Paper SD-11

The Keouk County CAFO Study: A Complementary Analysis Using Classification Trees in SAS® Enterprise Miner™

Leonard Gordon, University of Kentucky, Lexington, KY
Brian Pavilonis, University of Iowa, Iowa City, IA

ABSTRACT

There is an increase in larger specialized operations in the livestock production sector in the United States(US). The number of hogs raised in the US has been relatively constant but the number of producers has reduced resulting in concentrated animal feeding operations (CAFOs) which has not been without consequences. The amount of waste generated has had adverse effects on the environment. Studies have shown an association between CAFO exposure and adverse respiratory outcomes.

Classification trees (CT)- a non-parametric methodology- using SAS® Enterprise Miner™ are used for the analysis. Classification trees are underused in the public health literature and they have the ability to divide populations into meaningful subgroups which will allow the identification of vulnerable groups and enhance the provision of products and services.

Keywords: concentrated animal feeding operations (CAFO), classification trees, asthma

INTRODUCTION

In recent years there has been a considerable increase in larger more specialized operations across all sectors of agriculture in the United States (US). This includes the livestock production industry. Over the years that this change has taken place, the number of hogs raised in the United States has been relatively constant while the number of producers have decreased considerably (Donham et al 2006). This migration towards concentrated animal feeding operations (CAFOs) has enhanced production and decreased costs greatly. However, this has not been without consequences. One of the major consequences of this transition has been the large amount of animal waste generated by the CAFOs. This has had adverse effects on the environment (Cole et al 2000), the CAFO workers (Schenker et al 1998) and possibly nearby residents (Iowa State University and University of Iowa 2002).

There is a substantial amount of waste material produced by CAFOs. It is estimated that 100 million tons of feces and urine are produced annually by the 60 million hogs raised in the US (Cole et al 2000).The waste storage at swine CAFOs usually occurs in pits, located underneath the building, or lagoons, where it is stored until it can be applied as fertilizer or hauled away. While in the pit/lagoon, the waste undergoes fermentation which in turn produces high levels of harmful gases including but not limited to ammonia, hydrogen sulfide and other volatile organic compounds. Additionally, the waste also contains large quantities of endotoxin, which is highly inflammatory (Von Essen et al 2005). Epidemiologic studies conducted on workers at some of these AFO facilities have shown an increased prevalence of asthma and other respiratory related diseases as a result of exposure to these environments (Schenker et al. 1998). Presently, there is ongoing investigation of whether similar effects are experienced by those who live in close proximity to these facilities. This paper seeks to add to that body of literature and research.

Studies have shown a positive association between environmental exposure to CAFOs and adverse respiratory health outcomes (Kirkhorn and Garry 2000). A study by Merchant at al. (2005) examined the association between living on farms that raised swine and childhood asthma and concluded that children who lived on a hog farm have a significantly higher prevalence of at least one asthma outcome compared to children living on a farm that did not raise the animal (44.1% vs. 33.6%). The prevalence increased from 42.9% to 46% depending in whether the farm had less than 500 or 500 or more heads of swine. Mirabelli et al (2006) also found an association between estimated exposure to airborne pollution from swine CAFOs and adolescent wheezing symptoms, which a symptom for asthma. Asthma is a chronic respiratory disease that is characterized by reversible airflow obstruction and airway hyperresponsiveness. It is influenced by both genetic and environmental factors (Moorman et al 2007). Asthma is the leading chronic childhood illness in the US with an estimated 6.5 million cases (Akinbami and Schoendorf 2002). From 1980 to 1996, asthma prevalence among children ages 0-17 years more than doubled from 3.6% in 1980 to 7.5% at the peak of the trend in 1995.The evidence of the impact of the environment has been a major theory in the explanation of the increased prevalence of asthma (Gilmour et al 2006).
This study makes a novel contribution in that it employs an analytic strategy - classification trees- which are underused in the public health literature (Migongo, Charnigo et al 2011, Lemon et al 2003). Classification tree analysis is a nonparametric decision tree methodology that has the ability to divide populations into meaningful subgroups. It was developed by Breiman and colleagues (Breiman et al 1984). The main idea of a classification tree

is a statistician's version of the "20 questions" game. It is employed when there is some apprehension in the assumption that the relationship between the explanatory variables and the logit risk is linear. This is something that most researchers take for granted in regression analysis. Additionally, it takes into consideration interactions between the explanatory variables that are in the model that were not anticipated due to its hierarchical structure. Furthermore, a classification tree can handle situations in which the categorical response variable has more than two categories.

There are two aims for this analysis. First, this study aims to evaluate the association between proximity to swine CAFOs and childhood asthma. Children were chosen for the study because they are a unique and vulnerable population. Second, this paper tries to demonstrate the value of classification trees in the investigation of associations in public health data.

METHODS

The data were obtained from the Keokuk County Rural Health Study (KCRHS), a population-based prospective cohort study that began in 1994 to primarily assess the prevalence of injury and respiratory diseases in an agricultural population (Stromquist et al 2002). The county is rural and a large percentage of the area is used as farmland.

The study sample comprises of 565 children who range in age from birth up to 18 years. Children were selected by stratified random sampling with an oversampling of farms and rural non-farming households. Asthma status was determined from the question "Has a doctor ever diagnosed the child with asthma?"

Information was collected at a research facility in a centrally located town within the county by trained interviewers aided by computers. The majority of health information was provided by the child's mother or female legal guardian. Children older than twelve were eligible to participate in an occupational questionnaire that assessed certain farm tasks that the child might have performed. After the clinical assessment, an environmental audit of the property/home was also conducted.

Latitudinal and longitudinal coordinates of each CAFO and study participants homes were recorded using publicly available aerial photography. Additional information, including year of construction and total square footage, about the swine CAFOs was collected from the tax assessor records. The coordinates of the CAFOs and homes were imported into ArcGis version 9.2 and plotted. A three mile radius was drawn around each home and every CAFO within this vicinity was considered as having an influence on the home. This three mile buffer was chosen based on previous studies. Distance and direction from the home to every CAFO within its buffer zone was calculated. Any facility that was constructed after the child's clinic visit was excluded from analysis.

An exposure metric (the CAFO coefficient) was devised to account for the effect of multiple CAFOs on a single home. The metric is shown below in equation 1. The metric makes it possible to analyze the cumulative effect of all CAFOs within a 3 mile radius of the participant's home while accounting for distance of CAFOs from home, wind direction and speed and facility square footage. The square footage of the facility was used as a surrogate for the total amount of waste produced by the facility.

$$Equation\ 1: AFO\ Exposure\ Factor = \sum \frac{square\ footage\ of\ the\ AFO}{(distance\ of\ AFO\ to\ home)^2} * wind\ percentage$$

In a bid to reduce the number of variables in the model, some variables were combined into one variable. For example, the original data set contained indicator variables for whether the children had animals in confinement or free livestock animals on their property. These were combined to form a new variable with four categories that indicated whether the children did not have neither animals in confinement nor free livestock, whether they had only animals in confinement, whether they had only free livestock and whether they had both. For the purposes of this study, a farming household was defined as having an operation of 10 acres or more in agricultural production. The selection of 10 acres was used to differentiate between commercial and individuals farming for personal consumption.

The potential risk factors were initially analyzed for differences by asthma status and descriptive statistics were obtained. Means and standard deviations were obtained for the continuous variables and frequencies and percentages were obtained for categorical variables. Independent t-tests and chi-square tests were used to assess the significance of the differences by asthma status for the continuous and categorical variables respectively. Fisher's exact test was used in place of the chi-square test when the cell frequencies were exceptionally low.

The main analysis method that was used was CT as the primary outcome variable is a dichotomous indicator variable for asthma. However, it is worth noting that a classification tree can accommodate a categorical response that has

more than two categories. In the same way that we do not want too many variables in a logistic regression model, we do not want a lot of nodes or leaves in a CT. The Gini Index is the criteria used for splitting nodes and pruning the tree. The tree is constructed in such a way that this quantity is minimized. The Gini index is the summation over all nodes of the tree of the number of subjects in the node multiplied by the proportion of subjects with the outcome of interest and its complement. The minimum number of observations in a leaf was set to 5, the maximum number of branches from a node was set to 2 and the maximum depth of the tree was set to 6.
Significance was evaluated at the 5% level.
SAS and SAS Enterprise Miner were used for the analysis (SAS 2009).

RESULTS

Descriptive statistics for the study data are shown in table 1. The study contained 565 children from 277 households. The asthma prevalence for the study cohort was slightly lower than the general US population, 10.97% vs. 12.7% (Akinbami 2006). The mean age of children in the study was 9.6 years. Additionally, there is a difference in the mean CAFO coefficient by asthma status, but this difference is not significant due to the large standard deviations of the two groups. The CAFO coefficient means by asthma status were 153,039.4 vs. 50.6 for children with and without asthma respectively.

Figure 1 shows the results of the CT. There were 13 variables selected for use in the tree. Asthma was the target variable and gender, CAFO coefficient, smoking, respiratory illness, premature birth, education, income1, indoor pets, gas stove use, cockroaches, type of animals kept on the farm and whether the children work with the animals were the input variables. Of the twelve input variables selected, six were chosen for the decision rules for the tree. The six variables that were chosen were gender, CAFO coefficient, respiratory illness, premature birth, type of animals kept on the farm and whether the children worked with animals. The inclusion of some of these variables in the prediction of asthma is consistent with other studies that have observed these variables as strong factors contributing to the development of the disease. Several studies have identified gender (Akinbami et al 2009), respiratory illness (Mirabelli et al 2006) and premature birth (Dombkowski et al 2008) to be associated with asthma.

Before any questions were asked the predicted probability of having asthma for the study cohort was 10.97%. However, this probability increased to 26.6% for children without respiratory illness and decreased to 4.9% for children with respiratory illness after the first question was asked. For children without respiratory illness and a CAFO coefficient greater than 329715.99, the predicted probability of contracting the disease increased dramatically to 100%. Continuing on with the branches of the tree it was observed that there were four subgroups of children with predicted asthma probabilities greater than 50%. The first group of children that fell into this category were those without respiratory illness, a CAFO coefficient less than 329715.99, no premature birth and females. The next group of children in this category was those without respiratory illness, a CAFO coefficient less than 329715.99, premature birth and had neither confined animals or free livestock or just confined animals. The third group of children were those with no respiratory illness, a CAFO coefficient greater than 946.46 but less than 329715.99, no premature birth and females. The final group of children were those with no respiratory illness, a CAFO coefficient less than 329715.99, premature birth and had only free livestock or both confined animals and free livestock.

Table 2 shows the asthma status and predicted asthma status for the children in the study. The misclassification rate for the process was found to be 8%. The classification tree predicted asthma status correctly for 21 of the 62 children with the disease and for 496 of the 503 children without the disease. Therefore, 7% of the children had the disease and were classified as not having the disease and 1% of the children did not have the disease and were classified as having the disease. The predictions were made based on a 50% cut off for the estimated risk. The sensitivity and specificity of the model were found to be 51.2% and 98.6% respectively. However, the positive predictive value was 75% and the negative predictive value was 92.4%.

Variable	Asthma	No Asthma	p-value
Age (Mean and SD)	10.9(4.1)	9.4(4.7)	0.01
CAFO coefficient (Mean and SD)	153039.4(483008.9)	50616.1(243531.9)	0.1
Gender (No. and %)			
Female	13(20.97%)	244(48.51%)	<.0001
Male	49(79.03%)	259(51.49%)	
Farm (No and %)			
Yes	21(33.87%)	146 (29.03%)	0.43
No	41(66.13%)	347(70.97%)	
Income (No. and %)			
High	18(29.03%)	114(22.66%)	0.26
Low	44(70.97%)	389(77.34%)	
Education (No. and %)			
High	16(25.81%)	111(22.07%)	0.51
Low	46(74.19%)	392(77.93%)	
Premature Birth (No. and %)			
Yes	45(72.58%)	458(91.05%)	<.0001
No	17(27.42%)	45(8.95%)	
Respiratory Illness (No. and %)			
Yes	20(32.26%)	387(76.94%)	<.0001
No	42(67.74%)	116(23.06%)	
Indoor Pets (No. and %)			
Yes	33(53.23%)	263(52.29%)	0.9
No	29(46.77%)	240(47.71%)	
Smoking (No. and %)			
Yes	57(91.94%)	455(90.46%)	0.71
No	5(8.06%)	48(6.54%)	
Gas Stove (No. and %)			
Yes	46(74.19%)	305(60.64%)	0.04
No	16(25.81%)	198(39.36%)	
Cockroaches (No. and %)			
Yes	60(96.77%)	493(98.01%)	0.63
No	2(3.23%)	10(1.99%)	
Animal Type (No. and %)			
0	4(6.45%)	24(4.77%)	0.74
1	2(3.23%)	17(3.38%)	
2	6(9.68%)	72(14.31%)	
3	50(80.65%)	390(77.53%)	
Animal work (No. and %)			
0	3(4.84%)	17(3.38%)	0.17
1	5(8.06%)	15(2.98%	
2	9(14.52%)	96(19.09%	
3	45(72.58%)	375(74.55%)	

Table 1. Characteristics of children by Asthma Status

		Asthma Status	
		Positive	Negative
Predicted Status	Positive	21	7
	Negative	41	496

Table 2. Asthma Status and predicted asthma status for Study children

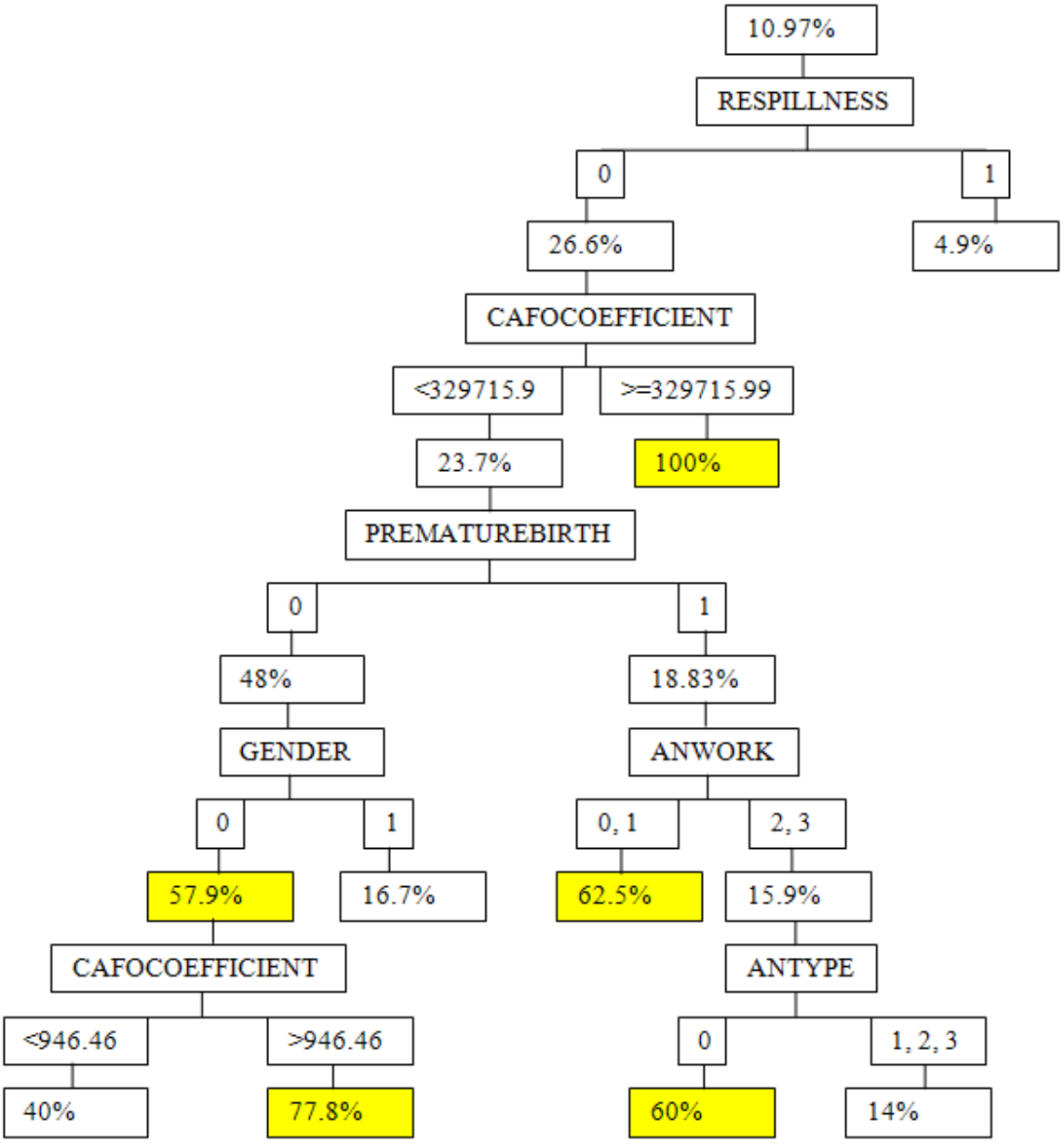

Figure 1. Classification tree for predicting Asthma status

DISCUSSION

One of the major strengths of this study is that it was able to collect a wide range of information on the participants for a single health outcome. Due to the complex nature of asthma, involving both environmental and genetic factors and possible interactions, controlling for all possible cofounders is essential to avoid biased results. However, there is also a trade-off between the amount of variables that can be entered in a model and the validity of the model taking into consideration the sample size of the study.

One of the limitations of the study is the underestimation of asthma prevalence due to the fact that asthma cases were limited to physician diagnosed asthma. It has been noted in the literature that asthma is frequently under diagnosed especially in rural communities (Chrischilles et al 2004). There is also the possibility of exposure misclassification given the time lapse between the end of the study and the collection of information on CAFOs. The former took place in 2004 while the latter took place in 2010.
One of the major advantages of classification trees s that they give a pictorial view of the results obtained. They support the age old adage that a picture is worth a thousand words.

One of the strengths of the trees analysis is that it has the ability to efficiently segment a population into meaningful subsets. This allows researchers to identify sectors of the population that are most likely to be involved with a particular health behavior and adequately target and maximize the distribution of public health resources (Lemon et al 2003). As a result, high risk populations are easily identified

Another advantage of the classification tree is the way it handles missing variables. Initially, the explanatory variable for which the subject has a missing value might not be encountered as one goes through the branches of the classification tree. Furthermore, if the explanatory variable is encountered, this can be resolved by using a surrogate question in case the original question cannot be answered. The surrogate question is chosen in such a way that that answers are similar to the answers that would have been obtained if the original question was used (Charnigo 2009). This surrogate split mechanism is specific to classification trees and is counterpart for continuous outcomes, regression trees.
In general classification trees offer a complimentary perspective to the traditional methods and they are further advantageous in that they can accommodate categorical responses with more than two categories without the complexities offered by other modeling techniques.
Finally, classification trees are non-parametric in that they do no assume a distribution for the data.

One of the limitations of the tree analysis is that sometimes one is not able to force variables into the model. This is a problem especially in epidemiological studies where we want to control for certain risk factors even when they are not necessarily significant in a model. If the tree process does not deem that risk factor important it might be omitted from the variables in the tree. As a result, they cannot be used for the estimation of average effects in which case the traditional regression methodology will be more appropriate. However, they are very useful to describe associations in the data as was the case in this study

Because of the ease with which classification trees can be constructed, it is easy to get carried away and perform "data dredging" by just entering all possible independent variables into the analysis (Lemon et al 2003). It is cautioned that classification trees be used with a theoretical consideration of which independent variables to consider.

CONCLUSION

This study suggests that that there is an association between very high CAFO coefficient and asthma status. It is observed that an increase in the CAFO coefficient is associated with an increased risk for asthma. More studies are needed to understand the dynamics by which asthma status is associated with the CAFO coefficient and the potential detrimental effects to children associated with farms and the animals on these farms.
Furthermore, it utilizes classification trees for a simple yet powerful analysis of a public health problem.

REFERENCES

Akinbami L.J., Moorman J. E., Garbe P.L. and Sondik E. J. (2009). Status of childhood asthma in the United States, 1980-2007. *Pediatrics* 123 (Suppl 3): S131-145.

Akinbami, L. J. (2006). The State of childhood asthma, United States, 1980-2005. Advance data from vital and health statistics. No. 381 Hyattsville, MD: National Center for Health Statistics.

Akinbami, Lara J., and Schoendorf, Kenneth C. (2002). Trends in Childhood Asthma: Prevalence, Health Care Utilization, and Mortality. *Pediatrics* 110(2): 315-322.

Breiman L., Friedman J.H., Olshen R.A. and Stone C. J. (1984). Classification and Regression Trees (2nd Ed.). Pacific Grove, CA; Wadsworth.

Charnigo, Richard.(2009). Data Mining in Public Health. http://www.richardcharnigo.net/CPH636S09/index.html. [Accessed May 2011].

Chrischilles, Elizabeth, Ahrens, Richard, Kuehl, Angela et al. (2004). Asthma prevalence and morbidity among rural Iowa schoolchildren. *Journal of Allergy and Clinical Immunology* 113(1): 66-71.

Cole, D., L. Todd and S. Wing (2000). Concentrated Swine Feeding Operations and Public Health: A Review of Occupational and Community Health Effects. *Environmental Health Perspectives* 108(8):685-699.

Dombkowski, Kevin J., Leung, Sonia W. and Gurney, James G. (2007). Prematurity as a predictor of Childhood Asthma among Low-Income Children. *Annals of Epidemiology* 18(4): 290-297.

Donham, Kelly J., Wing, Steven, Osterberg, Jan L. et al. (2007). Community Health and Socioeconomic Issues Surrounding concentrated animal feeding operations. *Environmental Health Perspectives* 115(2):317-320.

Gilmour, M. I., Jaakkola, M. S. et al. (2006). How exposure to environmental tobacco smoke, outdoor air pollutants, and increased pollen burdens influences the incidence of asthma. *Environmental Health Perspectives* 114(4): 627-633.

Iowa State University and University of Iowa. (2002). Iowa Concentrated Animal Feeding Operations Air Quality Study. Iowa City: The University of Iowa College of Public Health. http://www.ehsrc.uiowa.edu/cafo_air_quality_study.html. [Accessed May 2010].

Kirkhorn, S.R. and Garry V. F. (2000). Agricultural lung diseases. *Environmental Health Perspectives* 108 Suppl 4: 705-712.

Lemon, Stephenie C., Roy, Jason and Clark Melissa A. (2003). Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. *Annals of Behavioral Medicine* 26(3): 172-181.
Martinez, F. D. (2007). Recognizing early asthma. *Allergy* 54: 24-28.

Merchant, James A., Naleway, Allison L., Svendsen, Erik R. et al. (2005). Asthma and Farm Exposures in a Cohort of Rural Iowa Children. *Environmental Health Perspectives* 113(3): 350-356.

Migongo, Alice W., Charnigo, Richard, Love, Margaret M. et al. (2011). Factors Relating to Patient Visit Time With a Physician. *Medical Decision Making* (to appear).

Mirabelli, Maria C. , Wing, Steve, Marshall, Stephen W. and Wilcosky, Timothy C. (2006). Asthma Symptoms Among Adolescents Who Attend Public Schools That Are Located Near Confined Swine Feeding Operations. *Pediatrics* 118: e66-e75.

Moorman, J. E., Rudd, R. A., et al. (2007). Centers for Disease Control and Prevention (CDC), National surveillance for asthma- United States, 1980-2004. Morbidity and Mortality Weekly Report Surveillance Summary 56:1-54.

SAS Institute. (2009). SAS/STAT User's Guide. (Version 9.2). Cary: SAS Institute, Inc.

Schenker, M.B., Christiani, D., Cormier, Y. et al (1998). Respiratory health hazards in agriculture. *American Journal of Respiratory and Critical Care Medicine* 158(5).

Von Essen, Susanna G., Anderson, Colene I. and Smith Lynette M. (2005). Organic dust toxic syndrome: A noninfectious febrile illness after exposure to the hog barn environment. *Journal of Swine Health and Production* 13(5): 273-276.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Leonard Gordon
Enterprise: University of Kentucky
Address: 725 Rose Street
City, State ZIP: Lexington, KY 40536
Work Phone: 859-218-2097
Fax: 859-257-6430
E-mail: leonard.gordon@uky.edu
Web: N/A
Twitter: N/A