COMPARISON OF MACHINE LEARNING METHODS AND LINEAR REGRESSION MODELS IN IDENTIFYING IMPORTANT PREDICTOR VARIABLES FOR SERUM DIOXIN TEQ FOR A COMMUNITY IN MICHIGAN, USA

Chen Q[1], Lee S-Y[1], Hedgeman E[2], Ghosh D[1], Gillespie BW[1], Lepkowski J[3], Garabrant D[2]

[1]Department of Biostatistics, University of Michigan School of Public Health, 109 S Observatory, Ann Arbor, MI 48109; [2]Department of Environmental Health Sciences, University of Michigan School of Public Health, 109 S Observatory, Ann Arbor, MI 48109; [3]Survery Research Center, Institute for Social Research, University of Michigan, 426 Thompson, Ann Arbor, MI 48104;

Keywords: Blood, North America, Environmental samples, Dietary intake, Soil, Humans, TEQs.

Introduction and Study Goals

The primary goal of the University of Michigan Dioxin Exposure Study (UMDES) is to investigate whether known PCDD, PCDF and PCB(hereinafter collectively referred to as 'dioxins') contamination in the Tittabawassee River sediments downstream from the Dow Chemical plant are associated with elevated serum dioxin levels in the regional population. It is well known that age and body mass index (BMI) [1] can explain a part of human's serum dioxin levels. Therefore, we are interested in identifying which pathways of exposure and methods of elimination are more important in determining the serum dioxin levels after adjusting for age and BMI. The potential pathways include residential proximity to the river, property use, recreational activities in the contaminated region, consumption of food grown or raised in the river or region, consumption of fish and game from contaminated areas, various measures of soil contact, house dust dioxin levels, and occupational contact. Other factors that may affect the serum dioxin levels include weight change and smoking status for both sexes, and pregnancies and breast-feeding history for women. This study seeks to identify important pathways of exposure and methods of elimination for dioxin in serum.

Materials and Methods

In this study, a total of 946 subjects who have complete serum dioxin measures were included. Blood serum, house dust, and soil were analyzed by Alta Analytical Laboratory, Inc. Details of the serum, house dust, and soil sampling methods and analyses are reported elsewhere [2,3,4]. Additionally, the process of dealing with limit of detection issues for serum samples is reported elsewhere [5]. A 10 based logarithm transformation of the serum dioxin concentration expressed as the Toxic Equivalent (TEQ) was taken to reduce the right skewness of the distribution of the serum dioxin concentrations.

The exposure information was obtained from the UMDES questionnaire [6]. The respondent was asked to recall possible dioxin exposure pathways over their entire lifetime. These pathways included a full residential history, occupations, property use, recreational activities, and consumption of meat, fish, game, eggs, milk, other dairy products, and vegetables. Basic demographic (age, gender, race, education, income) and health questions (height, weight, weight loss and gain, smoking status, pregnancy, childbearing and months of breastfeeding for each child) were also included. The function forms of all potential predictor variables are reported elsewhere [7].

The statistical analyses were performed using two different approaches. First, we used linear regression method, in which we forced age, age square and BMI into the model, and used a backward selection procedure to identify the important predictors of the serum dioxin TEQ after adjusting for the effects of age and BMI. Details of the modeling procedure is reported elsewhere [7]. An advantage of linear regression modeling is that it provides estimates of effect magnitudes in addition to identify the significant explanatory variables. Second, we used machine learning methods, specifically, Random Forest [8] and tree-boosting [9] models. The residuals of logarithm transformed serum dioxin

concentration after adjusting for age, age square, and BMI were obtained as outcomes in the machine learning methods. The relative importance of each potential predictor variable to the overall model was calculated. The advantage of using the tree-boosting method is that there is no need to consider input variable transformations, since all tree-boosting procedures are invariant under all strictly monotone transformations of the individual input variables [9]. All the analyses were done using statistical software SAS 9.1 [10] and R 2.2.1 [11], respectively, for linear regression models and machine learning methods. The results of these two approaches for modeling the total serum dioxin TEQ were compared. The consistencies and inconsistencies of the two sets of results are reported.

Results and Discussion

Table 1 lists the important predictor variables identified in the linear regression models, and the ranks of the estimated importance of predictor variables using machine learning methods. For the machine learning methods, the higher the rank (highest=1), the more important is the predictor variable. Both methods agree that XXX are more important in predicting the blood dioxin levels. However, XXX are identified as important pathways in the linear regression models, but not by the machine learning models. XXX have high importance measures in machine learning models, but are not significant in the linear regression models.

Table 1: Important predictor variables identified using linear regression models and machine learning methods

Category	Variables	Linear Regression	Random Forest	Tree-boosting
Dust				
Soil				
Health				
Demographics				
Food				
Residence History				
Occupations				
Property Use				
Recreational Activities				

In addition, Figure 1 displays single variable partial dependence plots on those most important predictor variables identified by both methods for serum dioxin levels. The figures shows that XXX

Figure 1

In conclusion, XXX

Acknowledgements
The authors acknowledge the Dow Chemical Company for funding the study and Ms. Sharyn Vantine for her continued assistance.

References

1. Patterson Jr, DG, Patterson, D, Canady, R, Wong, L-Y, Lee, R, Turner, W, Caudill, S, Needham, L, Henderson, A. *Organohalogen Comp* 2004; 66:2878-2883.
2. Hedgeman E, Chen Q, Gillespie BW, Franzblau A, Knutson K, Zwica L, Sima C, Lee S-Y, Lepkowski J, Ward B, Ladronka K, Olson K, Sinibaldi J, Towey T, Adriaens P, Demond A, Chang S-C, Gwinn D, Swan S, Garabrant D. *Organohalogen Comp 2006* (Forthcoming)
3. Zwica L, Knutson K, Towey T, Hedgeman E, Franzblau A, Chen Q, Lee S-Y, Sima C, Gillespie BW, Adriaens P, Demond A, Lepkowski J, Ward B, Ladronka K, Olson K, Sinibaldi J, Chang S-C, Gwinn D, Swan S, Garabrant D. *Organohalogen Comp 2006* (Forthcoming)
4. Adriaens P, Demond A, Towey T, Chang S-C, Chen Q, Franzblau A, Gillespie BW, Gwinn D, Hedgeman E, Knutson K, Ladronka K, Lee S-Y, Lepkowski J, Olson K, Sima C, Sinibaldi J, Swan S, Ward B, Zwica L, Garabrant D. *Organohalogen Comp 2006* (Forthcoming)
5. Gillespie BW, Chen Q, Lee SY, Hong B, Garabrant D, Hedgeman E, Sima C, Lepkowski J, Olson K, Luksemburg W. *Organohalogen Comp* 2006 (forthcoming).
6. Olson, K[1], Lepkowski, J[1], Lohr-Ward, B[1], Ladronka, K[1], Sinibaldi, J[1], Garabrant, D[2], Franzblau, A[2], Adriaens, P[3], Gillespie, B[4], Bandyk, J[1], Chang, S-C[2], Chen, Q[2], Demond, A[3], Gwinn, D[4], Hedgeman, E[2], Hong, B[2], Knutson, K[2], Lee, S-Y[3], Sima, C[2], Towey, T[3], Wright, D[2], Zwica, L[2]. *Organohalogen Comp* 2006 (forthcoming).
7. Garabrant D[1], Franzblau A[1], Lepkowski J[2], Adriaens P[3], Demond A[3], Hedgeman E[1], Knutson K[1], Zwica L[1], Chen Q[4], Olson K[2], Ward B[2], Towey T[3], Ladronka K[2], Sinibaldi J[2], Chang S-C[3], Lee S-Y[4], Gwinn D[5], Sima C[4], Swan S[5], Gillespie BW[4]. *Organohalogen Comp* 2006 (forthcoming).
8. Breiman L. "Random Forests". *Machine Learning 2001*, 45: 5-32
9. Friedman J.H. "Greedy Function Approximation: A Gradient Boosting Machine". *The Annals of Statistics 2001*, 29: 1189-1232
10. SAS Institute. SAS/STAT User's Guide Version 8. Cary, NC: SAS Institute Inc.
11. The R Foundation for Statistical Computing, Version 2.2.1, 2005