Arch Public Health
2002, 60, 275-294

Classification Tree Analysis of Cervix Cancer Screening in the Belgian Health Interview Survey 1997

by

Hens N.[1], Bruckers L.[1], Arbyn M.[2], Aerts M.[1], Molenberghs G.[1]

Abstract

Objectives: *To outline an evidence-based health policy, one is often interested in the profiles of persons who are at risk to obtain certain diseases or who do not respond to prevention programs as e.g. cervix cancer screening via smears.*

Methods: *Statistical modelling can provide a tool to discover such profiles. In this paper the method of classification trees is described. The use of classification trees has advantages but also limitations with respect to their application in the survey domain. A closer look on the handling of missing data and weighting in this context will be given.*

Material: *The Belgian Health Interview Survey (HIS) was conducted in 1997. The Belgian communities are responsible for cervix cancer screening as a part of the preventional health care.*

Corresponding author: HENS N.[1]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium.

[2] Scientific Institute for Public Health, Rue J. Wytsman 14, B-1050 Brussels, Belgium.

Results: *There are no strong conclusions to be drawn with respect to the objective of determining a typical profile for women that underwent a screening. The application of the methods to the HIS data however provides insights and incitements for further investigation.*

Keywords

Cervix Cancer Screening, Classification Trees, Health Survey, Missing Data, Weights.

1. Introduction

According to the Nationaal Kankerregister (1), cervix cancer is the fifth most common cancer among women in Belgium in the period of 1993-1995. Therefore it is not surprising that, for health policy goals, cervix cancer is an important point of attention. In an early stage cervix cancer can already be detected by means of a simple smear. Because early detection decreases the mortality substantially, women between 25 to 64 years old should have a smear every 3 to 5 years, according to the European guidelines (2, 3, 4). However, in Belgium about three out of ten women in this risk group did not have a smear during the previous three years. In several Flemish provinces, there has been a call-recall attempt. In the Walloon region and the Brussels region there has been no such invitation procedure to undergo a screening. The inclusion of follow-up smears has little influence on the accuracy of the screening coverage. Much more important is the selection- and literature bias (5). Only one third of the women in the risk group, who received such an invitational letter, undergo a screening in reply to this letter. This number needs to be interpreted with some caution since also women without uterus and women that recently had a smear were included. From a health policy point of view it would be interesting to know what “type” of women do or do not go for a smear after receiving an invitational letter. Unfortunately we are not able to investigate this question based on the data of the health interview survey 1997, because there are not enough women eligible for screening that received a letter.

The question investigated in this contribution is in what respect the group of women, aged 25-64, not having a smear, is different from the group of women that did have a smear taken in the past three years.

Special interest goes out to whether an invitation letter increases the probability of undergoing screening.

In the next sections, the design of the HIS is specified; the logistic regression approach and the classification tree methodology are introduced. The results of these methods on cervix cancer screening in the HIS are given. Finally, discussions and conclusions are drawn.

2. Material and Methods

2.1. Design of the HIS

In the HIS a total sample of 10,000 interviews (0.1% of the Belgian population) was planned, equally spread over the year 1997. For the three regions of Belgium (Flemish region, Walloon region and Brussels region) the number of individuals to be successfully interviewed was preset at 3500, 3500 and 3000, respectively. An oversampling was planned for the German Community of Belgium (in the district Eupen-Malmédy), with 300 successful interviews. A detailed description of the sampling scheme used in the HIS was published elsewhere (6). The most important features are summarized in what follows. Sampling was based on a combination of stratification, multistage sampling, and clustering (7).

There were two stratification levels. First, stratification was done at the regional level, to ensure that the preset regional level could be reached. Secondly, stratification was conducted at the level of provinces, proportional to their size. Next, the individuals' sample is selected in three stages within each stratum. The first stage, yielding primary sampling units (PSU), consists of municipalities and sampling is carried out proportionally to (population) size via systematic sampling. Whenever a municipality is selected (and it can be more than once), a group of 50 persons is to be interviewed within this municipality. The next stage of random selection operates on households (HHs, secondary sampling units or SSU) according to a clustered systematic sampling procedure upon ordering of the HHs by statistical sector, size and age of the reference person. At this level, matching HHs are provided in case a HH refuses to participate. Finally, individuals or tertiary sampling units (TSU) are selected within HHs in such a way that 4 persons at most are interviewed in each HH and the reference person and his/her partner are automatically selected.

Since the design of the Health Interview Survey follows a complex multistage probability-sampling scheme, it is necessary to reflect these complex procedures in the statistical analysis. Individual weights, reflecting the stratification at provincial level and the differential selection probabilities within households were constructed. Furthermore, post-stratification for age, gender[1], and household size were applied.

The interpretation of the individual weight is that it indicates how many individuals the sampled subject represents. For example, in a simple random sample (2% of the population) each person in the sample represents 50 persons in the population; it can be said that each person has a weight of 50.

2.2. Introduction to the data

From the HIS data file, only women aged between 25 and 64, were selected. These 2893 subjects were all used in the analyses. The gen-

TABLE 1
General topics of which the explanatory variables were considered

Lifestyle	Physical Activity
	Nutritional Habits
	Alcohol Consumption
	Smoking
Health Problems	Subjective Complaints
	Chronical Conditions
	Mental Health
	Functional Limitations
Prevention and Health Promotion	Vaccination
	Cardiovascular Prevention
	Aids Prevention
Use of Health Care	Contacts with GP
	Contacts with specialists
	Contacts with dentist
	Paramedics
	Alternative Methods
	Hospital Admissions
	Use of Medication
Health and Society	Social Health
	Access to Health Care

[1] The analyses here are restricted to women, age 25-64 and so a post-stratification for gender is not necessary here.

eral topics of the explanatory variables are shown in Table 1[2]. (Women without uterus are excluded from the analysis.)

2.3. Methods to analyse the data

Due to the mathematical simplicity, logistic regression is the most commonly used statistical method for binary data. As a part of generalized linear models (8), logistic regression is a parametric method. Tree-based models provide a nonparametric alternative to linear and additive logistic models for classification problems. Trees are fitted using binary recursive partitioning whereby the data are successively split along coordinate axes of the predictor variables so that at any node, the split that maximally distinguishes the response variable in the left and the right branches is selected. Splitting continues until nodes are pure or data are too sparse. This results in a so-called saturated tree, which can be too large to be useful. Therefore this tree is pruned up from the bottom in order to find a useful tree with an acceptable predictive value. The selection of this final tree is based on a cost-complexity measure, which opposes a misclassification measure against the size of the tree. This will be explained further on.

Classification tree analysis is one of the main techniques used in data mining. The goal is to predict or explain responses on a categorical dependent variable, and as such, the available techniques have much in common with the techniques used in the more traditional methods of discriminant analysis, cluster analysis, nonparametric statistics, and nonlinear estimation. Classification trees are widely used in applied fields as diverse as medicine (diagnosis), computer science (data structures), botany (classification), and psychology (decision theory). Amenability to graphical display and ease of interpretation are perhaps partly responsible for the popularity of classification trees in applied fields, but two features that characterize classification trees more generally are their hierarchical nature and their flexibility. Let us first have a look at the logistic regression analyses and at the alternative classification tree methodology.

2.3.1. Logistic Regression

The standard statistical technique to investigate the relationship between the binary response variable “screening status” and a set of explanatory variables is logistic regression analysis. We applied this

[2] The full questionnaire can be consulted at http://www.iph.fgov.be/epidemio/epien/crospen/hisen/table.htm.

technique to predict the probability that a woman, aged between 25 and 65 years, underwent a cervix cancer screening examination in the past three years. The goal of this section is to discuss such an analysis in this context.

A logistic regression model can be formulated as a specific member of the parametric generalized linear models family. As for all parametric models, disadvantages are the strong model assumptions about the distribution of the data and about underlying regression relationships that have to be made. If such parametric assumptions do not hold, the result of the model fit is questionable.

With 85 explanatory variables it is almost impossible to investigate for each covariate the nature of the relationship (linear, quadratic, etc.). Moreover, there are 7140 possible two-way interactions, considering all of them is not feasible. We therefore limited ourselves to the investigation of the main effects. To find the model that is relatively the best of the competing models for the data, a model selection procedure was used. The likelihood equation for the most complex model, i.e. including all 85 main effects does not exist. The model has infinite parameters due to a complete separation of the sample points. As a consequence a backward selection procedure could not be applied to the data. The final models, obtained from a forward selection and stepwise procedure, are identical for the data at hand. This is however not always the case (9). In the analyses only the observations with nonmissing values for all independent variables and for the dependent variable were used. As a consequence more than 60% of the observations were ignored in the analyses, because of missingness in (at least) one of the explanatory variables. Of course one has to question the validity of the prediction model. Several techniques to analyse data in the presence of missingness are available. Multiple imputation could be a solution to handle the missingness and it is fairly robust to imputation model misspecification, which could be a problem. It is however not straightforward how the M results of the selection procedures for the M imputed data-files can be combined to present one final model (10).

To take in account the design features of the HIS a weighting scheme was developed. Incorporating these weights in the logistic analysis corrects for the fact that a complex sampling scheme was used.

2.3.2. The Methodology of Classification Trees

The classification tree methodology is a classification method where, following specific splitting rules; disjoint subsets of the data are con-

structed. These subsets are called nodes. Further splitting is repeated several times within these nodes. A node where a split is formed is called a parent node. The subsequent nodes are called child nodes. Terminal nodes are nodes that are not splitted further. The size of a tree is the number of parent nodes plus one. We focus on binary classification trees, where splitting occurs into exactly two child nodes.

This partitioning process results in a saturated tree. A tree is saturated in the sense that the offspring nodes subject to further division cannot be split. The saturated binary tree is then pruned to an optimal sized tree. This is the so-called pruning process. The final step is the selection process, which determines the final tree. In the following paragraphs a brief overview of the different processes is given.

The partitioning process is based on splitting rules. The splitting rules involve conditioning on predictor variables. The best split is the split with the largest reduction in impurity. The impurity of a node measures the homogeneity of this node. The least impure node has either all zeros or all ones as outcome. Whereas, the most impure node is characterized by 50% of zeros and 50% of ones. The impurity of a node is a goodness of fit measure. Not surprisingly, given the hierarchical nature of classification trees, these splits are selected one at a time, starting with the split at the root node, and continuing with splits of resulting child nodes until splitting stops, and the child nodes, which have not been split, become terminal nodes. The generally most used split-selection method is discussed here, i.e. an exhaustive search for univariate splits for categorical or ordered predictor variables. With this method, all possible splits for each predictor variable at each node are examined to find the split producing the largest improvement in goodness of fit and thus the lowest impurity for the nodes.

One common choice as a goodness of fit measure is the Gini index. This measure is computed as the sum of products of all pairs of class proportions for classes present at the node. For binary trees this Gini index can be expressed as a product p(1-p) where p is the probability to belong to the first of the two classes in the node considered. The Gini measure of node impurity reaches its maximum value when class sizes at the node are equal; and a value of zero is obtained when only one class is present at a node. The prevalence rate p at a node has to be available to compute the Gini index. In many applications this prevalence rate can be estimated empirically, at other times, additional prior information may be required. The priors are estimated from class sizes and equal misclassification costs. For example if it is known that the two

classes in the scoring dataset are e.g. distributed as 90% versus 10%, then this information will be used together with the probabilities as they are given in the training dataset using Bayes theorem as explained in Zhang et al. (12). The Gini measure is the most generally used measure of goodness of fit in computing packages.

Completing this partitioning process results in a saturated tree with the characteristic that if no limit is placed on the number of splits that are performed, eventually "pure" classification will be achieved. Each terminal node would contain only one class of observations. However, "pure" classification is usually unrealistic. The saturated tree is usually too large to be useful. Indeed, the terminal nodes are so small that no sensible inference can be made because such a tree has a small predictive value. Therefore, it is typically to set a minimum size of a node a priori. Splitting is stopped when a node is smaller than this minimum.

Breiman et al. (11) argue that depending on the stopping threshold, the partitioning tends to end too soon or too late. They propose to let the partitioning continue until it is saturated or nearly so. Beginning with this generally large tree it is pruned from the bottom up. The point is to find the subtree of the saturated tree that is most "predictive" of the outcome and least vulnerable to noise in the data.

They developed structured procedures for selecting the "right-sized tree". Selection of the "right-sized" tree is based on the cost complexity measure. This function is defined as the cost for the tree plus a complexity parameter times the tree size. In many typical applications, costs simply correspond to the proportion of misclassified observations, but other modifications are possible too (12). Here the proportion of misclassified observations is used to define costs.

The procedure generates a sequence of trees with a number of interesting properties. Trees are nested, because successively pruned trees contain all the nodes of the next smaller tree in the sequence. This sequence of trees is also optimally pruned, because for every size of a tree in the sequence, there is no other tree of the same size with lower costs.

It is well known that a classification tree computed from a learning sample in which the outcomes are already known, will not perform equally well in predicting outcomes in a second, independent test sample. Sometimes one better might use a smaller classification tree that does not classify perfectly in the learning sample, but which is expected to pre-

dict equally well in the test sample. V-fold cross-validation is useful when no test sample is available and the learning sample is too small to have the test sample taken from it. A specified V value for V-fold cross-validation determines the number of random subsamples, as equal in size as possible, that are formed from the learning sample. The classification tree of the specified size is computed V times, each time leaving out one of the subsamples from the computations, and using that subsample as a test sample for cross-validation. The CV costs computed for each of the V test samples are then averaged to give the V-fold estimate of the CV costs.

While there is nothing wrong with choosing the tree with the minimum cost as the "right-sized" tree, often there will be several trees with cross validation (CV) costs close to the minimum. Breiman et al. (11) make the reasonable suggestion that one should choose as the "right-sized" tree the smallest-sized (least complex) tree whose costs do not differ appreciably from the minimum costs. They proposed a "1 SE rule" for making this selection, i.e., chose the "right-sized" tree to be the smallest-sized tree whose costs do not exceed the minimum costs plus 1 times the standard error of the costs for the minimum costs.

2.3.3. Classification Trees and Missing Data

One attractive feature of tree-based methods is the ease with which missing values can be handled. The appropriateness of these methods is however not straightforward. In this section we will have a look at four methods proposed by Ripley (13).

A first approach is prediction on complete observations. Quinlan (14) suggests replacing missing values using the distribution within the class at that node when computing the expected value of a split. In his paper of 1993, Quinlan multiplies the impurity gain calculated on known observations by the proportion of missing values. This method has a major disadvantage when the number of complete observations in the node is quite small. Another disadvantage is that other available variables for this observation are neglected while they are possibly highly correlated with the missing one.

The second approach, Ripley (13) discusses, is the missing together approach (MT). Suppose that we attempt to split a node by a variable and that the measurement for that variable is missing for a number of observations. The MT approach forces all of these subjects to the same daughter node. If it is a nominal variable with several levels, the miss-

ing value is regarded as an additional level, so the variable has one level more. On the other hand, when the variable has a natural order, two copies are made. If a component is missing, the component in the first copy will be set on plus infinity and the corresponding component in the second copy will be given the value minus infinity. In this way, replacing the variable by its two variants, results in two possible splits such that the observations with missing measurement are sent to the same daughter node. The variant that gives the best split is chosen. This the key idea of the missing together approach (MT). The advantages of the MT approach are that it is very simple to implement and that a recursive partition algorithm that assumes no missing data can still be used without modification when the raw data contain missing values. Also the observations with missing information can easily be located in the tree structure. In contrast, both daughter nodes may contain some of these subjects by using surrogate splits instead. A major disadvantage of the MT approach is that imputation relies on the assumptions of simultaneous behaviour for subjects with a missing observation for the covariate of interest. Moreover, the most favourable split is chosen to be the best split to take, without considering the information in the other covariates. This can be circumvented by surrogate splits.

The third approach of surrogate splits is analogous to replacing a missing value in a linear model by regressing on the explanatory variable with a nonmissing value most highly correlated with it. However it is more robust because of no model assumptions. The surrogate split approach attempts to utilize the information in the other predictors to assist in making the decision to send a observation to the left or the right daughter node. One looks for the predictor that is most “similar” to the original predictor in classifying the observations. Similarity is measured by a measure of association. It is not unlikely that the predictor that yields the best surrogate split may also be missing. Then we have to look for the second best, and so on. In this way, all available information is used. If surrogate splits are used, the user should take full advantage of them.

A fourth possibility is to take missing as a further level of the attribute. This method allows multi-way splits, which are not appealing because making some values missing can increase the gain in impurity. This can be circumvented by allowing only binary splits, or by penalizing multi-way splits.

As a conclusion one can say that in most approaches tree construction is based on the observations without any missing values. Where

missing values are very frequent; as in large-scale surveys, this may be unacceptable or even impossible.

The practical implementation of the previous methods, handling missing data, is not an issue. The appropriateness of the chosen approach is. Especially the use of all available information by surrogate splits is appealing. Substantial improvements upon this method can be thought of although the practical implementation can be a drawback. All of these ideas have merits and demerits, depending on how common missing values are and whether they are missing at random are not (15).

3. Results and Discussion

3.1. Logistic Regression

In Table 2 the results of the weighted logistic regression are presented.

Most of the explanatory variables in the model are variables indicating an awareness of the patient towards his own health status, e.g. cholesterol control, heavy drinking moments, blood pressure control ... . Other predictor variables are of a demographic nature, e.g. age, income

TABLE 2
Output of the logistic regression

Variable	DF	Wald	P-value
Age Category	8	34.91	<0.0001
Income	4	10.25	0.0365
Household type	4	32.99	<0.0001
Consumed Bread	1	4.17	0.0411
Snack Eating	2	14.37	0.0008
Province	10	34.09	0.0002
Hospital Admission	2	16.04	0.0003
Blood Press. Control	1	7.07	0.0078
Profession	8	21.87	0.0052
Lack of Physical Act.	1	6.08	0.0137
BMI Category	5	17.64	0.0034
Cholesterol Control	1	8.14	0.0043
Heavy Drinking Mom.	5	12.57	0.0278
Milk Consummation	1	4.79	0.0286
Daily Drinker	1	4.74	0.0294
Medication	1	4.07	0.0438
Social	1	3.85	0.0498

and province. Appreciation of social relationships seems to have a small influence in the model. For health policy purposes the effect of a screening invitation is of great importance. This specific explanatory variable indicates whether a person received an invitation letter advising her to have a cervix cancer screening.

The predictive value of the model is given by the positive and negative predictive values. The positive predictive value is the probability that a woman really underwent screening when she is predicted to do so. The negative predictive value is defined in a similar way. Higher values for these probabilities are more desirable. The positive and negative predictive value depends on the cut-off point for classification on the logit scale. Based on the ROC-curve a cut-off point of 0.40 is chosen. This corresponds to a positive predictive value of 75% and a negative predictive value of 68%. These values are quite high but one can question the validity of them because we only use 40% of the data.

Fitting a logistic regression model without weights gives a model which retains the explanatory variables hospital admission, medication, age, household type, lack of physical activities, eating of snacks, sort of bread consummation, blood pressure -and cholesterol control and additional explanatory variables are screening invitation, dentist visits, preventive tooth control, region, HIV-protection knowledge, HIV screening and eating of breakfast. We see the same trend as above, where people with a greater awareness of the own health status will be predicted to undergo cervix cancer screening with a higher probability.

As is clear from this discussion, a logistic regression analysis is not the most optimal technique to formulate an answer to our research question due to (1) the large number of explanatory variables and (2) the missing values in the covariates. In order to overcome these problems the classification tree methodology will be used to analyse the data in the next section.

3.2. Classification Trees

There are several software packages that support the classification tree methodology. Some of them are especially developed for classification and regression trees. The most familiar one of this kind is CART (16) for windows. Well-known statistical computing packages like S-PLUS (17) and R (18) can handle classification trees. R has the main advantage that it can use weights in a straightforward manor, while SPLUS and

CART do not. The handling of missing data has some minor differences in all packages, but essentially it is of no concern to the user. R was used throughout this analysis.

In the construction of the classification tree for the "screening status" the Gini method was used together with the use of surrogates to handle missing values. In this first tree-based analysis no selection weights were included. Arbitrarily, a minimum of 20 observations had to be available in a parent node and a minimum of 10 observations had to be present in all nodes. To find the best split, a 10-fold cross-validation was used, the tree with the minimal cross validation relative error is chosen as the final tree. Figure 1 shows the cross-validation relative error, indicating a minimal value for a tree of size 7 corresponding with a cost complexity of 0.0094. This tree is shown in Figure 2.

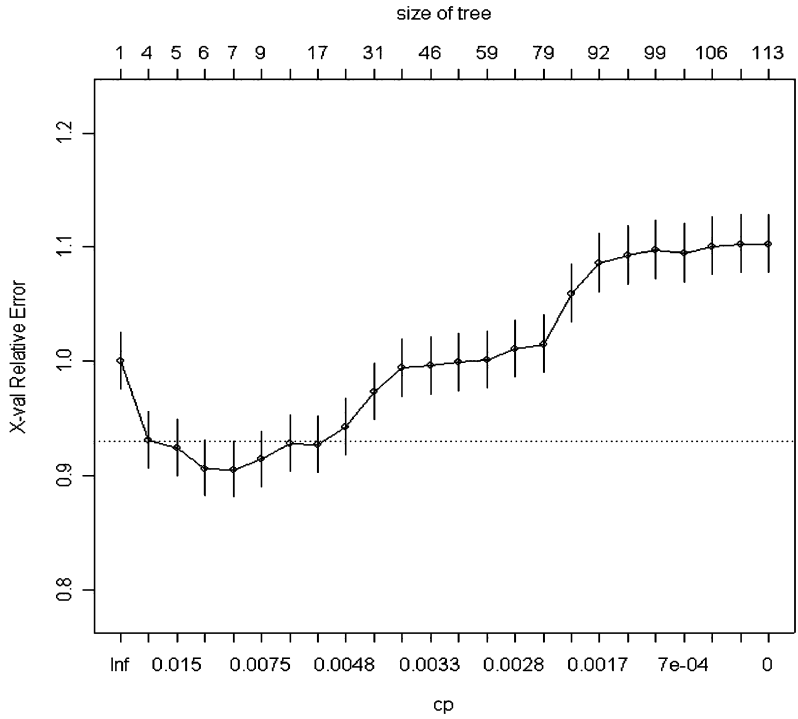

Fig. 1: The X-val relative error in function of the cost complexity parameter and the size of the tree

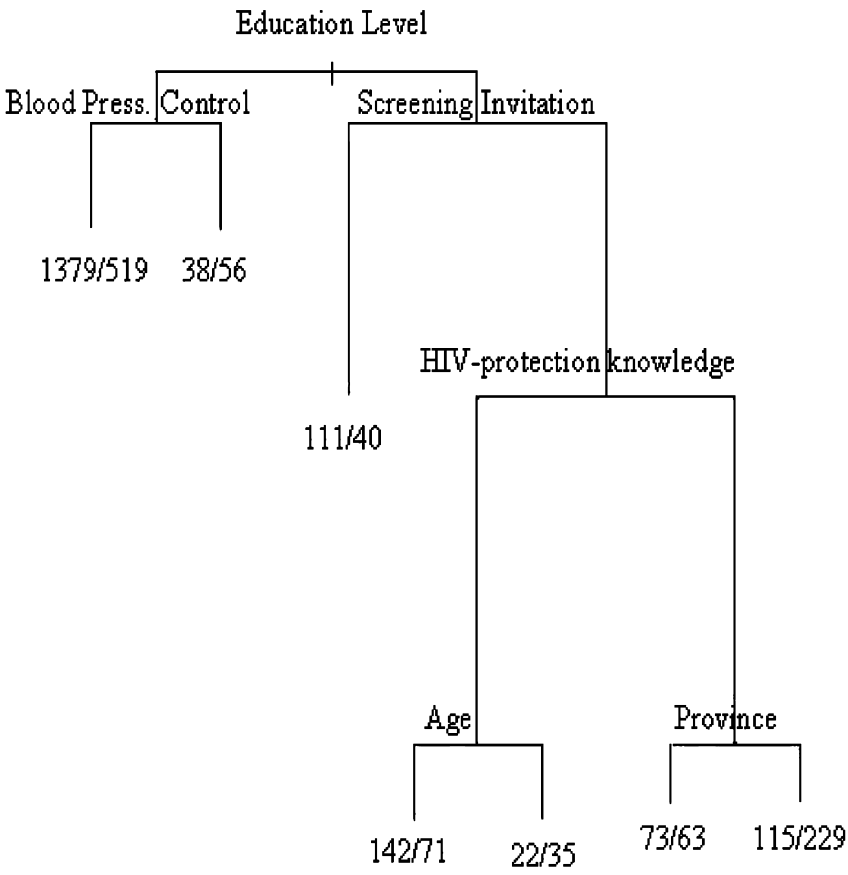

Fig. 2: The final tree with two numbers indicating the true class 1/class 2 of the response. Left child nodes are predicted to be of class 1, while right child nodes are predicted to be of class 2.

The complete output is given in Table 3.

The surrogate splits used in this tree are given in Table 4.

The explanatory variables that determine the final tree are comparable with those of the logistic regression analysis without weights (results not shown). Blood pressure control and HIV-protection knowledge could again be classified under the own health knowledge while age and province are demographic variables. The presence of the invitation to undergo a screening is not surprisingly to be of importance. The goodness of fit for the model can be represented by the positive predictive

TABLE 3
Output of the classification tree analysis without weights. Between brackets one finds the percentage correctly specified individuals of the corresponding outcome for the terminal nodes

Variable (* = terminal node)	Number of individuals	Splits left	Splits right
Education Level	2893	Superior sec education or higher	Up untill inferior sec education
Blood Pressure Control	1992	Yes* (0.73 1)	No* (0.60 2)
Screening Invitation	901	Yes* (0.74 1)	No
HIV-protection knowledge	750	Yes	No
Age	270	25-60 years old* (0.67 1)	60-70 years old* (0.61 2)
Province	480	Antwerp, East Flanders, Walloon Brabant and Luik*	Flemish Brabant, West Flanders, Limburg, Brussels, Henegouwen, Luxemburg and Namen* (0.67 2)

TABLE 4
Surrogate splits used in the classification tree analysis without weights. The three best surrogate splits, when available, are given from best to worse

Variable	Surrogate Splits
Education Level	Highest Professional Category Social Status Age
Blood Pressure Control	–
Screening Invitation	Province
HIV-protection knowledge	Age Excessive Drinking HIV-tranmission knowledge
Age	Vaccination for Influenza Social Status

value of 0.72 and the negative predictive value of 0.65. The predictive value of a tree without explanatory variables is about 0.65 since the prevalence rate is almost equal to 2/3. Comparing the result of our fit with this one, one can see that the performance is not so good. This is also clear from the estimate of the cross-validation relative error (0.91), which is rather high (Figure 1).

To incorporate the complex sampling scheme in the tree construction we fit the same model but now with weights. The final tree of the analysis with weights is given in Figure 3.

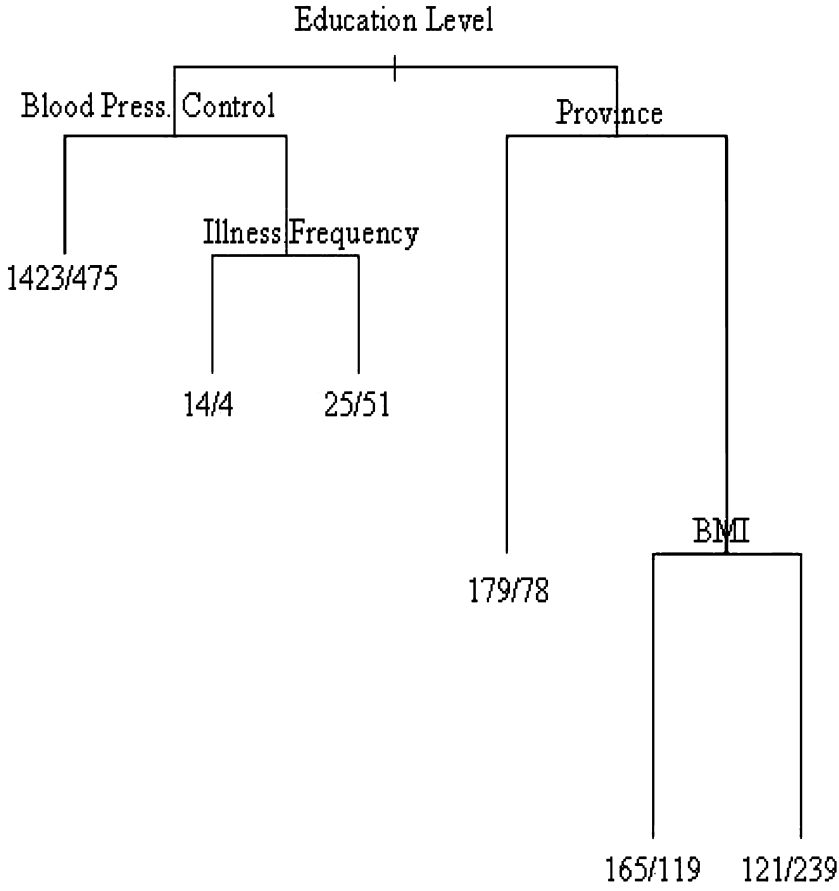

Fig. 3: The final tree with two numbers indicating the true class 1/class 2 of the response. Left child nodes are predicted to be of class 1, while right child nodes are predicted to be of class 2.

The size of the final tree based on the weights is 6 and thus smaller than the size of the tree without consideration of the weights. The variables in this tree are educational level, blood pressure control, the province, frequency of illness and the body mass index. The latter two variables were not present in the previous analysis but are fairly good indicators of the overall health status of an individual. The positive predictive value equals 0.72 and the negative predictive value equals 0.67,

indicating an acceptable model fit but the performance of this analysis is not much higher than what can be obtained from a tree with no explanatory variables. The complete output is given in Table 5.

TABLE 5
Output of the classification tree analysis with weights. Between brackets one finds the percentage correctly specified individuals of the corresponding outcome for the terminal nodes

Variable (* = terminal node)	Number of individuals	Splits left	Splits right
Education Level	2893	Superior sec education or higher	Up untill inferior sec education
Blood Pressure Control	1992	Yes* (0.75 1)	No (0.57 2)
Illness Frequency	94	0 or 1* (0.78 1)	2 or more* (0.67 2)
Province	901	Antwerp, Flemish Brabant, East Flanders, Limburg and Walloon Brabant	West Flanders, Brussels, Henegouwen, Luik, Luxemburg and Namen* (0.69 1)
Body Mass Index	644	<18, 20-25* (0.58 1)	18-20, >25* (0.66 2)

The surrogate splits used in this analysis are given in Table 6.

TABLE 6
Surrogate splits used in the classification tree analysis with weights

Variable	Surrogate Splits
Education Level	Highest Professional Category
	Physical Functional Score
	Social Status
	–
Blood Pressure Control	Limitations Due to Longterm Illness
Illness Frequency	Longterm Illness
	Hospital Admission
Province	Region
	Screening Invitation
	Medication Expenses
Body Mass Index	Tabac Consumption
	Smoking Behaviour
	Age

Comparing the results with the unweighted tree indicates that the weights have an effect on the outcome of the final tree. The logistic regression analyses confirm this finding, although considerations have to be made due to its limitations with respect to missingness and problems due to model specifications. There is quite some difference in the results between the logistic regression and the classification tree analysis. Explanatory variables as blood pressure control and province are mutual for all models. In the tree analyses the education level seems to be of great importance. The logistic regression analyses indicates that hospital admission, medication, age, household type, lack of physical activities, eating of snacks, sort of bread consummation, and cholesterol control are important.

The screening invitation seems to be important in both the analyses but disappears when weights are taken into account. The relation between the screening invitation and the weighting variable is significant (Wilcoxon-test, p-value <0.0001). Therefore the difference between the weighted and unweighted result is not all that surprising. The weighted tree analysis indicates that sending an invitational letter has no effect on the screening status. This weighted tree analysis is the correct analysis. Conclusions drawn from the unweighted tree can be misleading and incorrect.

3.3. A Combined Analysis

A combined analysis was performed by including the explanatory variables, obtained with the weighted tree analysis, in a weighted logistic regression model. The main advantage of such a combined analysis is that all subjects contribute to the selection of important explanatory variables. The performance of the obtained model is little bit better than the logistic regression model obtained after a forward selection procedure (section 2). The positive predictive value of the model equals 71%; the negative predictive value equals 75%. Particularly, women not undergoing a cervix screening examination can be better classified based on this model.

The implementation of the covariates in the logistic regression model has once more the disadvantage of neglecting the information contained in records with missing values. The predictive values have to be interpreted with caution, because they only rely on 50% of the data.

4. Conclusions

Logistic regression and classification tree analyses were used on the HIS data to investigate whether people with specific characteristics have to be encouraged to undergo cervix cancer screening. The predictive accuracies obtained with these two methodologies are comparable in this case. The variables used for the prediction differ. Classification tree methodology however has some advantages to use.

The classification tree methodology is a fully non parametric model that deals with the two major burdens of a parametric method as logistic regression, that is the model assumptions and the regression relationship. The methodology also nicely deals with the problem of missingness and uses all data in the construction of the tree, while in the parametric method missingness can lead to additional difficulties and loss of data.

The results according to the weighted classification tree, show that people with a lower educational level, that did not have a blood pressure control during the last five years, living in the Walloon region of Belgium or with obesitas or with overweight (BMI>25) are less likely to have underwent a cervix cancer screening. Policy actions should target these persons.

In our analyses, we did not focus on clustering (12), which could be of interest. How the ways to handle missing data relate to different missingness processes like missing completely at random, missing at random and missing not at random are not fully investigated here, but are possibly further research topics. We did not fully take in account the continuous nature of explanatory variables as age. This can be easily done but demands more computing time.

Currently a second health interview survey is taking place in Belgium. Together with the HIS of 1997, time trends could be investigated.

Samenvatting

Om een gefundeerd gezondheidsbeleid uit te zetten, is men geïnteresseerd in de profielen van subgroepen van personen, bijvoorbeeld zij die vatbaar zijn voor bepaalde ziektes of diegenen die niet reageren op preventie-programma's zoals baarmoederhalskanker-onderzoek via uitstrijkjes. Statistisch modelleren kan een oplossing bieden om zulke persoonsprofielen te zoeken. In deze tekst wordt de methode van classificatiebomen

toegepast. Deze methode heeft vele voordelen in vergelijking met logistische regressie. Er zijn echter ook enkele beperkingen tot het gebruik in het survey domein. Hoe deze methode omgaat met missing data en gewichten wordt van naderbij bekeken. De eerste Belgische Gezondheidsenquête (HIS) werd uitgevoerd in 1997. De Belgische gemeenschappen zijn verantwoordelijk voor het onderzoek naar baarmoederhalskanker als onderdeel van de preventieve gezondheidszorg. De HIS laat ons toe om profielen van personen te onderzoeken die aan zo'n onderzoek deelnemen. Mensen die niet voldoen aan dit patroon zouden aangemoedigd moeten worden om toch aan deze onderzoeken deel te nemen.

References

1. Haelterman, M. Kanker in België. National Cancer Registry, 1999.
2. Coleman D, Day N, Douglas G. et al. European Guidelines for Quality Assurance in Cervical Cancer Screening. Eu. J. Cancer, 29A, suppl. 4, 1-38, 1993.
3. Advisory Committee on Cancer Prevention. Recommendations on Cancer Screening in the European Union. Eur J Cancer 36: 1473-1478, 2000.
4. Arbyn M, Van Oyen H, Lynge E, Mickshe M. European Consensus on Cancer Screening Should be Applied Urgently by Health Ministers. BMJ 323: 396, 2001.
5. Arbyn M, Van Oyen H. Cervical Cancer Screening In Belgium. Eu. J. Cancer, 36, 17, 2191-7, 2000.
6. Quataert P, Van Oyen H, Tafforeau J. et al. Health Interview Survey 1997. Protocol for selection of the households and the respondents. SPH/Episerie No. 12, SPH 1998, Brussels.
7. Kish L. Survey Sampling. New York: Wiley, 1995.
8. McCullagh P, Nelder JA. Generalized Linear Models. London: Chapman and Hall, 1996.
9. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: Wiley, 1989.
10. Schafer JL. Analysis of incomplete multivariate data. London: Chapman and Hall, 1997.
11. Breiman L, Friedman JH, Olsen RA, Stone CJ. Classification and regression trees. The Wadsworth Statistics/Probability Series, Belmont, California 1984.
12. Zhang H, Singer B. Recursive Partitioning in the Health Sciences. New York: Springer-Verlag, 1999.
13. Ripley BD. Pattern Recognition and Neural Networks. Cambridge University Press, 1996.
14. Quinlan JR. Induction of decision trees. Machine Learning, 81-106. Reprinted in Shavlik & Dietterich (1990), 1986.
15. Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: Wiley, 1987.
16. Steinberg, Dan and Philip Colla. CART – Classification Trees San Diego, CA: Salford Systems 1997.
17. S-PLUS 6 Professional Edition Version 6.0.2 Release 1 for Microsoft Windows Copyright (c) 1988, 2001 Insightful Corp.
18. The R-Project 1.3.1 A language and Environment Copyright, 2001 The R Development Core Team.