Appendix 1: The CHAID Algorithm

In this appendix we describe the steps in the CHAID algorithm by applying it to a simple data set consisting just of a categorical response variable ALWPRIM (Y) and two explanatory variables ETHNIC (X1) and WELSH (X2). The variable Y has four categories, Y = 1,2,3,4; the variable X1 has four categories, X1 = 1,2,3,4; the variable X2 has three categories, X2 = 0,1,2. The steps in the CHAID algorithm are then:

1. Calculate the distribution of the response variable Y in the "root" node.

Cat.	%	n
1	35.00	35
2	8.00	8
3	35.00	35
4	22.00	22
Total	(100.00)	100

2. For each explanatory variable X, find that pair of categories of X that are least significantly different (that is has the largest p-value) with respect to the distribution of Y within this node. The method use to calculate this p-value depends on the measurement level of Y. In this example Y is categorical, and so a series of chisquare tests are performed:

   i. The relationship between the explanator ETHNIC (X1) and the response ALWPRIM (Y) within the node is given by the following crosstabulation:

X1/Y	**1**	**2**	**3**	**4**	***RowTotl***
1	23	5	19	4	*51*
2	12	2	15	13	*42*
3	0	1	0	1	*2*
4	0	0	1	4	*5*
Coltotl	*35*	*8*	*35*	*22*	*100*

Chi^2 = 25.63559 d.f.= 9 (p=0.002342955)

Since X1 has four categories, there are six 2×4 sub-crosstabulations to consider

X1/Y	**1**	**2**	**3**	**4**	***RowTotl***
1	23	5	19	4	*51*
2	12	2	15	13	*42*
Coltotl	*35*	*7*	*34*	*17*	*93*

Chi^2 = 9.193281 d.f.= 3 (p=0.02682849)

X1/Y	1	2	3	4	*RowTotl*
1	23	5	19	4	*51*
3	0	1	0	1	*2*
Coltotl	*23*	*6*	*19*	*5*	*53*

Chi^2 = 8.019281 d.f.= 3 (p=0.0456149)

X1/Y	1	2	3	4	*RowTotl*
1	23	5	19	4	*51*
4	0	0	1	4	*5*
Coltotl	*23*	*5*	*20*	*8*	*56*

Chi^2 = 19.72078 d.f.= 3 (p=0.0001939266)

X1/Y	1	2	3	4	*RowTotl*
2	12	2	15	13	*42*
3	0	1	0	1	*2*
Coltotl	*12*	*3*	*15*	*14*	*44*

Chi^2 = 7.23356 d.f.= 3 (p=0.0648145)

X1/Y	1	2	3	4	*RowTotl*
2	12	2	15	13	*42*
4	0	0	1	4	*5*
Coltotl	*12*	*2*	*16*	*17*	*47*

Chi^2 = 4.962482 d.f.= 3 (p=0.1745651)

X1/Y	2	3	4	*RowTotl*
3	1	0	1	*2*
4	0	1	4	*5*
Coltotl	*1*	*1*	*5*	*7*

Chi^2 = 3.08 d.f.= 2 (p=0.2143811)

ii. The algorithm then identifies the pair of categories of X1 with largest p-value above and compares this p-value to a prespecified alpha level, α_{merge} (= 0.05, default value). In this example, this is the pair defined by categories 3 and 4 of X1. Since the p-value for this pair (0.2143) is greater than α_{merge}, the two categories are merged to form a single compound category. As a result, a new set of categories of X1 is formed, and the sub-crosstabulation analysis of (i) is repeated. There are now three such sub-crosstabulations:

X1/Y	1	2	3	4	*RowTotl*
1	23	5	19	4	*51*
3,4	0	1	1	5	*7*
Coltotl	*23*	*6*	*20*	*9*	*58*

Chi^2 = 20.25577 d.f.= 3 (p=0.0001502343)

X1/Y	**1**	**2**	**3**	**4**	RowTotl
2	12	2	15	13	*42*
3,4	0	1	1	5	*7*
Coltotl	*12*	*3*	*16*	*18*	*49*

Chi^2 = 6.408565 d.f.= 3 (p=0.09333908)

X1/Y	**1**	**2**	**3**	**4**	***RowTotl***
1	23	5	19	4	*51*
2	12	2	15	13	*42*
Coltotl	*35*	*7*	*34*	*17*	*93*

Chi^2 = 9.193281 d.f.= 3 (p=0.02682849)

iii. Again, the pair that results in the largest p-value greater than $\alpha_{merge} = 0.05$ is merged. Here this corresponds to merging category 2 with the compound category 3,4. If further mergers were possible, this process would continue. However, since there are now only two (merged) categories remaining for X1, the merging process stops.

iv. The algorith now calculates an adjusted p-value for the set of merged categories of X1 and the categories of Y using a Bonferroni adjustment.

X1/Y	**1**	**2**	**3**	**4**	***RowTotl***
1	23	5	19	4	*51*
2,3,4	12	3	16	18	*49*
Coltotl	*35*	*8*	*35*	*22*	*100*

Chi^2 = 13.08861 d.f.= 3 (p=0.004448843)

The chisquare p-value above is adjusted using a Bonferroni multiplier. Since X1 is nominal, the Bonferroni multiplier is calculated as follows:

$$B_{free} = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{r!(r-i)!}$$

where c = number of original categories of X1 (4) and r = number of merged categories (2). This leads to an adjusted p-value of 0.0311 (= 0.00448843 × 7).

3. Steps (i) to (iv) above are now repeated, replacing X1 by X2 (WELSH).
   i. The crosstabulation of X2 by Y for the root node is

X2/Y	**1**	**2**	**3**	**4**	***RowTotl***
0	33	8	34	22	*97*
1	1	0	1	0	*2*
2	1	0	0	0	*1*
Coltotl	*35*	*8*	*35*	*22*	*100*

Chi^2 = 2.768778 d.f.= 6 (p=0.8372577)

ii. X2 has three categories, so there are only three 2×4 sub-crosstabulations:

X2/Y	**1**	**2**	**3**	**4**	RowTotl
0	33	8	34	22	*97*
1	1	0	1	0	*2*
Coltotl	*34*	*8*	*35*	*22*	*99*

Chi^2 = 0.8881097 d.f.= 3 (p=0.8282962)

X2/Y	**1**	**2**	**3**	**4**	RowTotl
0	33	8	34	22	*97*
2	1	0	0	0	*1*
Coltotl	*34*	*8*	*34*	*22*	*98*

Chi^2 = 1.901759 d.f.= 3 (p=0.5930452)

X2/Y	**1**	**3**	***RowTotl***
1	1	1	*2*
2	1	0	*1*
Coltotl	*2*	*1*	*3*

Chi^2 = 0.75 d.f.= 1 (p=0.3864762)

iii. Here, the pair of categories of X2 with largest p-value are categories 0 and 1 with a p-value of 0.8283. We therefore merge categories 0 and 1 to form a single compound category. The merging process for this variable now stops since there are just two categories left.

iv. The final p-value for the crosstabulation of X2 and Y is then calculated:

X2/Y	**1**	**2**	**3**	**4**	***RowTotl***
0,1	34	8	35	22	*99*
2	1	0	0	0	*1*
Coltotl	*35*	*8*	*35*	*22*	*100*

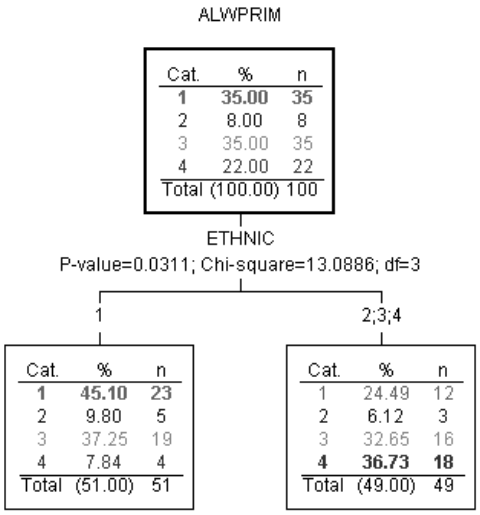
Chi^2 = 1.875902 d.f.= 3 (p=0.5985587)

The adjusted p-value in this case is 1 (since 0.5985587×3 is greater than 1).

4. The final step is to split the node on the basis of that "merged" explanator with smallest adjusted p-value less than asplit = 0.05. Here this means we split the root node into two subnodes on the basis of the merged categories of X1. That is, one subnode contains the 35 cases with X1 = 1 while the other contains the remaining 65 cases with X1 = 2, 3 or 4.

5. Continue to grow the tree until the stopping criteria are satisfied. Figure A1 shows the (two branch) tree constructed by CHAID in steps 1 - 4.

Figure A1


ALWPRIM
Cat. % n
1 35.00 35
2 8.00 8
3 35.00 35
4 22.00 22
Total (100.00) 100
ETHNIC
P-value=0.0311; Chi-square=13.0886; df=3
1
2;3;4
Cat. % n
1 45.10 23
2 9.80 5
3 37.25 19
4 7.84 4
Total (51.00) 51
Cat. % n
1 24.49 12
2 6.12 3
3 32.65 16
4 36.73 18
Total (49.00) 49