Human and Ecological Risk Assessment: An International Journal
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400879

Determinants of Serum PCBs in Adolescents and Adults: Regression Tree Analysis and Linear Regression Analysis

Eva Govarts[a]; Elly Den Hond[a]; Greet Schoeters[ab]; Liesbeth Bruckers[c]
[a] Environmental Risk and Health, Flemish Institute of Technological Research (VITO), Mol, Belgium [b] Department of Biomedical Sciences, University of Antwerp (UA), Antwerp, Belgium [c] University of Hasselt, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Diepenbeek, Belgium

Online publication date: 11 October 2010

To cite this Article Govarts, Eva , Den Hond, Elly , Schoeters, Greet and Bruckers, Liesbeth(2010) 'Determinants of Serum PCBs in Adolescents and Adults: Regression Tree Analysis and Linear Regression Analysis', Human and Ecological Risk Assessment: An International Journal, 16: 5, 1115 — 1132
To link to this Article: DOI: 10.1080/10807039.2010.512256
URL: http://dx.doi.org/10.1080/10807039.2010.512256

Human and Ecological Risk Assessment, 16: 1115–1132, 2010

ISSN: 1080-7039 print / 1549-7860 online
DOI: 10.1080/10807039.2010.512256

Exposure Assessment Articles

Determinants of Serum PCBs in Adolescents and Adults: Regression Tree Analysis and Linear Regression Analysis

Eva Govarts,[1] Elly Den Hond,[1] Greet Schoeters,[1,2] and Liesbeth Bruckers[3]
[1]Flemish Institute of Technological Research (VITO), Environmental Risk and Health, Mol, Belgium; [2]University of Antwerp (UA), Department of Biomedical Sciences, Antwerp, Belgium; [3]University of Hasselt, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Diepenbeek, Belgium

ABSTRACT

Regression tree analysis, a non-parametric method, was undertaken to identify predictors of the serum concentration of polychlorinated biphenyls (sum of marker PCB[1] 138, 153, and 180) in humans. This method was applied on biomonitoring data of the Flemish Environment and Health study (2002–2006) and included 1679 adolescents and 1583 adults. Potential predictor variables were collected via a self-administered questionnaire, assessing information on lifestyle, food intake, use of tobacco and alcohol, residence history, health, education, hobbies, and occupation. Relevant predictors of human PCB exposure were identified with regression tree analysis using ln-transformed sum of PCBs, separately in adolescents and adults. The obtained results were compared with those from a standard linear regression approach. The results of the non-parametric analysis confirm the selection of the covariates in the multiple regression models. In both analyses, blood fat, gender, age, body-mass index (BMI) or change in bodyweight, former breast-feeding, and a number of nutritional factors were identified as statistically significant predictors in the serum PCB concentration, either in adolescents, in adults or in both. Regression trees can be used as an explorative analysis in combination with multiple linear regression models, where relationships between the determinants and the biomarkers can be quantified.

Received 17 August 2009; revised manuscript accepted 4 November 2009.
Address correspondence to Eva Govarts, Flemish Institute of Technological Research (VITO), Environmental Risk and Health, Boeretang 200, 2400 Mol, Belgium. E-mail: eva.govarts@vito.be

[1]**ABBREVIATIONS:** BMI: body-mass index, CV: cross validation, ln: natural logarithm, ns: not significant, PCAHs: polychlorinated aromatic hydrocarbons, PCBs: polychlorinated biphenyls, R^2_a: adjusted coefficient of determination, VIF: variance inflation factor.

Key Words: human biomonitoring, biomarkers of exposure, polychlorinated biphenyls (PCBs), non-parametric analysis, regression trees, multiple linear regression.

INTRODUCTION

Human biomonitoring is applied to assess human exposures to environmental and workplace chemicals. Biomarkers of exposure take into account inter-individual differences in absorption, distribution, biotransformation, and excretion of a substance that may be associated with differences in age, gender, genetic constitution, height, weight, physiologic and nutritional status, duration of exposure, and so on (Lauwerys and Hoet 2001).

Between 2002 and 2006, a large human biomonitoring campaign was conducted in Flanders, the Dutch-speaking part of Belgium. In this campaign, biomarkers of exposure were measured in combination with biomarkers of effect, supplemented with individual health data from questionnaires. A biomarker of exposure is a proxy for the level of pollutant in the body, a biomarker of effect provides a measure for an early biological effect of altered function in the human body in reaction to environment pollutant exposure. As biomarker of exposure to polychlorinated aromatic hydrocarbons (PCAHs), we studied the serum concentration of polychlorinated biphenyls (sum of marker PCB 138, 153, and 180) in more than 3200 participants of the general population including 14- to 15-year-old teenagers (n = 1679) and adults between 50 and 65 years (n = 1583).

This study aims to compare the results obtained from the decision trees with the results of the former executed linear regression analyses (Den Hond *et al*. 2009) to determine the major factors of inter-individual variability of serum PCBs in those two age groups. Multiple linear regression techniques are often used to identify the factors that influence a biomarker. However, this parametric method may lead to unfaithful data descriptions when the underlying assumptions are not satisfied. Model interpretation can be problematic in the presence of higher-order interaction among potential predictors. Missing data may lead to serious loss of information. As a consequence, we may end up with imprecise or even false conclusions. In that case non-parametric methods such as regression tree analysis can be used as an alternative (Zhang and Singer 1999).

The results of the linear regression and regression tree analyses were compared in terms of identification, relative importance, and quantification of the relationship of the predictor variables with the biomarker values.

MATERIAL AND METHODS

Environment and Health Study

Between years 2002 and 2006 the Flemish Centre of Expertise on Environment and Health introduced a human biomonitoring network in Flanders. Three age groups were involved in the human biomonitoring study: newborns and their mothers, 14–15-year-old adolescents and adults between 50 and 65 years old. In total,

about 4500 participants were systematically recruited. This study will focus on adolescents and adults and will address only exposure markers.

Participants were enrolled at random within primary sampling units, that is, secondary schools (n = 42) for adolescents and communities (n = 46) for adults. Invitation letters were sent via the schools and by regular post. 71.6% of the adolescents and 47.5% of the adults replied to the invitation and respectively 85.7% and 75.3% of those who answered fulfilled the inclusion criteria and agreed to participate. The inclusion criteria were: being a resident of the study area for at least 5 years, be able to fill in a Dutch questionnaire, and give written informed consent. The study was approved by the medical ethical committee of the University of Antwerp.

Measurements of Marker PCBs in Serum

Marker PCB 138, 153, and 180 were analyzed by gas chromatography equipped with an electron capture detector using the method of Gomara *et al.* (2002). Chemical analyses were performed by two labs. Both laboratories participated in the AMAP proficiency testing scheme (Institut National de Santé Publique, Quebec, Canada). Precision (relative standard deviation) was estimated using results of the ClinChek and AMAP samples and ranged between 6.7% and 9.3% for all compounds. The limit of detection for all chlorinated compounds in serum was 0.02 μg/l. The total serum lipid concentration was determined gravimetrically. In case no value could be obtained gravimetrically, total lipid concentration was calculated on the basis of routinely measured triglycerides and total cholesterol by the following formula: total lipids = 1.33 × (triglycerides + cholesterol) + 50.5 mg/dl (Covaci *et al.* 2006). Both laboratories applied standard agreed quality control/quality assurance procedures.

Questionnaire Data

All participants completed an extensive self-administered questionnaire, assessing information on lifestyle, dietary intake, use of tobacco and alcohol, residence history, health, education, hobbies, and occupation (if applicable). Before the study, the questionnaire was pre-tested in a group of adults and adolescents with a lower educational level. Questions were adapted to the comments of the test group. Information from the questionnaire was carefully checked before being used in the statistical analysis. For the dietary questionnaire, inconsistent answers or extreme outlying values in the answers were checked by re-contacting the participant by telephone. Height and body weight were measured according to a standardized protocol (WHO 1995). Dietary intake was assessed via a semi-quantitative food frequency questionnaire (FFQ) as described in detail by Bilau *et al.* (2008). For each individual, dietary intake of fat from different sources was estimated in grams per day, for different sources of fat, that is, beef, pork, sheep, horse, chicken, turkey, cereals, yoghurt, milk, eggs, cheese, cooking and frying fats, seafood (shrimps and mussels), and fish (lean, fatty, smoked and canned). Participants were asked whether they regularly use local food products (meat or vegetables).

Statistical Analysis

Database management and statistical analyses were done with SAS software version 9.1 and JMP software version 8 (SAS Institute Inc., Cary, NC, USA).

Biomarker values below the detection limit (0.02 μg/l) were first replaced by half the detection limit, a method that is often applied in environmental biomonitoring (CDC 2005; Moore and McCabe 1999). Then, concentrations were expressed in molar units by using the following conversion factors: PCB congeners 138 and 153: 1 μg = 2.771 nmol and PCB congener 180: 1 μg = 2.530 nmol. Finally, the sum of marker PCB 138, 153, and 180 in molar units was calculated.

Both in adolescents and adults, the not normally distributed sum of marker PCBs, further referred to as PCBs, was ln-transformed (natural logarithm) and described by its geometric mean and 95% confidence interval (CI).

Based on a literature search and experience of the research group, 46 explanatory variables in adolescents, and 38 in adults, were selected as possibly related to the PCB measurement and were introduced in a regression tree and a multiple linear regression model.

Decision Trees

Classification and regression trees are used for predicting categorical dependent variables (classification) and continuous dependent variables (regression) (Breiman *et al*. 1984). In summary, a decision tree partitions data into smaller segments called terminal nodes or leaves that are homogeneous with respect to the target variable or the dependent variable. Partitions are defined in terms of other variables called input variables, thereby defining a predictive relationship between the inputs and the target. This partitioning continues until the subsets cannot be partitioned any further, according to user-defined stopping criteria. By creating homogeneous groups, one can predict with greater certainty how individuals in each group will behave in terms of the dependent variable.

The process of computing decision trees can be characterized by four basic steps: tree building, stopping tree building, tree pruning, and tree selection (Breiman *et al*. 1984; Lewis 2004).

Tree building begins at the root node, which includes all study subjects. Beginning with this node, the best possible predictor variable to split the node into two child nodes is searched for. In order to find the best variable, all possible splitting variables (called splitters), as well as all possible values of the variable to be used to split the node need to be considered. The Maximize Split Statistic is used to split and to select the best splitter. The Maximize Split Statistic splits based on the raw value of sum of squares (SAS Institute Inc. 2008). It uses the splitter that provides the largest reduction in the unexplained sum of squares of the parent node, that is the splitter that maximizes the between sum of squares of the resulting child nodes. It splits the data into segments that are as homogeneous as possible with respect to the dependent variable. A terminal node in which all cases have the same value for the dependent variable is a homogeneous, "pure" node.

Stopping Tree Building: The tree building process goes on until it is impossible to continue. The process is stopped when: (1) there is only one observation in each of the child nodes; (2) all observations within each child node have the identical

distribution of predictor variables; or (3) an external limit on the number of levels in the maximal tree has been set by the user. The "maximal" tree that is created is generally overfitting. In other words, the maximal tree also called saturated tree follows every idiosyncrasy in the dataset, many of which are unlikely to occur in a future human biomonitoring program. The later splits in the tree are more likely to represent overfitting than the earlier splits.

Tree Pruning: The objective is to find the subtree of the saturated tree that is most "predictive" of the outcome and least vulnerable to noise in the data.

In order to select a simpler tree, the appropriately fit final tree, the method of "cost-complexity" pruning is used. Starting from the terminal nodes, branches of the tree are pruned (cut-off) by comparing for each split the additional accuracy the split added to the entire tree as compared to the additional complexity. The accuracy was assessed by the 10-fold cross validation (10-fold CV) coefficient of determination and the complexity by the number of terminal nodes.

Ten-fold CV was performed for validating a procedure for model building. In 10-fold CV, the dataset is randomly split into 10 sections, stratified by the outcome variable of interest. One of these subsets of data is reserved for use as an independent test dataset, while the other 9 subsets are combined for use as the learning dataset in the tree building procedure. The entire tree building procedure is repeated 10 times, with a different subset of the data reserved for use as the test dataset each time (Lewis 2004). In this way, a 10-fold cross validation coefficient of determination was obtained. The relative importance of the predictors was determined by the contribution of each predictor to the total sum of squares of the final regression tree.

Multiple Linear Regression Models

Multiple linear regression models were built to identify the predictors of the biomarker levels, separately in adolescents and adults (Den Hond *et al*. 2009). This was done for serum PCBs expressed in nmol/l with blood fat (mg/dl) added as a separate explanatory variable in the model (volume-based units with adjustment for blood fat in the model). Possible covariates that determine inter-individual variation of the concentration of PCBs were listed based on a literature search. A multiple linear regression model was built including covariates that were significant at a 10% level in separate univariate analyses. Important covariates were identified by stepwise regression procedures in which we set the *p*-value at 0.10 for the independent variables to enter and at 0.05 to stay in the model. The adjusted R-square (coefficient of determination, R^2_a) of the obtained multiple linear regression models indicates the proportion of variability in the biomarker values that is accounted for by this model, penalizing for the number of explanatory variables in the model (Neter *et al*. 1996).

Variance inflation factors (VIFs) were used to analyze the effects of multicollinearity. If the VIFs were larger than 10, multicollinearity was concluded (Neter *et al*. 1996; Fox 1991).

The assumptions of normality, constancy of variance, independence (randomness), and linearity were checked with informal diagnostic plots and formal

tests (White's General test for constancy of variance (White 1980), Kolmogorov-Smirnov test for normality and the lack of fit test for linearity) (Neter *et al.* 1996).

Influential outlying cases, that is, cases that heavily influenced the fitted model, were identified. The impact of these outliers on the obtained model estimates was investigated. Models were fitted with and without influential cases and it was studied whether exclusion of the influential cases significantly changed the regression parameters. This has been reported more in detail by Den Hond *et al.* (2009).

The General Dominance Index was used for quantifying the relative importance of the predictors in the multiple linear regression model. This index is defined as the average increment in the coefficient of determination associated with predictor x across all possible sub-models (Chao *et al.* 2008). These sub-models are the p! orderings of how the p predictors can sequentially enter the model one-at-a-time. Quantitative relationships between the determinants and the biomarkers were calculated from the estimates of the beta coefficients of the multiple linear regression model, assuming that, when quantifying the relation of one covariate with the biomarker, all other covariates in the model are fixed at the population mean.

RESULTS

Descriptive statistics and mean exposure values for the sum of marker PCBs in the two study populations are given in Table 1.

Table 1. Descriptive statistics in the two age groups.

		Adolescents	Adults
N		1679	1583
Gender	female/male	792/887	808/775
Age (years)	mean ± SD	14.9 ± 0.5	57.6 ± 4.1
	range	13.8–16.5	49.8–65.3
BMI	mean ± SD	20.6 ± 3.1	26.9 ± 4.2
	range	13.7–36.6	15.1–48.1
Smoking	% non-smokers	86.4%	44.6%
	% former smokers	—	37.2%
	% current smokers	13.6%	18.3%
Blood fat in mg/dl	mean (95% CI)	444 (440–448)	612 (604–620)
Serum PCBs in nmol/l	geometric mean (95% CI)	0.781 (0.762–0.801)	5.485 (5.370–5.603)
Serum PCBs in pmol/g fat	geometric mean (95% CI)	178 (174–183)	919 (901–937)

PCBs: sum of marker PCB 138, 153, and 180.

Identification of Determinants of Serum PCBs

Decision trees

For the two age groups separately, we constructed regression trees for the ln-transformed PCBs expressed in nmol/l in order to identify the factors that determine inter-individual variability in the serum levels of PCBs. The results of a pruned tree for PCBs for the adolescent and adult dataset, respectively, are graphically displayed in Figures 1 and 2. The Maximize Split Statistic was used; it was specified that parent nodes consisted of at least 50 observations and child nodes of at least 20.

For both age groups the resulting tree has 16 terminal nodes.

In adolescents, the explanatory variables involved, constructing the pruned tree are: body-mass index (BMI), gender, being breast-fed as baby and duration of breast-feeding, blood fat concentration, and consumption of local meat (Figure 1).

In adults, the explanatory variables involved, constructing this tree are: blood fat concentration, change in bodyweight, age, gender, consumption of chicken, fish, dairy, and added fats (Figure 2).

Multiple linear regression analysis

For the two age groups separately, we constructed multiple linear regression models for the ln-transformed PCBs expressed in nmol/l in order to identify the factors that determine inter-individual variability in the serum levels of PCBs. The model explained 43% and 30% of the variability in respectively adolescents and adults. These models were checked for influential outliers. This has been reported more in detail by Den Hond *et al.* (2009). Fitting the model with or without the influential cases changed the adjusted R^2 of the model from 0.43 to 0.51 and 0.30 to 0.38, respectively. Identification and exclusion of influential cases in multiple regression models lead to better prediction models, with a higher proportion of the variability explained. Also, exclusion of influential outliers sometimes changed the regression coefficients (but not the sign) of the covariates and had an influence on the covariates that were retained as significant ($p < .05$) in the model. When we exclude influential cases, these models will be better applicable to the general population, and are therefore preferential for studies in which we want to extrapolate general guidelines for human biomonitoring. The linear regression model assumptions of normality, constancy of variance and independence were checked and were fulfilled for the final models without influential outliers both for adolescents and adults.

The determinants that were retained at the .05 level of significance in the model were BMI, gender, being breast-fed, blood fat concentration, education, age of mother at childbirth, area of residence, consumption of local meat and consumption of eel and dairy fat for adolescents. For adults these were change in body weight, blood fat concentration, age, consumption of local meat, area of residence, smoking status and consumption of eel, mussel, chicken, and vegetable fat.

Comparison regression tree and linear regression analysis

In adolescents, five predictors (BMI, gender, breast-feeding, blood fat content, and consumption of local meat) were identified both by the regression tree and the linear regression model. In the regression model, some additional explanatory

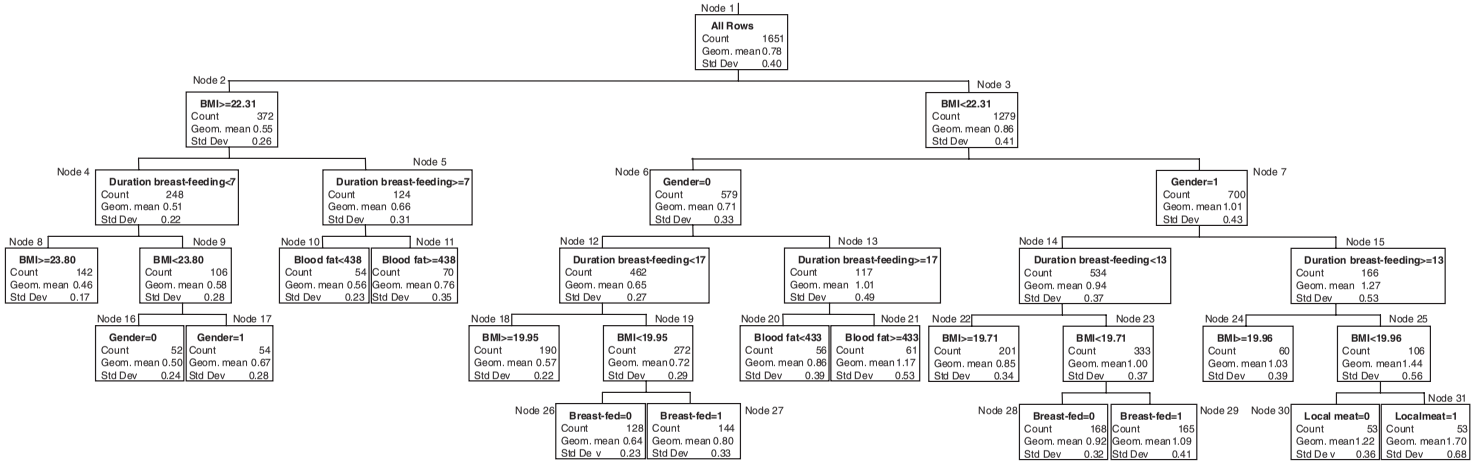

Figure 1. Final regression tree for the ln-transformed PCBs in nmol/l serum in Flemish adolescents.

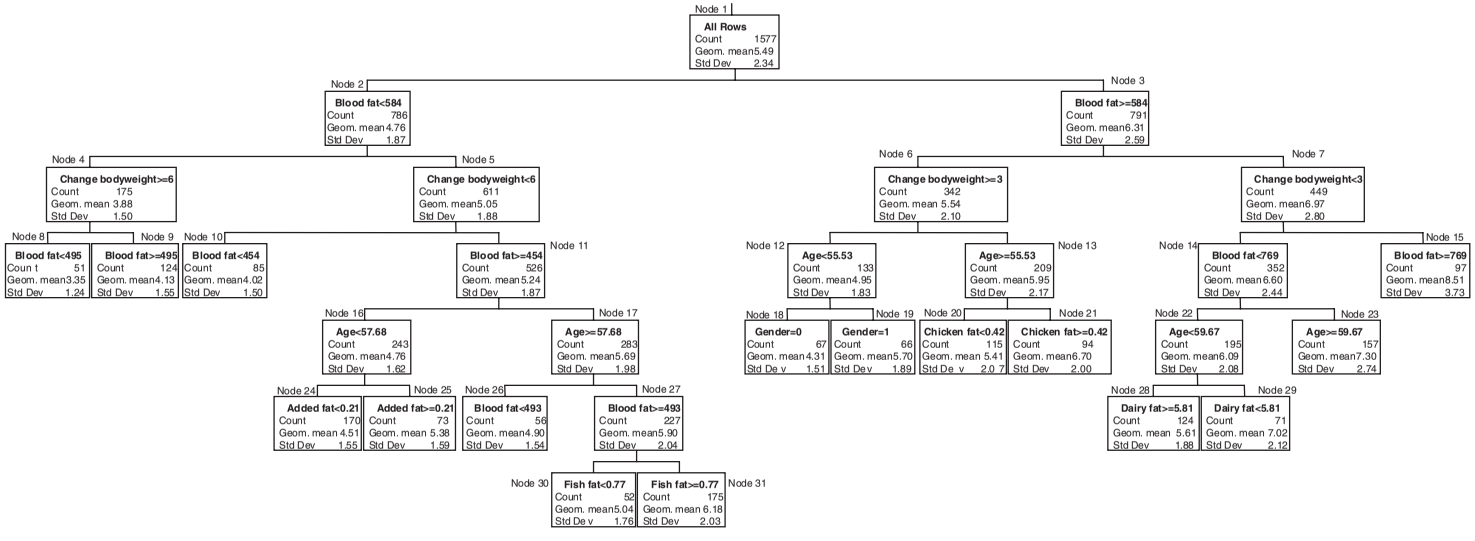

Figure 2. Final regression tree for the ln-transformed PCBs in nmol/l serum in Flemish adults.

variables were significant. Also in adults, five identical predictors (blood fat content, change in bodyweight, age, gender, and consumption of chicken fat) were identified both by the regression tree and the linear regression model. In each method, additional significant variables were selected.

Importance of Determinants of Serum PCBs

Decision trees

For the regression trees the 10-fold cross validation coefficient of determination indicates the proportion of variability in the biomarker values that is accounted for by this tree. In the Flemish adolescents and adults, respectively 36% and 28% of the variability in the PCB levels was explained by the final regression tree (Table 2). The relative importance of the predictors was determined by the contribution of each predictor to the total sum of squares of the final regression tree. In adolescents, the BMI contributed for about of half the 10-fold CV R^2, gender and breast-feeding for another large part, and small contributions of blood fat content and consumption of local meat. In adults, the blood fat content contributed for half the 10-fold CV R^2, change in bodyweight and age for almost the other half, and small contributions of gender and different sources of fat (chicken fat, dairy fat, fish fat, and added fat).

Multiple linear regression analysis

The adjusted R-square (coefficient of determination, R^2_a) of the obtained multiple linear regression models indicates the proportion of variability in the biomarker

Table 2. Contribution of the covariates§ and 10-fold CV R-square of the final regression tree for the ln-transformed PCBs in nmol/l serum in the two age groups.

		Adolescents	Adults
Predictors		Predictor Contribution§ (%)	
BMI	kg/m^2	16.56	ns
Breast-feeding	1: yes, 0: no weeks	9.11	—
Gender	1: male, 0: female	8.88	0.86
Blood fat	mg/dl	1.27	14.67
Local meat	1: yes, 0: no	0.65	ns
Change in bodyweight	kg	ns	6.47
Age	years	ns	3.23
Chicken fat	g fat/day	ns	0.78
Dairy fat	g fat/day	ns	0.75
Fish fat	g fat/day	ns	0.56
Added fat	g fat/day	ns	0.52
		10-fold CV R^2 (%)	
Model		36.46	27.83

ns: not significant; —: not applicable (information not available or not relevant).
§ The relative importance of the predictors was determined by the contribution of each predictor to the total sum of squares of the final regression tree.

Table 3. General Dominance Indexes§ of the covariates and Adjusted R-square of the final linear regression model for the ln-transformed PCBs in nmol/l serum in the two age groups.

		Adolescents	Adults
Predictors		General Dominance Index§ (%)	
BMI	kg/m^2	18.63	ns
Gender	1: male, 0: female	12.56	ns
Breast-feeding	1: yes, 0: no	6.28	—
Blood fat	mg/dl	2.70	18.42
Local meat	1: yes, 0: no	2.66	1.14
Education	1: primary school, 2: lower secondary school, 3: higher secondary school	1.98	ns
Dairy fat	g fat/day	1.52	ns
Age mother	years	0.98	—
Eel fat	g fat/day	0.68	1.84
Area	1: fruit area, 2: Albert Canal area, 3: waste incinerators, 4: Antwerp city, 5: Ghent city, 6: industrial area (harbors), 7: Olen area, 8: rural area	0.22	0.08
Age	years	ns	6.23
Change in bodyweight	kg	ns	5.08
Chicken fat	g fat/day	ns	1.04
Mussel fat	g fat/day	ns	1.01
Smoking status	1: never smoker, 2: past smoker, 3: current smoker	ns	0.85
Vegetable fat	g fat/day	ns	0.51
		Adjusted coefficient of determination (R^2_a) (%)	
Model		50.81	37.98

ns: not significant; —: not applicable (information not available or not relevant).
§ The General Dominance Index is defined as the average increment in the coefficient of determination associated with predictor x across all possible sub-models.

values that is accounted for by this model, penalizing for the number of explanatory variables in the model. In the Flemish adolescents and adults, respectively 51% and 38% of the variability in the PCB levels was explained by the model (Table 3). The General Dominance Index was used for quantifying the relative importance of the predictors in these multiple linear regression models. In adolescents, the BMI contributed the most to the adjusted R^2. Gender and breast-feeding contributed for another large part, and smaller contributions came from the blood fat content, consumption of local meat, educational level of the family, dairy fat, age of the mother at childbirth, eel fat, and area of residence. In adults, the blood fat content

contributed for half the adjusted R^2, change in bodyweight and age for another large part, with some small contributions of consumption of local meat, different sources of fat (eel fat, chicken fat, mussel fat, and vegetable fat), smoking status, and area of residence.

Comparison regression tree and linear regression analysis

In both age groups, the most important predictors identified by the two methods were identical, that is, the top five predictors (BMI, gender, breast-feeding, blood fat content, and consumption of local meat) in adolescents and the top three predictors (blood fat content, change in bodyweight, and age) in adults. Thereby defining an important variable as a predictor variable whose contribution in explaining the variability in the serum PCBs is relatively high.

Quantification and Interpretation of Determinants of Serum PCBs

Decision trees

The factors that contributed most to the variability of the level of serum PCBs were BMI, breast-feeding, and gender for adolescents and blood fat content, change in body weight, and age for adults (Table 2). Starting from the final regression trees, the interpretation of the results is straightforward.

In adolescents (Figure 1), node 31 contains 3.21% of the observations and on average had the highest PCB values (geometric mean = 1.70 nmol/l). This node contains the boys with a BMI less than 19.96, who consumed local meat and who were breast-fed for at least 13 weeks. Node 8 contains 8.6% of the observations and on average had the lowest PCB values (geometric mean = 0.46 nmol/l). This node contains the adolescents with a BMI higher than or equal to 23.80, who were breast-fed for less than 7 weeks.

In adults (Figure 2), node 15 contains 6.15% of the observations and on average had the highest PCB values (geometric mean = 8.51 nmol/l). This node contains the adults with a blood fat concentration higher than or equal to 769 mg/dl blood and with a weight gain less than 3 kg over the last 10 years. Node 8 contains 3.23% of the observations and on average had the lowest PCB values (geometric mean = 3.35 nmol/l). This node contains the adults with a blood fat concentration lower than 495 mg/dl blood and with a weight gain higher than or equal to 6 kg over the last 10 years.

Both in adolescents and in adults, we found a positive association between the blood fat content and variations in the serum PCB levels. In adults, the blood fat content was a major predictor of the PCB level, since it was the first splitter (node 1), and also further splits were based on the blood fat content (node 4, 5, 7, and 17). In adolescents, however, there was only a minor impact of the blood fat content on the serum concentrations of PCBs, since it was only selected as a splitter in a subgroup of the population (node 5 and 13), while for the remaining adolescents, no effect was found of the blood fat content on the serum PCB levels.

In adolescents, the duration of breast-feeding as a baby had a major impact on the serum concentration of PCBs of the teenager. The longer they were breast-fed the higher their serum PCB concentrations. Moreover, boys with a BMI lower than 19.71 kg/m^2 and girls with a BMI below 19.95 kg/m^2 who were breast-fed

(respectively, nodes 29 and 27) had higher serum concentrations of PCBs compared to those within the same BMI-class who only received formula (respectively, nodes 28 and 26).

Both in adolescents and in adults, we found an effect of gender on serum levels of PCBs (*i.e.,* levels were higher in males than in females). In adolescents, gender was an important predictor for almost the whole population, while in adults, gender showed only an effect in subgroups of the population, that is, in those younger than 55.53 years with a blood fat content higher than or equal to 584 mg/dl blood and a weight gain higher than or equal to 3 kg over the last 10 years (node 12).

In adolescents, a negative relationship was found between serum levels of PCB and BMI. The first split divided the adolescents into a group with BMI content higher than or equal to 22.31 kg/m^2 and a group lower than 22.31 kg/m^2, having a geometric mean PCB level of 0.55 and 0.86 nmol/l, respectively. On the other hand, an effect of changes in bodyweight was only found in adults, a weight gain resulted in lower serum PCB levels.

In adults, there was a positive association between the age of the respondent and the serum concentration of PCBs. As the age range in adolescents was only 2.7 years, it was not surprising that no relationship between PCBs and age was observed.

In adults, a small effect was found between serum PCBs on the one hand and chicken, dairy, fish, and added fats on the other hand, with higher fat consumption leading to higher PCB concentrations except for dairy fat. The effect of consumption of these fats was not observed for the whole population, only for some subgroups (nodes 13, 16, 22, and 27). In adolescents, a small effect of the consumption of local meat on the serum PCB concentration was found. Boys with a BMI lower than 19.96 kg/m^2 who were breast-fed for at least 13 weeks and who consumed local meat had a geometric mean PCB level of 1.70 nmol/l (node 31) compared to a geometric mean of 1.22 nmol/l for those who did not consume local meat (node 30).

Multiple linear regression analysis

The factors that contributed most to the variability of the level of serum PCBs were BMI, gender, and being breast-fed for adolescents and blood fat content, age, and change in body weight for adults (Table 3). Starting from the multiple linear regression models, estimates of the beta coefficients were used to quantify the relationship between PCB levels and significant determinants.

In both age groups, we found a relationship between changes in the blood fat content and variations in the serum concentrations of PCBs. An increase of the blood fat content with 100 mg/dl was associated with a 14% increase ($p < .0001$) in adolescents and a 11% increase ($p < .0001$) in adults in the concentration of PCBs.

In adults, there was a strong association between the age of the respondent and the serum concentration of PCBs. In 50- to 65-year-old adults, an increase of age with 5 years resulted in an increase in serum levels of PCBs with 11% ($p < .0001$). As the age range in adolescents was only 2.7 years, it was not surprising that no significant relationship between PCBs and age was observed in adolescents. On the other hand, the serum PCB concentration in adolescents was significantly ($p = .003$) influenced by the age of the mother at the adolescent's birth, with an increase of 5 years leading to an increase of 3% in the serum PCB concentration. Also, being breast-fed as a baby

had an impact on the serum concentration of PCBs of the teenager. Compared to adolescents who only received formula, those who were breast-fed had 26% higher serum concentrations of PCBs ($p < .0001$).

With respect to gender, we found only a significant effect in the adolescent group. Serum levels of PCBs were 40% higher in boys than in girls ($p < .0001$). Also for BMI, a significant relationship ($p < .0001$) was found in adolescents only. Serum levels of PCB were significantly negative related to BMI: the serum levels of PCBs decreased 40% when BMI was rising with 5 kg/m^2. On the other hand, changes in bodyweight were only found significant ($p < .0001$) in adults. A weight loss of 10 kg over the last 10 years was associated with a increase of 12% in the serum PCB levels.

Both in adolescents and in adults, a positive association was found between serum PCBs on the one hand and different sources of animal fat (dairy fat, eel fat, mussel fat, and chicken fat) on the other hand. In adults, a negative association was found between serum PCB levels and vegetable fat.

In both age groups, consumption of local meat was associated with higher serum PCB levels. Respectively, in adolescents and in adults, consumption of local meat was associated with 16% and 10% higher serum PCB values (all $p < .0001$).

In both age groups, area of residence was found a significant determinant of variations in serum PCB levels. This reflects the different types of pollution in the selected areas and has been reported more in detail by Schroijen *et al.* (2008).

Finally, a number of variables that are associated with the serum PCB levels, are probably not determining factors themselves but rather a dummy for other underlying factors that determine exposure to or metabolism of the persistent compounds. Smoking was identified as a statistically significant explanatory factor in adults, but the impact was relatively small. Possibly, smoking reflects other lifestyle factors such as nutrition, or social class. In adolescents, a higher educational level of the family was associated with higher serum levels of PCBs. It is likely that educational level is associated with nutritional habits (*e.g.,* fish or milk consumption), living conditions, or other lifestyle factors that may influence exposure to persisted compounds.

Comparison regression tree and linear regression analysis

The estimates of the regression coefficients of the final linear regression model can be used to quantify the relationship of the identified predictors with the biomarker values. For a given variation of the predictor variable, within the range of that predictor, the predicted change in the biomarker values can be calculated. This can not be attained by the decision trees. The trees only give information about the positive or negative influence of a certain predictor on the biomarker values, by focusing on each partition caused by this predictor. No predictions can be made on given variations of this explanatory variable. However, based on the subjects values for the predictor variables, the subjects can be attributed to a certain subgroup with a predicted geometric mean PCB value. These results can be extrapolated to new populations or individuals when their characteristics fall within the range of this study population in terms of age, BMI, PCB content in food, and so on.

DISCUSSION

Multiple linear regression techniques are often used to identify the factors that influence a biomarker. However, this parametric method may lead to unfaithful data descriptions when the underlying assumptions (*i.e.,* normality, constancy of variance, independence (randomness), and linearity) are not satisfied (Neter *et al*. 1996). In that case non-parametric methods can be used as an alternative. Decision trees are inherently non-parametric. In other words, no assumptions are made regarding the underlying distribution of values of the predictor variables.

Based on a literature search, 46 explanatory variables in adolescents, and 38 in adults, were selected as possibly related to the PCB measurement and were introduced in a multiple linear regression model. The analysis of such large numbers of variables causes problems for the standard statistical analyses. It is impossible to investigate for each covariate the nature of the relationship (linear, quadratic, *etc*.) with the biomarker of interest. There are 45 and 37 factorial two-way interactions, respectively in adolescents and adults, considering all of them is not feasible. We therefore limited ourselves to the investigation of the main effects. Decision trees automatically handle interactions between explanatory variables. There may be significant differences for certain effects (*e.g.,* food consumption) on the biomarker values between men/women, smokers/nonsmokers, and so on; these effects are known as variable interactions. Decision trees automatically deal with these interactions by partitioning the persons cases and then analyzing each group separately. On the other hand, additive structure is hard to detect and capture with regression trees (Hastie *et al*. 2001).

In multiple linear regression analysis, some of the main effects are correlated, which introduces the problem of multicollinearity into the model (Neter *et al*. 1996). In that case, one variable is put into the model and the other is omitted. But the omission of that variable in the final linear regression model should not be taken as evidence that the variable is not an important explanatory variable for the response. Since in regression trees, for each partition the best possible variable to split the node into two child nodes is searched for, correlated variables can be kept for testing. For regression trees, however, correlations among the predictor variables can make it hard to identify important interactions (see Sutton 2005 for additional remarks concerning this issue).

In the linear regression analyses, a model selection procedure was used to find the model that is relatively the best of the competing models for the data. The final models obtained from a forward, backward or stepwise selection procedure are not necessarily identical. For regression trees only some kind of forward selection called greedy algorithm is applied. This is an algorithm that always takes the best immediate, or local, solution while finding an answer. So, it should be noted that the produced regression trees are not guaranteed to be optimal. At each stage in the tree growing process, the split selected is the one that will immediately reduce the variation the most. It could be, however, that some other split would be better to set things up for further splitting to be effective (Sutton 2005). Also, different criteria are available to split the nodes. Depending on which criterion is chosen, a different tree can be generated. In this study, the maximize split statistic was used as criterion to split the nodes (SAS Institute Inc. 2008). As the method of binary

recursive partitioning was applied, only binary splits were considered to generate a tree. The term binary implies that each group of individuals, represented by a node in the tree, can only be split into two groups.

In the multiple linear regression analyses only the observations with non-missing values for all independent variables and for the dependent variable were used. As a consequence observations were ignored in the analyses, since there is missingness in (at least) one of the explanatory variables. For adolescents, 211 individuals were ignored in the obtained linear regression model for this reason, and for adults 71 individuals. One attractive feature of tree-based methods is the ease with which missing values can be handled. Surrogate splits are used to deal with missing data. For each node, the "primary splitter" is the variable that best splits the node, maximizing the purity of the resulting child nodes. When the primary splitting variable is missing for an individual observation, that observation is not discarded but, instead, a surrogate splitting variable is sought (Ripley 1996). A surrogate splitter is a variable whose pattern within the dataset, relative to the outcome variable, is similar to the primary splitter. Thus, the program uses the best available information in the face of missing values. In datasets of reasonable quality this allows all observations to be used. The procedure is analogous to replacing a missing value in a linear model by regressing on the parameter with a nonmissing value most highly correlated with it. However it is more robust because there are no model assumptions made.

The fitted linear regression models have to be checked for influential outliers. Identification and exclusion of influential cases in multiple regression models lead to better prediction models, with a higher proportion of the variability explained. Also, exclusion of influential outliers sometimes changed the regression coefficients (but not the sign) of the covariates and had an influence on the covariates that were retained as significant ($p < .05$) in the model. The splitting algorithm of decision trees will easily handle noisy data: outliers will be isolated in a separate node.

CONCLUSION

The results of the non-parametric analysis confirm the selection of the covariates in the multiple linear regression models. In both analyses, blood fat, gender, age, BMI or change in bodyweight, former breast-feeding in adolescents, and a number of nutritional factors were identified as significant predictors in the serum PCB concentration, either in adolescents, in adults, or in both. Also the ranking of importance of the different predictors was similar in the two methods. The way in which the effects of the determinants can be interpret is different for the two methodologies. Quantitative relationships between the determinants and the biomarkers can be calculated from the estimates of the beta coefficients of the multiple linear regression model, assuming that, when quantifying the relation of one covariate with the biomarker, all other covariates in the model are fixed at the population mean. In regression trees, nodes with their characteristics can be described by their geometric mean and standard deviation. The effect of a certain predictor on the biomarker values can be determined by focusing on each partition caused by this predictor and be described in terms of the difference in the geometric means, but nothing can be said of its statistical significance. As such, in multiple linear regression, effects are

described for the whole population, while in regression trees, effects are described for subgroups of the population. As trees are relatively simple for non-statisticians to interpret, they can be used as a sort of exploratory tool in combination with the multiple linear regression approach.

ACKNOWLEDGMENTS

This study was supported by a grant from the Long Range Research Initiative (LRI) of the European Chemical Industry Council Cefic (HETRA D2.2). The database was obtained with permission from the Flemish Environment and Health Study Group. The Flemish Environment and Health Study was commissioned, financed, and steered by the Flemish Community (Department of Science, Department of Public Health, Department of Environment), without any responsibility for the scientific content. The study was approved by the medical ethical committee of the University of Antwerp (July 4, 2002).

REFERENCES

Bilau M, Matthys C, Baeyens W, *et al.* 2008. Dietary exposure to dioxin-like compounds in three age groups: Results from the Flemish environment and health study. Chemosphere 70(4):584–92

Breiman L, Friedman JH, Olshen RA, *et al.* 1984. Classification and Regression Trees. Chapman & Hall, New York, NY, USA

CDC (Centers for Disease Control and Prevention). 2005. Third National Report on Human Exposure to Environmental Chemicals. Atlanta, GA, USA

Chao YC, Zhao Y, Kupper LL, *et al.* 2008. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. J Occup Environ Hyg 5(8):519–29

Covaci A, Voorspoels S, Thomsen C, *et al.* 2006. Evaluation of total lipids using enzymatic methods for the normalization of persistent organic pollutant levels in serum. Sci Total Environ 366(1):361–6

Den Hond E, Govarts E, Bruckers L, *et al.* 2009. Determinants of polychlorinated aromatic hydrocarbons in serum in three age classes—Methodological implications for human biomonitoring. Environ Res 109(4):495–502

Fox J. 1991. Regression Diagnostics: An Introduction. Sage, CA, USA

Gomara B, Ramos L, and Gonzalez MJ. 2002. Determination of polychlorinated biphenyls in small-size serum samples by solid-phase extraction followed by gas chromatography with micro-electron-capture detection. J Chromatography 766:279–87

Hastie T, Tibshirani R, and Friedman J. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 1st edit. Springer-Verlag, New York, NY, USA

Lauwerys R and Hoet P. 2001. Industrial Chemical Exposure: Guidelines for Biological Monitoring. Lewis Publishers, Boca Raton, FL, USA

Lewis MD. 2004. An Introduction to Classification and Regression Tree (CART) Analysis. Available at http://www.saem.org/download/lewis1.pdf

Moore DS and McCabe GP. 1999. Introduction to the Practice of Statistics. W.H. Freeman and Company, New York, NY, USA

Neter J, Kutner M, Nachtsheim C, *et al.* 1996. Applied Linear Statistical Models. McGraw-Hill, New York, NY, USA

Ripley BD. 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, UK

SAS Institute Inc. 2008. Recursive partitioning. In: JMP 8 Statistics and Graphics Guide, vol 1 and 2, pp 793–817. SAS Institute Inc, Cary, NC, USA

Schroijen C, Baeyens W, Schoeters G, *et al.* 2008. Internal exposure to pollutants measured in blood and urine of Flemish adolescents in function of area of residence. Chemosphere 71:1317–25

Sutton CD. 2005. Classification and regression trees, bagging, and boosting. In: Handbook of Statistics, vol 24, pp 303–29. Elsevier B.V., Amsterdam: North Holland

White H. 1980. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica 48:817–38

WHO (World Health Organization). 1995. Physical Status: The Use and Interpretation of Anthropomethry. World Health Organisation, Geneva, Switzerland

Zhang H and Singer B. 1999. Statistics for Biology and Health: Recursive Partitioning in Health Sciences. Springer, New York, NY, USA