EMPIRICAL COMPARISON OF APPROACHES TO FORMING STRATA:
USING CLASSIFICATION TREES TO ADJUST FOR
CONFOUNDING VARIABLES

BY

DANIEL A. BLOCH AND MARK R. SEGAL

TECHNICAL REPORT NO. 108
NOVEMBER 1985

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 R01 GM21215-11

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
EMPIRICAL COMPARISON OF APPROACHES TO FORMING STRATA:
USING CLASSIFICATION TREES TO ADJUST FOR
CONFOUNDING VARIABLES

By

Daniel A. Bloch and Mark R. Segal

TECHNICAL REPORT NO. 108
November 1985

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 R01 GM21215-11

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
EMPIRICAL COMPARISON OF APPROACHES TO FORMING STRATA:
USING CLASSIFICATION TREES TO ADJUST FOR
CONFOUNDING VARIABLES

Daniel A. Bloch and Mark R. Segal

Abstract

In medical studies, the computation of a meaningful incidence rate frequently demands adjustment for covariates. One means for achieving such adjustment is standardization. The (directly) standardized rate is calculated by using the appropriate weighted average of rates over strata determined by the covariates. Thus strata formation is crucial in obtaining good results. This paper compares three competing methods for stratification using data from a drug study in which the incidence rate of a certain side-effect is of interest. The techniques examined are (i) contingency table cell-aggregation, (ii) classification trees, and (iii) logistic regression. The relative merits of the three methods are discussed.

Key Words: Standardization, stratification, contingency-table, classification tree, logistic regression.

1. INTRODUCTION

A frequent aim of medical investigation is to evaluate the efficacy of a treatment in altering the course of a disease process. Outcomes of interest from a non-randomized study are usually affected not only by the treatment, but also by other variables. For example, this paper includes data from a recent study [1] investigating differences among several drug combinations used in the treatment of rheumatoid arthritis in their associated side-effect rates. Depending upon factors such as the sex of the patient, severity of disease, age, and the number of years
the patient has had the disease, purpura (bruising) side-effect rates vary from less than 5% to more than 60%. Older, sicker patients are in a high risk category. If one is interested in a specific drug combination and a disproportionate number of patients taking this combination are in a high risk category, then the side-effect rate will tend to be inflated because it is so often used in this group. Clearly, a proper analysis of side-effect rates will control for these "confounding" variables.

Standardization is one common method used to adjust for confounding variables. The directly standardized side-effect rate is determined by taking the rates of the treatment of interest in the various strata, as defined by confounding variables, and calculating what the overall side-effect rate would have been for the treatment if these rates had been applied to the target or standard population.

Central to the process of standardization is the formation of strata. For most situations involving binary outcomes, such as side-effect present versus no side-effect present, strata are formed by use of either a multiple linear logistic regression model or by considering contingency tables. A third, new, technique for constructing strata is to use recursive partitioning. An empirical comparison of these three approaches is the subject of this paper.

We compare the three approaches to forming strata using data on drug toxicity available from the ARAMIS post-marketing surveillance program. ARAMIS, The American Rheumatism Association Medical Information System, is a national data bank for rheumatic diseases. The data set contains information on 20 drugs and 47 different side-effects. Using data available on 1415 rheumatoid arthritis patients, we investigated differences in associated side-effect rates among several drug combinations commonly used for treatment, see Fries et al. (1985). For the purpose of this paper we confine attention to the side-effect purpura, and ask if patients treated with aspirin in combination with prednisone have significantly more purpura than those treated with aspirin alone. The eight potential confounding variables are sex, age, weight, duration
of disease (D.D.), patient disability (D.I.), number of sore joints (S.J.), amount of morning stiffness (M.S.), and amount of pain attributable to the disease (P.S.). Patient disability is measured by a disability index (D.I.) and takes values between 0.00 and 3.00; S.J. ranges from 0 to 68; M.S. is the duration of joint stiffness upon awakening, measured in hours; and pain is indexed on a scale (P.S.) from 0.00 to 3.00. The variables D.I., S.J., M.S., and P.S. measure different aspects of disease severity.

There are two decisions which must be made in order to calculate standardized side-effect rates. Firstly, however the strata are formed, the investigator must select the variables for which adjustments are to be made. For the drug combination study, selection was done by first listing all known potential confounding variables for which data had been collected. The mean value for each variable was calculated, and a variable was included if patients with values below the mean had a statistically significantly different average side-effect rate than those patients with values above the mean. Other variables were included if previous studies and/or the medical investigators' judgments indicated that they should be. Using this approach, 8 potential confounding variables were included from the original list of 50.

Secondly, the investigator needs to choose a standard population. For the drug combination study, the entire post-marketing surveillance data set of 1415 rheumatoid arthritis patients was used as the standardizing population.

2. FORMING STRATA BY AGGREGATING CELLS IN A CONTINGENCY TABLE

The approach discussed here is the "Pure-Aggregation Strata" procedure described by Moses et al. (1969). Let $X_1, X_2, \ldots, X_k$ be $k$ potential confounding variables. The range of each variable is divided into a number of classes: $C_1$ for variable 1, $C_2$ for variable 2, \ldots, $C_k$ for variable k. Altogether there are $C = C_1 \times C_2 \times \ldots \times C_k$ cells. The strata are formed by first calculating side-effect rates for each of the $C$ cells. Cells are then grouped, in side-effect rate order,
into a number of categories. These categories are the strata used for the standardization process. The number of strata that should be formed depends upon the number of confounding variables and the available sample size. If a large number of strata are formed, then the strata side-effect rate estimates will tend to be highly variable.

Since this procedure is implemented by hand, only 4 variables were included in the cell-aggregation approach to forming strata. The medical investigators felt that sex, age, duration of disease, and disability index were the most important variables to control for. The range of variate values were divided into 2 classes for sex (male, female), age (<55 or >55 years old) and duration of disease (<12 or >12 years), and 3 classes for disability index (0-1, 1-2, 2-3 units) (the number of cells, C, equals 2×2×2×3 = 24). Twenty-seven of the 1415 cases had at least one of these variate values missing; the sample size in this analysis is 1388.

The 24 cells formed from the confounding variables classes and their associated side-effect rates are displayed in Table 1. Four strata were formed; the circled numbers in Table 1 are the strata assignment numbers. Thus males <55 years old except for those with D.D. > 12 and D.I. > 2 and males >55 years old with D.D. ≤ 12 and D.I. < 1 are combined with females with D.D. ≤ 12 and D.I. < 1 to form strata 1, and so on. The stratum side-effect rates for the standardizing population and for the two drug groups of interest are displayed in Table 2.

It is easily verified that the direct standardized purpura side-effect rate for aspirin plus prednisone is .56 and for aspirin alone is .27 (the crude rates are .62 and .25, respectively). A test proposed by W. G. Cochran (1954) tests the null hypothesis that the two standardized rates are equal. Under the null hypothesis, the difference in the standardized side-effect rates divided by the standard error of the difference is approximately standard normally distributed (see e.g. Armitage (1971)). The value that was obtained is 3.85. The patients treated with aspirin plus prednisone had more purpura than those treated with aspirin alone at the .0001 level of significance.
3. CLASSIFICATION TREE APPROACH TO FORMING STRATA

The definitive reference for the technique of tree classifiers is "Classification and Regression Trees" (CART) by Breiman, Friedman, Olshen and Stone (1984). Here, a brief and simplified overview of the methodology will be presented, from the perspective of using the resultant tree for stratifying. The term classification is used throughout because the analysis involved a categorical (binary) response. However, the extensions of both the tree-building and stratifying techniques to the regression setting (continuous response) are straightforward.

There are four components necessary to construct a tree-classifier:

(i) a set of questions of the form: Is a case in a particular region \( A \) of the predictor space \( \mathcal{X} \)?

(ii) a goodness-of-split measure than can be used to determine the worth of any such split. Worth relates to node purity.

(iii) a mechanism for determining what sized tree is best.

(iv) a rule for assigning every terminal node of the selected tree to a specific class. This is more pertinent to description/prediction than stratification and consequently is not pursued here.

The set of possible splits in (i) is reduced to a computationally feasible number by constraining that:

(a) each split depends upon the value of only a single confounding variable [note: this restriction can be loosened; the software now permits splits on linear combinations of variables].

(b) for ordered variables (e.g., age) only splits resulting from questions of the form - Is age > c? are considered.

(c) for categorical confounding variables all possible splits into disjoint subsets of the categories are allowed.
The tree is grown as follows: for each node (the initial node contains the entire sample of 1415 cases)

(i) examine every allowable split on each confounding variable.
(ii) select and execute (create two new daughter nodes) the "best" of these splits.

Steps (i) and (ii) are then reapplied to each of the daughter nodes, and so on.

"Best" is assessed in terms of the goodness-of-split criterion, which was taken to be the Gini diversity index (see Chapter 4, CART). The splitting criterion is such that we recursively create smaller and smaller nodes of progressively increased purity. For example, the tree diagram for males (Figure 1) reveals that the first split was on age and occurred at 44.5 years. This split is more favorable in terms of increasing within node purity than any other split on age and any potential split on any of the other confounding variables. This purity increase is highlighted in the scatterplot (Figure 2) which also displays the subsequent split on disease duration.

It remains to resolve what is an appropriate sized tree. This is done via the following strategy: initially, grow a very large tree. Then, iteratively prune this tree all the way back up, thereby creating a sequence of trees. Finally, select the best tree from this sequence using cross-validated estimates of misclassification rates. This involved procedure was adopted since simpler methods, based on stopping criteria, proved unsatisfactory. Details are found in chapter 3, CART.

A further, especially pertinent issue is the treatment of missing values. As opposed to conventional regression procedures which discard cases with any missing components (see Discussion), considerable information is extracted from cases containing missing values using "surrogate" variables. The first surrogate for a given non-terminal node is the variable that best reproduces the actual split of that node. The second surrogate is the variable that does second best, and so on. Thus, at
every non-terminal node an ordered list of surrogates is compiled. These are used in dealing with missing values as follows: when examining a candidate splitting variable \( X_j \), determine the best split using cases not missing \( X_j \). Those cases missing values of the best splitting variable are sent to the left or right daughter node in accordance with the first surrogate on which they have a non-missing value. For example, in the all patient optimal tree (Figure 3), cases with \( \text{D.I.} > .67 \) and \( \text{D.D.} > 7.15 \) years were further split on the variable \( \text{S.J.} \). The first surrogate for \( \text{S.J.} \) is the variable age. Those cases missing \( \text{S.J.} \) are sent to the left or right child node according to whether age \( \leq 48.6 \) years or age \( > 48.6 \) years.

Another feature of the methodology warrants comment. In this study, the number of cases without purpura outnumber the number with purpura by 2 to 1. If equal numbers of purpura and non-purpura cases are misclassified, then the effect on the purpura misclassification rate will be much larger than on the non-purpura rate. If this is undesirable, or if there is some genuine cost discrepancy between the various sorts of misclassifications, then we can tune the Gini diversity index by incorporating priors and non-unit costs (see chapter 4, CART). For the trees displayed, the cost of misclassifying a purpura case was taken to be twice that of misclassifying a non-purpura case - this being a natural ratio. However, at least for this particular data set, the tree sequence produced is surprisingly insensitive to the values ascribed to costs and priors.

Having selected a tree, the set of terminal nodes form the final strata. The classification tree strata and the side-effect rates for the two drug groups of interest are summarized in Table 3.

The standardized purpura side-effect rates for aspirin plus prednisone and for aspirin alone are \( .50 \) and \( .27 \) respectively. The value of the normal test-statistic is \( 3.03 \); we conclude that aspirin plus prednisone results in more purpura than aspirin alone at the \( .0024 \) level of significance.
4. LOGISTIC REGRESSION APPROACH TO FORMING STRATA

A multiple linear logistic regression model (see e.g. Cox (1970)) describes how a side-effect is related to the confounding variables. There are four steps to forming strata. The first step is to choose the form of the model. We assumed that the logit of \( p_1 \) can be expressed as:

\[
\ln\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}
\]

where \( p_1 \) represents the predicted probability that the \( i \)-th patient has purpura, and \( x_{1i}, x_{2i}, \ldots, x_{ki} \) are the values of the confounding variables for the \( i \)-th patient. The magnitude of the effect of the confounding variables are represented by the parameters \( \beta_1, \beta_2, \ldots, \beta_k \). This particular form of the model does not contain interaction terms, but they could be included if so desired.

The second step is to fit the model to the data in order to obtain estimates of the parameters. We used the BMDP (1983) step-wise logistic regression program.

The third step is to use the fitted model to calculate a predicted probability of purpura occurring for each patient.

Finally, the patients are grouped into strata by combining those patients with the lowest range of predicted probabilities into the first stratum, those with the next lowest range into the second stratum, and so on.

Of the 1415 patients, only 384 had values recorded on all 8 of the potential confounding variables. Weight, number of sore joints, and amount of morning stiffness were not specifically asked for in the information gathering questionnaires. Six hundred fifty-two patients had values for all variables excluding weight, and 1350 patients had values for the 5 variables excluding weight, number of sore joints, and amount of morning stiffness. The 652 cases with weight excluded were entered into the step-wise logistic regression program. The fitted model had the form:
\[
\ln \left( \frac{p_i}{1-p_i} \right) = -2.613 + (.436)(\text{D.I.})_i + (.019)(\text{D.D.})_i + (.018)(\text{Age})_i \\
- (.277)(\text{Sex})_i + (.008)(\text{S.J.})_i
\]

where \((\text{D.I.})_i\) is the value of the \(i\)-th patient's disability index, and so forth.

There is a close correspondence between which confounding variables were selected by this approach and the classification tree approach. See Table 4.

Four strata were formed. Their make-up and the side-effect rates for the two drug treatments of interest are displayed in Table 5.

The standardized purpura side-effect rates for aspirin plus prednisone and for aspirin alone are \(.53\) and \(.40\), respectively. The difference is not statistically significant (value of normal test-statistic is \(1.18\)). This is surprising and appears to contradict our previous results. The standardized aspirin rate of \(.40\) is much higher than the rate of \(.27\) calculated by the other two approaches. The explanation is that a large number of non-purpura patients on aspirin alone had incomplete records. For example, the investigator who most often prescribed aspirin alone had complete records on all aspirin users with purpura. In contrast, only 35% of his non-purpura aspirin patients had complete records.

If weight, S.J., and M.S. are excluded, then there are 1350 complete patient records. Entering these 1350 records into a step-wise logistic program generates a second form of the logistic model. Four strata were formed, as before, and the standardized prevalence rates for aspirin plus prednisone and aspirin alone were calculated to be \(.55\) and \(.26\), respectively; the value of the normal test statistic is \(3.78\). The conclusion is now similar to those obtained using the other two approaches: aspirin plus prednisone results in more purpura than aspirin alone at the \(.0002\) level of significance. A summary of the various analyses is presented in Table 6.
5. DISCUSSION

Three out of four variables used in the cell-aggregation approach to forming strata are continuous. To use this approach, continuous variables must be categorized and those cases with missing data are either excluded from the analysis or estimated by some missing value procedure. Allowing more categories may lead to finer control of these variables, but one is limited by the size of the sample. These concerns do not arise with the classification tree approach to forming strata.

The logistic regression approach reached the same conclusion (that aspirin plus prednisone results in more purpura than aspirin alone) as the other two approaches when the 1350 cases (excluding weight, S.J., M.S.) were analyzed. One can test the accuracy of the predicted logistic probabilities from the model obtained by omitting these 3 variables. The proportion of cases with purpura in each stratum can be compared with the relevant average prediction for that stratum. We obtain a value of 8.72 for the classical chi-square "goodness of fit" test-statistic. This value is to be compared to the $\chi^2_3$ distribution and is found to be significant at the .05 level. Thus, we have evidence that the predictions are inaccurate, especially since the predicted values were calculated from the (same) sample that is observed. The difficulty here is that the logistic model (more generally, any model) accounts for bias provided that the fitted model is correct in form. The evidence of inaccurate predictions makes unsure all subsequent conclusions based on the logistic model.

The logistic regression results, obtained by including only the 652 complete (excluding weight) cases, clearly point out the danger of excluding cases from an analysis because of missing values. In retrospect, one can understand that physicians tend to have more complete records on patients with purpura than on patients without purpura. However, the standardized aspirin side-effect rate was inflated because of the disproportionately large number of non-purpura aspirin patients
excluded from this analysis. This was discovered only because the other two approaches to forming strata contradicted the logistic regression results.

Inability to satisfactorily contend with missing values is a liability afflicting all regression procedures. The standard solution is to use some missing value estimation procedure. Usually this involves replacing the missing value by the fitted value of a (stepwise) linear regression on the variables that best predict the variable containing the missings. This was attempted using BMPD-AM; however, the proportion of missings was too great for the program to run. This could not be circumvented by subdividing the problem. Still, even should such a strategy be employed, the results can be poor. This is attributable to biases introduced by missing value estimation giving rise to badly conditioned covariance matrices. The resultant regression coefficients can be highly unstable.

One often wants to develop strata that logically structure how variables influence the outcome of interest as well as control for confounding factors. Both the logistic regression and cell-aggregation approaches resulted in strata that are confusing and difficult to interpret, whereas the classification tree strata are easy to interpret via the tree structure. The tree structure can be used to identify potentially significant interactions among the variables. For example, when two-way interactions were allowed to be included in the logistic model, the interaction term for D.I. and age was found to be highly significant. This result is in accordance with the tree structure displayed in Figure 3.

A summary of some of the merits and liabilities for the three methods are tabled below:
Comparison of Three Approaches to Forming Strata

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Cell Aggregation</th>
<th>Logistic Regression</th>
<th>Classification Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requires software</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Allows continuous variables</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Requires strong model assumptions</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Handles missing data well</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Easy strata interpretation</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Handles large numbers of variables</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Exploits response surface continuity</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

REFERENCES


Daniel A. Bloch is a Statistician in the Divisions of Immunology and Biostatistics, Stanford University Medical Center, Stanford, CA 94305. Mark R. Segal is a Ph.D. student in the Department of Statistics, Stanford University, Stanford, CA 94305. This research was supported in part by Public Health Service Grant 5 R01 GM21215-10 and National Institute of Health Grant AM21393.

The authors are grateful to Byron Wm. Brown, Jr. and L. E. Moses for helpful discussions and to Karola Decleve for secretarial assistance.
Table 1

CELL AGGREGATION STRATA: CELL-SPECIFIC ASSOCIATED SIDE-EFFECT RATES AND STRATA ASSIGNMENT NUMBERS

<table>
<thead>
<tr>
<th></th>
<th>Age ≤ 55</th>
<th></th>
<th>Age &gt; 55</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D.I.*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-1</td>
<td>M:1/41=.02 1</td>
<td>M:4/21=.19 1</td>
<td>M:11/64=.17 1</td>
<td>M:11/35=.31 2</td>
</tr>
<tr>
<td></td>
<td>F:25/121=.21 1</td>
<td>F:24/81=.30 2</td>
<td>F:16/94=.17 1</td>
<td>F:25/80=.31 2</td>
</tr>
<tr>
<td>2-3</td>
<td>M:0/7=0.0 0 1</td>
<td>M:3/7=.43 4</td>
<td>M:5/11=.45 4</td>
<td>M:10/31=.32 2</td>
</tr>
<tr>
<td></td>
<td>F:8/25=.32 2</td>
<td>F:12/34=.35 3</td>
<td>F:24/51=.47 4</td>
<td>F:66/130=.51 4</td>
</tr>
</tbody>
</table>

*D.I. = disability index, D.D. = disease duration
M = male, F = female. Circled numbers are strata assignment numbers.

Table 2

CELL AGGREGATION STRATA AND ASSOCIATED PURPURA SIDE-EFFECT RATES

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Proportion of Standardizing Population</th>
<th>Side-Effect Rates</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standardizing Population</td>
<td>Aspirin Plus Prednisone</td>
<td>Aspirin Alone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.276</td>
<td>62/383=.16</td>
<td>3/10=.30</td>
<td>10/51=.20</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.210</td>
<td>89/291=.31</td>
<td>6/10=.60</td>
<td>9/34=.27</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.260</td>
<td>137/361=.38</td>
<td>10/16=.63</td>
<td>13/36=.36</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.254</td>
<td>177/353=.50</td>
<td>18/24=.75</td>
<td>6/23=.26</td>
<td></td>
</tr>
</tbody>
</table>
Table 3

CLASSIFICATION TREE STRATA AND ASSOCIATED PURPURA SIDE-EFFECT RATES

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Proportion of Standardizing Population</th>
<th>Side-Effect Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Standardizing Population</td>
</tr>
<tr>
<td>1</td>
<td>.305</td>
<td>83/431=.19</td>
</tr>
<tr>
<td>2</td>
<td>.188</td>
<td>72/266=.27</td>
</tr>
<tr>
<td>3</td>
<td>.100</td>
<td>39/142=.27</td>
</tr>
<tr>
<td>4</td>
<td>.105</td>
<td>79/149=.53</td>
</tr>
<tr>
<td>5</td>
<td>.302</td>
<td>201/427=.47</td>
</tr>
</tbody>
</table>

Table 4

ORDER OF VARIABLE SELECTION

<table>
<thead>
<tr>
<th>Approach</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>D.I.</td>
<td>D.D.</td>
<td>Age</td>
<td>Sex</td>
<td>S.J.</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>D.I.</td>
<td>D.D.</td>
<td>S.J.</td>
<td>Age</td>
<td>--</td>
</tr>
</tbody>
</table>
Table 5
LOGISTIC STRATA AND ASSOCIATED PURPURA SIDE-EFFECT RATES

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Range of Predicted Probabilities</th>
<th>Proportion of Standardizing Population</th>
<th>Side Effect Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Standardizing Population</td>
</tr>
<tr>
<td>1</td>
<td>0.0 - .24</td>
<td>.214</td>
<td>30/139=.22</td>
</tr>
<tr>
<td>2</td>
<td>.24 - .3</td>
<td>.161</td>
<td>23/104=.22</td>
</tr>
<tr>
<td>3</td>
<td>.3 - .4</td>
<td>.261</td>
<td>56/171=.33</td>
</tr>
<tr>
<td>4</td>
<td>.40 - 1.0</td>
<td>.364</td>
<td>128/238=.54</td>
</tr>
</tbody>
</table>

Table 6
SUMMARY OF ANALYSES

<table>
<thead>
<tr>
<th>Approach</th>
<th>Variables Entered</th>
<th>N</th>
<th>Variables Used</th>
<th>Aspirin + Prednisone</th>
<th>Aspirin Alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contingency Table</td>
<td>Sex, Age, D.D., D.I.</td>
<td>1388</td>
<td>Sex, Age, D.D., D.I.</td>
<td>.56</td>
<td>.27</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>All 8</td>
<td>1415</td>
<td>D.I., D.D., S.J., Age</td>
<td>.50</td>
<td>.27</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>7 (weight excluded)</td>
<td>652</td>
<td>D.I., D.D., Age, Dex, S.J.</td>
<td>.53</td>
<td>.40</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>5 (weight, S.J., M.S. excluded)</td>
<td>1350</td>
<td>D.I., D.D., Sex</td>
<td>.55</td>
<td>.26</td>
</tr>
</tbody>
</table>
Figure 1

Classification Tree - Males

*upper value = number of non-purpura cases
lower value = number of purpura cases
†purpura side-effect rate
Figure 2

Classification Tree - Males and Females

*941
474

D.I.
†(0.33)

≤ 67
> 67

348
83

(0.19)

**1

194
72

(0.27)

593
391

D.D.
(0.40)

≤ 7.15
> 7.15

399
319

S.J.
(0.44)

≤ 15.5
> 15.5

(0.47)

Surrogate,
Age: 48.6

226
201

173
118

(0.41)

≥ 56.7
> 56.7

103
39

(0.27)

3

70
79

(0.53)

4

*Upper value = number of non-purpura cases
lower value = number of purpura cases
†purpura side-effect rate
**stratum number