time. In this study, investigators probed how these workers may have inhaled, ingested, or absorbed (through contact with the skin) these pesticides. The frequency of exposure during a work week, the nature of the lawn care worker's interaction with the pesticide (during application, mixing, or loading it into tanks), and preventive safety measures used in handling the pesticides (e.g., eye protection, gloves, clean uniforms, hand washing, etc.) also were explored.

Study participants were asked to provide urine samples for multiple days during their work weeks. Each sample collected was the accumulated individual urine voids of a participant over a specific time period (which in most cases was a 24-hour block of time and in other cases was a 12-hour block of time). Study participants were asked to provide additional information on their work schedules, their pesticide handling practices, existing health issues, diet information, and demographics.

All urine samples were analyzed for six pesticide parent compounds and/or their metabolites. In addition, these samples also were tested for their creatinine concentration. The 'completeness' of the urine collections is instrumental to the estimation of the

absorbed dosage of pesticide. If 'complete' urine samples were not provided by the study participants, the absorbed doses of pesticide will be underestimated.

Variations in 'complete' urine collection do differ by individual and within an individual over time. It is important to develop methods to evaluate and if necessary, correct for, within and between individual variations over time. This will allow for more accurate dose assessment and human health risk assessment for pesticides.

Objectives

There are two main objectives for this thesis. The first is to create a predictive model for a total 24-hour urinary creatinine level using the data from the NIOSH-CDC "Pesticide Dose Monitoring in Turf Applicators" study. Possible factors to be included in this model are participant's age, body height, body weight, Body Mass Index (BMI), gender, tobacco smoking habits, medical conditions that affect kidney function (such as diabetes, congestive heart failure, high blood pressure, arteriosclerosis, glomerulonephritis, pyelonephritis, and urinary obstruction), participation in protein intensive diets (e.g., Atkins diet), alcohol consumption, usage of prescription medications, and usage of creatine supplements (to build muscle mass).

In any one person, when tested over several days, there may be multiple reasons for the variability observed in the subject's 24-hour urine creatinine levels. This makes building a predictive model more challenging. A key question arises: how much of the inherent variability in a person's creatinine level is a function of within individual variation over time? Short term variation (measured in days) in creatinine excretion is expected in an individual due to short term changes in diet or medication use. One also would expect decreases in 12 or 24-hour creatinine with missed urine samples, which may or may not be reported by the study subjects. Over the longer term (years) one would expect changes in creatinine excretion with changes in body weight, muscle mass,

and age (Welle et al. 1996; Proctor et al. 1999). Because this study had urine samples from the same subjects over multiple days and during multiple seasons, is there a way to identify the effect of time in the overall variance of urine creatinine level? How much of the variation in measured total 24-hour creatinine is explained by the within subject variation? This objective, then, is to develop a predictive model that best makes use of this within subject variation and its corresponding covariance structure.

The second objective is to compare the predictions for total 24-hour creatinine obtained from the model of Objective 1 to the predictions from other pre-existing models for subjects exhibiting the same characteristics. Other models to be considered include those from the research of Turner and Cohn (1975), Kawasaki et al. (1991), Harris et al. (2000), and other more recent predictive models for creatinine levels in 24-hour urine samples.

Other Models to Predict 24-Hour Urinary Creatinine

Twenty-four hour creatinine is used in the monitoring and assessment of kidney related diseases, in the estimation of age related muscle loss, in dietary and body composition studies, and in pesticide exposure research. This section gives an overview of some of the earlier research studies in these arenas and the resultant models that were used to predict 24-hour urinary creatinine levels.

Turner and Cohn

One of the earliest models to predict 24-hour urinary creatinine was developed by William J. Turner, M.D. and Stanton Cohn, Ph.D. in 1975. In their investigation of total body potassium and 24-hour creatinine excretion, Turner and Cohn compared the urinary creatinine levels for 33 healthy male subjects to the creatinine levels of 31 chronic schizophrenic subjects (Turner and Cohn 1975).

Each group of subjects provided at least three 24-hour urine samples. Diet was uncontrolled for both the control group of healthy males and the group of schizophrenic patients. On the days of collection, weights and heights were recorded between 10:00 am and 11:00 am on all subjects lightly clothed and without shoes. The control subjects' ages ranged from 24 to 66 years, their heights ranged from 160.8 to 193 centimeters, and their weights ranged from 61.8 to 124.1 kilograms. The schizophrenic subjects' ages ranged

from 22 to 65 years, their heights ranged from 162.6 to 192 centimeters, and their weights ranged from 53.5 to 123.6 kilograms.

Using subject height, weight, and age, Turner and Cohn developed a prediction equation for 24-hour creatinine for each group of subjects (i.e., one equation for the healthy group and a second equation for the schizophrenic patient group) and a prediction equation for total body potassium (for the control group). Their creatinine prediction equation for the healthy male subjects of the control group was:

$$CR = 0.0143 * height + 0.00975 * weight - 0.00734(age - 20) - 1.391$$

where *CR* was the predicted 24-hour urinary creatinine level (measured in grams), height was measured in centimeters, weight was measured in kilograms, and age was measured in years.

Moriyama et al.

Masaki Moriyama, Hiroshi Saito, Atsuhiro Nakano, Shoetsu Funaki, and Saburo Kojima investigated the effect of dietary protein levels on 24-hour creatinine excretion (Moriyama et al. 1988). They conducted a field survey on a group of 40 healthy Japanese adults (22 male, 18 female) living in Akita, Japan. For this study, 24-hour urinary creatinine excretions and urea-N were measured on each subject once every three months (on the 20th day of February, May, August, and November in 1983) for a total of four times during the year. Urinary urea-N was used as a marker of dietary protein level.

Subjects were not restricted on diet or level of activity. Each subject's height and weight also were recorded at these four seasonal times. The mean of the male subjects'

ages (in years) was 46.6 with a standard deviation (SD) of 7.8 years, the mean height was 166.4 (SD 5.0) centimeters, and the mean weight was 67.1 (SD 12.6) kilograms. The mean of the female subjects' ages was 44.5 (SD 8.3) years, the mean height was 152.6 (SD 4.8) centimeters, and the mean weight was 51.6 (SD 5.5) kilograms.

Using multiple regression analysis and the factors of age, height, and weight, Moriyama et al. built gender specific prediction equations of 24-hour creatinine for each three month season. They also created a summary measure equal to the average of the four seasonal measurements and then used this average as the response variable in building a gender specific prediction equation for a 24-hour creatinine level. For example, for males, the prediction equation for the average of the four seasonal measurements was:

$$CR = 211 - 6.4 * age + 2.5 * height + 18 * weight$$

where *CR* was the predicted 24-hour urinary creatinine level (measured in milligrams), height was measured in centimeters, weight was measured in kilograms, and age was measured in years.

When adding the urea-N measurement, as a marker of dietary protein level, in the creatinine prediction equation (which already included height, weight and age), Moriyama et al. found that the differences of creatinine caused by seasonal variability of dietary protein level to be 9.2% for males and 3.6% for females.

Kawasaki et al.

Another study that used repeated 24-hour creatinine measurements from the same subject was published by Kawasaki et al. (1991). Three to five days of 24-hour urine collections were done on 256 male and 231 female healthy subjects with ages of 20 to 84 years. Regression analysis was used to develop an equation to predict the 24-hour creatinine level. For a male subject, the resultant prediction equation was:

$$CR = -12.63 * age + 15.12 * weight + 7.39 * height - 79.90$$

where *CR* was the predicted 24-hour urinary creatinine level (measured in milligrams), height was measured in centimeters, weight was measured in kilograms, and age was measured in years.

Based on their model validation efforts, using 47 non-Japanese subjects, Kawasaki et al. suggested that their prediction equations also were applicable for predicting 24-hour creatinine in non-Japanese individuals.

Jones, Newstead, and Will

In 1996, Colin Jones, Charles Newstead and Eric Will conducted a study to determine if an estimation of creatinine clearance from serum creatinine, gender, age, and weight would decrease the number of 24-hour urine and dialysate collections needed to monitor the adequacy of dialysis on patients using continuous ambulatory peritoneal dialysis (CAPD) (Jones et al. 1997). As part of this study, a prediction of 24-hour excreted creatinine was needed to derive a creatinine clearance value that was then compared to the actual measured creatinine clearance value.

The Jones et al. study included 187 urine collections (for 24-hour periods) from 99 CAPD patients (55 male, 44 female). Collections on the same individual were separated by at least four months. The data for male and female subjects were analyzed separately. Using multiple regression analysis, the creatinine excretion prediction equation for use in this study, for a male subject, was:

$$CR = 60 * (50 - age + 2 * weight)$$

where *CR* was the predicted 24-hour urinary creatinine level (measured in μmol), weight was measured in kilograms, and age was measured in years. The predicted values from this equation were then used to derive creatinine clearance values for the remainder of their research.

Harris et al.

As noted earlier, the completeness of a urine sample is crucial to accurately assessing the absorbed dose of pesticides in humans. An incomplete urine sample could lead to under-estimating the absorbed dosage of the pesticide (Harris et al. 2000). Harris et al. (2000) developed a predictive model for 24-hour excreted creatinine to use in identifying incomplete urine collections.

Two consecutive 24-hour urine samples of 98 professional turf applicators (93 male, 5 female), from a previous pesticide dose prediction study (Harris 1999), were analyzed. The subjects' creatinine measures were first corrected for any missed urine voids during either of the 24-hour collection periods. Any missed voids were self reported by the study participants. The two corrected creatinine values were then

averaged and a predictive model, using multiple regression, was developed using this average creatinine value as its response variable. The resultant prediction equation, for male subjects) was:

$$CR = 647 + 372 * gender + 13.5 * weight - 10.8 * age - 1.47 * [(age - 28.4) * (weight - 80.1)]$$

where *CR* was the predicted 24-hour urinary creatinine level (measured in milligrams), gender equaled 1 if a male subject (or 0 if a female), height was measured in centimeters, weight was measured in kilograms, and age was measured in years (Harris et al. 2000). This resultant prediction equation was then used in the analysis of determining the impact of correcting for missed urine voids in the measurement of pesticide levels.

Tanaka et al.

Salt intake in Japan is higher than in western countries (INTERSALT 1988). It has been shown that excessive salt intake is a known factor for high blood pressure (INTERSALT 1988; Elliott 1989). It also has been reported that an increase in potassium intake may decrease blood pressure (Elliott et al. 1989). Hence, the Japanese government has set goals for both salt and potassium intake (Japan Health Promotion and Fitness Foundation 2000).

To aid in the evaluation of salt and potassium intake, 24-hour urine collections are often used instead of food intake questionnaires (Tanaka et al. 2002). In an investigation of a method to estimate 24-hour urinary sodium and potassium excretion using casual urine specimens, Tanaka et al. (2002) developed a model to predict 24-hour urinary creatinine that was then used in their urinary sodium and potassium excretion formulae.

The 591 subjects (295 male and 296 female), with ages of 20 to 59 years, for the Tanaka et al. efforts were randomly selected from a larger data base of subjects from the INTERSALT study. The INTERSALT study was an international study (across 32 countries) of electrolyte excretion and blood pressure (INTERSALT 1988).

The subjects for the Tanaka et al. study lived in three cities in Japan (Osaka, Toyama, and Tochigi) during the years of 1987 and 1988. The average age of the male subjects was 40.0 (SD 11.1) years, the average height was 166.8 (SD 6.7) centimeters and the average weight was 63.3 (SD 12.6) kilograms. The average age of the female subjects was 39.0 (SD 11.2) years, the average height was 153.9 (SD 5.6) centimeters and the average weight was 52.2 (SD 7.1) kilograms. Each subject provided one 24-hour urine sample.

A regression model was built to predict 24-hour urinary creatinine using the factors of age, height and weight. The resultant model was:

$$CR = -2.04 * age + 14.89 * weight + 16.14 * height - 2244.45$$

where *CR* was the predicted 24-hour urinary creatinine level (measured in milligrams), height was measured in centimeters, weight was measured in kilograms, and age was measured in years. It is important to note that this model had no indicator variable for gender. Predicted values from this model were then used by Tanaka et al. to develop formulae to estimate population mean levels of 24-hour sodium and potassium excretion.

Kamata and Tochikubo

In 2002, Kumiko Kamata and Osamu Tochikubo studied the estimation of 24-hour urinary sodium excretion using lean body mass and overnight urine collected by a pipe sampling method. Researchers often use 24-hour urine specimens to assess salt intake rather than using less reliable methods involving dietary recall by the subjects (Kamata and Tochikubo 2002).

The Kamata and Tochikubo study included 351 healthy subjects in total (126 men and 225 women) between the ages of 20 and 70 years. The average age of the men in the study was 38 (SD 20.3) years, the average height was 169 (SD 7.2) centimeters and the average weight was 65.0 (SD 8.6) kilograms. The average age of the female subjects was 50 (SD 16) years, the average height was 156 (SD 6.0) centimeters and the average weight was 52.7 (SD 7.0) kilograms. Body fat was measured using a fat meter.

Study participants were not regulated on their food intake and continued their normal daily activities during the sample collection period. One 24-hour urine sample was collected from each participant. A sub group of the study subjects (71 men and 78 women) used the pipe sampling method to collect the overnight urine portion of their 24-hour sample.

The prediction equation for the male subjects in the study was:

$$CR = 0.027 * LBM - 0.006$$

where *CR* was the predicted 24-hour urinary creatinine level (measured in grams) and LBM is the lean body mass (measured in kilograms), which was calculated as body weight minus body fat. This resultant prediction equation for urinary creatinine was then

used by Kamata and Tochikubo to predict a 24-hour sodium value for the remainder of their research.

Penie, Porben, and Silverio

In constructing an Index of Creatinine Excretion for Cuban subjects, standards derived from Anglo-Saxon subjects traditionally were used (Penie et al. 2003). Penie et al. believed that this practice could lead to diagnostic errors because of differences in diet and body composition of non-Anglo-Saxon populations.

Cuban subjects for this Penie et al. study were drawn retrospectively from the databases of the Section of Urinalysis, Service Clinical Laboratory, "Hermanos Ameijeiras" Hospital in La Habana, Cuba. Fifty-five percent of the Cuban patients in this database were associated with renal function studies, 20% were associated with arterial pressure studies, 15% were associated with nutritional evaluations and 10% were identified as miscellaneous reasons.

For their study, 103 male subjects and 112 female subjects (with ages between 19 and 58 years), were selected. The average age of the male subjects was 40.1 (SD 1.12) years, the average height was 170.9 (SD 0.75) centimeters, and the average weight was 68.8 (SD 0.69) kilograms. For female subjects, the average age was 37.5 (SD 1.05) years, the average height was 159.3 (SD 0.58) centimeters, and the average weight was 60.3 (SD 0.63) kilograms.

Penie et al. used regression analysis to derive gender specific prediction equations for 24-hour creatinine levels. They tested two models; one that only contained age as a

predictor and a second model that only contained height as a predictor. Their final model, for male subjects, used only height and is shown here:

$$CR = -1791.05 + 17.69 * height$$

where *CR* was the predicted 24-hour urinary creatinine level (measured in milligrams) and height is the subject height measured in centimeters.

Penie et al. noted that after considering the subject's weight, the urinary creatinine excretion for Cuban subjects (both male and female) was lower than that of their Anglo-Saxon counterparts. They recommended that their resultant table of values for urinary creatinine excretion, from this study, be used in future body composition and nutritional evaluations of Cuban subjects.

Summary

Table 1 summarizes the models mentioned above. It includes the means and standard deviations of the ages, heights, and weights of the subjects used in the original studies. It indicates which studies had repeated measurements for any one subject. It also includes what factors were included in the resultant prediction equations. It is important to note that all of these prediction equations for these studies were built using ordinary least squares regression analysis. If a summary measure (such as a mean of the repeated measures) was not used as the response variable, then the resultant model was built using the assumption that any repeated measures for any one subject were uncorrelated.

Table 1: Summary of Other Models used to predict 24-hour Urinary Creatinine Level (Male Subjects Unless Otherwise Noted)

Model	Predictors Used	Repeated Measures on one subject?	Age (yrs) $mean_{age}\ (SD)$	Height (cm) $mean_{height}\ (SD)$	Weight (kg) $mean_{weight}\ (SD)$
Harris[1]	gender, weight, age	Yes (2)	28 (7.4) *	178 (7.0) *	80 (12.0) *
Kamata	lean body mass	No	38 (20.3)	169 (7.2)	65 (8.6)
Kawasaki	height, weight, age	Yes (3-5)	20-84 *,**	N/A	N/A
Jones	age, weight	Yes	N/A	N/A	N/A
Moriyama[1]	height, weight, age	Yes (4)	46.6 (7.80)	166.4 (5.00)	67.1 (12.6)
Penie	height	N/A	40.1 (1.12)	170.9 (0.75)	68.8 (0.69)
Tanaka	height, weight, age	No	40.0 (11.1)	166.8 (6.70)	63.3 (12.6)
Turner & Cohn	height, weight, age	Yes (≥ 3)	24-66 **	160.8-193 **	61.8-124.1 **

* *For both male and female subjects*
** *Range of ages and/or heights and/or weights*
[1] *Model used the average of the repeated measures as the response.*
'N/A' indicates information not available
'SD' indicates standard deviation

Study Data

Data Collection

The data for this thesis project were collected as part of the NIOSH CDC funded "Pesticide Dose Monitoring in Turf Applicators" study. This study was approved by the VCU Institutional Review Board. Lawn care and tree and shrub workers employed by the TruGreen ChemLawn (part of Service Master Corporation) were the subjects for both the pilot phase and the follow-on phase of this national study. Participation in the study was strictly on a volunteer basis.

The initial pilot phase of this study was conducted in Richmond, Virginia in 2003 and consisted of three rounds of collections (one round in the summer, two rounds in the fall). The follow-on phase was conducted in five locations around the United States (Puyullap, Washington; Plainfield, Illinois; Plano, Texas; Salt Lake City, Utah; Sterling, Virginia). The follow-on phase had three rounds of collections: the first round was in the Spring of 2004, the second round was in the Summer of 2004 and the final round was completed in the Fall of 2004. Upon completion, there were 21 male subjects who participated in the pilot phase and 113 subjects (3 female, 110 male) who participated in the follow-on phase of the study.

For the pilot phase of the study, each subject was asked to collect their individual urine voids (into one container) for a 12-hour period. Each participant was asked to

repeat this collection process twice a day for five consecutive days (i.e., two 12-hour collection periods per day for five consecutive days, resulting in ten containers per subject after the five days). In the fall months of 2003, the same subjects were then asked to collect their individual urine voids for 24-hour collection periods each day for a continuous two week period (for a total of 14 consecutive days).

For the follow-on phase of the study, during the spring round of collections, each subject was asked to collect their individual urine voids into one container for a 24-hour period. Each participant was then asked to repeat the 24-hour collection process for the next day. During the summer round of this phase, the same study participants (i.e., those who had participated in the spring round and additional new hires to TruGreen ChemLawn) were asked to provide four consecutive 12-hour urine samples (i.e., two 12-hour collection periods per day for two consecutive days). During the fall round of this phase, the same participants were asked to provide two consecutive 24-hour urine samples.

In both phases, for a specified collection period (e.g., a 12-hour period or a 24-hour period), each participant was instructed to collect, into one container, all of their individual urine voids during that time frame. Participants may have occasionally forgotten to use their container during a specific collection period. This would have resulted in an incomplete sample for that period. When this occurred, participants were asked to acknowledge the missed collection and approximate the volume of any 'un-captured' urine voided during the time period in question.

During any one collection period, a participant's container was kept in a cooler with ice packs. This was done to offset any possible degradation in creatinine concentration caused by temperature extremes (Fuller and Elia 1988). Both the cooler and the ice packs were provided to the participant by the study's administering personnel. After the desired 12-hour or 24-hour collection period was completed, each participant returned their 'full' container to the study administrator. Participants were then given a new container and ice packs to use for their next collection time period.

Upon return of the 'full' collection container, each participant's 12-hour or 24-hour sample had a specific gravity measurement completed on site, when possible, using the Leica AR200 digital hand-held refractrometer. Since only one instrument was available to take into the field, specific gravity measurements were not collected at all sites due to the overlap in field visit schedules.

All collected urine samples were shipped express overnight to VCU for freezing and long term storage. Creatinine concentrations were measured by Scientific Testing Laboratories (STL) in Richmond, Virginia using an automated method based on the Jaffe reaction. For the pilot phase, collected urine volume and weight were measured at the Environmental Health Lab at VCU; for the follow-on phase, the collected urine volume was measured with graduated cylinders in mobile measurement laboratories set up in the field sites.

Any missed urine voids during a specific collection period were self reported by the participant on the questionnaire that was returned to the study administrator at the end of the sampling week. If needed, the urine volume in a particular collection period could

be corrected by the estimated volumes provided by the participant for any missed voids that may have occurred during the time period in question.

Actual study questions also were answered by each participant at the end of each round. Age, body weight, body height, gender, tobacco smoking habits, medical conditions that affect kidney function, adherence to a protein intensive diet, alcohol consumption, usage of specific prescription medications, and usage of creatine supplements were self reported by each participant.

Data Calculations

Using the creatinine concentration level and the total urine volume collected in a container, the total creatine level for a 24-hour period was computed using

$$\textit{24-hour total creatinine} = \textit{creatinine concentration}_{(24\text{-hour})}\,(mg/dL) * \textit{urine volume}_{(24\text{-hour})}\,(mL) * 1/100.$$

Similarly, the total creatinine level for a 12-hour period was computed using

$$\textit{12-hour creatinine} = \textit{creatinine concentration}_{(12\text{-hour})}\,(mg/dL) * \textit{urine volume}_{(12\text{-hour})}\,(mL) * 1/100.$$

If the collection period time frame was two consecutive 12-hour collection periods, then the 24-hour total creatinine for any one day was computed as

$$\textit{24-hour total creatinine} = \textit{12-hour creatinine}_{(\textit{first 12 hour period})} + \textit{12-hour creatinine}_{(\textit{second 12 hour period})}$$

If only one of these two 12-hour creatinine values were available, at the direction of the study's primary investigator, the missing 12-hour creatinine value was imputed. See Appendix 1 for details.

Body Mass Index (BMI) was calculated using the formula (Boeniger et al. 1993):

$$BMI = \frac{weight}{height^2}$$

where weight is expressed in kilograms (1 kilogram = 2.2 pounds) and height is expressed in meters (1 meter = 39.37 inches) (Selby 1975).

Participant Enrollments and Descriptive Statistics

Each round of the pilot phase and the follow-on phase of the study had a differing number of participants who actually enrolled and completed at least one collection period in the round. Table 2 indicates the number of subjects by city and round. This table does not reflect any missed urine samples from the participants within any one phase. A missed sample within a round would be considered missing data (Vonesh and Chinchilli 1997). Changes in enrollments for each season are due to recruiting of newly hired participants. The seasonal differences of completed versus enrolled are due to the inability to collect data for a specific season (e.g., Plainfield in the fall of 2004) and/or individuals not participating in the study because they were no longer employed by TruGreen ChemLawn. This thesis project does not probe into modeling the dropout process.

Table 2: Study Enrollments by Location and Phase

Pilot Phase City	Final Number Subjects Enrolled	Summer completed/enrolled	Fall Round 1 completed/enrolled	Fall Round 2 completed/enrolled
Richmond	21	21/21	16/21	16/21
Follow-on Phase City	**Final Number Subjects Enrolled**	**Spring completed/enrolled**	**Summer completed/enrolled**	**Fall completed/enrolled**
Sterling	33	29/31	19/31	22/33
Plano	14	14/14	14/14	14/14
Puyallup	19	13/13	17/17	13/19
Salt Lake	27	22/22	19/27	15/27
Plainfield	20	20/20	16/20	No data collected
Follow-on Phase Totals	113	98/100	85/109	64/93

The NIOSH CDC "Pesticide Dose Monitoring in Turf Applicators" study encouraged and accepted both male and female study participants. Of the 134 total participants in the study, there were only three female subjects (all of whom were in the follow-on phase). Due to such a small number of females in the overall study, the decision was made to use only the male participant data for this thesis project.

It also was decided to use the follow-on phase data as the model building set and the pilot phase data as a model validation data set. For each of the 21 subjects in the pilot phase of the study, up to 19 days of 24-hour creatinine measurements were possible; this would have resulted in a total number of possible observations of 399. Of the 399 possible, 108 observations were missing valid creatinine levels and another 16 observations did not have weight and/or height measurements associated with them. The

end result was 275 observations that had creatinine values, heights and weights and thus were potential candidates to generate a random sample of model validation records.

There were up to six creatinine measurements possible for each of the 110 male subjects in the model building data set (for a total number of 660 possible observations). Of the 660 possible, 186 observations did not have a valid 24-hour creatinine level and another 17 observations did not have weight and/or height measurements associated with them. The end result was 457 observations that had creatinine values, heights and weights and thus were usable in the model building phase.

The technique used for selecting the sample of model validation records and the associated descriptive statistics for the resultant model validation dataset are described later in the Model Validation section. Descriptive statistics for the male participants in the model building set are shown here in Table 3.

Table 3: Descriptive Statistics for Model Building Data

	N	Mean	Std Dev	Min	Max
Creatinine (mg/day)	457	1609.000	582.005	196.115	3929.000
Height (cm)	457	179.540	6.552	162.560	198.120
Weight (kg)	457	90.680	21.611	52.163	188.241
BMI (kg/m^2)	457	28.013	5.696	18.654	49.502
Age at Collection (yrs)*	453	33.651	9.235	18.000	60.000

**Of the 457 observations with valid creatinine, height, and weight values, only 453 observations included a subject's age at time of urine collection.*

Model

Methodology

In both phases of this study, multiple measurements of each participant's total 24-hour creatinine levels were taken over 2-3 growing seasons. Observations stemming from repeated measurements on the same subject will most likely be correlated (Khattree and Naik 1999). Hence ordinary least squares regression, which assumes each observation is uncorrelated with the other observations, is not an appropriate technique to be used for this analysis.

There are two methods for accommodating correlated data – generalized estimating equations (GEE) and random effects models (Vittinghoff et al. 2005). Vittinghoff et al. suggest using generalized estimating equations when there are a large number of subjects and relatively few time points. Vittinghoff et al. also cited that GEE is limited in that "it is restricted to a single level of clustering, it is not designed for inferences about the correlation structure, and it does not give predicted values for each cluster."

Complicating this analysis is that the data collected are unbalanced. Unbalanced data occurs when the number of observations in the cells (where a cell is defined by one level of each factor) is not equal (Searle 1987). Table 2 demonstrated why the data for this study may be considered to be unbalanced. The varying number of participants

enrolled and completing the study in each season would yield an unequal number of observations for a specific time period. It also is appropriate to note that the urine collections, for the follow-on phase, were not equally spaced chronologically across all locations. The lapse in time (measured in days) between the spring and summer rounds (as well as between the summer and fall rounds) was not the same across all locations in the follow-on phase. This also causes the data to be unbalanced (Vonesh and Chinchilli 1997).

Considering the unbalanced nature of the data and the goals stated in Objective 1, a Mixed Model methodology was used to fit a model to predict creatinine level and to determine the within subject variation. Age at time of collection, body height, body weight, Body Mass Index (BMI), tobacco smoking habits, medical conditions that affect kidney function, protein intensive diets, alcohol consumption, prescription medications, and usage of creatine diet supplements were considered as possible explanatory factors for this model. These factors were self reported by the subject at the start of each round; hence all of these factors are time varying factors (Khattree and Naik 1999).

The mixed model in matrix notation (Khattree and Naik 1999) can be written as:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\upsilon}_i + \boldsymbol{\varepsilon}_i$$

where: i = $1, 2, \ldots, n$

n = number of subjects

p_i = number of measurements made on the i^{th} subject

q = number of fixed effects

r = number of random effects

$\boldsymbol{y_i}$ = the p_i x 1 vector of repeated measures on the i^{th} subject

$\boldsymbol{X_i}$ = the known p_i x q matrix of constants that describe the structure of the study with respect to fixed effects (including treatment design, regression explanatory or predictor variables) for the i^{th} subject (Littell et al. 2006).

$\boldsymbol{\beta}$ = the fixed q x 1 vector of unknown parameters

$\boldsymbol{Z_i}$ = the known p_i x r matrix of constants that describe the study's structure with regard to random effects (including blocking design and explanatory variables in random coefficient designs) for the i^{th} subject (Littell et al. 2006).

$\boldsymbol{\nu_i}$ = the r x 1 vector of random effects for the i^{th} subject

$\boldsymbol{\varepsilon_i}$ = the p_i x 1 vector of random errors for the i^{th} subject

Expected values, variances, and covariances of these matrices / vectors are shown here (Khattree and Naik 1999):

$E(\boldsymbol{\nu_i}) = \mathbf{0}$; $Var(\boldsymbol{\nu_i}) = \sigma^2\boldsymbol{G_I}$ (where $\boldsymbol{G_I}$ is the covariance matrix of the random effects)

$E(\boldsymbol{\varepsilon_i}) = \mathbf{0}$; $Var(\boldsymbol{\varepsilon_i}) = \sigma^2\boldsymbol{R_i}$ (where $\boldsymbol{R_i}$ is the covariance matrix of the repeated measures on subject i)

$E(\boldsymbol{\nu}) = \mathbf{0}$; $E(\boldsymbol{\varepsilon}) = \mathbf{0}$;

For $\boldsymbol{i \neq j}$: $Cov(\boldsymbol{\nu_i}, \boldsymbol{\nu_j}) = \mathbf{0}$; $Cov(\boldsymbol{\varepsilon_i}, \boldsymbol{\varepsilon_j}) = \mathbf{0}$; $Cov(\boldsymbol{\nu_i}, \boldsymbol{\varepsilon_j}) = \mathbf{0}$

$$\text{Cov}(\boldsymbol{\nu}_i, \boldsymbol{\varepsilon}_i) = \mathbf{0}; \qquad \text{Cov}(\boldsymbol{\nu}, \boldsymbol{\varepsilon}) = \mathbf{0}$$

$$\text{Var}(\boldsymbol{\nu}) = \sigma^2 \boldsymbol{G} \text{ where } \boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{G}_1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \boldsymbol{G}_1 \end{bmatrix}$$

$$\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{R} \text{ where } \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{R}_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \boldsymbol{R}_n \end{bmatrix}$$

Due to the random vectors $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}$ in this mixed model, there are two possible distributions to consider -- the conditional distribution of $(\boldsymbol{Y} \mid \boldsymbol{\nu})$ and the marginal distribution $(\boldsymbol{Y})$ as shown here (Littell et al. 2006):

$$\text{E}(\boldsymbol{Y} \mid \boldsymbol{\nu}) = \boldsymbol{X\beta} + \boldsymbol{Z\nu}, \ \text{Var}(\boldsymbol{Y} \mid \boldsymbol{\nu}) = \sigma^2 \boldsymbol{R}$$

$$\text{E}(\boldsymbol{Y}) = \boldsymbol{X\beta}, \ \text{Var}(\boldsymbol{Y}) = \sigma^2 [\boldsymbol{ZGZ'} + \boldsymbol{R}]$$

By using the mixed model approach, it is possible to obtain insight into the within subject variation, over time, for the total 24-hour creatinine level. The within subject variation for subject i would be contained in the $\boldsymbol{R}_i$ covariance matrix referenced in the mixed model description shown above. The potential explanatory factors of age, height, weight, BMI, etc. are placed in the matrix of constants that describe the structure of the study with respect to fixed effects (i.e. the $\boldsymbol{X}_i$ matrix). No explanatory factors are placed in the model in the matrix of constants that describe the study's structure with regard to random effects (i.e., the matrix $\boldsymbol{Z}_i = \boldsymbol{0}$). Hence, for this study's model, $\boldsymbol{\nu}$ is a zero vector,

$\boldsymbol{G}$ is a zero matrix, no random effects will be estimated, and the variation will be modeled through the $\boldsymbol{R}$ matrix.

Covariance Structure Selection

A key element in this analysis was to identify the covariance structure of the within subject variation (matrix $\boldsymbol{R_i}$). Inadequate modeling of the covariance structure would result in biased estimates of variances of estimates of fixed effects (Littell et al. 2006). Potential correlation of the repeated measurements of the 24-hour creatinine level for one subject had to be investigated. To perform this investigation, several covariance matrix structures, typically used for analysis of repeated measures data, were considered. These included the unstructured, compound symmetric, first-order autoregressive, Toeplitz, ante-dependence, and spatial matrix structures (Jenrich and Schluchter 1986; Littell et al. 2006). The details of these covariance matrix structures and how they differ from an independent covariance structure may be found in Appendix 2.

To determine which covariance structure best fit the data, an approach similar to what is outlined by Wolfinger (1993) was utilized. Models, that included all possible fixed factors, using each of the potential covariance matrix structures, were generated using release 9.1.3 of the SAS® statistical software package (SAS Institute, Inc. 2003). The SAS code for these models may be found in Appendix 3. The results from these six covariance approaches (found in Appendix 5) were compared using lag time plots, likelihood ratio tests, and information criteria.

A lag time plot shows covariance as a function of lag time between pairs of observations. This plot shows if, over time, the covariances are constant or if they are decreasing (or increasing). Lag is calculated as equal to one for measurements that are adjacent to each other [i.e. for the i^{th} and the $(i + 1)^{th}$ measurement]; a lag of two is assigned for measurements two time units apart [i.e. the i^{th} and the $(i + 2)^{th}$ measurement], etc. The covariance estimates versus lag in time when the unstructured covariance matrix is used for the model building data are shown in Appendix 4; the plot is below in Figure 1.

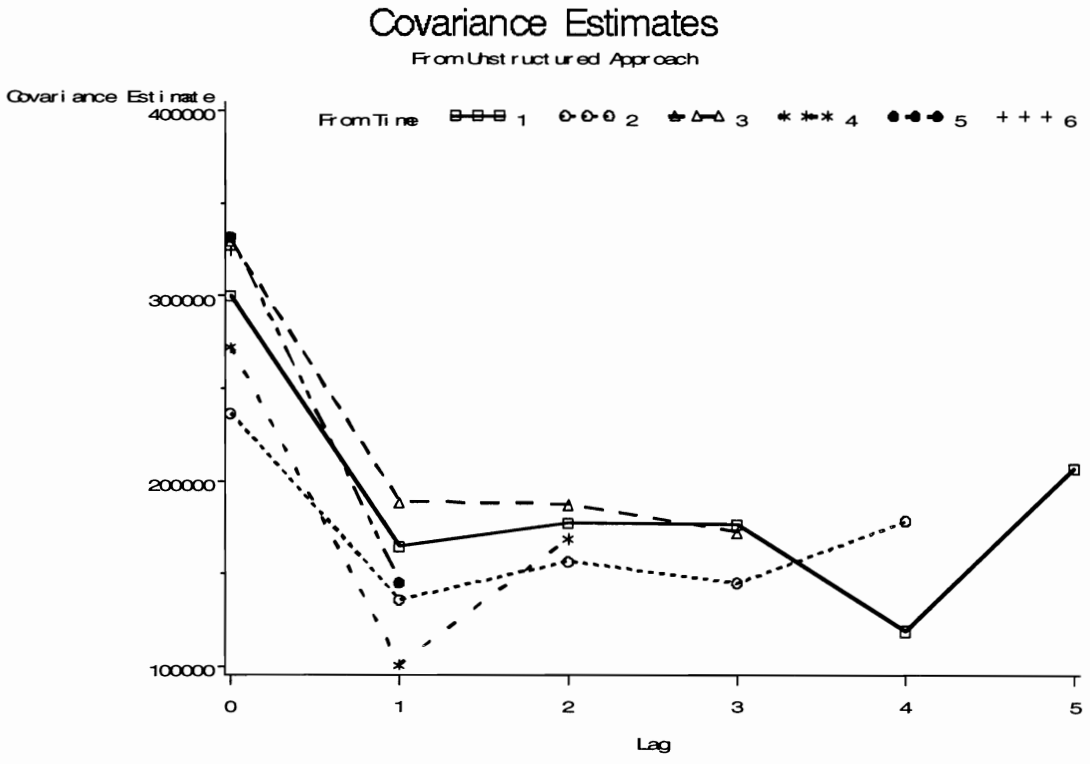

Figure 1: Lag Plot from Unstructured Covariance Matrix Results

The plot in figure 1 suggests that as lag increases, the covariance estimates tend to flatten out. This would suggest that a compound symmetric covariance approach, where the correlation between the i^{th} and j^{th} time is constant for a subject (Littell et al. 2006), may be appropriate.

The second step in the comparisons used likelihood ratio tests of the model under one covariance structure versus the same model under the unstructured covariance matrix. The appropriate likelihood ratio test statistic for the null hypothesis $\mathbf{H_o}$: $\boldsymbol{R_i}$ **has a specific covariance structure** is:

$$L = -2\ln\left[\frac{\max_{underHo}(likelihood \mid data)}{\max_{unstructured}(likelihood \mid data)}\right],$$

where maximum likelihood is derived from the Maximum Likelihood (ML) method (Khattree and Naik 1999). When n is large, under the null hypothesis, this test statistic approximately follows a χ^2 distribution with degrees of freedom = number of unknown parameters in the unstructured covariance matrix minus the number of unknown parameters in the covariance matrix specified in the null hypothesis (Khattree and Naik 1999).

The likelihood ratio test statistic for a null hypothesis of independent covariance matrix structure versus any of the other covariance structures considered was calculated by SAS; in all cases, this null hypothesis was rejected and it was concluded that the independent covariance matrix structure was not appropriate. Table 4 indicates the results of the likelihood ratio tests comparing the other covariance structures to the unstructured case.

Table 4: Likelihood Ratio Tests for Covariance Matrix Structures

Likelihood Ratio Test	Unknown Parameters	Test Statistic Value	χ^2 Critical Value*	Conclusion
H_0: R_i is Compound Symmetric H_A: R_i is Unstructured	CS = 2 UN = 21	24.1	χ^2_{19}= 30.14	Do Not Reject H_0
H_0: R_i is Autoregressive H_A: R_i is Unstructured	AR = 2 UN = 21	76.5	χ^2_{19}= 30.14	Reject H_0
H_0: R_i is Toeplitz H_A: R_i is Unstructured	TP = 6 UN = 21	20.7	χ^2_{15}= 25.00	Do Not Reject H_0
H_0: R_i is Ante-Dependence H_A: R_i is Unstructured	AN = 11 UN = 21	60.5	χ^2_{10}= 18.31	Reject H_0
H_0: R_i is Spatial H_A: R_i is Unstructured	SP = 2 UN = 21	77.9	χ^2_{19}= 30.14	Reject H_0

**Level of Significance of $\alpha = 0.05$*

The likelihood ratio tests suggest that, of the covariance structures considered, the compound symmetric approach and the Toeplitz approach are the most favorable candidates. However, as noted by Littell et al. (2006), the Toeplitz covariance structure is appropriate only when the repeated measurements are equally spaced chronologically. While it is true that the measures within one round were equally spaced (i.e., equal to a one day lapse in time across all locations), the lapse in time between rounds was not the same for all locations. Hence selecting the Toeplitz covariance structure over the compound symmetric structure, based on this likelihood ratio test, is deemed unwise.

There are three information criteria that may be used to identify which co-variance matrix is best (Littell et al. 2006). They are the Akaike Information Criterion

(AIC), the finite-population corrected criteria developed by Burnham and Anderson (Burnham and Anderson 1998) also known as the AICC, and the Schwarz's Bayesian Information Criterion (BIC). The calculations for these information criteria may be found in Appendix 6. The model (and hence the covariance structure used for it) that minimizes AIC or BIC is preferred (Littell et al. 2006).

Table 5 shows the information criteria for each of the covariance structures tested.

Table 5: Information Criteria Results for Covariance Matrix Structures

Approach	-2* Log Likelihood	AIC	AICC	BIC
Unstructured	7100.8	7244.8	7271.0	7439.3
Compound Symmetric	7124.9	**7230.9**	**7244.5**	**7374.0**
First Order Autoregressive	7177.3	7283.3	7296.9	7426.4
Toeplitz	7121.5	7235.5	7251.4	7389.4
Ante-Dependence	7161.3	7285.3	7304.3	7452.7
Spatial	7178.7	7284.7	7298.4	7427.9

Looking at this table, the compound symmetric approach represents the minimum value for each of these criteria and hence would be the covariance matrix of choice using these criteria.

Lastly, Figure 2 shows a graphical comparison of the covariance estimates versus lag in time for the unstructured (UN), compound symmetric (CS), first order autoregressive AR(1), Toeplitz (TOEP), and ante-dependence (ANTE) covariance approaches. (The covariance estimates for these approaches may be found in Appendix 4.) This graph suggests that the compound symmetric approach for the within subject errors is appropriate because it tracks closest to the covariance estimates from the unstructured approach.

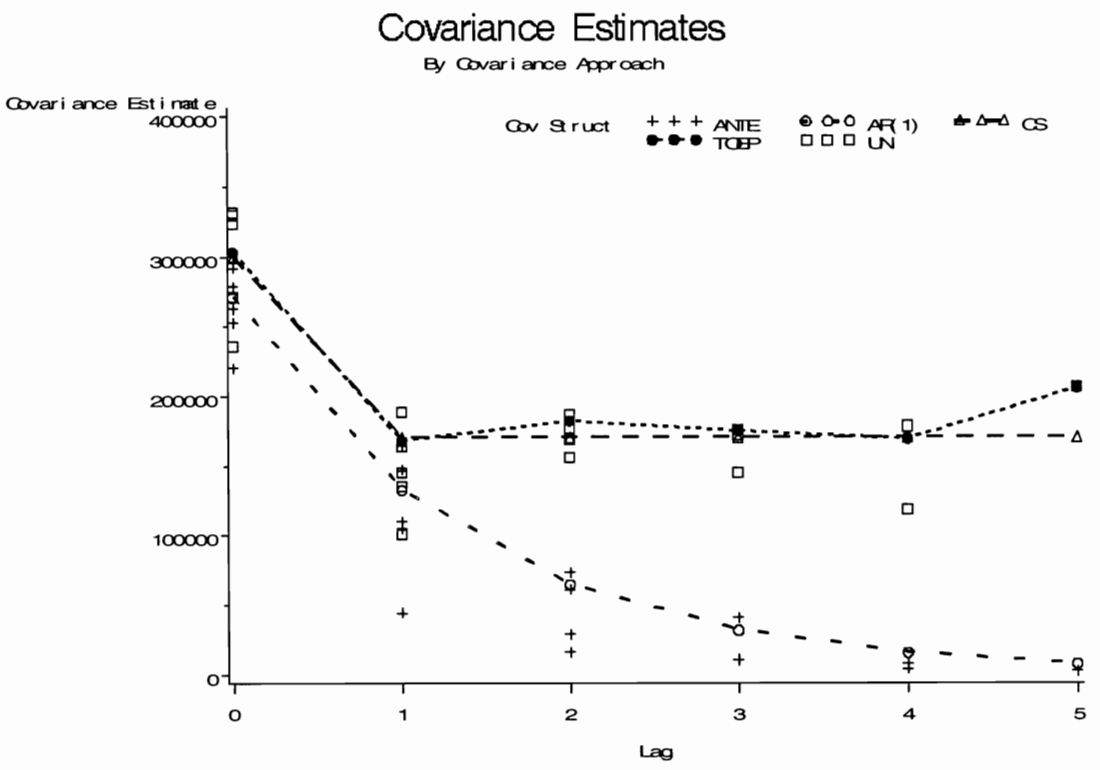

Figure 2: Lag Plot Comparison of Covariance Matrix Structure Approaches

Based on the graphical, likelihood ratio, and information criteria comparisons, and the fact that the repeated measurements were not equally spaced chronologically (hence, despite its likelihood ratio test result, the Toeplitz structure is not appropriate), it was decided that the compound symmetric covariance structure was the most appropriate covariance matrix structure to use in the ensuing model building effort.

Determination of Model Predictor Factors

To build the 24-hour creatinine model for this study, the MIXED procedure of the statistical software package SAS® (release 9.1.3) was used. Under the assumption of joint

multivariate normality for the random effects (ν) and the error term (ε) in the mixed model, this procedure uses the Restricted Maximum Likelihood (REML) methodology (as its default) to estimate the covariance parameters (i.e., the $\boldsymbol{G}$ and the $\boldsymbol{R}$ matrices) (Khattree and Naik 1999). As shown in Appendix 7, the covariance parameter estimates ($\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{R}}$) can then be used to obtain the fixed effect estimates via generalized least squares estimates (Littell et al. 2006).

A backwards elimination technique outlined by Kutner et al. (2005) was used to actually build the model. Beginning with a model containing all potential predictor variables, the predictor with the largest p-value was identified. If this 'maximum' p-value was greater than a predetermined value (in this case 0.05 was used), the predictor variable was dropped from the model. A new model containing all the predictor variables (except the one that was dropped) was then built. The process of identifying the predictor with the largest p-value and checking the p-value against the predetermined value was repeated until only those variables with a p-value less than the predetermined value remained.

As noted earlier, the potential explanatory factors of age, height, etc. were placed in the matrix of constants that describe the structure of the study with respect to fixed effects (i.e. the $\boldsymbol{X_i}$ matrix in the mixed model). The test statistic used for the p-value comparison in the backwards elimination was the resultant F-test statistic from the Type 3 test for the fixed effects. Based on the earlier covariance structure analysis, the compound symmetric structure was used for this model. The Kenward-Roger correction for the denominator degrees of freedom (Kenward and Roger 1997) for the Type 3 F-test

statistic was used to handle the possible bias introduced by using this non-independent covariance structure (Littell et. al 2006).

Location was not included in the initial model. As shown in Appendix 8, profile plots of 24-hour creatinine level, by location and day of measurement, did not show any trends by location and therefore did not provide any reason to suspect there was a location effect on the creatinine measurements. A multiple comparison of the mean creatinine measurements, by location, and box plots of creatinine measurements, by location, confirmed that there were no differences due to location.

The results for the initial model (with all possible predictor variables in it) may be found in Appendix 9. Predictor variables such as height, weight, BMI, and age at time of collection are considered to be covariates in the model. Predictor variables, to indicate that the subject is a smoker, takes a prescription medication, has a medical condition, etc., are represented as classification variables in the model. Based on the study questionnaire completed by the subjects, there are up to three possible levels for these classification variables: a value of '0' implies a negative response, a value of '1' implies a positive response, and a value of '9' implies 'no response.'

Possible interaction terms also were included in this initial model. However few interaction terms could be assessed by the software due to the sparseness of the data. The only non-sparse interactions included the kidney disease*kidney medicine interaction, the high blood pressure*blood pressure/ cholesterol medicine interaction, the diabetes*diabetes medicine interaction and the diet*Atkins diet interaction.

The final fitted model included the predictors of Body Mass Index, height and the classification variables for diabetes, for allergies, for a medical condition that affects kidney function (e.g., high blood pressure, glomerulonephritis), for the usage of a creatine supplement and for the taking of an anti-inflammatory medication. These results, along with the estimates for the fixed effects in the model, may be found in Appendix 10. (The SAS code for the final model may be found in Appendix 11). None of the interaction terms remained in the final model. The final model may be expressed as:

$$creatinine = -2457.5515 + 37.6721*BMI + 13.3569*height + diabetes + allergies + medcondition + creatine\_supp + anti\_inflammatory$$

where: BMI = Body Mass Index (kilograms per squared meter)

height = subject height (in centimeters)

diabetes = -496.5500 if subject had diabetes

allergies = -1135.3851 if subject did not have allergies

= -1430.2188 if subject had allergies

medcondition = 1893.0083 if subject did not have a medical condition that affected kidney function

= 1877.4174 if subject did have a medical condition

creatine_supp = 8.2967 if subject did not use creatine supplements

= -517.7042 if subject did use creatine supplements

anti_inflammatory = -599.3979 if subject did use an anti-inflammatory

Due to the compound symmetry, the intra-class correlation coefficient, which is the correlation between any two measures on the same subject, may be calculated as

$\rho = \frac{\sigma_S^2}{\sigma^2} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_B^2}$, where σ_S^2 is equivalent to the covariance between any two measures on the same subject and σ_B^2 is the residual variance component (Littell et al. 2006). In the final model's SAS output found in Appendix 10, σ_S^2 is labeled as covariance parameter 'CS' (and is equal to 148567) and σ_B^2 is labeled as covariance parameter 'Residual' (and is equal to 146147). Hence, this covariance structure gives an estimate for the error variance of $\sigma^2 = \sigma_S^2 + \sigma_B^2 = 294714$ and an intra-class correlation coefficient of $\rho = 0.5041$.

Residual Analysis and Influence Diagnostics

Residuals are used to examine both model assumptions and to detect possible outliers (Schabenberger 2004). To validate the assumptions for the error term, a residual analysis was performed on the residuals emanating from the final model.

For the mixed model, there are marginal residuals and conditional residuals. Schabenberger (2004) defines the marginal residual (r_{mi}) as "the difference between the observed data and the estimated marginal mean ($r_{mi} = y_i - x_i'\hat{\beta}$), and a conditional residual (r_{ci}) as the difference between the observed and the predicted value of the observation ($r_{ci} = y_i - x_i'\hat{\beta} - z_i'\hat{v}$)." He notes that in a mixed model without random effects, which is the model used by this thesis, the marginal and conditional residuals are the same (Schabenberger 2004).

Though calculated by the SAS® software, the raw marginal residuals, (r_{mi}), are not well suited to examine mixed model assumptions because their variances could differ (Schabenberger 2004). Studentized residuals address this unequal variance concern by dividing a raw residual by an estimate of its standard deviation (Schabenberger 2004). Studentized residual plots may be found in Appendix 12. A review of these plots suggested there were no violations of the assumptions surrounding the error term $\boldsymbol{\varepsilon} \sim \text{Normal}(\boldsymbol{0},\boldsymbol{R})$.

Using a yardstick of ±2.0 (Schabenberger 2004) for the Rstudent values (where Rstudent values are the externally studentized residuals), 23 observations were flagged as potential outliers (spanning 13 subjects). A review of these subjects' data indicated that five subjects acknowledged missing the collection of one or more voids for that day. The seven observations for these five subjects are flagged with an 'm' in the table of possible outliers found in Appendix 13.

Leverage values also were examined and sixteen observations were considered high leverage points (see Appendix 13). In a mixed model setting, leverage values should be not be interpreted as a measure of how unusual observations are in the original regressor space, but in terms of how unusual they are in the transformed space for the mixed model (Schabenberger 2004). However, it was noted that two of the suspected outliers (Subject 28 day 3 and Subject 127 day 4) did have the maximum values for two of the predictor variables (height and BMI respectively).

The influence diagnostics available in SAS® release 9.1.3 are considered experimental. There are two approaches (non-iterative and iterative) implemented by

SAS for calculating these diagnostics. Both are based on the concept of removing one or more observations from the analysis, computing new estimates for the model parameters, and then comparing the 'full' and 'reduced' data estimates to determine how much influence the observation(s) exerted on the analysis (Schabenberger 2004). The non-iterative approach holds the covariance parameters fixed and only re-calculates the estimates for $\boldsymbol{\beta}$ and σ^2 (Littell et al. 2006). The iterative approach re-computes the covariance parameters as well as the estimates for $\boldsymbol{\beta}$ and σ^2 (Littell et al. 2006). The iterative approach is recommended (Schabenberger 2004; Littell et al. 2006) and it was the approach used for the final model.

Another feature available with the SAS influential diagnostics is the ability to remove, from the 'full data set,' more than one observation at a time. One can remove a set of observations (e.g., all the observations for one subject) and then calculate the influence diagnostics for that set of observations (as one unit). This study focused on the single observation influence diagnostics only. These diagnostics may be found in Appendix 14 and are described below.

For a linear mixed model using the REML method, the Restricted Likelihood Distance (RLD) is used as a measure of overall influence (Cook and Weisberg 1982). The RLD is calculated by removing an observation (or multiple observations) from the 'full data set', computing the parameter estimates for the 'reduced data set,' and then assessing the height of the original restricted likelihood surface at the deleted data parameter estimates (Littell et al. 2006).

A plot showing the RLD measures for the observations in the final model analysis may be found in Appendix 14. This plot indicates that four of the observations for subject 45 (days 1, 3, 4, and 6), three observations for subject 63 (days 3, 4, and 6), and day 5 for subject 132 are all considered highly influential by this measure.

The DFFITS statistic was examined to determine the effect an observation had on the fitted values (Schabenberger 2004). The plot, also in Appendix 14, indicates that the day 3 and day 4 measures for subjects 45 and 63 had a large influence on the fit of the model. The Cook's Distance measure also was examined. The plot of Cook's Distance by observation (in Appendix 14) indicates that day 3 and day 4 measures for subjects 45 and 63 had a large influence on the fixed effect parameters ($\hat{\boldsymbol{\beta}}$).

Lastly, examination of the COVRATIO values was done to assess influence on the precision of the estimates (Littell et al. 2006). A plot of these values may be found in Appendix 14. A COVRATIO value of 1.0 implies no influence on the precision of the estimates, a value larger than 1.0 indicates increased precision for the 'full data' estimates, and a value less than 1.0 indicates higher precision for the 'reduced data' estimates (Littell et al. 2006). Because their COVRATIO values were much larger than 1.0, the day 3 and day 4 measures of subjects 45 and 63 were flagged as increasing the precision of the estimates.

As noted by Schabenberger (2004), the "goal of influence analysis is not primarily to identify observations for deletion, but rather to determine which cases are influential and the manner in which they are important." Using the table of outliers found in Appendix 13 and the influence diagnostics in Appendix 14, it is possible to identify

suspicious subject creatinine measurements and thereby identify potential misreported urine collections.

Model Validation

Validation of the final model was done using a subset of the pilot phase data (from Richmond, Virginia). The pilot phase data, like the data used to build the model, consisted of three rounds. The first round was five consecutive days with two 12-hour collections for each of the five days. The second and third rounds (several months later) consisted of 24-hour collection periods for two consecutive one week periods (resulting in a total of seven days for round two and the next seven days for round three).

As noted earlier, the model building data were from the follow-on phase of the study. The follow-on phase had three rounds of collections and the start of each round was separated by several months. However, only two consecutive days of collections existed in each round of this phase.

To create a model validation data set from the pilot phase data that closely resembled the lapses in time between the three rounds of the follow-on phase, days one and two from round one were included in the validation data set. The first two days from round two were also included in the validation data set. From the seven days in the third round, one day was selected at random (day i); both day i and day $(i + 1)$ were then included in the validation data set. Figure 3 shows which days of the pilot phase data were included in the validation set.

	Start Round 1					Start Round 2							Start Round 3						
Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	X	X				X	X								X	X			

Figure 3: Days of Pilot Phase used in Validation Data Set
An 'X' denotes day was included in the validation data set. Day 15 was the randomly selected day from Round 3 (i.e., day i = day 15).

For the 21 subjects in the pilot phase of the study, a maximum of 126 observations in the validation data set were possible. Of the possible observations, 91 had valid creatinine values, heights, and weights and thus were usable in the model validation phase. Table 6 shows the descriptive statistics for the model validation observations.

Table 6: Descriptive Statistics for Model Validation Data

	N	Mean	Std Dev	Min	Max
Creatinine (mg/day)	91	1720.000	612.038	472.640	3305.000
Height (cm)	91	179.865	5.529	170.180	193.040
Weight (kg)	91	91.646	14.808	72.575	134.717
BMI (kg/m^2)	91	28.279	4.113	23.571	40.281
Age at Collection (yrs)	91	32.538	5.502	25.000	48.000

Predicted 24-hour creatinine values, using the final model's parameter estimates, were calculated for each observation in the model validation data set. A comparison of the linear association between the predicted and actual 24-hour creatinine values (using the correlation coefficient) was done for both the model building data set and the model validation data set.

The hypothesis tested was $\mathbf{H_0}$: $\rho_{actual,predicted} = 0$ versus $\mathbf{H_A}$: $\rho_{actual,predicted} \neq 0$, where $\rho_{actual,predicted}$ is the population correlation coefficient for the predicted and actual creatinine values. A test statistic of $r_{actual,predicted} * \sqrt{\frac{n-2}{1-r^2_{actual,predicted}}}$, where $r_{actual,predicted}$ is the sample correlation coefficient for the predicted and actual values, has a t distribution with n-2 degrees of freedom (Anderson et al. 1993).

For the model building data, $r_{actual,predicted}$ was equal to 0.42763; for the model validation data, $r_{actual,predicted}$ was slightly lower with a value of 0.32169. It was concluded that both of these correlation coefficients were significantly different from zero, as the null hypothesis that $\rho_{actual,predicted}$ was equal to zero was rejected for both the model building data (with a p-value < 0.0001) and for the model validation data (with a p-value of 0.0019).

A plot of the observed 24-hour creatinine values versus the predicted creatinine values, for both data sets, is shown in Figure 4. If a model predicted accurately 100% of the time (i.e., the predicted value was equal to the actual value for all observations), then one would expect the points on a scatter plot of predicted values versus actual values to fall along a straight line whose slope is equal to 1 and whose y-intercept is the origin. The plot in Figure 4, along with the correlation coefficients for predicted versus actual creatinine values, suggests that the model's predictive behavior for the validation data is similar to its predictive behavior from the model building data.

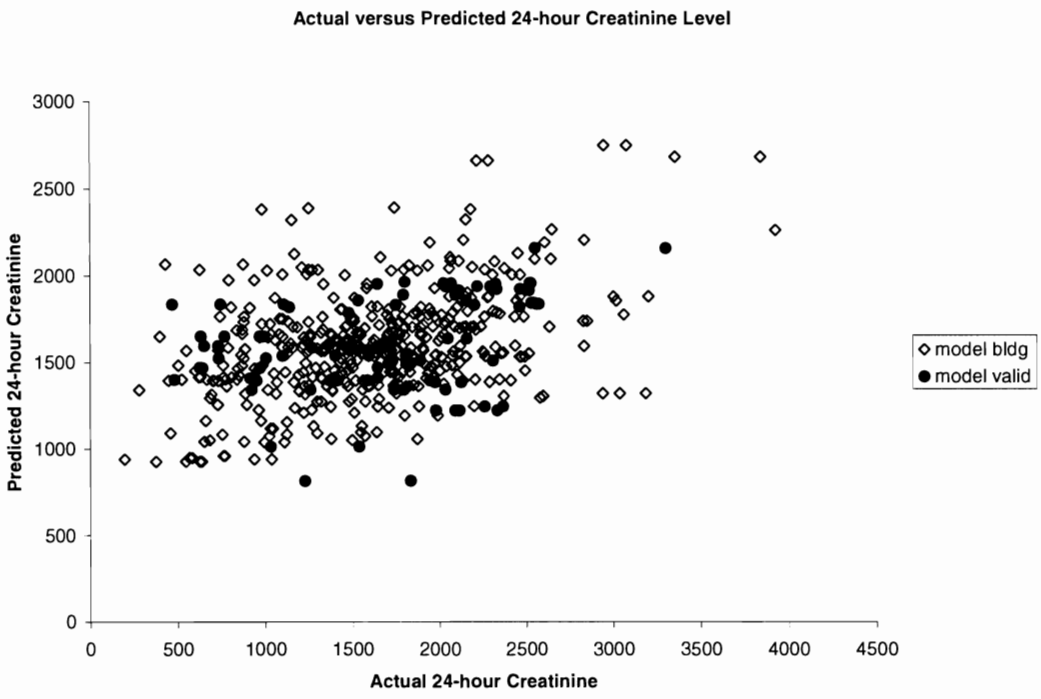

Figure 4: Predicted versus Actual 24-hour Creatinine Levels

Model Comparisons

The initial step for the model comparisons was a review of the participant age, height, and weight information for the studies mentioned earlier in the literature review of this thesis. As shown in Table 1 and Table 3, the ages of the male study participants for the NIOSH-CDC study are in the same range as the ages of participants from these earlier studies.

Applying Chebyshev's Theorem, where at least 75% of the items in any data set are within ± 2 standard deviations of the mean (Anderson et al. 1993), comparisons of the studies' subject heights and weights were done. As shown in the interval plots found in Figures 16 and 17 in Appendix 16, the range of heights and weights of the NIOSH-CDC study participants are similar to those from the Turner and Cohn (1975) and Harris et al. (2000) studies. The range of heights and weights of the participants in the Moriyama et al. (1988), Kamata and Tochikubo (2002), Tanaka et al. (2002), and Penie et al. (2003) studies are slightly lower than those for the NIOSH-CDC study participants.

Next, the models by Turner and Cohn (1975), Moriyama et al. (1988), Kawasaki et al. (1991), Harris et al. (2000), Tanaka et al. (2002), and Penie et al. (2003), and their parameter estimates, were used to calculate predicted values for the model building data set from the NIOSH-CDC "Pesticide Dose Monitoring in Turf Applicators" study. The models of Kamata and Tochikubo (2002) and of Jones et al. (1996) were not included in

this comparison due to the unavailability of subjects' lean body mass measurements (for the Kamata model) and due to all of the original study's subjects being CAPD patients (for the Jones model). A comparison of the other models' predictive capabilities versus the predictive capability of the model developed for Objective 1 of this thesis was performed.

Some of the earlier models were built using a single measurement of creatinine for each subject (e.g., Tanaka et al.). A few of the other earlier models (Harris et al., Turner et al., Kawasaki et al., Moriyama et al.) had repeated creatinine measurements for the same subject over time. The Harris et al. and the Moriyama et al. models first aggregated the observations for one subject into a summary measure (e.g., mean) and then used this summary measure for the response variable.

All of the models mentioned above used an ordinary least squares regression approach in their model building process. This approach assumed that the creatinine measurements (even if there were multiple measurements for the same subject) were uncorrelated and had a constant error variance. Vonesh and Chinchilli (1997), Khattree and Naik (1999), Vittinghoff et al. (2005), and Littell et al. (2006) note that repeated measures, taken on the same subject over time, very often are correlated. Hence, using an ordinary least squares model, in this situation, and its assumption of uncorrelated observations for one subject would not be appropriate.

In contrast, the model built for Objective 1 of this thesis identified the covariance structure for the repeated measures data of this study to be compound symmetric and therefore took into account that the observations for each subject were correlated. Usage

of this covariance structure also allowed the estimation of both the within subject covariance and the intra-class correlation coefficient (for the repeated measurements on the same subject).

To compare the predictive capability of the regression models (that assumed the uncorrelated observations) to the model developed for Objective 1 (that used a covariance structure that allowed observations within each individual to be correlated), the mean squared prediction error (MSPR) was calculated. MSPR is calculated by:

$$MSPR = \left[\frac{\sum_{i=1}^{n^*} (y_i - \hat{y}_i)^2}{n^*} \right],$$

where y_i is the value of the response in the i^{th} validation case, $\hat{y}_i$ is the predicted value of the i^{th} case based on the existing regression model, and n^* is the number of cases in the 'validation data set' (Kutner et al. 2005). For this comparison effort, the 'validation data set' in the above MSPR definition is the model building set for this NIOSH-CDC study. The SAS code for calculating the predicted values ($\hat{y}_i$) for each comparison model is found in Appendix 15. If this MSPR is close to the Mean Square Error (MSE) of the regression model derived from the original model building data set (used by Kawasaki, Turner, or whomever), then the MSE for the original model is not seriously biased and it gives an appropriate indication of the original model's predictive ability (Kutner et al. 2005).

Using the NIOSH-CDC model building data that were used in Objective 1 of this thesis, MSPR values were calculated for the fitted values emanating from the regression models by Turner and Cohn (1975), Moriyama et al. (1988), Kawasaki et al. (1991),

Harris et al. (2000), Tanaka et al. (2002), and Penie et al. (2003). MSPR was also calculated for the model built for Objective 1. The MSE values for the original regression models built by Kawasaki et al., Turner et al., etc. were not published, hence making the comparison of their MSPR values to their original MSE values impossible. However, an ordering of the MSPR values is possible and an ordered list (in descending order) of the MSPR results is shown in Table 7.

The model, listed as 'Objective 1 Revised,' was developed using the model building data set from the NIOSH-CDC study data. It was fitted using the REML method for a mixed model and a compound symmetric covariance structure. The Kawasaki, Moriyama, Tanaka, Harris, Turner, and Penie models listed in this table did not have data for all of the potential predictors (e.g., to indicate usage of creatine supplements, having allergies, etc.) available for their study subjects as did the final model for Objective 1. The model listed as 'Objective 1 Revised' was built to illustrate the predictive performance of a model using a compound symmetric covariance structure with only the same predictors that the comparison models had available to them at the time of their studies (i.e., subject height, weight, and age at collection).

Intuitively, a smaller MSPR is better. The MSPR provides a model specific summary measure of the model's 'error in prediction' for all of the applicable observations in the NIOSH-CDC study model building data set. From Table 7, it is clear that the model developed for Objective 1 has the smallest MSPR for this NIOSH-CDC study data.

Table 7: MSPR Results for Models

Model	Predictors Used	MSPR using the NIOSH-CDC Study Data
Kawasaki	height, weight, age	658860
Moriyama	height, weight, age	542089
Tanaka	height, weight, age	430680
Harris	gender, weight, age	429227
Turner	height, weight, age	426030
Penie	height	393139
Objective 1 Revised	height, BMI*	289638
Objective 1 Model	height, BMI*, diabetes, creatine supplement, anti-inflammatory, medical condition, allergies	279184

**Body Mass Index (BMI) was calculated from height and weight.*

A second measure to assess the predictive capability of these published models, versus the predictive capabilities of the model from Objective 1 was the correlation coefficient between the predicted creatinine values and the actual creatinine values for the NIOSH-CDC study model building data. Table 8 displays the correlation coefficients for these models. This table indicates that the model from Objective 1 has the largest correlation between the predicted creatinine values and the actual creatinine values. Therefore, both the correlation results and the ordering of the MSPR measures suggest that the model built for addressing Objective 1 has better predictive capabilities of a 24-hour creatinine level than the other models listed.

Table 8: Correlation Coefficients ($r_{actual,predicted}$) for Predicted and Actual Creatinine Values by Model

Model	$r_{actual,predicted}$*
Harris	0.17906
Penie	0.20921
Turner	0.35315
Kawasaki	0.36488
Tanaka	0.36544
Moriyama	0.37827
Objective 1 Revised	0.37875
Objective 1 Model	0.42763

**In testing H_o: $\rho_{actual,predicted} = 0$ versus H_A: $\rho_{actual,predicted} \neq 0$, all are significantly different than zero at a level of significance of $\alpha = 0.05$ (p-value <0.0001).*

Conclusions

Creatinine is a metabolic waste product, removed from the blood by the kidneys, and excreted in the urine. The measurement of creatinine is used in the assessment and monitoring of many medical conditions as well as in the area of pesticide research where the 'completeness' of the urine collections is instrumental to the analysis of the absorbed dosage of pesticide. If complete urine samples were not provided by study participants, the absorbed doses of pesticide will be underestimated. Hence, the ability to identify potentially 'incomplete' or suspect urine samples, is extremely important in this type of research. By identifying potential outliers and influential data observations, statistical models to predict 24-hour creatinine can be used to help identify suspicious urine samples provided by pesticide study participants.

As noted earlier in this thesis, numerous studies have been conducted and there are multiple statistical based models that can be used to predict 24-hour urinary creatinine levels. The models investigated for this thesis share a common ground in that they were all developed using a ordinary least squares regression approach. A key element of this approach is the assumption that the observations used to build the model are uncorrelated. However, many of these studies had repeated creatinine measurements for each of their subjects. As noted earlier by Khattree and Naik (1999), Vittinghoff et al. (2005), and Littell et al. (2006), repeated measures on the same subject frequently are correlated and

therefore, this would violate the uncorrelated assumption used by the earlier studies' regression analyses.

The approach used in this thesis, to build a predictive model for 24-hour urinary creatinine, was a mixed model methodology. This method allowed for the identification and the specification of a covariance structure that would allow the observations within one individual to be correlated. The resultant model contained the covariates of height and body mass index, and the classification variables to indicate if the study participant had allergies, diabetes, or another medical condition that affected kidney function. It also included the classification variables that indicated if the participant was taking an anti-inflammatory prescription medication or was using a creatine dietary supplement. Using this model, specific subjects' observations were flagged as 'suspect' (i.e., misreported) urine collections.

Caution should be taken in interpreting this model's results as 'causal.' The data for the NIOSH-CDC "Pesticide Dose Monitoring in Turf Applicators" study were not collected in a designed experiment fashion. Random assignment of the classification variables' levels to the study participants could not occur. The variables of interest were data collected about the study participants at multiple points in time during their participation in the study. Hence, this study is a prospective observational study. It cannot directly demonstrate a cause and effect relationship between the participant's data and the participant's response (the 24-hour total creatinine level); it can only suggest an association between this response and the participant's data (Kutner et al. 2005).

However, it is important to note that the predictive performance of this model, as evaluated by the MSPR (mean square prediction error) and the correlation coefficient (for the predicted versus observed creatinine values), for the NIOSH-CDC pesticide data, was better than the predictive performance of any of the competing models by Turner and Cohn. (1975), Moriyama et al.(1988), Kawasaki et al. (1991), Harris et al. (2000), Tanaka et al. (2002), and Penie et al. (2003).

Why the difference in predictive performance of the final model built for this thesis versus the predictive performance of the other models? Two possible reasons come to mind. The first reason focuses on the similarity in age and body size (i.e., height and weight) of the NIOSH-CDC study participants to the older studies' participants. Kutner et al. (2005) recommends caution on inferences using the fitted regression function for values of predictor variables that are far beyond the range of the original predictors used when the model was built. As noted in the Model Comparisons chapter of this thesis, the range of ages was comparable for both the NIOSH-CDC study participants and the other studies' subjects. The body size characteristics for the NIOSH-CDC participants were slightly larger than the body size characteristics for the participants in the Moriyama et al. (1988), Tanaka et al. (2002), and Penie et al. (2003) studies. This may explain some of the difference in predictive capabilities between these models and the final model built for this thesis. However, the NIOSH-CDC study participants were more similar in height and weight to those for the Turner and Cohn (1975) and Harris et al. (2000) studies, so dissimilarity in body size is not as likely to be the reason for the difference in predictive capability.

The second possible reason for the improved predictive capability of the final model, versus the others, is that all the other models assumed that the measurements within one individual were uncorrelated. Littell et al. (2006) noted that ignoring important correlation by using a 'too simple' covariance model can risk underestimating standard errors. The work of Guerin and Stroup (2000), in documenting the effects of various covariance modeling decisions using the SAS mixed model procedure for repeated measures data, also showed that "inference is severely compromised by a poor choice of covariance model." The improved predictive capability of this thesis' final model, over the others, may very well be attributed, at least in part, to its identification and usage of a covariance structure that allowed the repeated measurements for any one individual to be correlated.

References

Anderson, D.R., D.J. Sweeney, T.A. Williams. *Statistics for Business and Economics*. 5th ed. St. Paul, MN: West Publishing Company, 1993.

Boeniger, M.F., L.K. Lowry, J. Rosenberg. "Interpretation of Urine Results Used to Assess Chemical Exposure with Emphasis on Creatinine Adjustments: A Review." *American Industrial Hygiene Association Journal*, Volume 54, Number 10 (1993): 615-627.

Burnham, K.P., D.R. Anderson. *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag, 1998.

Cook, R.D., S. Weisberg. *Residuals and Influence in Regression*. New York: Chapman and Hall, 1982.

Elliott, P. "The INTERSALT study: an addition to the evidence on salt and blood pressure, and some implications." *Journal of Human Hypertension*, Volume 3 (1989): 289-298.

Elliott, P., A. Dyer, R. Stamler, on behalf of the INTERSALT Cooperative Research Group. "The INTERSALT study: results for 24 hour sodium and potassium, by age and sec." *Journal of Human Hypertension*, Volume 3 (1989): 323-330.

Fuller, N.J., M. Elia. "Factors influencing the production of creatinine: implications for the determination and interpretation of urinary creatinine and creatine in man." *Clinica Chimica Acta*, Volume 175 (1988): 199-210.

Guerin, L., W. Stroup. "A Simulation Study to evaluate PROC MIXED Analysis of Repeated Measures Data." *Proceedings of the Twelfth Annual Conference on Applied Statistics in Agriculture*. Manhattan: Kansas State University (2000).

Gurka, M.J. "Selecting the Best Linear Mixed Model Under REML." *The American Statistician*, Volume 60, Number 1 (February 2006):19-26.

Harris, S.A. "The Development and Validation of a Pesticide Dose Prediction Model." PhD dissertation, University of Toronto, 1999.

Harris, S.A., J.T. Purdham, P.N. Corey, A.M. Sass-Kortsak. "An evaluation of 24-hour urinary creatinine excretion for use in identification of incomplete urine collections and adjustment of absorbed dose of pesticides." *American Industrial Hygiene Association Journal*, Volume 61, Number 5 (2000): 649-657.

Henderson, C.R. *Applications of Linear Models in Animal Breeding*. Guelph: University of Guelph Press, 1984.

Heymsfield, S.B., C. Artega, C. McManus, J. Smith, S. Moffitt. "Measurement of muscle mass in humans: validity of the 24-hour creatinine method." *The American Journal of Clinical Nutrition*, Volume 37, Number 3 (March 1983): 478-494.

INTERSALT Cooperative Research Group. INTERSALT: an international study of electrolyte excretion and blood pressure. Result for 24 hour urinary sodium and potassium excretion. *British Medical Journal*, Volume 297 (1988):319-328.

Japan Health Promotion and Fitness Foundation. Kenkou Nippon 21. Kouken Shuppan: Tokyo, 2000 (in Japanese).

Jennrich, R.I., M.D. Schluchter. "Unbalanced repeated-measures models with structured covariance matrices." *Biometrics,* Volume 42, Number 4 (1986): 805-820.

Jones, C.H., C.G. Newstead, E.J. Will. "Estimation of total daily creatinine clearance in CAPD from serum creatinine concentration." *Peritoneal Dialysis International*, Volume 17, Number 3 (May-June 1997): 250-254.

Johnson, C.A., A.S. Levey, J. Coresh., A. Levin, J. Lau., G. Eknoyan. "Clinical practice guidelines for chronic kidney disease in adults: Part II. Glomerular filtration rate, proteinuria, and other markers." *American Family Physician*, Volume 70, Number 6 (September 15, 2004): 1091-1097.

Kamata, K. and O. Tochikubo, "Estimation of 24-h urinary sodium excretion using lean body mass and overnight urine collected by a pipe-sampling method." *Journal of Hypertension*, Volume 20, Number 11 (November 2002): 2191-2197.

Katzung, B.G. *Basic & Clinical Pharmacology*. 9th ed. New York: Lange Medical Books/McGraw Hill, 1989.

Kawasaki, T., K. Uezono, K. Itoh, M. Ueno. "Prediction of 24-hour urinary creatinine excretion from age, body weight and height of an individual and its application." *Nippon Koshu Eisei Zasshi*, Volume 38, Number 8 (August 1991): 567-574 (in Japanese).

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

Kesteloot, H.E., J.V. Joossens, "Relationship Between Dietary Protein Intake and Serum Urea, Uric Acid and Creatinine, and 24-Hour Urinary Creatinine Excretion: The BIRNH Study." *Journal of the American College of Nutrition*, Volume 12, Number 1 (1993): 42-46.

Khattree, R., D.N. Naik. *Applied Multivariate Statistics with SAS® Software*. 2nd ed. Cary, N.C.: SAS Institute, Inc., 1999.

Kutner, M.H., C.J. Nachtsheim, J. Neter, W. Li. *Applied Linear Statistical Models*. 5th ed. Boston: McGraw-Hill Irwin, 2005.

Letteri, J.M., S.N. Asad, R. Caselnova, K.J. Ellis, S.H. Cohn. "Creatinine excretion and total body potassium in renal failure." *Clinical Nephrology*, Volume 4, Number 2 (August 1975): 58-61.

Littell, R.C., G.A. Milliken, W.W. Stroup, R.D. Wolfinger, O. Schabenberger. *SAS® for Mixed Models*. 2nd ed. Cary, N.C.: SAS Institute, Inc., 2006.

Moriyama, M., H. Saito, A. Nakano, S. Funaki, S. Kojima. "Estimation of Urinary 24-hr Creatinine Excretion by Body Size and Dietary Protein Level: A Field Survey Based on Seasonally Repeated Measurements for Residents Living in Akita, Japan." *Tohuku Journal of Experimental Medicine*, Volume 156 (1988): 55-63.

Penie, J.B., S.S. Porben, Y.D.C. Silverio. "Local Reference Intervals for the Excretion of Creatinine in Urine for an Adult Population." *Nutrición Hospitalaria*, Volume 18 (2003): 65-75 (in Spanish).

Poortmans, J.R., H. Auquier, V. Renaut, A. Durussel, M. Saugy, G.R. Brisson. "Effect of short-term creatine supplementation on renal responses in men." *European Journal of Applied Physiology & Occupational Physiology*, Volume 76, Number 6 (1997): 566-567.

Proctor, D.N., P.C. O'Brien, E.J. Atkinson, K.S. Nair. "Comparison of techniques to estimate total body skeletal mass in people of different age groups." *American Journal of Physiology*, Volume 277, Number 3 Pt 1 (September 1999): E489-E495.

SAS Institute, Inc. *SAS® Software for Windows*, Release 9.1.3 Service Pack 3. Cary, N.C.: SAS Institute, Inc., 2003.

Schabenberger, O. "Mixed Model Influence Diagnostics." *Proceedings of the Twenty-ninth Annual SAS® Users Group International Conference*, Cary, N.C.: SAS Institute, Inc., (2004): 1-17.

Searle, S.R. *Linear Models for Unbalanced Data*. New York: Wiley, 1987.

Selby, S.M. *CRC Standard Mathematical Tables*. Cleveland, OH: CRC Press Inc., 1975.

Tanaka, T., T. Okamura, K. Miura, T. Kadowaki, H. Ueshima, H. Nakagawa, T. Hashimoto. "A simple method to estimate populational 24-h urinary sodium and potassium excretion using a casual urine specimen." *Journal of Human Hypertension*, Volume 16, Number 2 (February 2002): 97-103.

Turner, W.J., S. Cohn. "Total Body potassium and 24-hour creatinine excretion in healthy males." *Clinical Pharmacology and Therapeutics*, Volume 18, Number 4 (1975): 405-412.

Vittinghoff, E., D.V. Glidden, S.C. Shiboski, C.E. McCulloch. *Regression Methods in Biostatistics : Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer, 2005.

Vonesh, E.F., V.M. Chinchilli. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker Inc., 1997.

Welle, S., C. Thornton, S. Totterman, G. Forbes. "Utility of creatinine excretion in body composition studies of healthy men and women older than 60-y." *The American Journal of Clinical Nutrition*, Volume 63 (1996): 151-156.

Wolfinger, R.D. "Covariance structure selection in general mixed models." *Communications in Statistics - Simulation and Computation*. Volume 22, Number 3 (1993):1079-1106.

Wolfinger, R.D. "Heterogeneous variance-covariance structures for repeated measures." *Journal of Agricultural, Biological, and Environmental Statistics*, Volume 1, Number 2 (1996): 205-230.

Appendix 1

Imputing of 12-hour Creatinine Values for Follow-on Phase Data

Linear regression models (using ordinary least squares) were developed using a best subsets approach. If three of the four 12-hour creatinine values were available for the subject, then all three values were initially considered in the subsets approach. Similarly, if only two of the four 12-hour creatinine values were available for the subject, then these two values were initially considered in the subsets approach. The resultant model selected to be used was the model that had all parameter estimates identified as significant (at $\alpha = 0.05$) and did not violate any of the assumptions associated with the linear regression model used. Table 9 shows which subjects' 12-hour creatinine values were imputed and what equation was used.

Table 9: Imputing Equations

Subjects	Day 3 am	Day 3 pm	Day 4 am	Day 4 pm	Impute Equation
38, 50, 71, 117, 122	X	X	X		day4_pm = 127.9721 + 4.2051*day3_am + 3.7477*day3_pm
83, 128		X	X	X	day3_am = 241.8805 + 4.4221*day4_am + 3.2657*day4_pm
41	X	X		X	day4_am = 350.7363 + 5.5381*day3_am
33, 44, 69	X		X	X	day3_pm = 327.5140 + 6.2250*day4_pm
35, 52, 54	X		X		day4_pm = 532.8776 + 3.1329*day4_am day3_pm = 510.6217 + 3.3952*day3_am
58, 109		X		X	day4_am = 629.8250 + 2.0998*day4_pm day3_am = 643.1244 + 2.2160*day3_pm
27			X		day4_pm = 532.8776 + 3.1329*day4_am

An 'X' indicates a 12-hour creatinine value was originally available.

Appendix 2

Covariance Matrix Structures' Descriptions

Note: All covariance matrices (Σ) shown here are symmetric so that only the upper triangle entries are given.

Independent: Within-subject error correlations are zero (i.e. repeated measures for a subject are uncorrelated) (Littell et al. 2006).

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 \\ & & & & 1 & 0 \\ & & & & & 1 \end{bmatrix} = \sigma^2 I$$

Where Σ is the covariance matrix for one subject and $\boldsymbol{I}$ is the 6x6 identity matrix.

Unstructured (UN): Each pair of measurements has its own unique correlation (Littell et al. 2006).

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} \\ & & \sigma_3^2 & \sigma_{34} & \sigma_{35} & \sigma_{36} \\ & & & \sigma_4^2 & \sigma_{45} & \sigma_{46} \\ & & & & \sigma_5^2 & \sigma_{56} \\ & & & & & \sigma_6^2 \end{bmatrix}$$

Where Σ is the covariance matrix for one subject

σ_i^2 is the variance for the i^{th} measurements

σ_{ij} is the covariance between the i^{th} and j^{th} measurements ($i \neq j$)

Compound Symmetric (CS): Is also known as the intra-class correlation structure (Khattree and Naik 1999). Compound Symmetric structure has equal variances at all times and equal covariances between observations on the same subject at all pairs of times ($Cov[Y_{ij}, Y_{ij'}] = \rho\sigma^2$) (Littell et al. 2006).

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho & \rho & \rho \\ & 1 & \rho & \rho & \rho & \rho \\ & & 1 & \rho & \rho & \rho \\ & & & 1 & \rho & \rho \\ & & & & 1 & \rho \\ & & & & & 1 \end{bmatrix}$$

Where Σ is the covariance matrix for one subject

σ^2 is the error variance

$\rho = \rho_{ij}$ is the correlation between the i^{th} and j^{th} measurements ($i \neq j$)

First Order Autoregressive (AR(1)): Is based on the idea that correlation between observations is a function of their lag in time—adjacent observations are more highly correlated than observations further apart in time (Littell et al. 2006). In an **AR(1)** covariance matrix, the correlation between adjacent measures is the same ρ regardless if it is the 1st and 2nd, the 3rd and 4th, or the 5th and 6th measures. The correlation between measures two time units apart is ρ^2, between three time units apart is ρ^3, etc. (Littell et al. 2006).

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & & 1 & \rho & \rho^2 & \rho^3 \\ & & & 1 & \rho & \rho^2 \\ & & & & 1 & \rho \\ & & & & & 1 \end{bmatrix}$$

Where Σ is the covariance matrix for one subject

σ^2 is the error variance

$\rho = \rho_{1,2}$ is the correlation between the first and second measurements

Toeplitz (TOEP): Is similar to First Order Autoregressive in that pairs of within subject errors separated by a common lag share the same correlation. However, errors *d* units apart have correlation ρ_d instead of ρ^d (Littell et al. 2006).

$$\Sigma = \sigma_o^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 \\ & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ & & 1 & \rho_1 & \rho_2 & \rho_3 \\ & & & 1 & \rho_1 & \rho_2 \\ & & & & 1 & \rho_1 \\ & & & & & 1 \end{bmatrix}$$

Where Σ is the covariance matrix for one subject

σ_o^2 is the error variance

ρ_i is the correlation for two measurements that are *i* time units apart

NOTE: AR(1) and **Toeplitz** models are generally inappropriate if the observation times are not equally spaced chronologically (Littell et al. 2006).

Ante-Dependence (ANTE): Allows for unequal spacing between measurements and changes in variance over time. The correlation between a pair of observations is the product of the correlations between adjacent times (between the observations) (Littell et al. 2006; Wolfinger 1996).

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 & \sigma_1\sigma_4\rho_1\rho_2\rho_3 & \sigma_1\sigma_5\rho_1\rho_2\rho_3\rho_4 & \sigma_1\sigma_6\rho_1\rho_2\rho_3\rho_4\rho_5 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho_2 & \sigma_2\sigma_4\rho_2\rho_3 & \sigma_2\sigma_5\rho_2\rho_3\rho_4 & \sigma_2\sigma_6\rho_2\rho_3\rho_4\rho_5 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho_3 & \sigma_3\sigma_5\rho_3\rho_4 & \sigma_3\sigma_6\rho_3\rho_4\rho_5 \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho_4 & \sigma_4\sigma_6\rho_4\rho_5 \\ & & & & \sigma_5^2 & \sigma_5\sigma_6\rho_5 \\ & & & & & \sigma_6^2 \end{bmatrix}$$

Where Σ is the covariance matrix for one subject

σ_i^2 is the variance for the measurements at time point *i*

σ_i is the standard deviation for the measurements at time point *i*

ρ_i is the correlation for two measurements (for one subject) that are *i* time units apart.

Spatial: Used for unequally spaced longitudinal measurements (Khattree and Naik 1999). The covariance between two measurements at times t_i and t_j is:

$COV[Y_{t_i} Y_{t_j}] = \sigma^2 \rho_{ij}^{d_{ij}}$ (Littell et al. 2006).

Where $d_{ij} = |t_i - t_j|$ is the time elapsed between the i^{th} and j^{th} repeated measure for a subject (Khattree and Naik 1999)

σ^2 is the error variance

ρ_{ij} is the correlation between the i^{th} and j^{th} measurements

$\rho_{ij} = 1$ for $i = j$ (Khattree and Naik 1999)

Appendix 3

SAS Code for Covariance Matrix Structure Tests

```
title1 'UNSTRUCTURED...ML RESULTS';
proc mixed data=my.nonrichmondfixed covtest method=ML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed;

model creatinine_sum =  time1 n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed
      / ddfm=kr  ;
repeated /subject=subjectid type=un
r=2 rcorr=2;
run;

title1 'COMPOUND SYMMETRIC   ML RESULTS';
proc mixed data=my.nonrichmondfixed covtest method=ML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed;

model creatinine_sum =  time1 n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed
      / ddfm=kr  ;
repeated /subject=subjectid type=cs
r=2 rcorr=2; run;
```

```
title1 'TOEPLITZ   ML RESULTS';
proc mixed data=my.nonrichmondfixed covtest method=ML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed;

model creatinine_sum =  time1 n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed
      / ddfm=kr  ;
repeated /subject=subjectid type=toep
r=2 rcorr=2;
run;

title1 AUTO REGRESSIVE1...ML RESULTS';
proc mixed data=my.nonrichmondfixed covtest method=ML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed;

model creatinine_sum =  time1 n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed
      / ddfm=kr  ;
repeated /subject=subjectid type=ar(1)
r=2 rcorr=2;
run;

title1 'ANTE DEPENDENCE...ML RESULTS';
proc mixed data=my.nonrichmondfixed covtest method=ML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed;

model creatinine_sum =  time1 n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
```

```
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed
      / ddfm=kr  ;
repeated /subject=subjectid type=ANTE(1)
r=2 rcorr=2;
run;

data newdayadd;
set my.nonrichmondfixed;
newday = time1;
run;

title1 'SPATIAL POWER...ML RESULTS';
proc mixed data=newdayadd covtest method=ML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed;

model creatinine_sum =  time1 n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed
      / ddfm=kr  ;
repeated /subject=subjectid type=sp(pow) (newday)
r=2 rcorr=2;
run;
```

Appendix 4

Covariance Estimates by Type of Covariance Matrix Structure

firsttime	secondtime	lag	covtype	CovParm	estimate
1	1	0	ANTE	an(1,1)	278502
1	2	1	ANTE	an(1,2)	147088
1	3	2	ANTE	an(1,3)	73212
1	4	3	ANTE	an(1,4)	40814
1	5	4	ANTE	an(1,5)	7175.52
1	6	5	ANTE	an(1,6)	2586.88
2	1	1	ANTE	an(2,1)	147088
2	2	0	ANTE	an(2,2)	220430
2	3	1	ANTE	an(2,3)	109718
2	4	2	ANTE	an(2,4)	61165
2	5	3	ANTE	an(2,5)	10753
2	6	4	ANTE	an(2,6)	3876.77
3	1	2	ANTE	an(3,1)	73212
3	2	1	ANTE	an(3,2)	109718
3	3	0	ANTE	an(3,3)	295498
3	4	1	ANTE	an(3,4)	164734
3	5	2	ANTE	an(3,5)	28962
3	6	3	ANTE	an(3,6)	10441
4	1	3	ANTE	an(4,1)	40814
4	2	2	ANTE	an(4,2)	61165
4	3	1	ANTE	an(4,3)	164734
4	4	0	ANTE	an(4,4)	253009
4	5	1	ANTE	an(4,5)	44481
4	6	2	ANTE	an(4,6)	16036
5	1	4	ANTE	an(5,1)	7175.52
5	2	3	ANTE	an(5,2)	10753
5	3	2	ANTE	an(5,3)	28962
5	4	1	ANTE	an(5,4)	44481
5	5	0	ANTE	an(5,5)	292112
5	6	1	ANTE	an(5,6)	105311
6	1	5	ANTE	an(6,1)	2586.88
6	2	4	ANTE	an(6,2)	3876.77
6	3	3	ANTE	an(6,3)	10441
6	4	2	ANTE	an(6,4)	16036
6	5	1	ANTE	an(6,5)	105311
6	6	0	ANTE	an(6,6)	263034

firsttime	secondtime	lag	covtype	CovParm	estimate
1	1	0	CS	cs(1,1)	299410
1	2	1	CS	cs(1,2)	170652
1	3	2	CS	cs(1,3)	170652
1	4	3	CS	cs(1,4)	170652
1	5	4	CS	cs(1,5)	170652
1	6	5	CS	cs(1,6)	170652
1	1	0	AR(1)	ar(1,1)	271066
1	2	1	AR(1)	ar(1,2)	132583
1	3	2	AR(1)	ar(1,3)	64848
1	4	3	AR(1)	ar(1,4)	31718
1	5	4	AR(1)	ar(1,5)	15514
1	6	5	AR(1)	ar(1,6)	7588.07
1	1	0	TOEP	tp(1,1)	303324
1	2	1	TOEP	tp(1,2)	168253
1	3	2	TOEP	tp(1,3)	182338
1	4	3	TOEP	tp(1,4)	175259
1	5	4	TOEP	tp(1,5)	169608
1	6	5	TOEP	tp(1,6)	205953
1	1	0	UN	UN(1,1)	300157
2	1	1	UN	UN(2,1)	164735
2	2	0	UN	UN(2,2)	236397
3	1	2	UN	UN(3,1)	177383
3	2	1	UN	UN(3,2)	135932
3	3	0	UN	UN(3,3)	329659
4	1	3	UN	UN(4,1)	176321
4	2	2	UN	UN(4,2)	156387
4	3	1	UN	UN(4,3)	188623
4	4	0	UN	UN(4,4)	272166
5	1	4	UN	UN(5,1)	119125
5	2	3	UN	UN(5,2)	145037
5	3	2	UN	UN(5,3)	187253
5	4	1	UN	UN(5,4)	100790
5	5	0	UN	UN(5,5)	332034
6	1	5	UN	UN(6,1)	206765
6	2	4	UN	UN(6,2)	178500
6	3	3	UN	UN(6,3)	172166
6	4	2	UN	UN(6,4)	168709
6	5	1	UN	UN(6,5)	145104
6	6	0	UN	UN(6,6)	324452

Appendix 5

Output from Covariance Analysis

UNSTRUCTURED ..ML RESULTS...

The Mixed Procedure

Model Information

Data Set	MY.NONRICHMONDFIXED
Dependent Variable	creatinine_sum
Covariance Structure	Unstructured
Subject Effect	SubjectID
Estimation Method	ML
Residual Variance Method	None
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Class Level Information

Class	Levels	Values
SubjectID	110	27 28 29 30 31 32 33 34 35 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 104 105 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 150 151 152 153 154
time1	18	1 2 50 51 57 58 78 79 141 142 162 163 176 177 197 198 211 212
n_diabetes	2	1 9
n_HBP	2	1 9
n_kidneystnephr	2	1 9
n_allergies	3	0 1 9
n_drinker	3	0 1 9
n_smoker	3	0 1 9
n_ondiet	3	0 1 9
n_onatkinsdiet	3	0 1 9
n_liftwts	3	0 1 9
n_onpresc	3	0 1 9

n_medcondition	3	0 1 9
n_creatine_supp	3	0 1 9
n_antidepress	2	1 9
n_analgesia	2	1 9
n_allergymed	2	1 9
n_bloodmed	2	1 9
n_kidneymed	2	1 9
n_antibiotic	2	1 9
n_GERDmed	2	1 9
n_asthmamed	2	1 9
n_diabetesmed	2	1 9
n_antiinflam	2	1 9
n_thyroidmed	2	1 9
n_unknownmed	2	1 9

Dimensions

Covariance Parameters	21
Columns in X	76
Columns in Z	0
Subjects	110
Max Obs Per Subject	6

Number of Observations

Number of Observations Read	660
Number of Observations Used	474
Number of Observations Not Used	186

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	7254.42743229	
1	3	7110.96711465	0.00456758
2	2	7103.37735332	0.00068123
3	1	7100.97261193	0.00004547
4	1	7100.82273177	0.00000043
5	1	7100.82138246	0.00000000

Convergence criteria met.

/* By using the R and RCORR options in the REPEATED statement of PROC MIXED, the user can get the Estimated Covariance Matrix (R Matrix) and the Correlation Matrix displayed for the person specified in these options. In the SAS code used for this case, the R and RCORR options were set = 2 which corresponded to Subject ID 28. This was done solely for example purposes. ***/**

Estimated R Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	300157	164735	177383	176321	119125	206765
2	164735	236397	135932	156387	145037	178500
3	177383	135932	329659	188623	187253	172166
4	176321	156387	188623	272166	100790	168709
5	119125	145037	187253	100790	332034	145104
6	206765	178500	172166	168709	145104	324452

Estimated R Correlation Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.6184	0.5639	0.6169	0.3773	0.6626

2	0.6184	1.0000	0.4869	0.6165	0.5177	0.6445
3	0.5639	0.4869	1.0000	0.6297	0.5660	0.5264
4	0.6169	0.6165	0.6297	1.0000	0.3353	0.5677
5	0.3773	0.5177	0.5660	0.3353	1.0000	0.4421
6	0.6626	0.6445	0.5264	0.5677	0.4421	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	SubjectID	300157	47882	6.27	<.0001
UN(2,1)	SubjectID	164735	35210	4.68	<.0001
UN(2,2)	SubjectID	236397	38194	6.19	<.0001
UN(3,1)	SubjectID	177383	43196	4.11	<.0001
UN(3,2)	SubjectID	135932	37206	3.65	0.0003
UN(3,3)	SubjectID	329659	52941	6.23	<.0001
UN(4,1)	SubjectID	176321	38654	4.56	<.0001
UN(4,2)	SubjectID	156387	33794	4.63	<.0001
UN(4,3)	SubjectID	188623	40518	4.66	<.0001
UN(4,4)	SubjectID	272166	45115	6.03	<.0001
UN(5,1)	SubjectID	119125	53763	2.22	0.0267
UN(5,2)	SubjectID	145037	48712	2.98	0.0029
UN(5,3)	SubjectID	187253	51827	3.61	0.0003
UN(5,4)	SubjectID	100790	47603	2.12	0.0342
UN(5,5)	SubjectID	332034	65337	5.08	<.0001
UN(6,1)	SubjectID	206765	53310	3.88	0.0001
UN(6,2)	SubjectID	178500	45370	3.93	<.0001
UN(6,3)	SubjectID	172166	51880	3.32	0.0009
UN(6,4)	SubjectID	168709	49877	3.38	0.0007
UN(6,5)	SubjectID	145104	52657	2.76	0.0059
UN(6,6)	SubjectID	324452	69713	4.65	<.0001

Fit Statistics

-2 Log Likelihood	7100.8
AIC (smaller is better)	7244.8
AICC (smaller is better)	7271.0
BIC (smaller is better)	7439.3

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
20	153.61	<.0001

COMPOUND SYMMETRIC...ML RESULTS

Model Information

Data Set	MY.NONRICHMONDFIXED
Dependent Variable	creatinine_sum
Covariance Structure	Compound Symmetry
Subject Effect	SubjectID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Dimensions

Covariance Parameters	2
Columns in X	76
Columns in Z	0
Subjects	110
Max Obs Per Subject	6

Number of Observations

Number of Observations Read	660
Number of Observations Used	474
Number of Observations Not Used	186

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	7254.42743229	
1	2	7130.08285559	0.00118217
2	1	7125.61078450	0.00020714
3	1	7124.88660474	0.00000883
4	1	7124.85815422	0.00000002
5	1	7124.85809410	0.00000000

Convergence criteria met.

/* By using the R and RCORR options in the REPEATED statement of PROC MIXED, the user can get the Estimated Covariance Matrix (R Matrix) and the Correlation Matrix displayed for the person specified in these options. In the SAS code used for this case, the R and RCORR options were set = 2 which corresponded to Subject ID 28. This was done solely for example purposes. ***/**

Estimated R Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	299410	170652	170652	170652	170652	170652
2	170652	299410	170652	170652	170652	170652
3	170652	170652	299410	170652	170652	170652
4	170652	170652	170652	299410	170652	170652
5	170652	170652	170652	170652	299410	170652
6	170652	170652	170652	170652	170652	299410

Estimated R Correlation Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.5700	0.5700	0.5700	0.5700	0.5700
2	0.5700	1.0000	0.5700	0.5700	0.5700	0.5700
3	0.5700	0.5700	1.0000	0.5700	0.5700	0.5700
4	0.5700	0.5700	0.5700	1.0000	0.5700	0.5700
5	0.5700	0.5700	0.5700	0.5700	1.0000	0.5700
6	0.5700	0.5700	0.5700	0.5700	0.5700	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
CS	SubjectID	170652	30543	5.59	<.0001
Residual		128758	9807.14	13.13	<.0001

Fit Statistics

-2 Log Likelihood	7124.9
AIC (smaller is better)	7230.9
AICC (smaller is better)	7244.5
BIC (smaller is better)	7374.0

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	129.57	<.0001

TOEPLITZ...ML RESULTS...

Model Information

Data Set	MY.NONRICHMONDFIXED
Dependent Variable	creatinine_sum
Covariance Structure	Toeplitz
Subject Effect	SubjectID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Dimensions

Covariance Parameters	6
Columns in X	76
Columns in Z	0
Subjects	110
Max Obs Per Subject	6

Number of Observations

Number of Observations Read	660
Number of Observations Used	474
Number of Observations Not Used	186

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	7254.42743229	
1	2	7126.74574240	0.00196163
2	1	7122.23335054	0.00021267
3	1	7121.51065052	0.00000604
4	1	7121.49134690	0.00000001

Convergence criteria met.

/* By using the R and RCORR options in the REPEATED statement of PROC MIXED, the user can get the Estimated Covariance Matrix (R Matrix) and the Correlation Matrix displayed for the person specified in these options. In the SAS code used for this case, the R and RCORR options were set = 2 which corresponded to Subject ID 28. This was done solely for example purposes. ***/**

Estimated R Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	303324	168253	182338	175259	169608	205953
2	168253	303324	168253	182338	175259	169608
3	182338	168253	303324	168253	182338	175259
4	175259	182338	168253	303324	168253	182338
5	169608	175259	182338	168253	303324	168253
6	205953	169608	175259	182338	168253	303324

Estimated R Correlation Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.5547	0.6011	0.5778	0.5592	0.6790

2	0.5547	1.0000	0.5547	0.6011	0.5778	0.5592
3	0.6011	0.5547	1.0000	0.5547	0.6011	0.5778
4	0.5778	0.6011	0.5547	1.0000	0.5547	0.6011
5	0.5592	0.5778	0.6011	0.5547	1.0000	0.5547
6	0.6790	0.5592	0.5778	0.6011	0.5547	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
TOEP(2)	SubjectID	168253	31348	5.37	<.0001
TOEP(3)	SubjectID	182338	32135	5.67	<.0001
TOEP(4)	SubjectID	175259	32974	5.32	<.0001
TOEP(5)	SubjectID	169608	36656	4.63	<.0001
TOEP(6)	SubjectID	205953	39195	5.25	<.0001
Residual		303324	31212	9.72	<.0001

Fit Statistics

-2 Log Likelihood	7121.5
AIC (smaller is better)	7235.5
AICC (smaller is better)	7251.4
BIC (smaller is better)	7389.4

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
5	132.94	<.0001

AUTO REGRESSIVE1 ..ML RESULTS...

Model Information

Data Set	MY.NONRICHMONDFIXED
Dependent Variable	creatinine_sum
Covariance Structure	Autoregressive
Subject Effect	SubjectID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Dimensions

Covariance Parameters	2
Columns in X	76
Columns in Z	0
Subjects	110
Max Obs Per Subject	6

Number of Observations

Number of Observations Read	660
Number of Observations Used	474
Number of Observations Not Used	186

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	7254.42743229	

1	2	7177.33796707	0.00000614
2	1	7177.31852443	0.00000000

Convergence criteria met.

/* By using the R and RCORR options in the REPEATED statement of PROC MIXED, the user can get the Estimated Covariance Matrix (R Matrix) and the Correlation Matrix displayed for the person specified in these options. In the SAS code used for this case, the R and RCORR options were set = 2 which corresponded to Subject ID 28. This was done solely for example purposes. ***/**

Estimated R Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	271066	132583	64848	31718	15514	7588.07
2	132583	271066	132583	64848	31718	15514
3	64848	132583	271066	132583	64848	31718
4	31718	64848	132583	271066	132583	64848
5	15514	31718	64848	132583	271066	132583
6	7588.07	15514	31718	64848	132583	271066

Estimated R Correlation Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.4891	0.2392	0.1170	0.05723	0.02799
2	0.4891	1.0000	0.4891	0.2392	0.1170	0.05723
3	0.2392	0.4891	1.0000	0.4891	0.2392	0.1170
4	0.1170	0.2392	0.4891	1.0000	0.4891	0.2392
5	0.05723	0.1170	0.2392	0.4891	1.0000	0.4891
6	0.02799	0.05723	0.1170	0.2392	0.4891	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
AR(1)	SubjectID	0.4891	0.04949	9.88	<.0001
Residual		271066	21906	12.37	<.0001

Fit Statistics

-2 Log Likelihood	7177.3
AIC (smaller is better)	7283.3
AICC (smaller is better)	7296.9
BIC (smaller is better)	7426.4

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	77.11	<.0001

ANTE DEPENDENCE...ML RESULTS...

Model Information

Data Set	MY.NONRICHMONDFIXED
Dependent Variable	creatinine_sum
Covariance Structure	Ante-dependence
Subject Effect	SubjectID
Estimation Method	ML
Residual Variance Method	None
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville

Degrees of Freedom Method Kenward-Roger

Dimensions

Covariance Parameters	11
Columns in X	76
Columns in Z	0
Subjects	110
Max Obs Per Subject	6

Number of Observations

Number of Observations Read	660
Number of Observations Used	474
Number of Observations Not Used	186

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	7254.42743229	
1	2	7161.83884820	0.00016244
2	1	7161.27658827	0.00000548
3	1	7161.25889241	0.00000001

Convergence criteria met.

/* By using the R and RCORR options in the REPEATED statement of PROC MIXED, the user can get the Estimated Covariance Matrix (R Matrix) and the Correlation Matrix displayed for the person specified in these options. In the SAS code used for this case, the R and RCORR options were set = 2 which corresponded to Subject ID 28. This was done solely for example purposes. ***/**

Estimated R Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	278502	147088	73212	40814	7175.52	2586.88
2	147088	220430	109718	61165	10753	3876.77
3	73212	109718	295498	164734	28962	10441
4	40814	61165	164734	253009	44481	16036
5	7175.52	10753	28962	44481	292112	105311
6	2586.88	3876.77	10441	16036	105311	263034

Estimated R Correlation Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.5936	0.2552	0.1538	0.02516	0.009558
2	0.5936	1.0000	0.4299	0.2590	0.04238	0.01610
3	0.2552	0.4299	1.0000	0.6025	0.09858	0.03745
4	0.1538	0.2590	0.6025	1.0000	0.1636	0.06216
5	0.02516	0.04238	0.09858	0.1636	1.0000	0.3799
6	0.009558	0.01610	0.03745	0.06216	0.3799	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Var(1)	SubjectID	278502	44863	6.21	<.0001
Var(2)	SubjectID	220430	36149	6.10	<.0001
Var(3)	SubjectID	295498	49059	6.02	<.0001
Var(4)	SubjectID	253009	44179	5.73	<.0001
Var(5)	SubjectID	292112	57821	5.05	<.0001
Var(6)	SubjectID	263034	53379	4.93	<.0001
Rho(1)	SubjectID	0.5936	0.07106	8.35	<.0001
Rho(2)	SubjectID	0.4299	0.1189	3.62	0.0003
Rho(3)	SubjectID	0.6025	0.07348	8.20	<.0001

Rho(4)	SubjectID	0.1636	0.1832	0.89	0.3719
Rho(5)	SubjectID	0.3799	0.1190	3.19	0.0014

Fit Statistics

-2 Log Likelihood	7161.3
AIC (smaller is better)	7285.3
AICC (smaller is better)	7304.3
BIC (smaller is better)	7452.7

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
10	93.17	<.0001

SPATIAL POWER...ML RESULTS..

Model Information

Data Set	WORK.NEWDAYADD
Dependent Variable	creatinine_sum
Covariance Structure	Spatial Power
Subject Effect	SubjectID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Dimensions

Covariance Parameters	2
Columns in X	76
Columns in Z	0
Subjects	110
Max Obs Per Subject	6

Number of Observations

Number of Observations Read	660
Number of Observations Used	474
Number of Observations Not Used	186

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	7254.42743229	
1	2	7178.72862657	0.00000000

Convergence criteria met.

/* By using the R and RCORR options in the REPEATED statement of PROC MIXED, the user can get the Estimated Covariance Matrix (R Matrix) and the Correlation Matrix displayed for the person specified in these options. In the SAS code used for this case, the R and RCORR options were set = 2 which corresponded to Subject ID 28. This was done solely for example purposes. ***/**

Estimated R Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	259035	136161				
2	136161	259035				
3			259035	136161		

4			136161	259035		
5					259035	136161
6					136161	259035

Estimated R Correlation Matrix for SubjectID 28

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.5256				
2	0.5256	1.0000				
3			1.0000	0.5256		
4			0.5256	1.0000		
5					1.0000	0.5256
6					0.5256	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
SP(POW)	SubjectID	0.5256	0.04716	11.15	<.0001
Residual		259035	18920	13.69	<.0001

Fit Statistics

-2 Log Likelihood	7178.7
AIC (smaller is better)	7284.7
AICC (smaller is better)	7298.4
BIC (smaller is better)	7427.9

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	75.70	<.0001

Appendix 6

AIC, AICC, BIC Calculations

For the mixed model, Var($\boldsymbol{Y}$) = $\sigma^2[\boldsymbol{ZGZ'} + \boldsymbol{R}]$. Let $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$. Using the maximum likelihood methodology in the SAS mixed model procedure (PROC MIXED), the information criteria (Gurka 2006) are calculated as:

$$AIC = -2l_{ML} + 2(q + k)$$

$$AICC = -2l_{ML} + 2(q + k)\left(\frac{\sum_{1}^{n} p_i}{((\sum_{1}^{n} p_i) - (q + k) - 1)}\right);$$

$$BIC = -2l_{ML} + (q + k)(\log n)$$

where the log-likelihood (l_{ML}) is

$$l_{ML} = -\frac{1}{2}ln|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}) - \frac{n}{2}ln(2\boldsymbol{\pi})$$ (Littell et al. 2006)

and: q = number of fixed effects (in the $\boldsymbol{\beta}$ vector)

k = number of covariance parameters being estimated for both the $\boldsymbol{G}$ and $\boldsymbol{R}$ covariance matrices.

i = 1, 2, 3, ..., n

n = number of subjects

p_i = number of measurements for the i^{th} subject

Appendix 7

Estimation of Fixed and Random Effects in the Mixed Model when using REML

As noted earlier, for the mixed model, $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ and $\text{Var}(\boldsymbol{Y}) = \sigma^2[\boldsymbol{ZGZ'} + \boldsymbol{R}]$. The Estimated Best Linear Unbiased Estimator (EBLUE) of the fixed effects ($\hat{\boldsymbol{\beta}}$) and the Estimated Best Linear Unbiased Predictor (EBLUP) of the random effects ($\hat{\boldsymbol{v}}$) can be obtained by letting $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$, 'plugging in' the REML estimates of the $\boldsymbol{G}$ and $\boldsymbol{R}$ covariance matrices ($\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{R}}$ respectively) into V to get $\hat{\boldsymbol{V}} = \boldsymbol{Z\hat{G}Z'} + \hat{\boldsymbol{R}}$, and then solving this system of mixed model equations (Henderson 1984; Littell et al. 2006) shown here:

$$\begin{bmatrix} \boldsymbol{X'\hat{R}^{-1}X} & \boldsymbol{X'\hat{R}^{-1}Z} \\ \boldsymbol{Z'\hat{R}^{-1}X} & \boldsymbol{Z'\hat{R}^{-1}Z} + \hat{\boldsymbol{G}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{v}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'\hat{R}^{-1}Y} \\ \boldsymbol{Z'\hat{R}^{-1}Y} \end{bmatrix}$$

The solutions (Khattree and Naik 1999) are:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{v}} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{X'\hat{V}^{-1}X})^{-1}\boldsymbol{X'\hat{V}^{-1}Y} \\ \boldsymbol{\hat{G}Z'\hat{V}^{-1}}(\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X'\hat{V}^{-1}X})^{-1}\boldsymbol{X'\hat{V}^{-1}Y}) \end{bmatrix} = \begin{bmatrix} (\boldsymbol{X'\hat{V}^{-1}X})^{-1}\boldsymbol{X'\hat{V}^{-1}Y} \\ \boldsymbol{\hat{G}Z'\hat{V}^{-1}}(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}}) \end{bmatrix}$$

Appendix 8

Profile Plots, Box Plots and Multiple Comparisons by Location

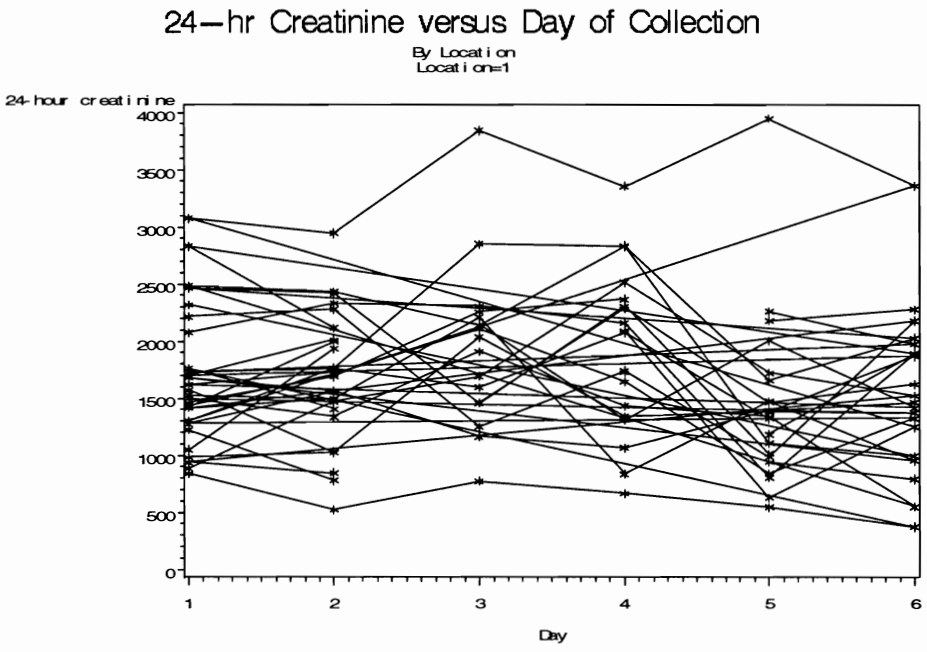

Figure 5: Plot of Creatinine versus Day of Collection for Location 1

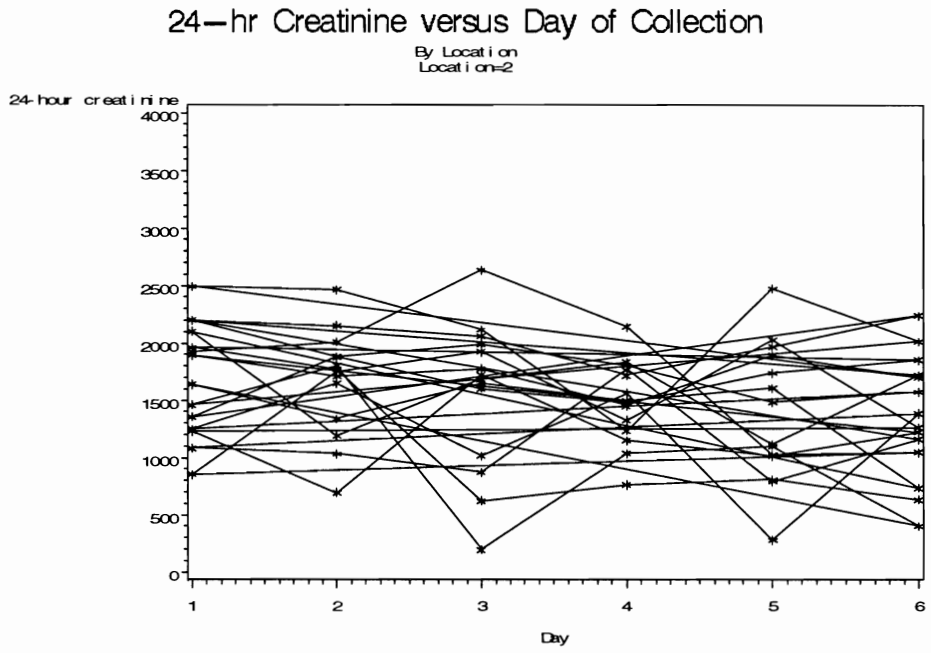

Figure 6: Plot of Creatinine versus Day of Collection for Location 2

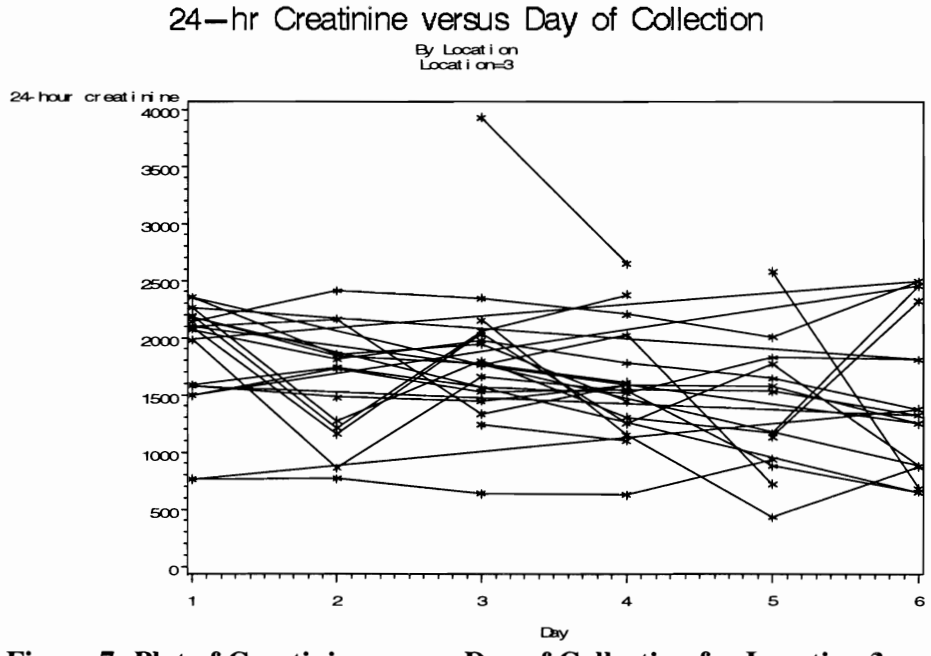

Figure 7: Plot of Creatinine versus Day of Collection for Location 3

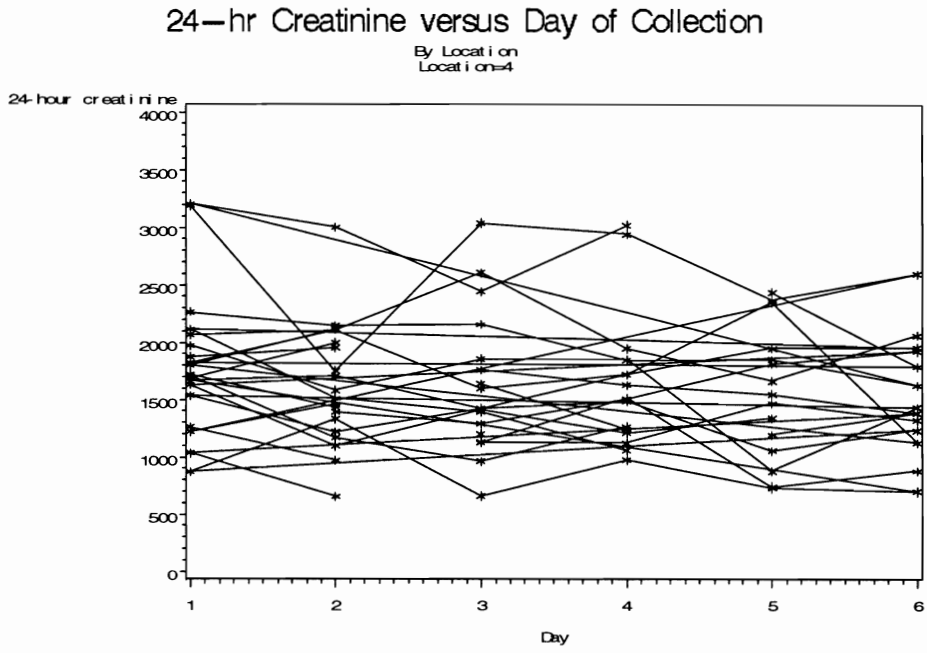

Figure 8: Plot of Creatinine versus Day of Collection for Location 4

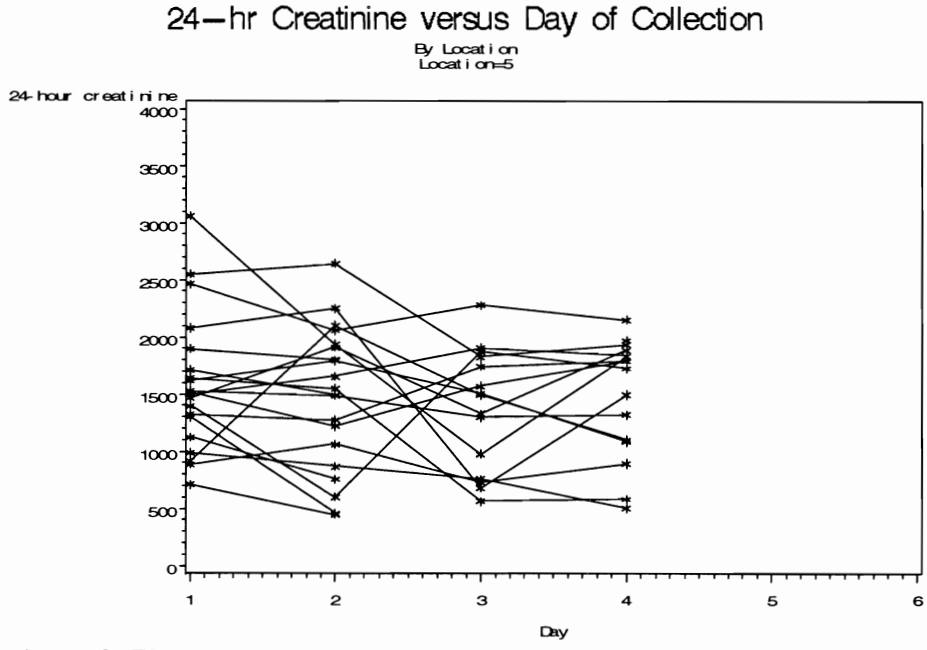

Figure 9: Plot of Creatinine versus Day of Collection for Location 5

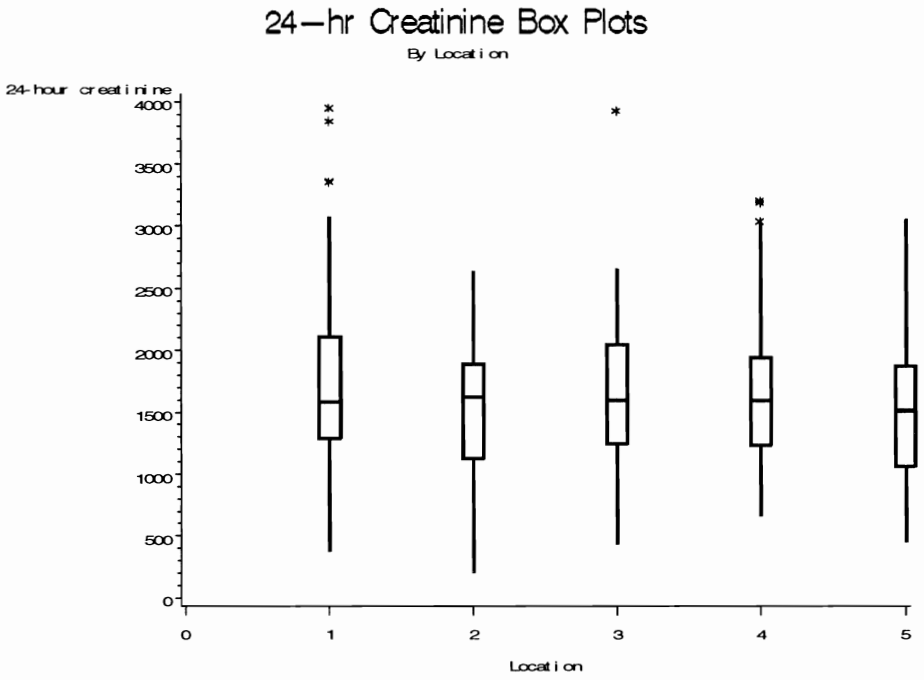

Figure 10: Box Plot of Creatinine by Location

24-hr Creatinine By Location

Tukey-Kramer Comparison with All

Level	Compared	Root MSE DF = 469	sqrt(2)q* = 591.74	= 3.8727	p
		Mean Diff	Lower Limit	Upper Limit	Value
1	2	164.3471	-61.8222	390.5164	0.2725
1	3	44.2620	-181.9072	270.4313	0.9836
1	4	41.5799	-171.4604	254.6202	0.9837
1	5	200.1677	-38.3196	438.6550	0.1472
2	1	-164.3471	-390.5164	61.8222	0.2725
2	3	-120.0851	-370.1246	129.9545	0.6819
2	4	-122.7672	-360.9969	115.4625	0.6207
2	5	35.8206	-225.4138	297.0550	0.9958
3	1	-44.2620	-270.4313	181.9072	0.9836
3	2	120.0851	-129.9545	370.1246	0.6819
3	4	-2.6822	-240.9119	235.5475	1.0000
3	5	155.9057	-105.3287	417.1401	0.4762
4	1	-41.5799	-254.6202	171.4604	0.9837
4	2	122.7672	-115.4625	360.9969	0.6207
4	3	2.6822	-235.5475	240.9119	1.0000
4	5	158.5879	-91.3662	408.5419	0.4120
5	1	-200.1677	-438.6550	38.3196	0.1472
5	2	-35.8206	-297.0550	225.4138	0.9958
5	3	-155.9057	-417.1401	105.3287	0.4762
5	4	-158.5879	-408.5419	91.3662	0.4120

Appendix 9

Initial Model Output

The Mixed Procedure
Model Information

Data Set	MY.NONRICHMONDFIXED
Dependent Variable	creatinine_sum
Covariance Structure	Compound Symmetry
Subject Effect	SubjectID
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Class Level Information

Class	Levels	Values
SubjectID	106	27 28 29 30 31 32 33 34 35 37 38 39 40 41 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 104 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 150 151 152 153
time1	18	1 2 50 51 57 58 78 79 141 142 162 163 176 177 197 198 211 212
n_diabetes	2	1 9
n_HBP	2	1 9
n_kidneystnephr	2	1 9
n_allergies	3	0 1 9
n_drinker	2	0 1
n_smoker	3	0 1 9
n_ondiet	3	0 1 9
n_onatkinsdiet	3	0 1 9
n_liftwts	3	0 1 9
n_onpresc	3	0 1 9
n_medcondition	3	0 1 9

n_creatine_supp	3	0 1 9
n_antidepress	2	1 9
n_analgesia	2	1 9
n_allergymed	2	1 9
n_bloodmed	2	1 9
n_kidneymed	2	1 9
n_antibiotic	2	1 9
n_GERDmed	2	1 9
n_asthmamed	2	1 9
n_diabetesmed	2	1 9
n_antiinflam	2	1 9
n_thyroidmed	2	1 9
n_unknownmed	2	1 9

Dimensions

Covariance Parameters	2
Columns in X	80
Columns in Z	0
Subjects	106
Max Obs Per Subject	6

Number of Observations

Number of Observations Read	660
Number of Observations Used	453
Number of Observations Not Used	207

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	6400.61362658	
1	2	6309.91829239	0.00040637
2	1	6308.61870799	0.00003151
3	1	6308.52646824	0.00000024
4	1	6308.52580027	0.00000000

Convergence criteria met.

/* By using the R and RCORR options in the REPEATED statement of PROC MIXED, the user can get the Estimated Covariance Matrix (R Matrix) and the Correlation Matrix displayed for the person specified in these options. In the SAS code used for this case, the R and RCORR options were set = 48 which corresponded to Subject ID 76. This was done solely for example purposes. ***/**

Estimated R Matrix for SubjectID 76

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	308194	161271	161271	161271	161271	161271
2	161271	308194	161271	161271	161271	161271
3	161271	161271	308194	161271	161271	161271
4	161271	161271	161271	308194	161271	161271
5	161271	161271	161271	161271	308194	161271
6	161271	161271	161271	161271	161271	308194

Estimated R Correlation Matrix for SubjectID 76

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.5233	0.5233	0.5233	0.5233	0.5233
2	0.5233	1.0000	0.5233	0.5233	0.5233	0.5233
3	0.5233	0.5233	1.0000	0.5233	0.5233	0.5233
4	0.5233	0.5233	0.5233	1.0000	0.5233	0.5233
5	0.5233	0.5233	0.5233	0.5233	1.0000	0.5233
6	0.5233	0.5233	0.5233	0.5233	0.5233	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
CS	SubjectID	161271	32726	4.93	<.0001	0.05	97129	225413
Residual		146923	11733	12.52	<.0001	0.05	126389	172931

Fit Statistics

-2 Res Log Likelihood	6308.5
AIC (smaller is better)	6312.5
AICC (smaller is better)	6312.6
BIC (smaller is better)	6317.9

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	92.09	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Intercept	1	117	0.69	0.4062
age_at_collect	1	89.6	1.09	0.2998
calc_bmi	1	123	0.50	0.4824
calc_cm_ht	1	116	0.74	0.3899
calc_kg_wt	1	122	0.11	0.7400
n_diabetes	1	171	0.74	0.3911
n_HBP	1	171	0.24	0.6282
n_kidneystnephr	1	140	0.00	0.9923
n_allergies	2	337	15.32	<.0001
n_drinker	1	323	0.40	0.5293
n_smoker	2	190	0.01	0.9888
n_ondiet	2	402	0.73	0.4848
n_onatkinsdiet	2	302	0.30	0.7395
n_liftwts	2	401	1.09	0.3381
n_onpresc	2	398	1.51	0.2211
n_medcondition	2	244	10.50	<.0001
n_creatine_supp	2	397	8.63	0.0002
n_antidepress	1	409	3.14	0.0771
n_analgesia	1	147	0.02	0.8815
n_allergymed	1	411	0.06	0.8015
n_bloodmed	1	305	0.96	0.3282
n_kidneymed	1	160	0.33	0.5670
n_antibiotic	1	201	0.70	0.4022
n_GERDmed	1	257	0.14	0.7066
n_asthmamed	1	116	0.21	0.6456

n_diabetesmed	1	90.8	0.08	0.7845
n_antiinflam	1	187	3.69	0.0563
n_thyroidmed	1	92.3	0.00	0.9733
n_unknownmed	1	274	3.09	0.0800
n_ondiet*n_onatkinsd	2	294	0.59	0.5524
n_diabete*n_diabetes	1	339	1.12	0.2913
n_HBP*n_bloodmed	1	346	5.90	0.0156
n_kidneys*n_kidneyme	1	155	0.06	0.8026

Appendix 10

Final Model Output

The Mixed Procedure
Model Information

Data Set	MY.NONRICHMONDFIXED
Dependent Variable	creatinine_sum
Covariance Structure	Compound Symmetry
Subject Effect	SubjectID
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Class Level Information

Class	Levels	Values
SubjectID	107	27 28 29 30 31 32 33 34 35 37 38 39 40 41 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 104 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 150 151 152 153
time1	18	1 2 50 51 57 58 78 79 141 142 162 163 176 177 197 198 211 212
n_diabetes	2	1 9
n_HBP	2	1 9
n_kidneystnephr	2	1 9
n_allergies	3	0 1 9
n_drinker	2	0 1
n_smoker	3	0 1 9
n_ondiet	3	0 1 9
n_onatkinsdiet	3	0 1 9
n_liftwts	3	0 1 9
n_onpresc	3	0 1 9
n_medcondition	3	0 1 9

n_creatine_supp	3	0 1 9
n_antidepress	2	1 9
n_analgesia	2	1 9
n_allergymed	2	1 9
n_bloodmed	2	1 9
n_kidneymed	2	1 9
n_antibiotic	2	1 9
n_GERDmed	2	1 9
n_asthmamed	2	1 9
n_diabetesmed	2	1 9
n_antiinflam	2	1 9
n_thyroidmed	2	1 9
n_unknownmed	2	1 9

Dimensions

Covariance Parameters	2
Columns in X	16
Columns in Z	0
Subjects	107
Max Obs Per Subject	6

Number of Observations

Number of Observations Read	660
Number of Observations Used	457
Number of Observations Not Used	203

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	6891.89017725	
1	2	6777.60640768	0.00009905
2	1	6777.28526039	0.00000218
3	1	6777.27865416	0.00000000

Convergence criteria met.

/* By using the R and RCORR options in the REPEATED statement of PROC MIXED, the user can get the Estimated Covariance Matrix (R Matrix) and the Correlation Matrix displayed for the person specified in these options. In the SAS code used for this case, the R and RCORR options were set = 48 which corresponded to Subject ID 76. This was done solely for example purposes. ***/**

Estimated R Matrix for SubjectID 76

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	294714	148567	148567	148567	148567	148567
2	148567	294714	148567	148567	148567	148567
3	148567	148567	294714	148567	148567	148567
4	148567	148567	148567	294714	148567	148567
5	148567	148567	148567	148567	294714	148567
6	148567	148567	148567	148567	148567	294714

Estimated R Correlation Matrix for SubjectID 76

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	1.0000	0.5041	0.5041	0.5041	0.5041	0.5041
2	0.5041	1.0000	0.5041	0.5041	0.5041	0.5041
3	0.5041	0.5041	1.0000	0.5041	0.5041	0.5041
4	0.5041	0.5041	0.5041	1.0000	0.5041	0.5041

5	0.5041	0.5041	0.5041	0.5041	1.0000	0.5041
6	0.5041	0.5041	0.5041	0.5041	0.5041	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
CS	SubjectID	148567	27924	5.32	<.0001	0.05	93836	203298
Residual		146147	11283	12.95	<.0001	0.05	126329	171051

Fit Statistics

-2 Res Log Likelihood	6777.3
AIC (smaller is better)	6781.3
AICC (smaller is better)	6781.3
BIC (smaller is better)	6786.6

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	114.61	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Intercept	1	127	6.27	0.0135
calc_bmi	1	122	27.20	<.0001
calc_cm_ht	1	129	4.80	0.0303
n_diabetes	1	374	6.78	0.0096
n_allergies	2	368	13.04	<.0001
n_medcondition	2	428	9.04	0.0001
n_creatine_supp	2	434	10.14	<.0001
n_antiinflam	1	206	6.63	0.0107

Fixed Effect Estimates:

Effect	n_diabetes	n_allergies	n_medcondition	n_creatine_supp	n_antiinflam	Estimate	StdErr	DF	tValue	Probt
Intercept						-2457.5515	1154.78	140.32	-2.13	0.04
calc_bmi						37.6721	7.22	121.68	5.22	0.00
calc_cm_ht						13.3569	6.10	129.13	2.19	0.03
n_diabetes	1					-496.5500	190.75	374.40	-2.60	0.01
n_diabetes	9					0.0000				
n_allergies		0				-1135.3851	349.34	406.62	-3.25	0.00
n_allergies		1				-1430.2188	358.44	416.49	-3.99	0.00
n_allergies		9				0.0000				
n_medcondition			0			1893.0083	445.26	431.11	4.25	0.00
n_medcondition			1			1877.4174	452.76	434.91	4.15	0.00
n_medcondition			9			0.0000				

n_creatine_supp			0		8.2967	212.02	424.04	0.04	0.97
n_creatine_supp			1		-517.7042	241.46	439.73	-2.14	0.03
n_creatine_supp			9		0.0000				
n_antiinflam				1	-599.3979	232.81	205.55	-2.57	0.01
n_antiinflam				9	0.0000				

Confidence Intervals for Fixed Effect Estimates (alpha = 0.05)

Effect	n_diabetes	n_allergies	n_medcondition	n_creatine_supp	n_antiinflam	Estimate	StdErr	Lower	Upper
Intercept						-2457.5515	1154.78	-4740.5688	-174.5343
calc_bmi						37.6721	7.22	23.3728	51.9714
calc_cm_ht						13.3569	6.10	1.2939	25.4200
n_diabetes	1					-496.5500	190.75	-871.6178	-121.4823
n_diabetes	9					0.0000			
n_allergies		0				-1135.3851	349.34	-1822.1295	-448.6407
n_allergies		1				-1430.2188	358.44	-2134.7874	-725.6501
n_allergies		9				0.0000			
n_medcondition			0			1893.0083	445.26	1017.8606	2768.1560
n_medcondition			1			1877.4174	452.76	987.5556	2767.2793
n_medcondition			9			0.0000			
n_creatine_supp				0		8.2967	212.02	-408.4489	425.0424
n_creatine_supp				1		-517.7042	241.46	-992.2699	-43.1384
n_creatine_supp				9		0.0000			
n_antiinflam					1	-599.3979	232.81	-1058.3974	-140.3985
n_antiinflam					9	0.0000			

Appendix 11

SAS Code for Final Model

```
/* to get residual plots and create an output file with the predicted
values */

ods output SolutionF;

ods html;
ods graphics on;

proc mixed data=my.nonrichmondfixed cl covtest method=REML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed _unknownmed;

model creatinine_sum =   calc_bmi calc_cm_ht
      n_diabetes
      n_allergies
      n_medcondition n_creatine_supp
      n_antiinflam
      /ddfm=kr s intercept cl htype=3  residual outp=predictedvalues;

repeated /subject=subjectid type=cs r=48 rcorr=48;
run;

ods html close;
ods graphics off;

/* to get influence diagnostics */

ods output SolutionF influence;

title1 'REML; ITER Influence diagnostics; used default ddfm not KR';
proc mixed data=my.nonrichmondfixed cl covtest  method=REML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
```

```
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed;

      model creatinine_sum =  calc_bmi calc_cm_ht
      n_diabetes
      n_allergies
      n_medcondition n_creatine_supp
      n_antiinflam
/ s intercept htype=3

/* For single observation case deletion use: */

influence (iter=5);

/* iter forces refit of model & covariance parameters; however must use
default degrees of freedom; cannot use dfddfm=kr when running influence
diagnostics*/

/* For all of one subject's observations to be deleted, use:*/

influence (effect=subjectid iter=5);

repeated /subject=subjectid type=cs;
run;
```

Appendix 12

Residual Plots for Final Model

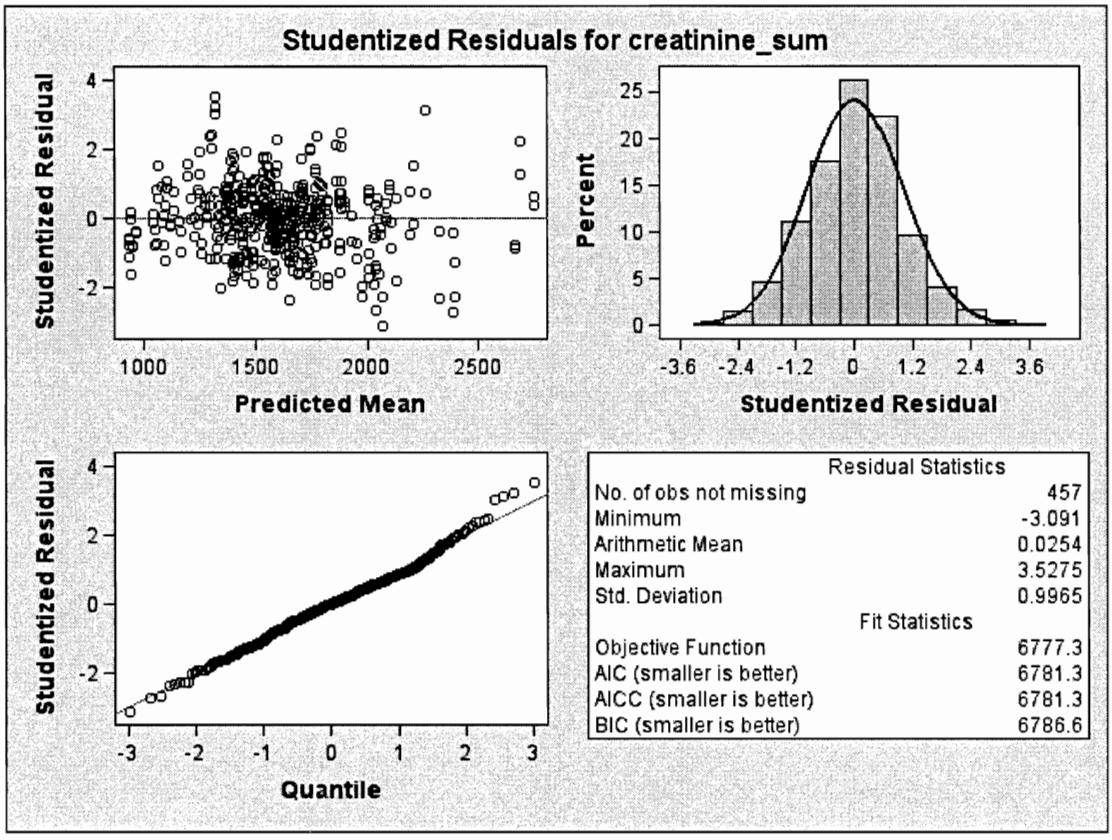

Figure 11: Residual Plots for Final Model

Appendix 13

Outliers and High Leverage Observations

Potential Outliers:

Sub	Day	Observed	Predicted	Residual	Leverage	Student	RMSE	RStudent	Miss
27	5	627.500	2034.587	-1407.1	0.0145	-2.621	381.983	-2.648	
28	3	3845.300	2680.445	1164.9	0.0396	2.273	381.343	2.285	
32	1	2833.700	1594.073	1239.6	0.0107	2.302	381.839	2.322	
35	3	2851.835	1738.008	1113.8	-0.0013	2.068	380.651	2.075	
35	4	2829.546	1738.008	1091.5	-0.0013	2.027	380.793	2.033	
38	3	1252.000	2387.668	-1135.7	0.0472	-2.209	382.118	-2.216	m
38	5	984.810	2380.311	-1395.5	0.0422	-2.698	380.836	-2.715	m
57	6	794.760	1975.003	-1180.2	0.0280	-2.221	383.030	-2.236	
63	6	400.680	1650.712	-1250.0	0.0035	-2.315	379.633	-2.327	m
100	1	3205.440	1878.336	1327.1	0.0129	2.483	382.693	2.496	
100	2	3004.260	1878.336	1125.9	0.0129	2.107	383.038	2.112	
100	4	3018.695	1854.237	1164.5	0.0111	2.177	383.001	2.184	
102	1	3187.700	1319.349	1868.4	0.0155	3.527	381.706	3.553	
102	3	3038.910	1319.349	1719.6	0.0155	3.246	382.348	3.264	
102	4	2941.185	1319.349	1621.8	0.0155	3.061	382.653	3.075	
102	5	2370.960	1303.758	1067.2	0.0488	2.038	382.633	2.035	
102	6	2597.430	1303.758	1293.7	0.0488	2.470	383.032	2.473	
126	1	3062.230	1772.943	1289.3	0.0094	2.388	378.294	2.404	m
127	4	1153.310	2321.098	-1167.8	0.0575	-2.259	383.003	-2.266	
127	5	431.600	2065.633	-1634.0	0.0067	-3.090	381.368	-3.121	m
127	6	875.520	2065.633	-1190.1	0.0067	-2.251	383.000	-2.259	m
128	3	3928.726	2260.905	1667.8	0.0323	3.150	379.589	3.200	
132	5	2580.480	1294.171	1286.3	0.0150	2.397	375.617	2.402	m

High Leverage Observations (leverage > 0.1):

Sub	Day	Observed	Predicted	Residual	Leverage
45	3	2141.080	2203.041	-61.961	0.5000
45	4	2837.000	2203.041	633.960	0.5000
45	5	1110.435	1039.815	70.620	0.2342
45	6	996.810	1039.815	-43.005	0.2342
47	5	547.760	928.099	-380.340	0.1539
47	6	376.300	928.099	-551.800	0.1539
50	5	1725.840	1569.270	156.570	0.2204
50	6	1524.400	1569.270	-44.870	0.2204
63	3	196.115	940.354	-744.240	0.3690
63	4	1037.700	940.354	97.346	0.3690
64	1	1641.640	1585.633	56.007	0.2150
64	2	1339.560	1585.633	-246.070	0.2150
107	1	1512.000	1209.664	302.340	0.1329
107	2	1221.660	1209.664	11.996	0.1329
108	1	1124.240	1083.792	40.448	0.2137
108	2	757.610	1083.792	-326.180	0.2137

Appendix 14

Influence Diagnostics by Observation

Influential Observations based on Restricted Likelihood Distance (RLD) >0.3:

Sub	Day	RLD
45	1	0.36143
45	3	1.18525
45	4	1.18525
45	6	0.31279
63	3	0.71372
63	4	1.12781
63	6	0.39782
132	5	0.30113

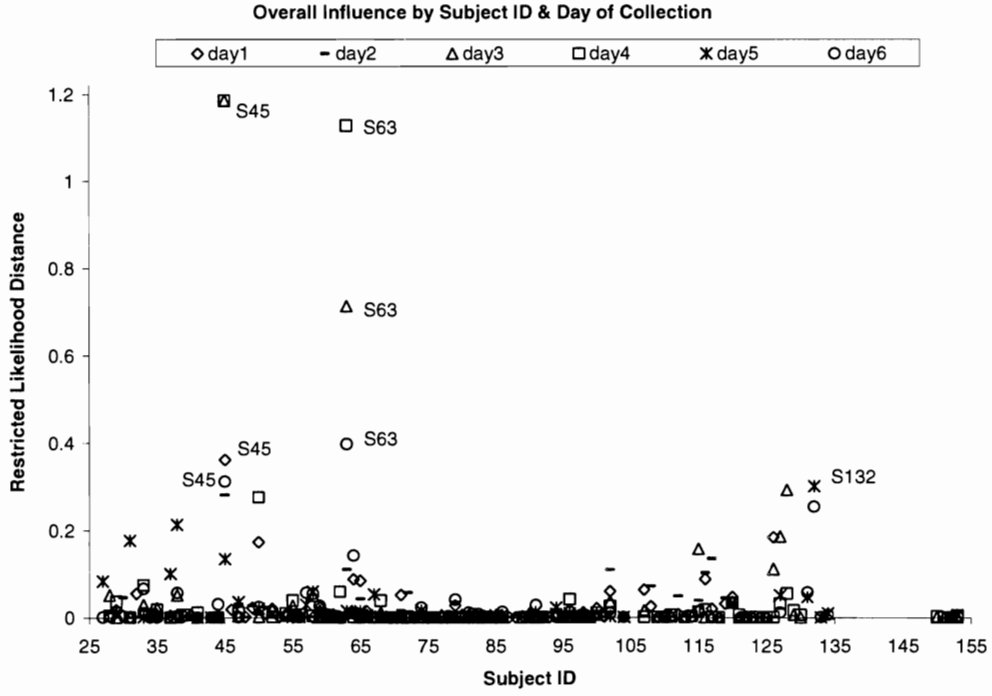

Figure 12: Plot of Overall Influence by Subject and Day of Collection

|DFFITS| >0.8

Sub	Day	DFFITS
45	3	-0.95571
45	4	0.95679
63	3	-0.83313
63	4	1.03904

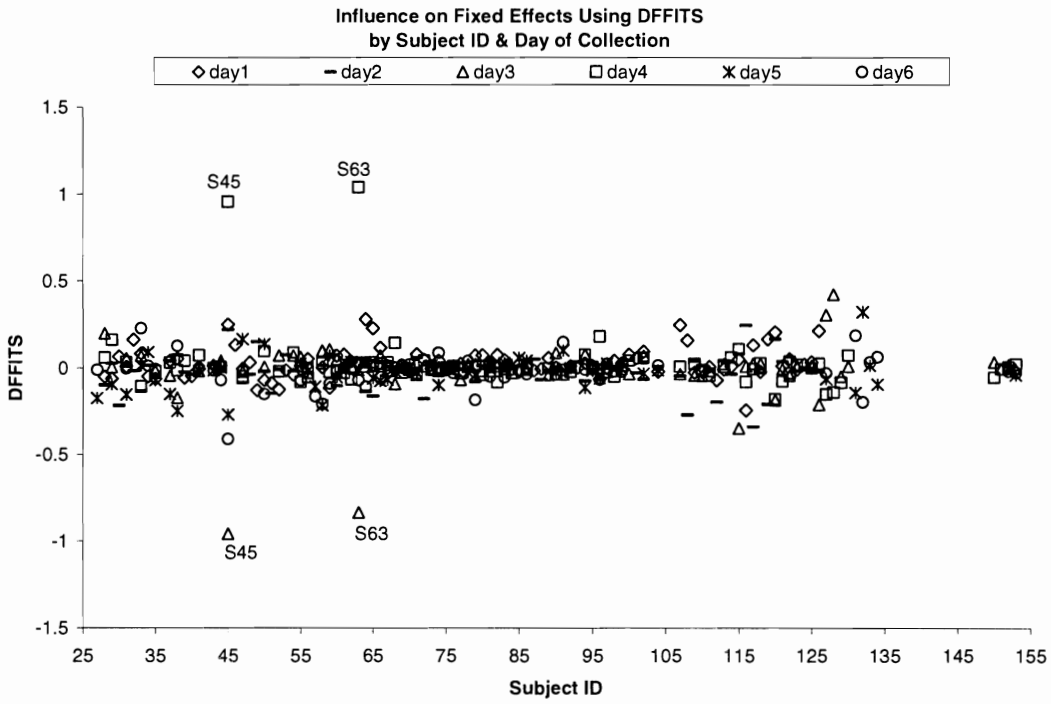

Figure 13: Plot of DFFITS by Subject and Day of Collection

Cook's Distance (>0.04)

Sub	Day	Cook's Distance
45	3	0.107385
45	4	0.107602
63	3	0.064955
63	4	0.101095

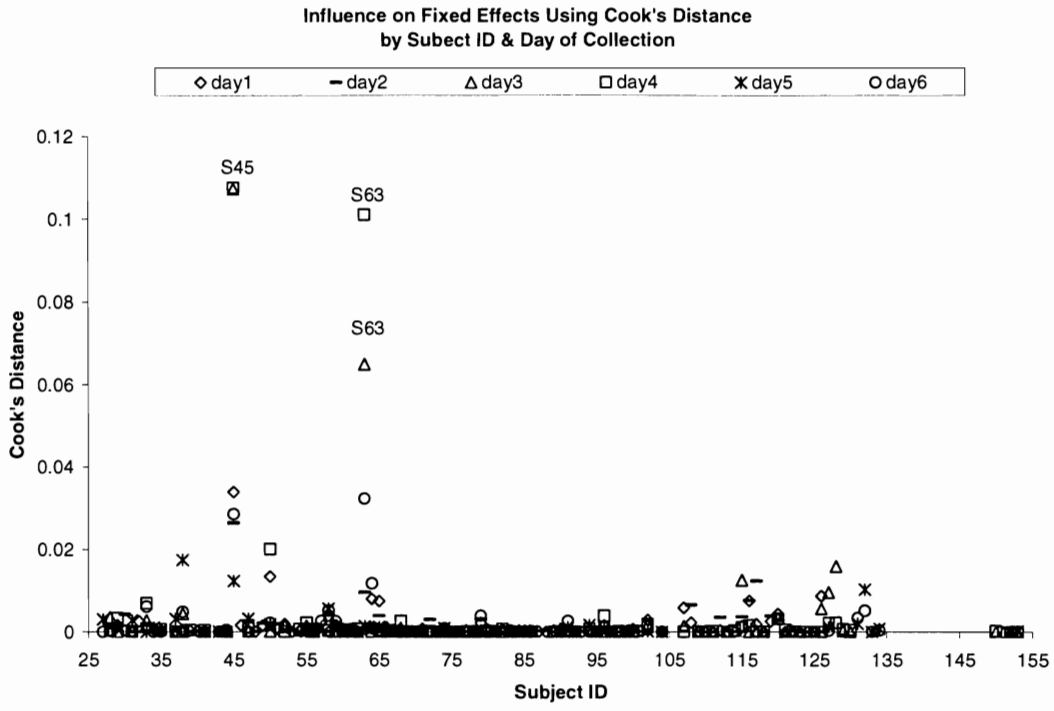

Figure 14: Plot of Cook's Distance by Subject and Day of Collection

COVRATIO

Sub	Day	CovRatio
45	3	1.698945
45	4	1.698945
63	3	1.443698
63	4	1.432325

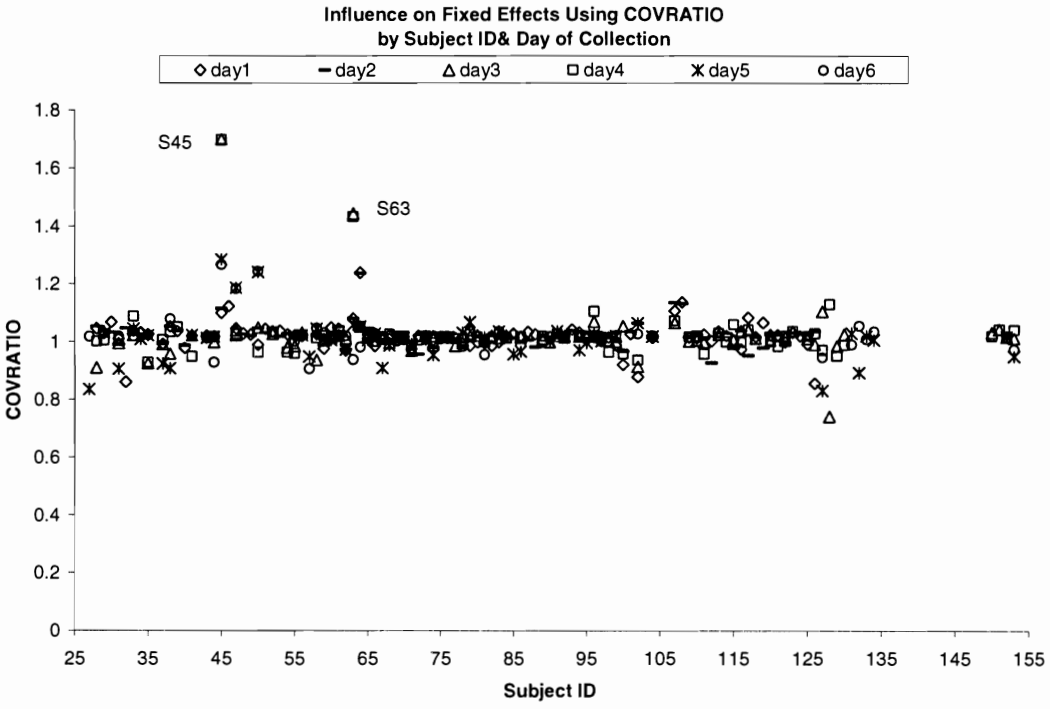

Figure 15: Plot of COVRATIO by Subject and Day of Collection

Appendix 15

SAS Code for Calculating Other Models' Predicted Values

```
libname my "C:\Documents and Settings\Owner\Desktop\thesis\";
run;

data my.othermodels;
set my.nonrichmondfixed;

peniepredict = -1791.05 + 17.69 * calc_cm_ht;
penieresid = creatinine_sum - peniepredict;
penieresidsq=penieresid**2;

turnerpredict = (0.0143*calc_cm_ht + 0.00975*calc_kg_wt -
                 0.00734*(age_at_collect - 20.0) - 1.391) *1000.0;
turnerresid = creatinine_sum - turnerpredict;
turnerresidsq = turnerresid**2;

kawasakipredict = (-12.63)*age_at_collect + 15.12*calc_kg_wt +
7.39*calc_cm_ht -79.9;
kawasakiresid = creatinine_sum - kawasakipredict;
kawasakiresidsq = kawasakiresid**2;

bodysurface= 0.02350 * (calc_cm_ht**0.42246)*(calc_kg_wt**0.51456);
dodgepredict= (0.1457*bodysurface + 0.2888*bodysurface*bodysurface -
0.0215) * 1440;
dodgeresid = creatinine_sum - dodgepredict;
dodgeresidsq=dodgeresid**2;

harrispredict = 647 + 372*1 + 13.5*calc_kg_wt - 10.8*age_at_collect
                - 1.47*((age_at_collect-28.4)*(calc_kg_wt-80.1));
harrisresid = creatinine_sum - harrispredict;
harrisresidsq = harrisresid**2;

moriyamapredict = 211 - 6.4*age_at_collect +2.5*calc_cm_ht +
18.0*calc_kg_wt;
moriyamaresid = creatinine_sum - moriyamapredict;
moriyamaresidsq = moriyamaresid**2;

tanakapredict = -2.04*age_at_collect + 14.89 * calc_kg_wt +
16.14*calc_cm_ht -2244.45;
tanakaresid = creatinine_sum - tanakapredict;
tanakaresidsq = tanakaresid**2; run;
```

```
/* now create a model with just height and bmi in it using compound
symmetric covariance; use to compare with other models            */

ods output SolutionF;
proc mixed data=my.nonrichmondfixed cl covtest  method=REML;
class subjectid time1
      n_diabetes n_HBP n_kidneystnephr
      n_allergies n_drinker n_smoker n_ondiet n_onatkinsdiet
      n_liftwts n_onpresc n_medcondition n_creatine_supp
      n_antidepress n_analgesia n_allergymed n_bloodmed n_kidneymed
      n_antibiotic n_GERDmed
      n_asthmamed n_diabetesmed n_antiinflam n_thyroidmed n_unknownmed;
      model creatinine_sum =   calc_bmi calc_cm_ht
      / ddfm=kr  s INTERCEPT  cl  HTYPE=3  residual
      outp=predictedvalues   ;
repeated /subject=subjectid type=cs ;
run;

data my.modelhtbmionly;
set predictedvalues;
htbmipred=pred;
htbmiresid=resid;
htbmiresidsq=resid*resid;
keep subjectid round day calc_cm_ht calc_kg_wt calc_bmi creatinine_sum
htbmipred htbmiresid htbmiresidsq;
run;
```

Appendix 16

Comparisons of Other Studies' Participants

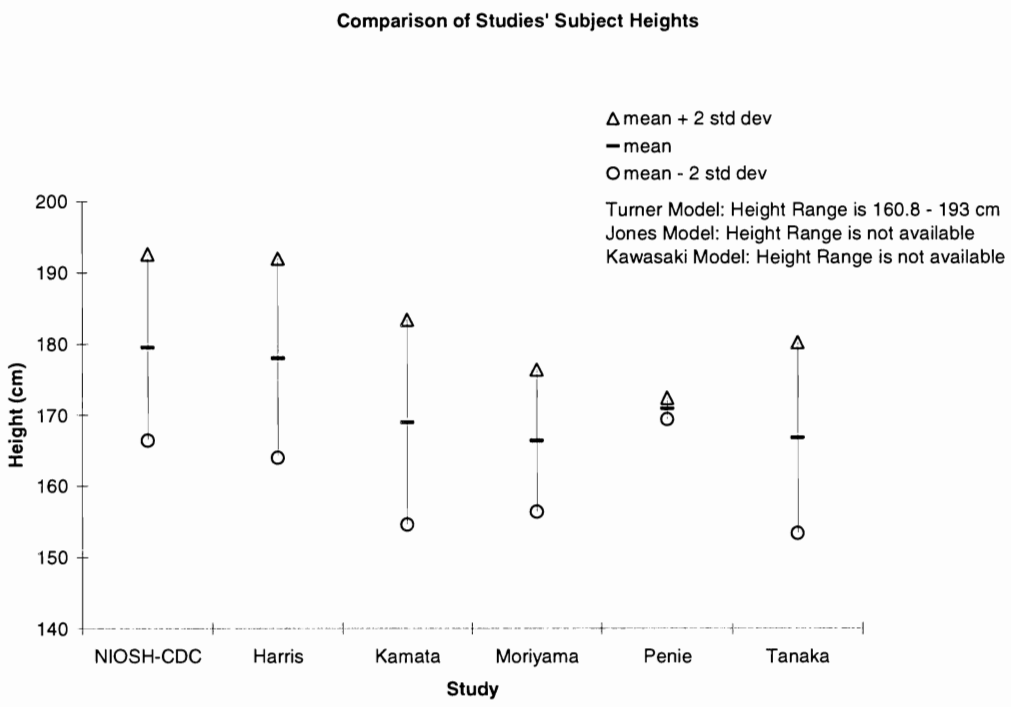

Figure 16: Comparison of Studies' Subject Heights

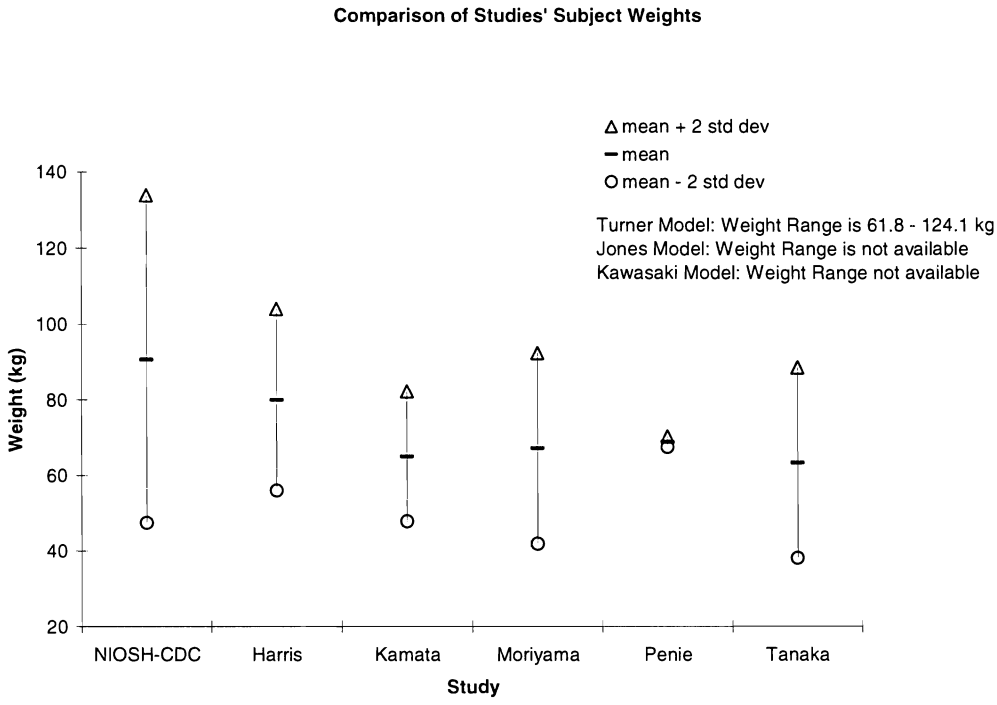

Figure 17: Comparisons of Studies' Subject Weights

Vita

Donna S. Kroos was born in Buffalo, New York and is a citizen of the United States of America. She graduated from the Pennsylvania State University in 1980 with a Bachelor of Science Degree in Mathematics. Early in her professional career, Ms. Kroos worked in the Information Systems and Product Software Engineering disciplines at the Eastman Kodak Company in Rochester, New York. She received a United States Patent in 1992 for her software in the Photobooth Compositing Apparatus (U.S. Patent Number 5,117,283). In 1994, she decided to pursue a career change and became a full-time graduate student at the Rochester Institute of Technology. In 1996, Ms. Kroos graduated from the Rochester Institute of Technology with a Master of Business Administration degree majoring in Quantitative Decision Making. Since that point in time, Ms. Kroos has held the professional positions of business research analyst (at Eastman Kodak Company), senior research associate (at Harris Interactive, Rochester, New York), and project manager/senior analyst (at Circuit City Stores Incorporated, Richmond, Virginia). Ms. Kroos currently is pursuing a Master of Science degree, at the Virginia Commonwealth University, in Mathematical Sciences with an emphasis in Statistics.