



## **A Simple Tool for Detecting Unrecognized Prediabetes and Diabetes**

**Kenneth E. Heikes PhD**

**David M. Eddy MD PhD**

**October 30, 2006**

## Objective

The goal is to build a simple paper-and-pencil screening tool that indicates the risk of prediabetes and diabetes based on health information readily known to the average person. The screening tool might take the form of a test consisting of check boxes for the answers to a series of true/false questions, where the sum or weighted sum of the answers map to a risk of having prediabetes or diabetes. Alternatively, the test could take the form of a decision tree navigated from root to terminal node through a series of branching internal nodes, where the path followed depends on the answers to true/false questions at each internal node encountered along the path. The final terminal node arrived at determines the risk of prediabetes and diabetes.

Two possible approaches are considered as a foundation for the paper-and-pencil screening tool: a logistic regression or a classification tree. Both approaches have advantages and disadvantages and, therefore, both are considered. The traditional approach – logistic regression – is well grounded in statistical theory, but the resulting probability equation does not easily translate to a simple check-box model. A classification tree is amenable to either a check-box or flowchart format, (but does not have the theoretical underpinning of logistic regression).

## Method

### Definitions

The definitions of prediabetes and diabetes used in this study are based on thresholds of fasting plasma glucose (mg/dL), obtained after an 8-24 h fast, and 2-h postprandial plasma glucose (mg/dL), as shown in [Table 1](#). Thus, the definitions depend on both fasting plasma glucose (FPG) and oral glucose tolerance test (OGTT) data. This study uses data from NHANES III (1988-94) and NHANES 1999-2004. An oral glucose tolerance test was not administered during NHANES 1999-2004 ([1](#)) and, therefore, there is no OGTT data available in that dataset. Nevertheless, it was deemed important to include

---

<sup>1</sup> Diabetes Mellitus Interagency Coordinating Committee Meeting, National Institutes of Health Campus, Natcher Conference Center, Conference Room A Bethesda, Maryland, February 25, 2003. The Metabolic Syndrome, Summary Minutes.

*The oral glucose tolerance test (OGTT) was collected in NHANES III but when the yearly NHANES began, it was dropped, largely to tie in with the ADA recommendations for using fasting blood glucose to diagnose diabetes. With the increasing prevalence of type 2 diabetes and the DPP and other data related to the risk factors for CVD, the decision to use the 2-hour OGTT is being reconsidered.*

OGTT in the definitions. Harris (2) showed that total glucose intolerance in NHANES III was higher when abnormal OGTT values were included, compared to results based on FPG alone. Therefore, the tool is developed based on NHANES III, which includes OGTT data. NHANES 1999-2004 data is used to indicate the approximate validity of the tool, although the definitions of prediabetes and diabetes differ for the two datasets in that the NHANES 1999-2004 data does not include OGTT data.

**Table 1: Prediabetes and diabetes definitions based on FPG and OGTT**

	OGTT < 140	140 ≤ OGTT < 200	OGTT ≥ 200
FPG < 100	Normal	Prediabetes	Diabetes
100 ≤ FPG < 126	Prediabetes	Prediabetes	Diabetes
FPG ≥ 126	Diabetes	Diabetes	Diabetes

### Data

Prediabetes and diabetes.

During NHANES III, examinees aged ≥ 12 years were instructed to fast for 10-16 hours prior to the morning examination. The instructions were not followed uniformly. Laboratory test results and the duration of the fast are included on the data file regardless of the examinee's fasting compliance. An OGTT was conducted on participants aged 40-74 years who attended the mobile examination part of the survey. Participants were randomly assigned to receive an OGTT in the morning after an overnight fast, or they were assigned to an afternoon or evening OGTT. Approximately half of the OGTT examinees received the morning OGTT. This subsample most closely conforms to the World Health Organization 1995 criteria for OGTT testing to identify diabetes.

The prediabetes and diabetes variables used to develop the screening tool are derived from the NHANES III morning subsample records with an overnight fast of 8-24 hours

<sup>2</sup> Harris MI, Flegal KM, Cowie CC, et al. Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults: The Third National Health and Nutrition Examination Survey, 1988--1994. *Diabetes Care* 1998; 21:518--24.

and age  $\geq 20$  years ( $N = 7056$ ). Over half of these records (56%) are missing OGTT data, mainly because OGTT data was not collected for people age  $< 40$  or age  $\geq 75$  years. Even for ages 40-74 years, OGTT data is missing for some records. The reasons for incomplete OGTT for this age group are shown in [Table 2](#). Only a small percentage of the data is incomplete for reasons related to diabetes.

<b>Table 2: Reasons for an incomplete OGTT among NHANES III participants ages 40-74 years assigned to receive an OGTT in the morning after an overnight fast</b>		
<b>Reason</b>	<b>Frequency</b>	<b>Percent</b>
Chemotherapy within 4 weeks	44	15.8
Diabetic on insulin	6	2.2
Refused venipuncture	22	7.9
Ill/faint during test	40	14.3
Venipuncture unsuccessful	2	0.7
Physician canceled test	66	23.7
Refused glucose challenge	99	35.5

In order to use as many records as possible, disease status (normal, prediabetes, or diabetes) is determined by FPG alone when OGTT data is missing. For ages 40-74 years, where OGTT data is not missing, [Table 3](#) shows how the percentage of observations defined as prediabetes and diabetes vary depending on whether OGTT data is, or is not, incorporated in the definitions. Based on FPG alone, the prevalences of normal, prediabetes and diabetes observations are 58.2, 33.4, and 8.5%, respectively. Including OGTT increases the prevalence of prediabetes to 37.1% and increases the prevalence of diabetes to 11.7%. Therefore, the prevalences of prediabetes and diabetes in NHANES 1999-2004, which are based on FPG alone, are expected to be lower than in NHANES III.

Table 3: Sample weighted distribution of NHANES III observations by FPG and OGTT ranges used in the definitions of prediabetes and diabetes for persons age 40-74 receiving the morning OGTT				
	OGTT < 140	140 ≤ OGTT < 200	OGTT ≥ 200	Total
FPG < 100	51.2	6.4	0.6	58.2
100 ≤ FPG < 126	21.8	8.9	2.6	33.4
FPG ≥ 126	0.6	1.2	6.7	8.5

Undiagnosed diabetes.

The purpose of the screening tool is to inform people whether or not they are at risk for prediabetes and diabetes. Obviously, the tool is of little utility for a person who has already been diagnosed with diabetes. Therefore, one target variable of the tool is undiagnosed diabetes. The other target variable is prediabetes, which is not routinely diagnosed. Undiagnosed diabetes is defined as the presence of actual diabetes, based on FPG and/or OGTT, and the absence of a person having been told they have diabetes.

Table 4 shows the prevalence (%) of disease states for three age groups based on sample weighted NHANES III data. The prevalence of undiagnosed diabetes is higher than that of diagnosed diabetes for all age groups. Of the population diagnosed with diabetes, 0.99% but do not meet the criteria for diabetes. These cases may have diabetes that is under control with medication or they may be misdiagnosed. The logistic regression and classification tree analyses that follow are based only on the data in the first three rows of Table 4.

<b>Table 4: Prevalence (%) of prediabetes and undiagnosed diabetes for three age groups based on sample weighted NHANES III data</b>				
<b>Disease Level</b>	<b>Age</b>			<b>Total prevalence</b>
	<b>20-44</b>	<b>45-64</b>	<b>65+</b>	
<b>Normal</b>	84.33	54.89	39.59	66.03
<b>Prediabetes</b>	14.79	33.53	41.32	26.14
<b>Diabetes</b>				
<b>Undiagnosed</b>	0.56	6.28	9.52	4.16
<b>Diagnosed</b>	0.15	4.03	6.86	2.69
<b>Diagnosed but does not meet criteria</b>	0.17	1.27	2.72	0.99
<b>Percent in age group</b>	45.9	38.5	15.6	

Explanatory variables.

The explanatory (independent) variables selected to predict the risks of prediabetes and diabetes consist of health information easily known by most people. Most are dichotomous variables resulting from yes/no answers to questions about, for example, family history of diabetes or having had a diagnosis of high blood pressure. Some, such as waist circumference, are continuous variables. BMI and waist-to-hip ratio are not necessarily known by a person, but can be determined from a lookup table using information that likely is known. Not all explanatory variables are ultimately used – only those that are good predictors of the target variables appear in the final tool. The candidate explanatory variable names, descriptions, and basic statistics are shown in [Table 5](#).

**Table 5: Target and explanatory variables, number of records without and with missing fields, number of distinct values, unweighted mean, and minimum and maximum values for NHANES III, age  $\geq 20$  years**

Variable	Description	N	Missing	Distinct	Mean	Min	Max
WTPFHSD6	Weights for morning fasting subsample	6620	0	N/A	59917	446.5	2.91E+05
UNDIAGNOSEDDIABETES	Undiagnosed diabetes	6620	0	2	0.04316	0	1
PREDIABETES	Prediabetes	6212	408	2	0.28358	0	1
EXTENDEDPREDIABETES	Undiagnosed diabetes or prediabetes	6620	0	2	0.3145	0	1
DISEASELEVEL	Normal, prediabetes, or diabetes	6620	0	3	0.35766	0	2
BMI	Body mass index (kg/m <sup>2</sup> )	6608	12	341	26.377	13.3	68.5
HGT	Height (cm)	6613	7	515	168.62	126.9	201.7
WGT	Weight (kg)	6609	11	1628	75.319	31.24	202.47
WST	Waist circumference (cm)	6406	214	706	91.297	58.6	168.8
WHR	Waist-to-hip ratio	6397	223	58	0.90404	0.55	1.52
AGE	Age (years)	6620	0	71	43.776	20	90
SEX	Gender	6620	0	2	1.5248	1	2
RAC	Race/ethnicity	6620	0	4	1.4357	1	4
BPM	Taking blood pressure medication	6314	306	2	1.113	1	2
BCM	Taking cholesterol medication	3320	3300	2	1.0502	1	2

Table 6 continued							
Variable	Description	N	Missing	Distinct	Mean	Min	Max
GDB	Had gestational diabetes	6620	0	2	1.0046	1	2
HBP	High blood pressure	6319	301	2	1.2221	1	2
HBC	High cholesterol	3325	3295	2	1.3401	1	2
HST	History of diabetes (any blood relative)	6517	103	2	1.4398	1	2
HDM	History of diabetes (parent or sibling)	6517	103	2	1.2282	1	2
HDP	History of diabetes (parent)	6517	103	2	1.1817	1	2
HDS	History of diabetes (sibling)	6517	103	2	1.0727	1	2
EXC	Exercise compared to peers	6493	127	3	2.1398	1	3

Missing values of the explanatory variables limit the number of records available for the logistic analysis. Logistic regression requires that values for all dependent variables be present for every record. Classification tree analysis is able to compensate for missing fields by substituting a surrogate variable for a missing field. Table 5 shows that approximately half of the records are missing values for the two cholesterol variables HBC and BCM. However, test regressions and classification trees that included the cholesterol variables showed that they do not significantly contribute to the prediction and they are dropped from the set of candidate variables. The value of the gestational diabetes variable is taken as 1 (had gestational diabetes = no) for men. Only 0.5% of women report having had gestational diabetes.

## Results

### Logistic regression

The probability  $p$  of prediabetes or undiagnosed diabetes is modeled by a logistic, or log-odds, transformation

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (1)$$

where the  $x_i$  are the continuous or dichotomous explanatory variables and the coefficients  $\beta_i$  are the regression coefficients estimated using the method of maximum likelihood, with the response variable assumed to have a binomial distribution. The probability may be rewritten as

$$p = \left( 1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)} \right)^{-1} \quad (2)$$

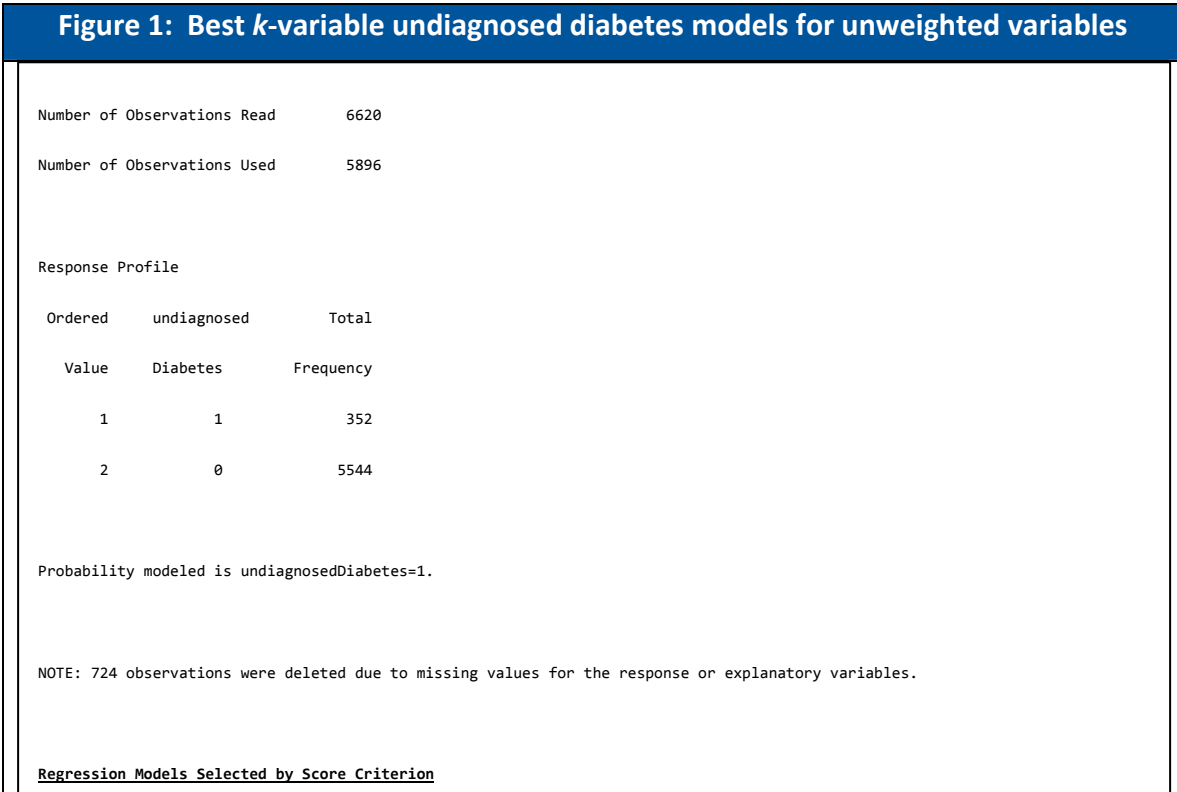
Logistic regressions were generated with SAS 9.1 <sup>(3)</sup>.

### Techniques

The search for an optimal logistic regression model is approached in two ways. One approach finds the best model that includes exactly  $k$  explanatory variables ( $k$ -variable model). The approach finds all models for  $k$  varying from 0 to  $n$ , where  $n$  is the total number of explanatory variables. **Figure 1** shows the explanatory variables included in models for undiagnosed diabetes based on the full set of explanatory variables, excluding HBC, BCM, and RAC (see **Table 5** for definitions). The cholesterol variables are excluded for reasons already discussed relating to missing values and insignificance as a predictor. Race/ethnicity is omitted because its four categories (white, black, mexican american, and other) have no natural order. The value assigned to a particular category of race/ethnicity is arbitrary. The category number is not necessarily identified with increased or decreased risk of prediabetes or diabetes. The only way to include race/ethnicity is to develop separate regressions for each category but that increases complexity, contrary to the aim of developing a simple model.

---

<sup>3</sup> SAS Institute Inc., Cary, NC, USA



The chi-square score (**Figure 1**) is a goodness-of-fit test based on the likelihood that the dependent variable is predicted by the explanatory variables. Larger values of chi-square score indicate better predictions of the dependent variable. Likelihood is a probability that varies from 0 to 1. A good model is one in which the likelihood probability is high or, equivalently, minus 2 times the log of the likelihood (-2LL) is low, since -2LL varies from 0 to infinity. The difference in -2LL between two models is the likelihood ratio. Accounting for degrees of freedom, which depend on the number of parameters in each model, the likelihood ratio approximates a chi-square distribution. The chi-square distribution thus provides a statistical measure of the significance of the likelihood ratio.

Adding explanatory variables to the model rapidly improves the chi-square score at first. After about eight variables, the rate of increase in chi-square score slows as additional variables are added, indicating that about eight variables are responsible of most of the predictive power of the regression. Among the best explanatory variables are age (AGE), waist circumference (WST), waist-to-hip ratio (WHR), body mass index (BMI), blood pressure medication (BPM), waist circumference (WST), family history of diabetes in a parent (HDP) or sibling (HDS), and gestational diabetes (GDB).

The second approach in the search for an optimal logistic regression model is stepwise logistic regression (Figure 2). In the stepwise approach, a variable enters or leaves the model depending on a chi-squared statistic based on the difference in log likelihood of models with and without the variable. After adding a new variable, all variables already in the model are rechecked and successively dropped if they do not provide significant improvement to the model. The procedure continues until some cutoff level for variable significance is reached (e.g., until the step at which all variables not in the model have a significance higher than 0.05). The stepwise approach arrives at a model similar to the 8-variable model found by the best  $k$ -variable model approach. The one difference between the two models is that blood pressure medication (BPM) in the best  $k$ -variable model is replaced by high blood pressure (HBP) in the stepwise model.

The fact that undiagnosed diabetes is negatively correlated with waist circumference is likely an attempt by the model to compensate for an overly positive correlation of undiagnosed diabetes with waist-to-hip ratio (Figure 2). The problem arises because waist-to-hip ratio and waist circumference are correlated. In fact, most of the explanatory variables are correlated. This illustrates a problem with the logistic regression approach. It is difficult to assign physical meaning to the coefficients of the explanatory variables, making it difficult to translate them into a simple paper-and-pencil model. This was the motivation for considering the classification tree approach.

Using unweighted records, the undiagnosed diabetes model in Figure 2 gives a sensitivity and specificity of 79.0% and 70.5%, respectively, and the area under the ROC curve is 0.824. Using weighted records with the same explanatory variables gives a sensitivity and specificity of 85.5% and 64.9%, respectively, and an area under the ROC curve of 0.818.

Cross products of the explanatory variables were also considered. Not all possible cross product models can, in practice, be considered for the best  $k$ -variable model approach because, for fifteen explanatory variables, the number of possible models increases from  $2^n$  (32768) without cross products to  $2^{n + n(n+1)/2}$  ( $4 \times 10^{40}$ ) with cross products, where  $n$  is the number of explanatory variables. Furthermore, the results of the logistic regression with cross products are even more difficult to interpret than the results without cross product. It was decided not to pursue this aspect of the logistic regression approach.

**Figure 2: Stepwise undiagnosed diabetes model for unweighted variables**

<u>Summary of Stepwise Selection</u>							
Step	Effect		DF	Number		Score	Variable
	Entered	Removed		In	Chi-Square		
1	whr		1	1	231.1254	<.0001	Waist-to-hip ratio
2	age		1	2	70.3527	<.0001	Age at interview
3	bmi		1	3	99.3156	<.0001	Body mass index
4	hdp		1	4	34.6491	<.0001	Parent had diabetes
5	gdb		1	5	16.0226	<.0001	Diabetes only when pregnant
6	hbp		1	6	11.4114	0.0007	Told had high blood pressure
7	hds		1	7	10.1595	0.0014	Sibling had diabetes
8	wst		1	8	8.5084	0.0035	Waist circumference (cm)

<u>Analysis of Maximum Likelihood Estimates</u>						
Parameter	DF	Estimate	Standard		Wald	
			Error	Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-17.1570	0.9727	311.0895	<.0001	0.000
age	1	0.0377	0.00391	93.2915	<.0001	1.038
wst	1	-0.0411	0.0141	8.4762	0.0036	0.960
whr	1	9.0164	1.2129	55.2624	<.0001	8236.696
bmi	1	0.1527	0.0279	29.8888	<.0001	1.165
hbp	1	0.4026	0.1222	10.8648	0.0010	1.496

### Prediction equations

The goal of the logistic regression approach is to produce a single equation capable of predicting both prediabetes and undiagnosed diabetes. The idea is to produce a prediabetes model with maximal sensitivity and specificity, and then apply it to both prediabetes and undiagnosed diabetes. Probability thresholds are selected that give

approximately 80% sensitivity for prediabetes and, at a higher threshold, 80% sensitivity for diabetes, regardless of specificity.

For the purpose of calculating the prediabetes model, any observation in the dataset that exceeds the FPG and/or OGTT criteria for prediabetes or diabetes is classified as “prediabetes.” This “prediabetes” variable is termed *elevated plasma glucose* to signify that it includes some records that would otherwise be classified as diabetes. The purpose of this reclassification is to use the entire dataset when deriving the prediabetes logistic regression, rather than dropping observations classified as diabetes.

Table 7: Elevated plasma glucose maximum likelihood estimates for the  $\beta$  coefficients based on sample weighted data and age  $\geq 20$  years. No additional explanatory variables meet the 0.05 significance level for entry into the model

Parameter	Description	Estimated $\beta$
Intercept	Intercept	-21.6343
AGE	Age at interview	0.0402
SEX	Gender	-0.5042
WGT	Weight (kg)	-0.0829
HGT	Standing height (cm)	0.0730
WHR	Waist-to-hip ratio	5.3827
BMI	Body mass index	0.2947
HBP	Told had high blood pressure	0.3449
HDP	Parent had diabetes	0.3981

shows the estimated coefficients  $\beta$  (Equation (2)) for the elevated plasma glucose model based on sample weighted NHANES III prediabetes and undiagnosed diabetes data for people age  $\geq 20$  years ( $N = 6009$ ). Gender (SEX) has values male=1 and female=2. The dichotomous variables HBP and HDP have values no=1 and yes=2.

**Table 7: Elevated plasma glucose maximum likelihood estimates for the  $\beta$  coefficients based on sample weighted data and age  $\geq 20$  years. No additional explanatory variables meet the 0.05 significance level for entry into the model**

Parameter	Description	Estimated $\beta$
Intercept	Intercept	-21.6343
AGE	Age at interview	0.0402
SEX	Gender	-0.5042
WGT	Weight (kg)	-0.0829
HGT	Standing height (cm)	0.0730
WHR	Waist-to-hip ratio	5.3827
BMI	Body mass index	0.2947
HBP	Told had high blood pressure	0.3449
HDP	Parent had diabetes	0.3981

Table 8 shows that for a sensitivity of 79.9%, the elevated plasma glucose model has a specificity of 63.6% and the area under the ROC curve is 0.793. The elevated plasma glucose model is used to predict undiagnosed diabetes by setting the probability threshold for diabetes to be 0.453. Thus probabilities from 0.254 to  $<0.453$  are true prediabetes and probabilities  $\geq 0.453$  are classified as undiagnosed diabetes. For a sensitivity of 79.8%, the elevated plasma glucose model applied to undiagnosed diabetes has a specificity of 76.4% and the area under the ROC curve is 0.856. Figure 3 shows the ROC curve for elevated plasma glucose model in Table 6 and for the elevated plasma glucose model applied to undiagnosed diabetes.

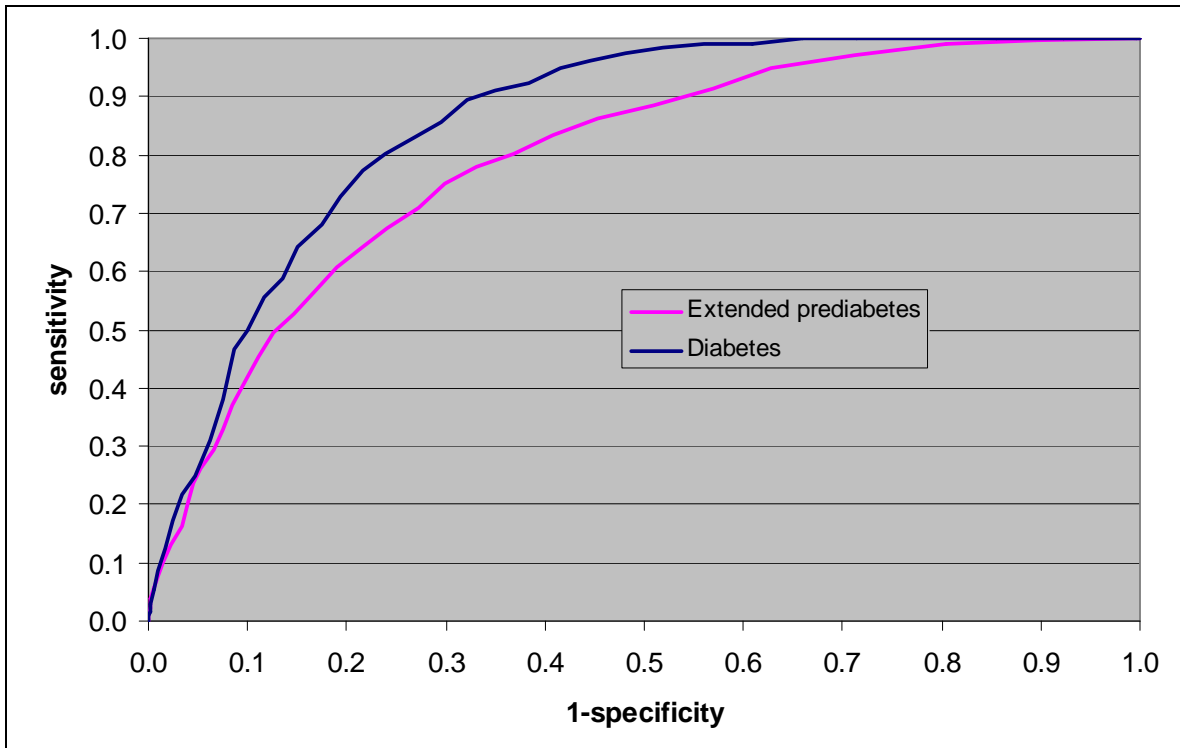
Table 8: Sensitivity, specificity, and area under ROC curve for elevated plasma glucose model in				
Table 7: Elevated plasma glucose maximum likelihood estimates for the $\beta$ coefficients based on sample weighted data and age $\geq 20$ years. No additional explanatory variables meet the 0.05 significance level for entry into the model				
Parameter	Description	Estimated $\beta$		
Intercept	Intercept	-21.6343		
AGE	Age at interview	0.0402		
SEX	Gender	-0.5042		
WGT	Weight (kg)	-0.0829		
HGT	Standing height (cm)	0.0730		
WHR	Waist-to-hip ratio	5.3827		
BMI	Body mass index	0.2947		
HBP	Told had high blood pressure	0.3449		
HDP	Parent had diabetes	0.3981		
and for the elevated plasma glucose model applied to undiagnosed diabetes				
	Probability threshold	Sensitivity	Specificity	ROC
Elevated plasma glucose	0.254	79.9%	63.6%	0.793
Undiagnosed diabetes	0.453	79.8%	76.4%	0.856

**Figure 3: ROC curve for elevated plasma glucose model in**

**Table 7: Elevated plasma glucose maximum likelihood estimates for the  $\beta$  coefficients based on sample weighted data and age  $\geq 20$  years. No additional explanatory variables meet the 0.05 significance level for entry into the model**

<b>Parameter</b>	<b>Description</b>	<b>Estimated <math>\beta</math></b>
Intercept	Intercept	-21.6343
AGE	Age at interview	0.0402
SEX	Gender	-0.5042
WGT	Weight (kg)	-0.0829
HGT	Standing height (cm)	0.0730
WHR	Waist-to-hip ratio	5.3827
BMI	Body mass index	0.2947
HBP	Told had high blood pressure	0.3449
HDP	Parent had diabetes	0.3981

**and for the elevated plasma glucose model applied to undiagnosed diabetes**



Whereas the elevated plasma glucose model given in

Table 7: Elevated plasma glucose maximum likelihood estimates for the  $\beta$  coefficients based on sample weighted data and age  $\geq 20$  years. No additional explanatory variables meet the 0.05 significance level for entry into the model

Parameter	Description	Estimated $\beta$
Intercept	Intercept	-21.6343
AGE	Age at interview	0.0402
SEX	Gender	-0.5042
WGT	Weight (kg)	-0.0829
HGT	Standing height (cm)	0.0730
WHR	Waist-to-hip ratio	5.3827

BMI	Body mass index	0.2947
HBP	Told had high blood pressure	0.3449
HDP	Parent had diabetes	0.3981

is based on all available data, validation tests were conducted on split datasets where the model is “trained” on 2/3 of the data, selected at random, and tested on the remaining 1/3 of the data. Validation tests were repeated for different random selections of training and test data. These tests all produced elevated plasma glucose models closely approximating that given in

Table 7: Elevated plasma glucose maximum likelihood estimates for the  $\beta$  coefficients based on sample weighted data and age  $\geq 20$  years. No additional explanatory variables meet the 0.05 significance level for entry into the model

Parameter	Description	Estimated $\beta$
Intercept	Intercept	-21.6343
AGE	Age at interview	0.0402
SEX	Gender	-0.5042
WGT	Weight (kg)	-0.0829
HGT	Standing height (cm)	0.0730
WHR	Waist-to-hip ratio	5.3827
BMI	Body mass index	0.2947
HBP	Told had high blood pressure	0.3449
HDP	Parent had diabetes	0.3981

and performed nearly as well on test data as on training data. We conclude that the model is robust and will perform well on previously unseen data.

Attempts were made to improve the prediabetes model (

Table 7: Elevated plasma glucose maximum likelihood estimates for the  $\beta$  coefficients based on sample weighted data and age  $\geq 20$  years. No additional explanatory variables meet the 0.05 significance level for entry into the model

Parameter	Description	Estimated $\beta$
Intercept	Intercept	-21.6343
AGE	Age at interview	0.0402
SEX	Gender	-0.5042
WGT	Weight (kg)	-0.0829
HGT	Standing height (cm)	0.0730
WHR	Waist-to-hip ratio	5.3827
BMI	Body mass index	0.2947
HBP	Told had high blood pressure	0.3449
HDP	Parent had diabetes	0.3981

) by including products of the explanatory variables or by generating separate equations for different race/gender combinations. Those tests succeeded in producing small gains in predictive power. However, the product terms are difficult to assign physical meaning. The small improvements resulting from race/gender combinations or product terms are not judged worth sacrificing the simplicity of model in

Table 7: Elevated plasma glucose maximum likelihood estimates for the  $\beta$  coefficients based on sample weighted data and age  $\geq 20$  years. No additional explanatory variables meet the 0.05 significance level for entry into the model

Parameter	Description	Estimated $\beta$
Intercept	Intercept	-21.6343
AGE	Age at interview	0.0402

SEX	Gender	-0.5042
WGT	Weight (kg)	-0.0829
HGT	Standing height (cm)	0.0730
WHR	Waist-to-hip ratio	5.3827
BMI	Body mass index	0.2947
HBP	Told had high blood pressure	0.3449
HDP	Parent had diabetes	0.3981

#### Classification and regression tree (CART<sup>4</sup>)

Classification trees have advantages over logistic regression for the creation of a simple paper-and-pencil screening tool for the prediabetes and undiagnosed diabetes. They provide a solution that is simple to understand and interpret and require no mathematical calculations like those for logistic regression, e.g., Equation (2). Data need little preprocessing such as normalization, creation of dummy variables, or removal of records with missing fields, and both nominal and categorical variables are handled. Model training and testing make use of the full data set though a technique known as v-fold cross-validation (<sup>5</sup>).

A classification tree separates data into mutually exclusive groups that concentrate a particular class of the target variable. Here the target variable has a value of 0 or 1 depending on whether undiagnosed diabetes (or prediabetes) is absent or present, respectively. The value of a target variable is referred to as its class. Starting at the tree root, the data is split into two groups conditional on whether an explanatory variable, or a linear combination of explanatory variables, is greater than some value. The particular explanatory variable value selected for the split is the one that best separates the target classes, 0 or 1, at the root node into two child nodes. If the primary splitter variable

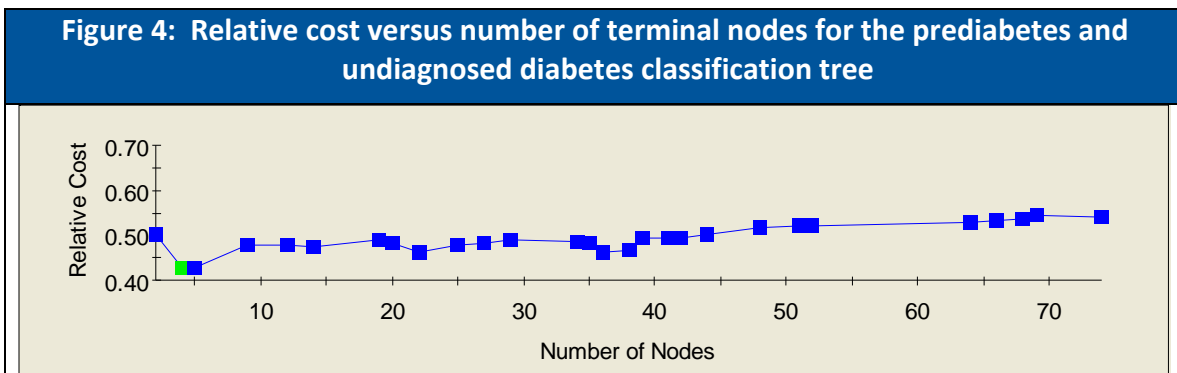
<sup>4</sup> All classification trees were created using CART software v5.0, Salford Systems, San Diego, CA 92123.

<sup>5</sup> Breiman L, Friedman J, Olshen RA, Stone CJ. *Classification And Regression Trees*. 1998, Chapman & Hall / CRC, Boca Raton, FL 33431.

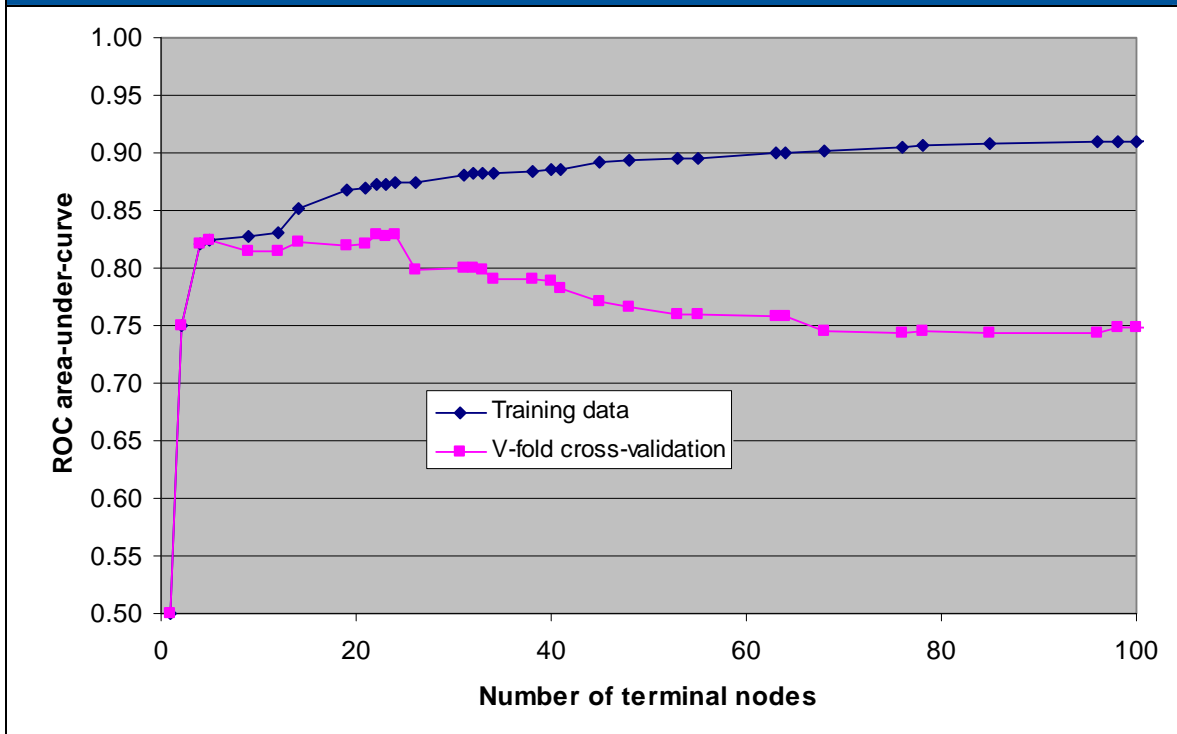
field is missing, a surrogate splitter variable is instead used. The process then repeats for each of the child nodes. Subsequent splits may involve another explanatory variable or a different value of a previously used variable. A node that is not further split is referred to as a terminal node. Each terminal node is assigned to a target class conditional on whether target class prevalence exceeds threshold prevalence. The same threshold prevalence is applied to all terminal nodes and determines overall tree sensitivity and specificity.

The classification tree is grown to its maximum size and then pruned based on a criterion that balances the number of terminal nodes (*complexity*) and tree accuracy (*misclassification rate*), termed *misclassification cost*. The largest tree for a given dataset is one in which every unique combination of target and explanatory variables constitutes a terminal node. This tree perfectly reproduces the training data, but its complexity precludes insight and understanding into the predictive structure of the data and it often does not perform well on new test data. Therefore, it is pruned to a tree where the number of terminal nodes is nearly equal to the number for the minimum misclassification cost.

The tree selected for the undiagnosed diabetes model is one with a low relative cost and a high ROC integral value for test data. It has 14 terminal nodes. **Figure 4** shows that relative cost is a minimum for 4 terminal nodes (green square). **Figure 5** shows that the ROC integral value based on test data is a maximum for 24 terminal nodes. Because ROC integral value as a function of number of terminal nodes is relatively flat over a large range, accuracy is only slightly compromised by selecting a tree of slightly lower ROC integral value than the maximum, and lower relative cost tree, to reduce complexity.



**Figure 5: Value of ROC integral for undiagnosed diabetes versus number of terminal nodes**



### Undiagnosed diabetes

The classification tree described here is trained to predict undiagnosed diabetes and then applied to the prediction of prediabetes. Recall that in the logistic regression approach described above, the regression model, originally trained to predict prediabetes, is applied to the prediction of undiagnosed diabetes by using a different probability threshold. A similar technique allows a single classification tree to predict both prediabetes and undiagnosed diabetes by using different threshold prevalences for each disease level.

Since one of the goals is to create a simple model, body mass index (BMI) and waist-to-hip ratio (WHR) are dropped from the list of candidate explanatory variables (Table 5) in favor of weight (WGT), height (HGT), and waist circumference (WST), which require no calculation or lookup tables. As in the logistic regression analysis, the cholesterol variables HBC and BCM are eliminated from consideration because of the large number of missing fields and low predictive value. History of diabetes in any blood relative (HST)

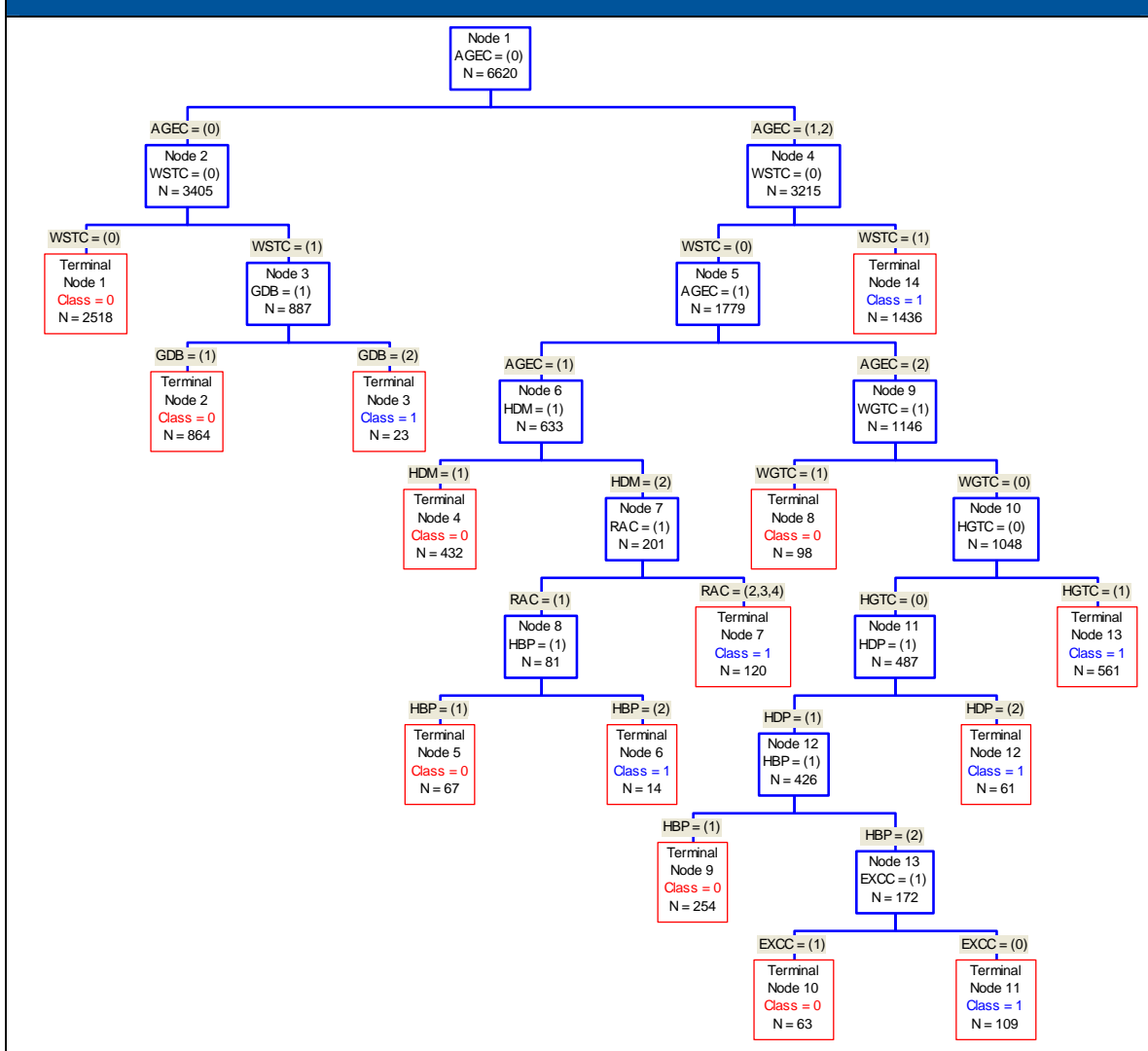
is eliminated in favor of the more specific diabetes history variables – history of diabetes in a parent or sibling (HDM), a parent only (HDP), or sibling only (HDS).

A problem with the classification trees initially created for this analysis is that nodes near the root node split multiple times on only a few of the continuous explanatory variables before other candidate predictor variables are incorporated into the tree. It is necessary to grow the tree to a large, unmanageable number of nodes before nodes begin to split on the full range of explanatory variables. However, the desire is to incorporate a reasonable number of different explanatory variables in the model and, at the same time, avoid complexity.

A method found to achieve this goal is to first grow the tree with all explanatory variables (except those dropped from consideration), including the continuous variables. Once a node splits on the value of a continuous variable, redefine the variable to be categorical based on the split value. This approach is extended, as in the case of AGE, to allow nodes to split on a second value of the variable and creating three categories based on the first and second split values. Converting a continuous variable to a categorized variable prevents further splitting on that variable and forces the tree to incorporate other variables to handle subsequent node splits. The naming convention used for a new variable formed by categorizing a continuous variable is to append the letter C to the original variable name, e.g., categorized AGE becomes AGE<sub>C</sub>.

Figure 6 shows the 14-node classification tree for undiagnosed diabetes based on sample weighted NHANES III data. Blue boxes denote internal nodes and red boxes denote terminal nodes. Internal and terminal nodes are numbered separately. Starting with the root node (Node 1), the data ( $N=6620$  records) is split based on the value of categorized age AGE<sub>C</sub>=0. Data that satisfy this condition ( $N=3405$ ) are assigned to the left child node (Node 2) and those that do not ( $N=3215$ ) (AGE<sub>C</sub>=1 or 2) are assigned to the right child node (Node 4). Sample weighted observation frequency for each node is not shown in this view. Gray highlighted equalities above each child node indicate the value(s) of the explanatory variable passed to the child node from the parent node. Each terminal node shows the class assigned to that node. For example, terminal node 1 is designated class 0, i.e., no undiagnosed diabetes. For prediabetes, nodes 2, 5, 8, and 9 are switched to class = 1.

Figure 6: Classification tree for undiagnosed diabetes (14 terminal nodes, 9 variables)



The categorization boundaries for continuous explanatory variables are shown in [Table 9](#). Relative importance of the primary splitter variables is shown in [Error! Reference source not found.](#). Aside from those variables eliminated from consideration, the only candidate explanatory variable that does not appear in the tree is gender (SEX).

<b>Table 9: Categorization boundaries for continuous explanatory variables</b>	
<b>Variable</b>	<b>Value</b>
HGTC	160.35
WGTC	76.52
WSTC	97.55
AGEC=0	43.5
AGEC=1	56.5
EXCC	2

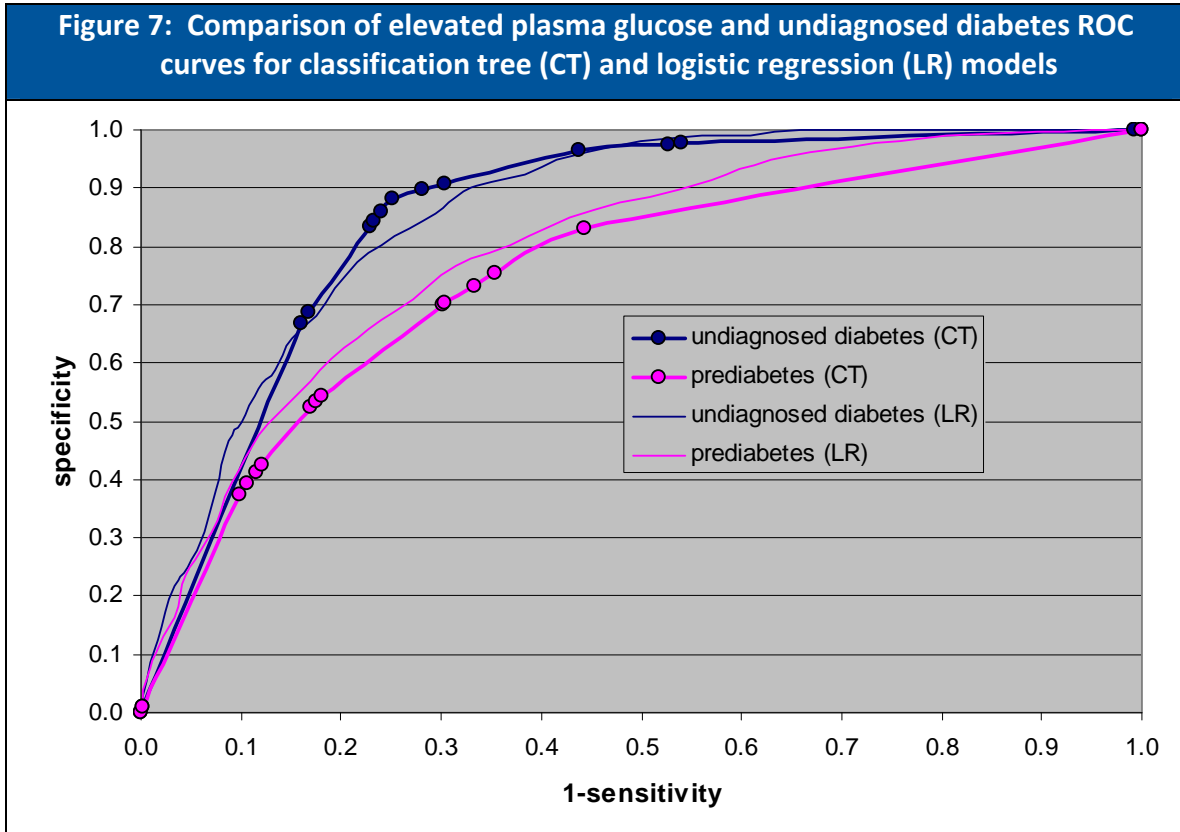
<b>Table 10: Relative importance of primary splitter variables</b>	
<b>Variable</b>	<b>Score</b>
AGEC	100.00
WSTC	27.83
HDM	4.31
WGTC	2.89
GDB	2.75
HGTC	1.77
EXCC	1.54
HBP	1.37
HDP	1.24
RAC	0.70

As discussed above, sensitivity and specificity of the classification tree depend on a threshold prevalence (or probability) of the target class applied to all terminal nodes. In the logistic regression analysis, threshold probability was set to give a sensitivity of about 80%. Here the threshold prevalence is set to the sample weighted prevalence of target class 1 for the dataset, for which the resulting sensitivity is not necessarily 80%.

**Table 11** shows measures of predictive accuracy for the prediabetes and undiagnosed diabetes classification trees based on training data, and when applied in v-fold cross-validation tests and to NHANES 1999-2004 test data. The decrease in sensitivity and specificity when applied to test data is not extreme and the model appears to be robust. Positive predictive value (PPV) indicates the fraction of observations predicted to have the condition that actually have it. PPV for undiagnosed diabetes is much lower for NHANES 1999-2004 than for NHANES III data because the prevalence of undiagnosed diabetes is much lower in the NHANES 1999-2004 dataset. The classification tree performs about the same as the logistic regression for undiagnosed diabetes and slightly worse than logistic regression for elevated plasma glucose (**Table 8** and **Figure 7**). Nevertheless, the classification tree is a viable predictor of prediabetes and undiagnosed diabetes and more easily translates to a simple paper-and-pencil screening test than does logistic regression.

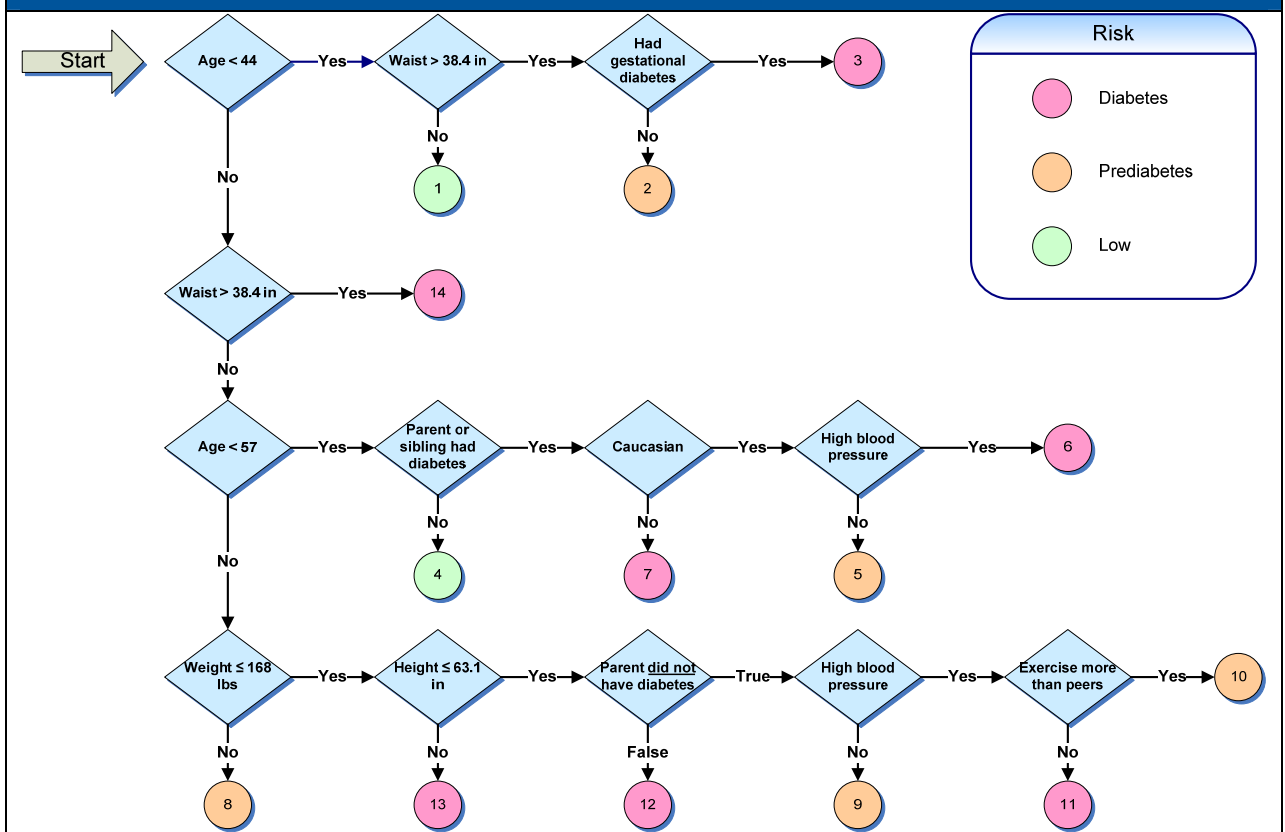
**Table 11: Specificity, sensitivity, positive and negative predictive values (PPV, NPV) and ROC for tree in Figure 6 applied to elevated plasma glucose and undiagnosed diabetes**

Variable	Sensitivity (%)	Specificity (%)	PPV	NPV	ROC
<b>Undiagnosed diabetes</b>					
Training	88.16	74.92	0.1369	0.9929	0.8508
V-fold cross-validation	78.22	74.13			0.8219
NHANES 1999-2004	81.02	66.81	0.0627	0.9923	0.7685
<b>Prediabetes or undiagnosed diabetes</b>					
Training	75.36	64.59	0.4940	0.8511	0.7503
NHANES 1999-2004	77.65	51.36	0.4053	0.8433	0.6991



The classification tree in [Figure 6](#) is translated to the form of a screening test shown in [Figure 8](#). A screening test participant starts at “Start” and, at each decision point, moves right for an affirmative or response of true, or down for a negative or response of false, to the question in the blue diamond-shaped box. The participant eventually arrives at one of the red, orange or green circles. The color of the circle is indicative of the primary risk, be it prediabetes or undiagnosed diabetes. The actual prevalence and relative risk are found by referring to [Table 12](#), arranged in order of increasing relative risk of undiagnosed diabetes. Relative risk compares the prevalence of prediabetes or undiagnosed diabetes to the prevalence in the general US population of people age  $\geq 40$  years.

Figure 8: Simplified classification tree in the form of an easily understandable screening test



**Table 12: Prevalence and relative risk of prediabetes or undiagnosed diabetes corresponding to colored circles in Figure 8**

Circle number	Prevalence		Relative risk	
	Undiagnosed diabetes	Prediabetes	Undiagnosed diabetes	Prediabetes
10	0.002	0.429	0.039	1.581
1	0.002	0.119	0.054	0.440
8	0.003	0.527	0.065	1.942
4	0.006	0.280	0.138	1.032
2	0.018	0.354	0.418	1.306
5	0.020	0.295	0.465	1.088
9	0.025	0.295	0.572	1.089
11	0.082	0.453	1.890	1.670
7	0.091	0.369	2.097	1.358
6	0.096	0.542	2.215	1.998
13	0.099	0.379	2.283	1.398
12	0.118	0.393	2.740	1.449
14	0.156	0.478	3.625	1.760
3	0.264	0.097	6.114	0.358

## Conclusions

Prediabetes and undiagnosed diabetes prediction models are developed for a set of simple, health-related questions known to the average person. Logistic regression and classification tree analysis of the corresponding data provide similar accuracy in predicting the target variables. The advantage of the classification tree is that it easily translates to a paper-and-pencil screening test that is understandable by a screening participant. The model final screening test presented here performs well on training data with little loss of predictive power when applied to test.